Universidade de Aveiro Departamento de Matemática
2020

Marta Luísa Santos Maltez

Novas Abordagens na Deteção de *Outliers* em Dados Composicionais

Universidade de Aveiro Departamento de Matemática
2020

Marta Luísa Santos Maltez

Novas Abordagens 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 - Ramo de Estatística e Otimização, realizada sob orientação científica de Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Dedico este trabalho à minha avó Adelina, por ser a pessoa que mais me inspirou na vida e que me ensinou que nenhuma adversidade é tão grande que não possa ser ultrapassada.

«It always seems impossible until its done.»

Nelson Mandela

O júri

Presidente

Doutor Eugénio Alexandre Miguel Rocha
Professor Auxiliar Universidade de Aveiro

Vogais

Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Auxiliar, Universidade de Aveiro

Doutor Marco André da Silva Costa
Professor Adjunto, Universidade de Aveiro

Agradecimentos

Gostaria de, em primeiro lugar, manifestar a minha gratidão à minha orientadora, a professora doutora Adelaide Freitas, que desde o primeiro momento, se prontificou a esclarecer todas as dúvidas que surgiram ao longe deste trabalho, incentivou-me a fazer mais e melhor, aconselhou-me e sempre teve uma palavra de carinho e de atenção em todos os momentos. Foi, sem dúvida, uma honra trabalhar sob a sua orientação e, por isso, o meu eterno obrigado.

Agradeço também à investigadora Cristiana Silva e à doutora Anabela Cachada pela gentileza na autorização do uso dos conjuntos de dados que permitiram a elaboração desta dissertação. Agradeço também aos meus pais que, apesar de não entenderem esta paixão pela Matemática, nunca me impediram de ir atrás daquilo que gostava e à minha irmã Sílvia, que sempre foi uma presença constante e motivadora no meu percurso académico.

Uma palavra de carinho e um obrigado a todos que conheci na cidade de Aveiro que, nestes últimos dois anos, me acolheu, em especial ao Miguel Ramos e à Mariana Pinto, companheiros e amigos desta jornada e que foram de um apoio fundamental e à Mafalda Campos, uma irmã mais nova que Aveiro me deu, que tornaram esta etapa especial.

Uma menção especial também a todos os meus amigos e colegas de licenciatura por todos os ensinamentos e palavras de apoio ao longo dos anos.

Agradeço, igualmente, ao meu amigo de longa data Ruben Antunes por ter sido um ponto de abrigo ao longo dos últimos dez anos.

Em último, mas não menos importante, menciono o meu apreço por todos os professores que tive ao longo destes anos, em particular alguns que me marcaram profundamente, a todos os meus explicandos e aos seus pais, pela compreensão que sempre tiveram e que me permitiu conciliar, mesmo que com algum sacrifício, tudo o que gosto de fazer e a todos aqueles com quem colaboro em outros projetos e que contribuíram para o meu desenvolvimento pessoal.

O meu mais sincero e profundo obrigada!

Palavras-chave dados composicionais, observações atípicas, distância de Mahalanobis robusta, abordagem comediana, estimador de Stahel-Donoho, Atipicidade Ajustada, dados epidemiológicos, qualidade dos solos

Resumo

Dados composicionais são um caso especial de dados multivariados que representam informação relativa na forma de log-razões entre as componentes. Os vetores são constituídos por componentes estritamente positivas, que têm como propriedades fundamentais a invariância de escala, a invariância de permutação e a coerência subcomposicional. As composições têm a sua representação num subespaço designado de simplex, sobre o qual se define a chamada Geometria de Aitchison.

Os *outliers*, ou observações atípicas, são dados que parecem desviar-se substancialmente das demais observações da amostra da qual este faz parte e sempre despertaram o interesse dos estatísticos. Os métodos de deteção de *outliers* são geralmente classificados em dois tipos: os métodos baseados em distância robusta e os métodos não tradicionais. Os primeiros baseiam-se em calcular estimativas para a média e covariância dos dados e depois calcular a distância robusta dessas observações e os segundos evitam o uso da distância e optam por fazer uma abordagem com mapas não lineares, uso dos vetores próprios ou projeções, entre outros.

Até ao momento, os métodos numéricos e gráficos para detetar *outliers* em dados composicionais baseiam-se na distância de Mahalanobis robusta.

Neste trabalho, propõem-se duas outras abordagens, também baseadas em distâncias robustas, para a deteção de *outliers* em dados composicionais. O primeiro método é a Abordagem Comediana (*Comedian Approach*) e o segundo método é a Atipicidade Ajustada (*Adjusted Outlyingness*), que se baseia no Estimador de Stahel-Donoho, não pressupondo qualquer tipo de distribuição a respeito dos dados. Pretende-se então, aplicar esses métodos a dois conjuntos de dados: um de dados epidemiológicos – a SIDA em Cabo Verde – e um outro de qualidade de solos em Lisboa, Portugal, e tentar perceber se, numa perspetiva composicional, existem observações atípicas ou não.

Keywords compositional data, outliers, Mahalanobis robust distance, comedian approach, Stahel-Donoho estimator, adjusted outlyingness, epidemiological data, soil quality

Abstract

Compositional data are a special case of multivariate data which represent relative information in the form of log-ratios between the components. The vectors are constituted by components strictly positive with fundamental properties the scale invariance, permutation invariance and subcompositional coherence. Compositions are represented in a Euclidian subspace named simplex where the so-called Aitchison Geometry is applied.

Outliers, or atypical observations, are data which seems to be substantially deviated from the other observations in the same dataset. Outlier detection methods are usually classified into two types: robust distance-based methods and non-traditional methods. The former are based on the Mahalanobis distance calculated using robust estimates for the mean and the covariance matrix. The later avoid the use of distance and prefer to use non-linear maps, eigenvectors or projections, between others.

Until now, outlier detection methods in compositional data are based on robust distance and Minimum Covariance Determinant for estimating the covariance matrix. Besides numerical approach, these methodologies are also depicted on the graphical representations.

In this work, two other approaches are proposed to detect outliers in compositional data. The first method is the Comedian Approach and the second is the Adjusted Outlyingness. The last one is a modification of the Stahel-Donoho Estimator and any type of distribution about the data is assumed. These methods are applied on two real datasets: epidemiological data related to AIDS in Cape Verde and geochemical data related to soil quality in Lisbon (Portugal). Results show the existence of atypical observations, in a compositional perspective.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Dados Composicionais

O interesse inicial da análise de dados composicionais surgiu no final do século XIX, com Karl Pearson (1857-1936). No seu artigo *'On a Form of Spurious Correlation which May Arise when Indices Are Used in the Measurement of Organs'*, Pearson evidenciou alguns perigos com a interpretação de correlações entre rácios cujos numeradores e denominadores têm partes comuns afetando, consequentemente, o estudo de dados composicionais. Por diversas vezes, a aplicação de métodos estatísticos *standard*, sem levar em conta a estrutura composicional dos dados tem conduzido a resultados erróneos, o que alertou os cientistas e estatísticos para lidarem com maior cuidado com este tipo de dados.

Apenas em 1986 com John Aitchison e a publicação do seu livro *'The Statistical Analysis of Compositional Data'*, que é tido como a principal referência sobre a análise de dados composicionais, foram introduzidos os conceitos fundamentais no que à estrutura de dados composicionais diz respeito e, desse modo, a abordagem correta à sua análise.

Já em 2018, Peter Filzmoser, Karel Hron e Matthias Templ publicaram o livro *'Applied Compositional Data Analysis with Worked Examples in R'* onde, além de enunciarem os conceitos inerentes aos dados composicionais, apresentam exemplos práticos de dados composicionais, utilizando o *software* R e diversas *packages* desenvolvidas para fazer o tratamento e análise deste tipo de dados, algumas das quais desenvolvidas pelos próprios autores do livro.

Os dados composicionais são um tipo de observações multivariadas que traduzem informação de modo relativo, isto é, descreve como cada unidade amostral (indivíduo, objeto) está constituída indicando como cada parte da composição - designada de componente - contribui para o seu todo. Usualmente, este tipo de dados pode ser visto como percentagens ou proporções, mas também em outro tipo de unidades como, por exemplo, mg/kg, mg/l ou partes por milhão (ppm). Apesar de proporções ou percentagens remeterem para que a soma das partes seja uma constante (1 ou 100, respetivamente), isso não se traduz numa obrigatoriedade, mas apenas numa conveniência

[1; 2].

Definição 1.1.1 (Dados Composicionais)**.** *Seja* $\boldsymbol{x} = (x_1, \ldots, x_d)$ *um vetor composto por* d *partes de um todo,* $d \geq 2$*, com* $x_i > 0$*,* $i \in \{1, \ldots, d\}$*. Então* $\boldsymbol{x}$ *pode ser visto como uma observação composicional se* $x_1 + \cdots + x_d = k$*,* $k \in \mathbb{R}^+$*.*

De modo particular, a constante k toma o valor 1 quando se trata de proporções, e 100 quando se trata de percentagens.

Apesar de percentagens ou proporções remeterem, de modo intuitivo, para informação relativa, pode-se definir, de modo equivalente, dados composicionais como *ratios* entre componentes (partes) [1].

Analisar dados sob a perspetiva composicional, onde cada variável representa uma parte de um todo, pode ser de extremo interesse. São diversas as áreas nas quais a perspetiva de contribuições relativas se apresenta como uma mais valia na análise de dados: na Geologia pode ser empregue no estudo dos constituintes dos solos, na Economia para a análise das componentes nos gastos das famílias, na Medicina pode ser aplicada à composição do corpo (ex: gordura, ossos, músculos), na Indústria Alimentar à composição dos alimentos, entre outras. Esta abordagem deverá ser utilizada sempre que o interesse na análise dos dados recaia sobre a informação relativa em vez da informação absoluta. Contudo, este tipo de análise pode ser complementada por uma perspetiva absoluta dos dados, dependendo da sensibilidade de quem está a estudar os dados, bem como o objetivo do estudo em questão.

Como os dados composicionais têm uma estrutura muito particular, os métodos estatísticos de dados multivariados, podem conduzir a análises e interpretações inadequadas. A solução que se utiliza para que os métodos estatísticos convencionais se possam aplicar é trabalhar numa estrutura algébrico-geométrica que siga os princípios dos dados composicionais (como sejam, invariante quanto à escala, invariante quanto à permutação e coerência subcomposicional) que é conhecida, nos nossos dias, por Geometria de Aitchison, por ter sido introduzida no seu livro [3] e que permite interpretar dados expressos como observações composicionais como se de coordenadas reais se tratassem e, desse modo, as metodologias estatísticas comuns podem ser aplicadas diretamente. Essas coordenadas são formadas por *logratios* de pares de partes composicionais e as suas agregações, nascendo, dessa forma, a metodologia designada por *logratio*.

Apesar das vantagens que as técnicas baseadas em transformações *logratios* apresentam para a análise de dados composicionais, essas técnicas não alcançaram o sucesso que se esperava no seio dos estatísticos. Tal facto pode, possivelmente, ser explicado devido à tendência natural de interpretar e analisar resultados em termos absolutos direcionando a investigação para este tipo de estudo e, consequentemente, a uma menor à vontade na abordagem segundo a perspetiva relativa, necessitando que o raciocínio seja direcionado para análises em termos de razões [4; 5].

Contudo, essa reticência inicial já não se verifica e, atualmente, o estudo de dados composicionais está em expansão. No momento, existe, por exemplo, um grupo de

investigação forte na Universidade de Girona, Espanha, um grupo em crescimento na República Checa, em torno de Karel Hron, e um outro grupo forte à volta de Peter Filzmoser em Viena, Áustria. Peter Filzmoser foi orientador de doutoramento de Karel Hron. Importa ainda destacar os contributos de Michael Greenacre na análise de *logratios*.

Hodiernamente, a análise de dados composicionais pode ser basicamente descrita por três etapas ([3; 4; 5]):

- representação de dados em coordenadas *logratios*;
- uso de técnicas de análise estatística multivariada sobre os dados em coordenadas *logratios* transformadas;
- interpretação dos resultados, quer nas coordenadas transformadas, quer nas coordenadas originais.

O estudo aprofundado das transformações utilizadas, bem como toda a geometria envolvendo dados composicionais será explorada no Capítulo 2.

1.2 Observações Atípicas

Observações atípicas, vulgarmente designadas em Estatística por *outliers*, podem ser definidas, de forma concisa, como dados que se diferenciam drasticamente dos demais. A definição de *outlier*, em particular de *outlier* em dados multivariados, não é fácil sendo que podem ser encontradas na literatura especializada várias formas de caraterizar uma observação atípica. Algumas das mais comuns são:

> "[An outlier is a point such that] *in observing a set of observations in some practical situation one (or more) of the observations 'jars' stands out in contrast to other observations, as extreme value.*"[6; 7]

> "*An outlying observation, or 'outlier', is one that appears to deviate markedly from other members of the sample in which it occurs.*"[6; 8]

> "*An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.*"[6; 9]

Das definições anteriores pode-se concluir que um *outlier* é caracterizado pela sua relação com as restantes observações que fazem parte de uma amostra, isto é, a sua definição requer que a observação esteja integrada numa amostra. O seu distanciamento em relação a essas observações é fundamental para se fazer a sua caracterização [10]. Todavia, nem sempre é fácil fazer a sua caraterização e termos como "destaca-se", "parece desviar-se" ou "levanta suspeitas", muitas vezes utilizados para caraterizar uma observação que se desmarca das restantes, envolvem alguma subjetividade ou ideias pré-concebidas dos dados por parte de quem os está a tratar [6].

A procura pelo incomum, não representativo, atípico, aberrante, estranho, anormal sempre recebeu um grande interesse dos analistas de dados que desenvolveram métodos, não só para simplesmente identificar as observações atípicas, mas também para

acomodar estas numa variedade de situações. Os dois motivos para a procura de observações desviantes das demais são o interesse por elas próprias e a possibilidade dessas observações desviantes poderem influenciar o resto dos dados e, consequentemente, as conclusões da análise efetuada. Juntamente com a identificação ou acomodação dos *outliers*, ter uma ideia de porquê e como eles surgem é importante. Barnett e Lewis, no seu artigo [7], classificaram os tipos de variação dos dados em três grupos:

i. variabilidade inerente - é a variabilidade natural de qualquer conjunto de dados;

ii. erro de medição - inclui as limitações do instrumento de medição, bem como os de gravação/digitação dos dados feito por quem recolheu os dados;

iii. erro de execução - inclui situações com observações que não são do interesse da população ou situações onde uma amostra enviesada ou inadequada é utilizada.

Deste modo, uma observação tida como atípica deve permanecer no conjunto de dados se a variabilidade da mesma se dever à inerente variação dos dados; caso se deva a erros de medição ou execução, ela deverá ser devidamente identificada da amostra [6].

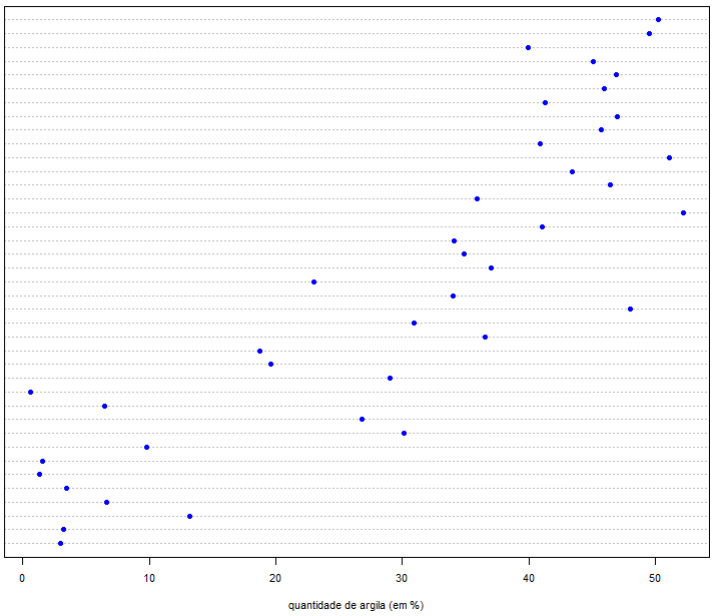
1.2.1 Outliers Univariados

Em dados univariados, isto é, em que apenas uma variável está em estudo, o conceito de observação atípica parece ser relativamente fácil de definir: todos as observações que estejam "bastante afastadas" da maioria dos dados e que podem, potencialmente, não seguir o mesmo modelo, poderão ser *outliers*. Quer o tamanho da amostra seja elevada ou não, usualmente recorre-se a metodologias gráficas para observar a dispersão e forma dos dados, bem como a algumas estatísticas sumárias (mínimo, máximo, média, mediana, moda, desvio padrão ou variância e quantis que se considerem pertinentes para a análise) para caracterizar a estrutura dos dados.

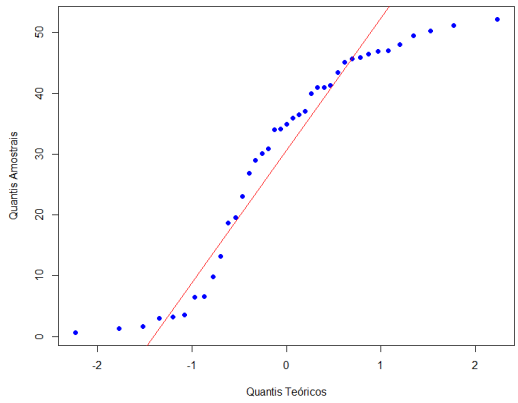
No que diz respeito a metodologias gráficas, as mais comuns recorrem a representações gráficas da função de frequências (absoluta ou relativa) como sejam, por exemplo, o *dotplot* (gráficos de pontos), diagramas de caule-e-folhas, ou ainda, à avaliação gráfica do ajustamento de um dado modelo aos dados, como sejam os QQ-*plots* ou à visualização da distribuição dos dados como seja o *box plot*. Qualquer que seja a preferência pela representação gráfica, todas elas são de interpretação relativamente fácil e podem indicar observações como sendo, eventualmente, atípicas. Nas figuras 1.1a, 1.1b, 1.1c e 1.1d são exibidos exemplos das metodologias gráficas citadas usando a variável Argila do conjunto de dados 'Arctic Lake' contendo 39 observações o qual foi retirado do livro de J. Aitchison [3].

```
0 | 11233477
1 | 039
2 | 0379
3 | 01445677
4 | 011135666778
5 | 0012
```

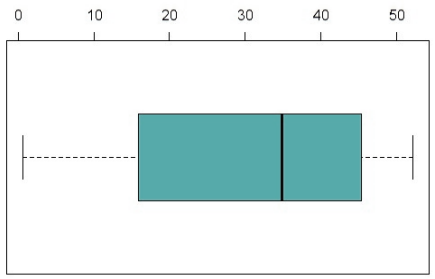
(a) Diagrama de caule e folhas (*steam and leaf plot*)



(b) Gráfico de pontos (*dotplot*)



(c) *QQ-plot* da distribuição Normal



(d) *Box plot*

Figura 1.1: Diversas representações gráficas do conjunto de dados 'Arctic Lake', para a variável Argila.

De todas as metodologias gráficas enunciadas, a mais popular é o *box plot*. A popularização do uso deste gráfico deve-se a John Tukey [**?**], apesar de haver indícios que apontam para o seu uso em 1933, mais de 30 anos antes de Tukey, por parte de P. R. Crowe [13].

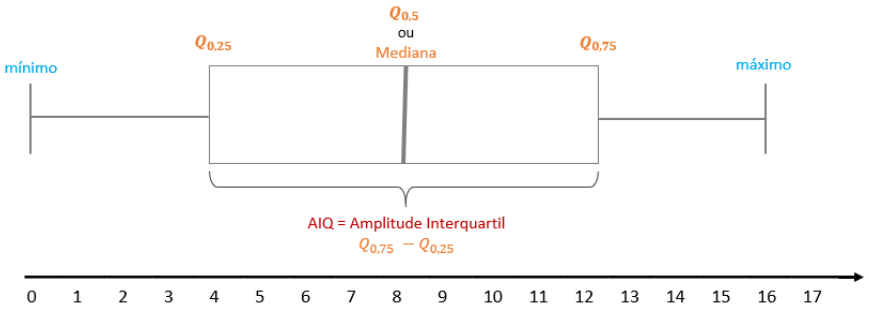


Figura 1.2: Box plot Clássico (ou de Tukey)

O *box plot* é uma representação gráfica de cinco medidas sumárias, de acordo com a figura 1.2 e que permite traduzir informação acerca da zona de concentração ou dispersão dos dados e simetria da distribuição dos dados. Na sua forma mais simples (figura 1.2), os traços verticais (bigodes) que limitam os traços horizontais assinalam o mínimo e o máximo da amostra (valores extremos). Na caixa, que vai do quantil de ordem $1/4$, $Q_{0.25}$, até ao quantil de ordem $3/4$, $Q_{0.75}$, estão contidas 50% das observações mais centrais. A largura (i.e., altura) desta caixa é arbitrária. No interior dessa caixa coloca-se, em geral com uma espessura maior do que a usada para os limites da caixa, um segmento de reta sobre o valor de abcissa igual à mediana [14].

Com a ajuda do *box plot*, a dispersão dos dados pode ser avaliada recorrendo, para tal, a duas medidas de dispersão:

- a amplitude do intervalo de variação:

$$H = \max - \min \tag{1.1}$$

- a amplitude interquartil:

$$AIQ = Q_{0.75} - Q_{0.25} \tag{1.2}$$

Geralmente, a preferência recai sobre a amplitude interquartil, uma vez que, ao contrário da amplitude do intervalo, é insensível à presença de *outliers* e, desse modo, mais robusta.

Para estudar a simetria da distribuição, geralmente comparam-se a diferença entre quartis consecutivos. Assim:

- se $Q_{0.75} - Q_{0.5} \approx Q_{0.5} - Q_{0.25}$, então é de crer que a distribuição seja aproximadamente simétrica;
- se $Q_{0.75} - Q_{0.5} \gg Q_{0.5} - Q_{0.25}$, então é de crer que a distribuição seja assimétrica à direita;
- se $Q_{0.75} - Q_{0.5} \ll Q_{0.5} - Q_{0.25}$, então é de crer que a distribuição seja assimétrica à esquerda.

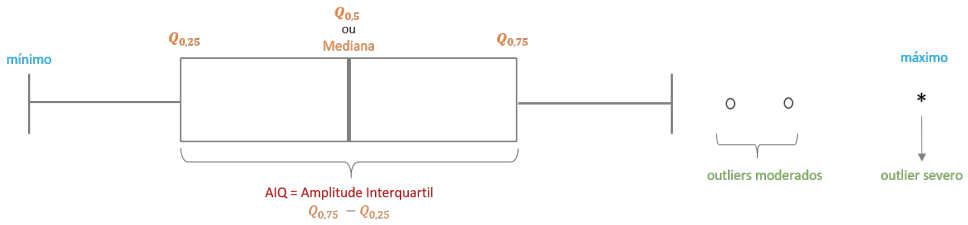


Figura 1.3: *Box plot* clássico (ou de Tukey) com representação de *outliers*.

O *box plot* da figura 1.3 contempla possíveis *outliers* - observações do conjunto de dados que estão relativamente distantes da maioria das observações. Após calculada a amplitude interquartil, podem ser definidos os limites que permitem classificar as observações mais afastadas das demais como *outliers* moderados ou como *outliers* severos. Concretamente, são classificados como *outliers* severos se estiverem fora das barreiras externas definidas como se segue:

- barreira externa inferior:
$$\text{BEI} = Q_{0.25} - 3\text{AIQ}, \tag{1.3}$$
isto é, observações inferiores a este valor são considerados *outliers* severos;
- barreira externa superior:
$$\text{BES} = Q_{0.75} + 3\text{AIQ}, \tag{1.4}$$
isto é, observações superiores a este valor são considerados *outliers* severos.

Estando dentro das barreiras externas, são classificados como *outliers* moderados se estiverem fora das barreiras internas, sendo estas definidas como se segue:

- barreira interna inferior:
$$\text{BII} = Q_{0.25} - 1.5\text{AIQ}, \tag{1.5}$$
isto é, observações inferiores a este valor são considerados *outliers* moderados;
- barreira interna superior:
$$\text{BIS} = Q_{0.75} + 1.5\text{AIQ}, \tag{1.6}$$
isto é, observações superiores a este valor são considerados *outliers* moderados.

Resumindo, dada uma amostra univariada $(x_1, \ldots, x_n)$, uma observação x_i, para algum $i \in \{1, \ldots, n\}$:

- não é considerada *outlier* se BII $< x_i <$ BIS;
- é considerada *outlier* moderado se BEI $< x_i \leq$ BII ou BIS $\leq x_i <$ BES;

- é considerada *outlier* severo se $x_i \leq$ BEI ou $x_i \geq$ BES.

Há *softwares* que distinguem os dois tipos de *outliers* acima mencionados e outros que não. Usualmente, em caso de distinção, os moderados são representados por um círculo, $\circ$, e os severos por um asterisco ou uma pequena estrela, $*$ ou $\star$ [14], de acordo com a representação da figura 1.3.

Uma vez que estas regras não são dependentes do tamanho da amostra, isto é, o tamanho da amostra não é usado diretamente no cálculo dos *outliers*, a probabilidade de serem declaradas observações atípicas quando, na verdade, não existe nenhuma, muda com o número de observações e, por isso mesmo, a necessidade muitas vezes referida na literatura da representatividade da amostra. Desse modo, o diagrama de extremos e quartis difere das regras padrão de identificação de *outliers*, que são definidas com α probabilidade de identificar observações discrepantes quando, efetivamente, não existem [6], como acontece, por exemplo, com os testes de hipóteses.

Hoaglin *et al.* em [15], concluíram que, apesar da sua fácil representação, a verdade é que, na prática, o *box plot* tem uma probabilidade de 0.5 de declarar pelo menos um *outlier* numa amostra de tamanho 75 gerada aleatoriamente de uma distribuição Normal. A tentativa de melhorar a eficácia na identificação de potenciais observações atípicas, vários investigadores introduziram ligeiras alterações ao critério acima descrito de modo a reduzir a probabilidade de declarar, erroneamente, uma observação como atípica (no sentido gaussiano) quando não é. Na verdade, uns tentaram levar em conta o tamanho da amostra, fazendo-o depender do número de observações (ex: Hoaglin e Iglewicz, [16]), outros tomaram em conta possíveis enviesamentos das distribuições (ex: Kimber, [17]), entre outras abordagens que incluem regressão dos valores observados e modelos de funções de distribuição (ex: Van der Loo, [18]). No Capítulo 3 serão introduzidas outras metodologias que permitem utilizar a ideia do *box plot*, aplicado a outras distribuições e, desse modo, avaliar quais as observações que são, de facto, atípicas, tomando um modelo mais adequado aos dados.

1.2.2 Outliers Multivariados

Em dados multivariados, onde cada observação é constituída por d partes, onde $d \geq 2$, a deteção de observações atípicas torna-se consideravelmente mais difícil, uma vez que as metodologias gráficas funcionam para $d = 2$ e, possivelmente para $d = 3$, mas para dimensões $d > 3$, há a impossibilidade de se representar graficamente os dados como um todo. Por vezes, é comum fazerem-se representações de cada variável separadamente ou, então, representações por pares de variáveis [6]. Contudo, como num espaço d-dimensional, $d \geq 2$, há um infinidade de direções que cada observação pode tomar, mesmo com aquelas alternativas visuais, não é possível ter-se uma real percepção do que é um *outlier*, pois, ao contrário do que acontece nos dados univariados, em que só há duas direções possíveis para uma dada observação se afastar do conjunto dos dados, um sem número de direções de um espaço d-dimensional dificulta esse processo.

Não sendo possível a visualização de um espaço de dimensão $d > 3$, uma das propostas mais abordada na literatura especializada para a identificação de observações atípicas em dados multivariados com $d > 2$ é baseada na distância de Mahalanobis [19], reduzindo o problema a um espaço univariado. Esta distância, que recebeu este nome em homenagem ao matemático indiano Prasanta Chandra Mahalanobis, é invariante à escala, ou seja, não depende da escala das medições [20], ao contrário da distância Euclidiana, que se encontra ilustrada na figura 1.4.

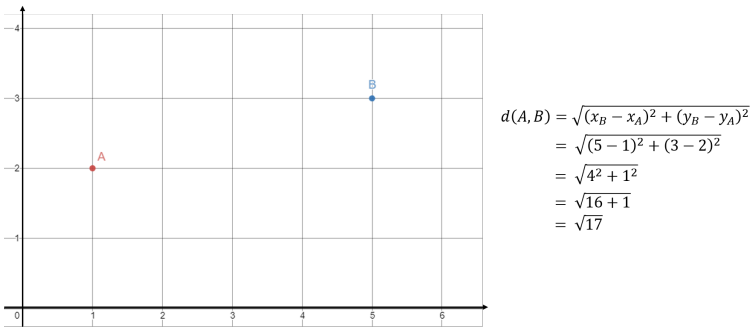


Figura 1.4: Distância Euclidiana entre dois pontos do plano.

Definição 1.2.1 (Distância de Mahalanobis)**.** *Seja $\boldsymbol{X} = [x_{ij}]$ uma matriz com n observações d-dimensionais. A distância de Mahalanobis da observação $\boldsymbol{x}_i$ é dada por:*

$$DM(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})} \tag{1.7}$$

onde $\bar{\boldsymbol{x}}$ e $\boldsymbol{S}$ são o vetor de médias e a matriz de covariâncias amostrais do conjunto de dados $\boldsymbol{X}$, respetivamente.

Exemplo 1.2.1. *Considere-se o conjunto de dados 'Arctic Lake', usando as variáveis Lodo e Argila, ambas expressas em percentagens. Na figura 1.5 visualiza-se este conjunto de dados através de um gráfico de dispersão.*

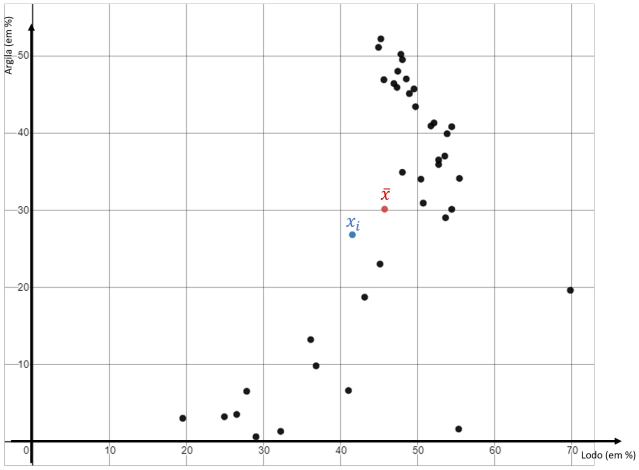


Figura 1.5: Representação gráfica das variáveis Lodo e Argila do conjunto de dados 'Arctic Lake'.

Para ilustração do cálculo da distância de Mahalanobis, tome-se a observação assinalada a azul na figura 1.5 ; a vermelho está representado o vetor da média amostral, $\bar{\boldsymbol{x}}$.
Sabendo que, nesta amostra:

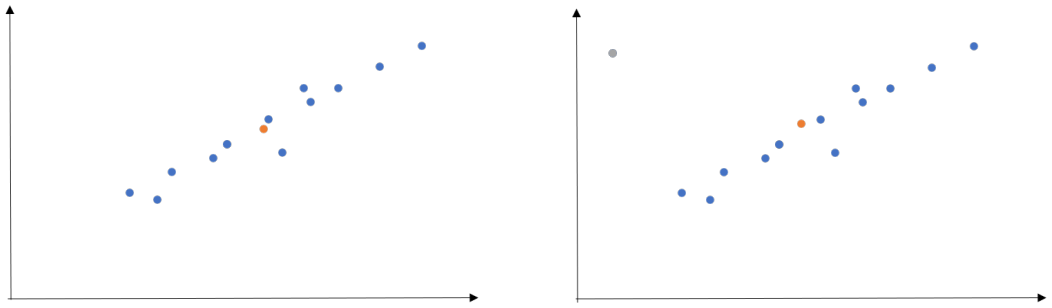
- $\boldsymbol{x}_i = [41.5 \quad 26.8]^T$
- $\bar{\boldsymbol{x}} = [45.69 \quad 30.11]^T$
- $\boldsymbol{S} = \begin{bmatrix} 102.453 & 103.077 \\ 103.077 & 294.205 \end{bmatrix}$

pode-se então proceder ao cálculo da distância de Mahalanobis:

$$\begin{aligned} DM(\boldsymbol{x}_i) &= \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})} \\ &= \sqrt{\left(\begin{bmatrix} 41.5 \\ 26.8 \end{bmatrix} - \begin{bmatrix} 45.69 \\ 30.11 \end{bmatrix} \right)^T \begin{bmatrix} 102.453 & 103.077 \\ 103.077 & 294.205 \end{bmatrix}^{-1} \left(\begin{bmatrix} 41.5 \\ 26.8 \end{bmatrix} - \begin{bmatrix} 45.69 \\ 30.11 \end{bmatrix} \right)} \\ &= \sqrt{\begin{bmatrix} -4.19 \\ -3.31 \end{bmatrix}^T \begin{bmatrix} 102.453 & 103.077 \\ 103.077 & 294.205 \end{bmatrix}^{-1} \begin{bmatrix} -4.19 \\ -3.31 \end{bmatrix}} \\ &= \sqrt{\begin{bmatrix} -4.19 \\ -3.31 \end{bmatrix}^T \begin{bmatrix} 0.0150 & -0.005 \\ -0.005 & 0.005 \end{bmatrix} \begin{bmatrix} -4.19 \\ -3.31 \end{bmatrix}} \\ &= \sqrt{0.176} \\ &\approx 0.42 \end{aligned}$$

A distância de Mahalanobis da observação $\mathbf{x}_i$ indica o quão longe ela está do centro do conjunto de dados, ou seja, do vetor de médias, levando em conta a estrutura dos dados no espaço d- dimensional, isto é, a matriz de covariâncias. Todavia, é bem conhecida na literatura que esta abordagem sofre do efeito de máscara, isto é, múltiplas observações atípicas não apresentam, necessariamente, valores elevados para a distância de Mahalanobis. Isto deve-se ao facto de o vetor de médias e a matriz de covariâncias amostrais não são estatísticas suficientemente robustas: um pequeno número de observações desviantes irá atrair o vetor de médias (uma vez que todas as observações são tidas com o mesmo peso) e inflacionar a matriz de covariância amostral na sua direção [19], como se ilustra no seguinte exemplo.

Exemplo 1.2.2. *Considere-se o conjunto original de 12 dados bidimensionais representados no gráfico de dispersão da figura 1.6a. A vermelho assinala-se a média amostral.*



(a) Conjunto de dados.

(b) Conjunto de dados ao qual foi adicionada uma observação atípica (assinalada a cinzento).

Figura 1.6: Influência de *outliers* num conjunto de dados. O conjunto de dados originais representado a azul, o ponto que representa a média a laranja e o *outlier* introduzido a cinzento.

Como é possível observar na figura 1.6b, a simples introdução de uma observação atípica faz com que o vetor de médias se desloque, sendo atraído pelo outlier. Numericamente, a introdução desta nova observação afetou o vetor de médias e a matriz de covariância amostral. Deste modo, o vetor de médias e a matriz de covariâncias sofreram alterações passando de (figura 1.6a):

$$m_{(a)} = \begin{bmatrix} 1.97 \\ 4.83 \end{bmatrix}, \quad S_{(a)} = \begin{bmatrix} 0.393 & 0.831 \\ 0.831 & 1.888 \end{bmatrix}$$

a (figura 1.6b):

$$m_{(b)} = \begin{bmatrix} 1.86 \\ 4.99 \end{bmatrix}, \quad S_{(b)} = \begin{bmatrix} 0.516 & 0.540 \\ 0.540 & 2.078 \end{bmatrix}$$

Assim, tomando a observação $x = (3.1, 7.2)$, pertencente ao conjunto de dados, e calculando a sua distância de Mahalanobis nos dois casos obtêm-se:

$$DM_{(a)}(x) = 1.803$$
$$DM_{(b)}(x) = 1.879$$

Uma vez que as estatísticas média amostral e matriz de covariâncias amostral são influenciáveis pela introdução de observações atípicas, é necessário conhecer estimadores que sejam menos afetados por "pertubações" nos dados ou modelos, nomeadamente que sejam robustas.

Definição 1.2.2 (Estatísticas robustas)**.** *As estatísticas robustas são uma extensão das estatísticas paramétricas clássicas que especificamente levam em consideração o facto que os assumidos modelos paramétricos usados pelos investigadores são apenas uma aproximação [21].*

Exemplo 1.2.3. *Sejam $x_1 = \{2, 4, 5, 8\}$ e $x_2 = x_1 \cup \{15\} = \{2, 4, 5, 8, 15\}$, conjunto obtido a partir de x_1 com a adição de uma nova observação, dois conjuntos de dados. Deste modo obter-se-á:*

→ *média:* $\bar{x}_1 = 4.75$*;* → *média:* $\bar{x}_2 = 6.8$*;*

→ *mediana:* $\tilde{x}_1 = 4.5$ → *mediana:* $\tilde{x}_2 = 5$

Pelo exemplo anterior, pode-se concluir que a média não é um estimador robusto do parâmetro de localização central da população onde foi recolhida a amostra, uma vez que a inserção de mais um elemento na amostra implicou um maior desvio na estimativa da média, enquanto que a mediana pode ser considerada uma estatística robusta, uma vez que a inserção de uma nova observação em pouco altera o valor da estimativa.

Associado às estatísticas robustas, importa referir outros conceitos-chave como o ponto de rutura (do inglês *breakdown point*) [6].

Definição 1.2.3 (Ponto de rutura)**.** *O ponto de rutura é uma medida usada para descrever a resistência dos estimadores robustos na presença de outliers. É dada pela fração de contaminação arbitrária de valores extremos, quer elevados, quer pequenos, que podem ser introduzidos na amostra antes de o valor do estimador se tornar arbitrariamente grande.*

Definição 1.2.4 (Ponto de rutura para a média amostral)**.** *Seja* $\hat{\mu}$ *um estimador para a média populacional, função do conjunto de* n *observações,* $\boldsymbol{X}$*. O breakdown point para* $\hat{\mu}$ *é dado por:*

$$\epsilon_n^*(\hat{\mu}, \boldsymbol{X}) = \min_m \left\{ \frac{m}{n}, \sup_{\tilde{\boldsymbol{X}}} ||\hat{\mu}(\tilde{\boldsymbol{X}}) - \hat{\mu}(\boldsymbol{X})|| = \infty \right\} \tag{1.8}$$

onde $\tilde{\boldsymbol{X}}$ *é o conjunto de observações corrompidas, obtidas ao substituir as observações verdadeiras por valores arbitrários.*

Pela definição anterior, pode-se concluir que o ponto de rutura para um estimador da média (populacional) é a menor fração da amostra que pode ser corrompida em que a distância entre a verdadeira média amostral e a média da amostra corrompida se tornar arbitrariamente grande.

De modo análogo pode ser definido o ponto de rutura para o estimador da matriz de covariância.

Definição 1.2.5 (Ponto de rutura para o estimador da matriz de covariância)**.** *Seja* $\hat{\Sigma}$ *o estimador para a covariância do conjunto de observações* $\boldsymbol{X}$*. O ponto de rutura para* $\hat{\Sigma}$ *é dado por:*

$$\epsilon_n^*(\hat{\Sigma}, \boldsymbol{X}) = \min_m \left\{ \frac{m}{n}, \sup_{\tilde{\boldsymbol{X}}} D\Big(\hat{\Sigma}(\tilde{\boldsymbol{X}}), \hat{\Sigma}(\boldsymbol{X})\Big) = \infty \right\} \tag{1.9}$$

onde $\tilde{\boldsymbol{X}}$ *é o conjunto de observações corrompidas, obtidas ao substituir as observações verdadeiras por valores arbitrários e* $D(\boldsymbol{A}, \boldsymbol{B}) = \max\{|\lambda_1(\boldsymbol{A}) - \lambda_1(\boldsymbol{B})|, |\lambda_d(\boldsymbol{A})^{-1} - \lambda_d(\boldsymbol{B})^{-1}|\}$ *, com* $\lambda_i(\boldsymbol{A})$ *o valor do* i*-ésimo valor próprio ordenado de* $\boldsymbol{A}$*.*

Pela definição de ponto de rutura para o estimador da matriz de covariâncias, pode-se concluir que corresponde à menor fração da amostra que pode ser contaminada por observações atípicas antes que a diferença entre os maiores valores próprios da verdadeira matriz de covariâncias estimada e da matriz de covariâncias da amostra corrompida se torne arbitrariamente grande ou, de modo equivalente, a diferença entre os

menores valores próprios das duas estimativas se torne muito próxima de zero.

Combinando as definições anteriores, será adequado ter estimativas para as estatísticas relevantes cujo ponto de rutura seja elevado e, desse modo, menos sensível a *outliers*. Estimativas com essas características dizem-se com um elevado ponto de rutura, preferencialmente o mais próximo de 50%, que corresponde ao limite teórico. Todavia, os estimadores média e matriz de covariâncias amostrais clássicos têm um ponto de rutura de $\frac{1}{n}100\%$, onde n é o tamanho da amostra, e que é bastante inferior ao expetável (para $n > 2$), ao valor do limite teórico, tendendo para zero quando o tamanho da amostra aumenta indefinidamente.

Outra importante medida de robustez é a função de influência (do inglês *influence function*). A função de influência é uma medida de dependência no valor de uma das observações da amostra, quando apenas essa observação é contaminada.

O ideal será que uma contaminação infinitesimal numa determinada observação da amostra tenha uma reduzida influência no estimador, isto é, se se tiver um estimador robusto, a função de influência será limitada.

Além das propriedades desejáveis já enunciadas para um estimador robusto falta destacar a invariância afim (do inglês *affine equivariance*). Esta propriedade envolve a transformação nos dados, permitindo que, caso a origem ou a escala da variável original seja alterada (por exemplo, por uma centralização dos dados), não haja perda de coerência nas conclusões, isto é, elas são as mesmas face à nova origem ou escala.

Definição 1.2.6 (Estimador de localização invariante afim)**.** *Um estimador de localização $\hat{\mu} \in \mathbb{R}^d$ diz-se invariante afim se e apenas se, para qualquer $\boldsymbol{b} \in \mathbb{R}^d$ e qualquer matriz não singular $\boldsymbol{A}_{(d\times d)}$,*

$$\hat{\mu}(\boldsymbol{AX + b}) = \boldsymbol{A}\hat{\mu}(\boldsymbol{X}) + \boldsymbol{b} \tag{1.10}$$

onde $\boldsymbol{X}$ se refere à matriz dos dados d-dimensional.

Definição 1.2.7 (Estimador de escala invariante afim)**.** *Um estimador de escala $\hat{\Sigma}$, matriz semidefinida positiva diz-se invariante afim se e apenas se, para qualquer $\boldsymbol{b} \in \mathbb{R}^d$ e qualquer matriz não singular $\boldsymbol{A}_{(d\times d)}$,*

$$\hat{\Sigma}(\boldsymbol{AX + b}) = \boldsymbol{A}\hat{\Sigma}(\boldsymbol{X})\boldsymbol{A}^T \tag{1.11}$$

onde $\boldsymbol{X}$ se refere à matriz dos dados.

Geometricamente, pode-se concluir que aplicar uma transformação linear ou uma rotação à matriz dos dados não terá qualquer efeito no estimador invariante afim, no sentido em que este estimador capta a transformação realizada nos dados tornando-se indistinguível usar o estimador sobre os dados transformados ou efetuar a transformação sobre o estimador aplicado aos dados originais.

Ao longo dos anos, diversos métodos têm sido propostos para detetar observações atípicas. Estes métodos são, geralmente, divididos em dois grupos:

- métodos baseados em distância robusta - usam estimativas robustas para estimar o vetor de médias e a matriz de covariâncias populacional e, de seguida, é calculada a distância de Mahalanobis usando essas estimativas robustas. As observações cuja a sua distância exceda o valor crítico ou de corte (no caso dos dados serem provenientes de uma distribuição Normal, esse valor é obtido por uma distribuição χ^2);
- métodos não tradicionais - estes métodos evitam o uso da distância robusta de Mahalanobis e optam por utilizar técnicas de *clustering*, projeções, entre outras.

1.3 *Software* Utilizado

Os cálculos e a maioria dos gráficos apresentados foram realizados com recurso ao *software* R, versão 3.6.1, no ambiente RStudio versão 1.2.1335.

As *packages* utilizadas foram: `robCompositions`, `matlib`, `robustbase`, `mvoutlier`, `aplpack`, `StatDA`, `foreign`, `VIM`, `dplyr`, `ggplot2`, `mrfDepth`, `Matrix`, `gridExtra`, `maps`, `mapdata`, `rworldmap`, `ggmap`, `car`.

Algumas imagens foi também elaboradas com recurso à aplicação e ao *website* Desmos (`https://www.desmos.com/`) e ao Microsoft Power Point.

Capítulo 2

Dados Composicionais

2.1 Noções Básicas de Dados Composicionais

Vetores com componentes, todas positivas, a expressar proporções ou percentagens são dados composicionais, isto é, observações que traduzem informação relativa [1], [23]. Apesar de serem observações multivariadas, onde uma análise absoluta pode ser feita, deve ver-se como dados composicionais sempre que a preferência da análise delas recai sobre o relativo.

Definição 2.1.1 (Dados Absolutos)**.** *Dados absolutos referem-se à contagem dos dados brutos, sem nenhuma operação que não seja a contagem ou a de medida [24].*

Definição 2.1.2 (Dados Relativos)**.** *Dados relativos são dados obtidos através da combinação de dados absolutos com uma determinada operação que permita a sua comparação com outros dados do mesmo tipo [24].*

Exemplo 2.1.1. *Segundo dados do website PORDATA, em 2019* [1]*, os dados por sexo dos alunos matriculados no ensino superior eram:*

Dados Absolutos:

Total	***Masculino***	***Feminino***
385247	176660	208587

Dados Relativos:

Total	***Masculino***	***Feminino***
100%	45.86%	54.14%

Assim, na comparação a informação absoluta interpreta-se usando a diferença entre os valores. Neste caso, em 2019, havia mais 31927 pessoas do sexo feminino no ensino superior do que do sexo masculino; contudo, em termos relativos, há mais $8.28\%(= 54.14\% - 45.86\%)$ de pessoas do sexo feminino que do sexo masculino matriculadas no ensino superior.

Se a informação está a ser analisada em termos relativos, é irrelevante se os dados estão representados como percentagens ou proporções: os rácios entre as componentes mantêm-se inalterados [1]. Contudo, a análise estatística dos dados deve levar em consideração o espaço amostral onde os dados estão inseridos [5].

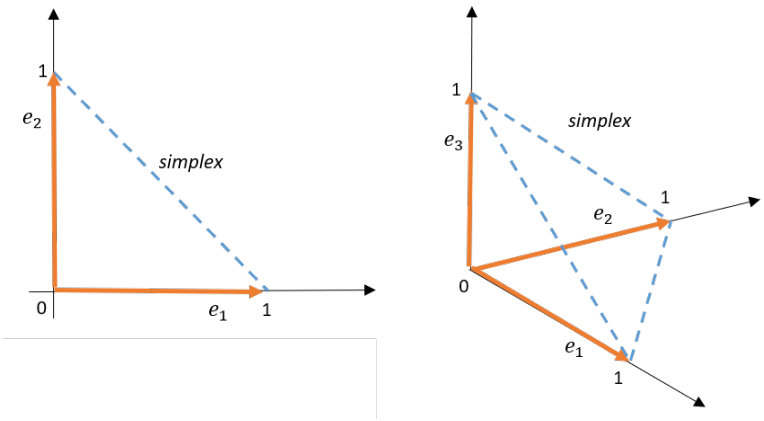
[1] Fontes/Entidades: DGEEC/ME-MCTES, PORDATA Última atualização: 2019-10-01

Um dos argumentos geralmente usados para a necessidade de uma abordagem diferente na análise com os dados composicionais deve-se ao facto que este tipo de dados não ser coerente com a geometria Euclidiana e, desse modo, o uso de técnicas estatísticas usuais numa estrutura de dados, que pertence a uma parte restrita do espaço real com as respetivas operações associadas, pode levar a resultando incoerentes e que não estão de acordo com a realidade dos dados [5]. Assim, os dados composicionais seguem a designada Geometria de Aitchison do simplex.

Definição 2.1.3 (Simplex). *Seja $\boldsymbol{x} = (x_1, \ldots, x_d)$ uma composição com d partes. O simplex de dimensão d é definido como:*

$$\mathcal{S}^d = \{\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_i > 0, \sum_{i=1}^{d} x_i = 1\} \tag{2.1}$$

Na figura 2.1, é ilustrado o conceito de simplex para composições de dimensão $d = 2$ e $d = 3$.



(a) (b)

Figura 2.1: Representação do simplex (a) de dimensão 1 em $\mathbb{R}^2$ e (b) de dimensão 2 em $\mathbb{R}^3$, delimitado pelas linhas a azul.

Assim, os d vértices do simplex correspondem aos vetores unitários $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)$, $\ldots$, $\mathbf{e}_d = (0, \ldots, 0, 1)$, que são vetores em $\mathbb{R}^d$. O simplex $\mathcal{S}^d$, definido em (2.1), corresponde a um subconjunto de dimensão $d - 1$. Este espaço é de particular interesse na análise em dados composicionais, uma vez que é onde se encontra contido o espaço amostral destes dados [1].

Apesar de definição de simplex ser dada como a soma das d componentes, estritamente maiores que zero, somarem a unidade, ela pode ser generalizada para qualquer constante positiva k.

$$\mathcal{S}^d = \{\mathbf{x} = (x_1 \ldots, x_d) \in \mathbb{R}^d : x_i > 0, \sum_{i=1}^{d} x_i = k\} \tag{2.2}$$

De facto, como será mostrado de seguida, a escolha de k, a constante de soma, é irrelevante, uma vez que é invariável quanto à escala [1]. Por exemplo, é análogo dizer que no ensino superior, no ano 2019, havia 45.86% de pessoas do sexo masculino no ensino superior (onde $k = 100$) ou que a proporção de pessoas do sexo masculino é de 0.4586 (onde $k = 1$). Assim, a ideia a reter deste pequeno exemplo é que a mudança de escala ou a multiplicação por uma constante positiva não modifica o conteúdo da informação. Isto sugere que a mesma informação pode ser traduzida como classes de equivalência de vetores proporcionais [4].

Definição 2.1.4 (Composições como classes de equivalência)**.** *Sejam $\boldsymbol{x}, \boldsymbol{y}$ dois vetores em $\mathbb{R}_+^d$, onde $\mathbb{R}_+^d$ denota o espaço real $d-$dimensional com componentes positivas; logo, $x_i, y_i > 0, \forall i \in \{1, \ldots, d\}$.*
Os vetores $\boldsymbol{x}$ e $\boldsymbol{y}$ dizem-se composicionalmente equivalentes se existir um escalar $\lambda \in \mathbb{R}_+$ tal que $\boldsymbol{x} = \lambda \boldsymbol{y}$, isto é, as composições $\boldsymbol{x} = (\lambda y_1, \ldots, \lambda y_d)$ e $\boldsymbol{y} = (y_1, \ldots, y_d)$ contêm a mesma informação relativa, $\forall \lambda \in \mathbb{R}_+$.

A figura 2.2 ilustra o conceito de composições como classe de equivalência.

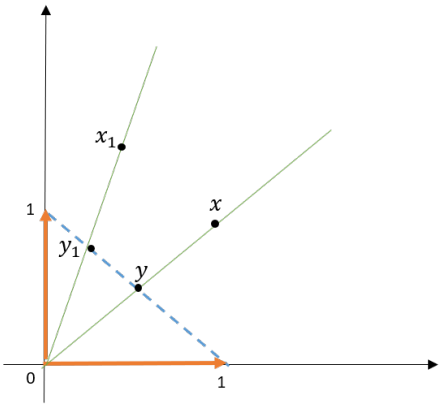


Figura 2.2: Composição de duas partes, onde $\mathbf{x}_1$ e $\mathbf{y}_1$ são equivalentes, bem como $\mathbf{x}$ e $\mathbf{y}$, pois estão localizados na mesma linha de projeção. $\mathbf{y}$ e $\mathbf{y}_1$ correspondem às projeções no simplex.

Desse modo, qualquer que seja o vetor escolhido de uma classe de equivalência, ele pode ser usado como representante dessa mesma classe, utilizando um fator de escala adequado. De modo a facilitar a análise, é escolhido um representante da classe de equivalência, pela normalização dos vetores, de modo a que a soma das componentes seja igual a um determinada constante k, podendo k tomar qualquer valores em $\mathbb{R}_+$. Assim, pode ser introduzido o conceito de operação de fecho (do inglês *closure operator*) [5], [4].

Definição 2.1.5 (Operação de fecho). *Seja $\boldsymbol{x} = (x_1, \ldots, x_d)$ um vetor em $\mathbb{R}^d$. O fecho de $\boldsymbol{x}$, para uma dada constante $k \in \mathbb{R}_+$, é definido por:*

$$\mathcal{C}_k(\boldsymbol{x}) = \left(\frac{k \times x_1}{\sum_{i=1}^{d} x_i}, \ldots, \frac{k \times x_d}{\sum_{i=1}^{d} x_i} \right) \tag{2.3}$$

Ao aplicar a operação de fecho a uma dada composição **x**, cria-se num novo vetor composicional, com o mesmo número de elementos e cuja soma dos mesmos perfaz k, o que faz com que se duas composições, **x** e **y**, são equivalentes então $\mathcal{C}_k(\mathbf{x}) = \mathcal{C}_k(\mathbf{y})$ [5]. De facto, se **x** e **y** são composições equivalentes, então $\mathbf{x} = \lambda\mathbf{y}$ e, portanto, o fecho de **x** para a constante k é dada por:

$$\begin{aligned}
\mathcal{C}_k(\mathbf{x}) &= \mathcal{C}_k(\lambda\mathbf{y}) \\
&= \left(\frac{k \times \lambda y_1}{\sum_{i=1}^{d} \lambda y_i}, \ldots, \frac{k \times \lambda y_d}{\sum_{i=1}^{d} \lambda y_i} \right) \\
&= \left(\frac{\lambda \times k \times y_1}{\lambda \sum_{i=1}^{d} y_i}, \ldots, \frac{\lambda \times k y_d}{\lambda \sum_{i=1}^{d} y_i} \right) \\
&= \left(\frac{k y_1}{\sum_{i=1}^{d} y_i}, \ldots, \frac{k y_d}{\sum_{i=1}^{d} y_i} \right) \\
&= \mathcal{C}_k(\mathbf{y})
\end{aligned}$$

Desse modo, tem-se $\mathcal{C}_k(\mathbf{x}) = \mathcal{C}_k(\lambda\mathbf{y})$.

Exemplo 2.1.2. *Tome-se o conjunto de dados 'Arctic Lake', disponível em [3].*
Considere-se a observação que corresponde à profundidade de $10.4m$, constituída por 3 sedimentos: Areia (x_1), Lodo (x_2) e Argila (x_3).
Seja $\boldsymbol{x} = (77.5, 19.5, 3)$. A soma das 3 partes é $x_1 + x_2 + x_3 = 77.5 + 19.5 + 3 = 100$. Então o fecho da composição $\boldsymbol{x}$ para $k = 1$ é:

$$\begin{aligned}
y &= \mathcal{C}_1(\boldsymbol{x}) \\
&= \left(\frac{1 \times 77.5}{100}, \frac{1 \times 19.5}{100}, \frac{1 \times 3}{100} \right) \\
&= (0.775, 0.195, 0.03)
\end{aligned}$$

onde a composição resultante $\boldsymbol{y}$ satisfaz a condição $y_1 + y_2 + y_3 = 1 = k$. Tem-se que $\boldsymbol{x}$ e $\boldsymbol{y}$ são composições equivalentes já que pode escrever-se $\boldsymbol{x} = 100\boldsymbol{y}$.

Esta caraterística de soma constante confere aos dados composicionais particularidades que os distingue de outros dados multivariados e, por esse motivo, é que estes dados se inserem no simplex.

Uma composição que seja resultante de uma operação de fecho é equivalente a que lhe deu origem, pois está na mesma classe de equivalência. Contudo, por vezes apenas algumas partes da composição é que tem o verdadeiro interesse para quem a está analisar, o que é designado por subcomposição.

Definição 2.1.6 (Subcomposição). *Seja $\boldsymbol{x} = (x_1, \ldots, x_d)$ uma composição com d partes e seja $S = \{s_1, \ldots, s_p\}$ um conjunto de índices que indica as partes selecionadas para a subcomposição.*

Uma subcomposição $\boldsymbol{x}_s$, com p partes, é obtida pela aplicação da operação de fecho ao subvetor $(x_{s_1}, \ldots, x_{s_p})$ de $\boldsymbol{x}$.

Exemplo 2.1.3. *Retomando a composição do exemplo 2.1.2, pode-se formar uma subcomposição apenas selecionando as partes de interesse: areia e argila. Assim, a composição original é $\boldsymbol{x} = (0.775, 0.195, 0.03)$. O conjunto dos índices corresponde a $S = \{1, 3\}$. Como $x_1 + x_3 = 0.775 + 0.030 = 0.805$, então a subcomposição será dada por:*

$$\begin{aligned}\boldsymbol{x}_s &= \mathcal{C}(x_1, x_3) \\ &= \Big(\frac{0.775}{0.775 + 0.030}, \frac{0.030}{0.775 + 0.030}\Big) \\ &= \Big(\frac{0.775}{0.805}, \frac{0.030}{0.805}\Big) \\ &= (0.963, 0.037)\end{aligned}$$

2.2 Princípios da Análise Composicional

Conforme visto na seção anterior, os dados composicionais e as suas operações não se encontram no espaço Euclidiano usual e sim no simplex. Para conseguir operar com estes dados, foi necessário conceber o conceito geométrico que permitisse isso: a designada Geometria de Aitchison [1] .

Assim, é necessário estabelecer os princípios que qualquer método estatístico deve cumprir de modo a que possa ser aplicado a composições [2]. Esses três princípios são:

- invariância de escala;
- invariância de permutação;
- coerência subcomposicional.

2.2.1 Invariância de Escala

Um dos princípios da análise composicional é a invariância de escala. Nos dados composicionais, o valor absoluto das partes da composição não é relevante, uma vez que composições na mesma classe de equivalência contém, essencialmente, a mesma informação [4]. Um exemplo disso é a composição de uma amostra de solo: a mesma amostra pode ser vista como a percentagem dos constituintes (onde a soma das partes é 100) ou como as miligramas de cada sedimento numa amostra de 1 grama (onde a soma das partes é 1000). Assim, se se designar por $\mathbf{x}$ a composição em termos de percentagens e por $\mathbf{y}$ a composição do peso em miligramas dos constituintes tem-se que $\mathbf{y} = 10\mathbf{x}$. Assim, estas duas composições diferem apenas de um fator de escala, $\lambda = 10$, mas a sua análise será idêntica, independente do valor de λ. Essa propriedade é designada por invariância de escala (do inglês *scale invariance*).

Definição 2.2.1. *Seja f uma função definida em $\mathbb{R}^d_+$. Diz-se que f é uma função invariante quanto à escala se, $\forall \lambda \in \mathbb{R}_+$ e para qualquer composição $\boldsymbol{x}$ em $\mathcal{S}^d$, satisfaz:*

$$f(\lambda \boldsymbol{x}) = f(\boldsymbol{x}) \tag{2.4}$$

Assim, a função é invariante quanto à escala se a imagem de vetores composicionalmente equivalentes por meio de f é sempre a mesma [4]. Deste modo, se as partes estiverem escritas na forma de *log-ratios*, então f é invariante quanto à escala [2].

2.2.2 Invariância de Permutação

Outro dos princípios da análise composicional é a invariância de permutação, isto é, as conclusões que se obtêm de uma análise composicional não devem diferir se a ordem das partes da composição for alterada. Assim, qualquer permutação das partes deverá ter o mesmo resultado e em nada afetará os resultados organizar as partes do modo mais conveniente possível.

Exemplo 2.2.1. *Considere-se novamente a observação que corresponde à profundidade de* 10.4 *m, constituída por 3 sedimentos: areia* (x_1)*, lodo* (x_2) *e argila* (x_3)*, escrita como:* $\boldsymbol{x} = (x_1, x_2, x_3) = (77.5, 19.5, 3)$*. Se, por conveniência, for necessário permutar as partes da composições, para que argila esteja em primeiro, ter-se-á uma nova composição,* $\boldsymbol{x}_A = (x_3, x_1, x_2) = (3, 77.5, 19.5)$*. Esta permutação das partes em nada altera a informação que se obtém do vetor composicional.*

2.2.3 Coerência Subcomposicional

O terceiro e último princípio da análise composicional é a coerência subcomposicional: subcomposições (como estabelecido na definição 2.1.6) devem-se comportar como as projeções ortogonais fazem na análise real convencional [2].Concretamente, a análise sobre um conjunto de partes de uma composição - subcomposição - não deverá depender das partes que não foram selecionadas, pelo que o resultado da análise de uma subcomposição não pode contradizer os resultados obtidos com a composição [4], isto é, os resultados da subcomposição deverão ser coerentes com os da composição completa.

As duas principais implicações práticas deste princípio são:

- a distância entre duas composições deve ser maior, ou pelo menos igual, à distância entre duas quaisquer subcomposições - dominância subcomposicional;
- caso uma parte não informativa seja retirada, os resultados deverão manter-se.

Exemplo 2.2.2. *Sejam* $\boldsymbol{x} = (77.5, 19.5, 3)$ *e* $\boldsymbol{y} = (9.5, 53.5, 37)$ *duas observações do conjunto de dados 'Artic Lake' à profundidade de* 10.4 *e* 47.1 *metros, respetivamente, que correspondem a duas composições do solo.*
Tome-se as subcomposições constituídas pelos sedimentos areia e argila. Assim, tem-se: $\boldsymbol{x}_A = (77.5, 3)$ *e* $\boldsymbol{y}_A = (9.5, 37)$*. Assim:*

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(77.5 - 9.5)^2 + (19.5 - 53.5)^2 + (3 - 37)^2} \approx 82.28$$

$$d(\boldsymbol{x}_A, \boldsymbol{y}_A) = \sqrt{(77.5 - 9.5)^2 + (3 - 37)^2} \approx 76.03$$

O mesmo aconteceria com quaisquer composições que fossem selecionadas e respetivas subcomposições.

Levando todos estes princípios em consideração e usando técnicas que sejam adequadas a dados composicionais, tem-se a garantia que a escolha de uma subcomposição não altera a relação entre as partes, uma vez que a ordem pela qual as partes estão representadas não interessa, mas sim a proporção que cada parte representa [4].

2.3 Geometria de Aitchison no Simplex

Como foi explicado na seção 2.1, o espaço amostral onde se inserem os dados composicionais é o simplex e a estrutura geométrica das composições é designada por Geometria de Aitchison [1]. O objetivo agora é definir uma estrutura de espaço vetorial no simplex e, para tal, é necessário definir algumas operações básicas. Essas operações correspondem à adição de dois vetores e à multiplicação de um vetor por um escalar no espaço simplex. Para distinguir estas operações das da Geometria Euclidiana, é usada uma notação distinta [1].

Definição 2.3.1 (Perturbação)**.** *Sejam $\boldsymbol{x} = (x_1, \ldots, x_d)$ e $\boldsymbol{y} = (y_1, \ldots, y_d)$ duas composições em $\mathcal{S}^d$. A perturbação de $\boldsymbol{x}$ por $\boldsymbol{y}$ é definida por:*

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \ldots, x_d y_d) \tag{2.5}$$

onde $\mathcal{C}(.)$ refere-se à operação de fecho.

Exemplo 2.3.1. *Sejam $\boldsymbol{x} = (0.775, 0.195, 0.03)$ e $\boldsymbol{y} = (0.095, 0.535, 0.37)$ duas composições. A perturbação de $\boldsymbol{x}$ por $\boldsymbol{y}$ é dada por:*

$$\begin{aligned}
\boldsymbol{x} \oplus \boldsymbol{y} &= (0.775, 0.195, 0.03) \oplus (0.095, 0.535, 0.37) \\
&= \mathcal{C}(0.775 \times 0.095, 0.195 \times 0.535, 0.03 \times 0.37) \\
&= \mathcal{C}(0.074, 0.104, 0.011) \\
&= \Big(\frac{0.074}{0.074 + 0.104 + 0.011}, \frac{0.104}{0.074 + 0.104 + 0.011}, \frac{0.011}{0.074 + 0.104 + 0.011}\Big) \\
&= \Big(\frac{0.074}{0.189}, \frac{0.104}{0.189}, \frac{0.011}{0.189}\Big) \\
&= (0.392, 0.55, 0.058)
\end{aligned}$$

Em geral, como se pode observar no exemplo 2.3.1, duas composições no simplex, quando aplicando a perturbação, resulta numa composição no mesmo espaço. No exemplo, $0.392 + 0.55 + 0.058 = 1$.

Definição 2.3.2 (Potenciação)**.** *Sejam $\boldsymbol{x} = (x_1, \ldots, x_d)$ uma composição em $\mathcal{S}^d$ e α uma constante em $\mathbb{R}$. A potenciação de $\boldsymbol{x}$ por α é uma composição definida como:*

$$\alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, \ldots, x_d^\alpha) \tag{2.6}$$

onde $\mathcal{C}(.)$ refere-se à operação de fecho.

Exemplo 2.3.2. *Seja $\boldsymbol{y} = \mathcal{C}(0.095, 0.535, 0.37)$ uma composição e $\alpha = 3$ um escalar. A potenciação de $\boldsymbol{x}$ por α é dada por:*

$$\begin{aligned}
\alpha \odot \boldsymbol{y} &= (0.095^3, 0.535^3, 0.37^3) \\
&= \mathcal{C}(0.0009, 0.1531, 0.0507) \\
&= \Big(\frac{0.0009}{0.0009 + 0.1531 + 0.0507}, \frac{0.1531}{0.0009 + 0.1531 + 0.0507}, \frac{0.0507}{0.0009 + 0.1531 + 0.0507}\Big) \\
&= \Big(\frac{0.0009}{0.2047}, \frac{0.1531}{0.2047}, \frac{0.0507}{0.2047}\Big) \\
&= (0.0044, 0.7479, 0.2477)
\end{aligned}$$

Em geral, como se pode observar no exemplo 2.3.2, a potenciação de uma composição no simplex resulta numa composição no mesmo espaço. No exemplo $0.0044 + 0.7479 + 0.2477 = 1$.

As operações de perturbação e potenciação são, na verdade, suficientes para definir no simplex um espaço vetorial e, consequentemente, verificar-se-ão as propriedades usuais (comutatividade, associatividade, distributividade).

Teorema 2.3.1. *$(\mathcal{S}^d, \oplus, \odot)$ é um espaço vetorial.*

Demonstração. $(\mathcal{S}^d, \oplus)$ tem a estrutura de grupo comutativo, isto é, para $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in \mathcal{S}^d$, as seguintes propriedades verificam-se:

1. propriedade comutativa: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$
   Tome-se $\mathbf{x} = (x_1, \ldots, x_d)$, $\mathbf{y} = (y_1, \ldots, y_d)$. Então:

$$\begin{aligned}
\mathbf{x} \oplus \mathbf{y} &= (x_1, \ldots, x_d) \oplus (y_1, \ldots, y_d) \\
&= \mathcal{C}(x_1 y_1, \ldots, x_d y_d) \\
&= \mathcal{C}(y_1 x_1, \ldots, y_d x_d) \\
&= \mathbf{y} \oplus \mathbf{x}
\end{aligned}$$

2. propriedade associativa: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$
   Tome-se $\mathbf{x} = (x_1, \ldots, x_d)$, $\mathbf{y} = (y_1, \ldots, y_d)$, $\mathbf{z} = (z_1, \ldots, z_d)$. Então:

$$\begin{aligned}
(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} &= \Big((x_1, \ldots, x_d) \oplus (y_1, \ldots, y_d)\Big) \oplus (z_1, \ldots, z_d) \\
&= \mathcal{C}(x_1 y_1, \ldots, x_d y_d) \oplus (z_1, \ldots, z_d) \\
&= (x_1 y_1, \ldots, x_d y_d) \oplus (z_1, \ldots, z_d) \\
&= \mathcal{C}\Big((x_1 y_1) z_1, \ldots, (x_d y_d) z_d\Big) \\
&= \mathcal{C}\Big(x_1 (y_1 z_1), \ldots, x_d (y_d z_d)\Big) \\
&= (x_1, \ldots, x_d) \oplus \mathcal{C}(y_1 z_1, \ldots, y_d z_d) \\
&= (x_1, \ldots, x_d) \oplus \Big((y_1, \ldots, y_d) \oplus (z_1, \ldots, z_d)\Big) \\
&= \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})
\end{aligned}$$

3. existência de elemento neutro: **n**, definido como:

$$\mathbf{n} = \mathcal{C}(1, \ldots, 1) = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$$

que verifica a igualdade $\mathbf{n} \oplus \mathbf{x} = \mathbf{x} \oplus \mathbf{n} = \mathbf{x}$:

$$\begin{aligned}
\mathbf{n} \oplus \mathbf{x} &= \left(\frac{1}{d}, \ldots, \frac{1}{d}\right) \oplus (x_1, \ldots, x_d) \\
&= \mathcal{C}\left(\frac{x_1}{d}, \ldots, \frac{x_d}{d}\right) \\
&= \left(\frac{x_1}{d}, \ldots, \frac{x_d}{d}\right) \\
&= \frac{1}{d}(x_1, \ldots, x_d) \\
&= \mathcal{C}(x_1, \ldots, x_d) \\
&= \mathbf{x}
\end{aligned}$$

onde **n** corresponde ao baricentro do simplex e é único, visto que a operação $\oplus$ é comutativa. A igualdade $\mathbf{x} = \mathbf{x} \oplus \mathbf{n}$ verifica-se de modo análogo.

4. existência do inverso de um elemento **x**, $\mathbf{x}^{-1}$:
Tome-se $\mathbf{x}^{-1}$ como a composição que resulta da operação de potenciação da composição **x** pelo escalar $\alpha = -1$. Desse modo, tem-se:

$$\begin{aligned}
\alpha \odot \mathbf{x} &= (-1) \odot \mathbf{x} \\
&= \mathcal{C}(x_1^{-1}, \ldots, x_d^{-1})
\end{aligned}$$

Além disso, pela propriedade do elemento neutro, tem-se:

$$\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n} = \mathbf{x}^{-1} \oplus \mathbf{x}$$

Na verdade, tem-se:

$$\begin{aligned}
\mathbf{x} \oplus \mathbf{x}^{-1} &= (x_1, \ldots, x_d) \oplus (x_1^{-1}, \ldots, x_d^{-1}) \\
&= (x_1, \ldots, x_d) \oplus (\frac{1}{x_1}, \ldots, \frac{1}{x_d}) \\
&= \mathcal{C}(x_1 \frac{1}{x_1}, \ldots, x_d \frac{1}{x_d}) \\
&= \mathcal{C}(\frac{x_1}{x_1}, \ldots, \frac{x_d}{x_d}) \\
&= \mathcal{C}(1, \ldots, 1) \\
&= \mathbf{n}
\end{aligned}$$

A igualdade $\mathbf{x}^{-1} \oplus \mathbf{x} = \mathbf{n}$ verifica-se de modo análogo.

A operação de potenciação satisfaz as propriedades de um produto externo. Para **x**, $\mathbf{y} \in \mathcal{S}^d$ e $\alpha, \beta \in \mathbb{R}$, verificam-se as seguintes propriedades:

1. propriedade associativa: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \odot \beta) \odot \mathbf{x}$
   Tome-se $\mathbf{x} = (x_1, \ldots, x_d)$. Então:

$$\begin{aligned}
\alpha \odot (\beta \odot \mathbf{x}) &= \alpha \odot \mathcal{C}(x_1^\beta, \ldots, x_d^\beta) \\
&= \alpha \odot \mathcal{C}\Big(\frac{x_1^\beta}{\sum x_i^\beta}, \ldots, \frac{x_d^\beta}{\sum x_i^\beta}\Big) \\
&= \mathcal{C}\left(\frac{x_1^{\beta\alpha}}{(\sum x_i^\beta)^\alpha}, \ldots, \frac{x_d^{\beta\alpha}}{(\sum x_i^\beta)^\alpha} \right) \\
&= \left(\frac{x_1^{\beta\alpha}}{(\sum x_i^\beta)^\alpha}, \ldots, \frac{x_d^{\beta\alpha}}{(\sum x_i^\beta)^\alpha} \right) \\
&= \frac{1}{(\sum x_i^\beta)^\alpha} \left(x_1^{\beta\alpha}, \ldots, x_d^{\beta\alpha} \right) \\
&= \mathcal{C}(x_1^{\alpha\beta}, \ldots, x_d^{\alpha\beta}) \\
&= (\alpha \odot \beta) \odot (x_1, \ldots, x_d) \\
&= (\alpha \odot \beta) \odot \mathbf{x}
\end{aligned}$$

2. propriedade distributiva em relação à perturbação: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$
   Tome-se $\mathbf{x} = (x_1, \ldots, x_d)$, $\mathbf{y} = (y_1, \ldots, y_d)$. Então:

$$\begin{aligned}
\alpha \odot (\mathbf{x} \oplus \mathbf{y}) &= \alpha \odot \Big((x_1, \ldots, x_d) \oplus (y_1, \ldots, y_d)\Big) \\
&= \alpha \odot \mathcal{C}(x_1 y_1, \ldots, x_d y_d) \\
&= \alpha \odot \Big(\frac{x_1 y_1}{\sum x_i y_i}, \ldots, \frac{x_d y_d}{\sum x_i y_i}\Big) \\
&= \Big(\frac{x_1 y_1}{\sum x_i y_i}^{\alpha}, \ldots, \frac{x_d y_d}{\sum x_i y_i}^{\alpha}\Big) \\
&= \mathcal{C}\Big((x_1 y_1)^\alpha, \ldots, (x_d y_d)^\alpha\Big) \\
&= \mathcal{C}\Big(\frac{(x_1 y_1)^\alpha}{\sum (x_i y_i)^\alpha \frac{\sum (x_i y_i)^\alpha}{\sum (x_i y_i)^\alpha}}, \ldots, \frac{(x_d y_d)^\alpha}{\sum (x_i y_i)^\alpha \frac{\sum (x_i y_i)^\alpha}{\sum (x_i y_i)^\alpha}}\Big) \\
&= \mathcal{C}(x_1^\alpha y_1^\alpha, \ldots, x_d^\alpha y_d^\alpha) \\
&= (x_1^\alpha, \ldots, x_d^\alpha) \oplus (y_1^\alpha, \ldots, y_d^\alpha), \quad \text{por definição de } \oplus \\
&= \Big(\alpha \odot (x_1, \ldots, x_d)\Big) \oplus \Big(\alpha \odot (y_1, \ldots, y_d)\Big), \quad \text{por definição de } \odot \\
&= (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})
\end{aligned}$$

3. propriedade distributiva em relação à adição: $(\alpha+\beta)\odot\mathbf{x}=(\alpha\odot\mathbf{x})\oplus(\beta\odot\mathbf{x})$
   Tome-se $\mathbf{x}=(x_1,\ldots,x_d)$. Então:

$$\begin{aligned}(\alpha+\beta)\odot\mathbf{x}&=(\alpha+\beta)\odot(x_1,\ldots,x_d)\\&=\mathcal{C}(x_1^{\alpha+\beta},\ldots,x_d^{\alpha+\beta})\\&=\mathcal{C}(x_1^{\alpha}x_1^{\beta},\ldots,x_d^{\alpha}x_d^{\beta})\\&=\mathcal{C}(x_1^{\alpha},\ldots,x_d^{\alpha})\oplus\mathcal{C}(x_1^{\beta},\ldots,x_d^{\beta})\\&=\Big(\alpha\odot(x_1,\ldots,x_d)\Big)\oplus\Big(\beta\odot(x_1,\ldots,x_d)\Big)\\&=(\alpha\odot\mathbf{x})\oplus(\beta\odot\mathbf{x})\end{aligned}$$

4. existência de elemento neutro: $1\odot\mathbf{x}=\mathbf{x}$.
   O elemento neutro é único e corresponde a $\alpha=1$, já que:

$$1\odot\mathbf{x}=\mathcal{C}(x_1^1,\ldots,x_d^1)=\mathbf{x}$$

■

Uma estrutura de espaço vetorial Euclidiano pode ser obtido ao definir o conceito de norma, produto interno e de distância entre composições do simplex $\mathcal{S}^d$. Apresenta-se, de seguida, essas definições no sentido de Aitchison [1].

Definição 2.3.3 (Produto Interno de Aitchison)**.** *Sejam $\boldsymbol{x}=(x_1,\ldots,x_d)$ e $\boldsymbol{y}=(y_1,\ldots,y_d)$ duas composições em $\mathcal{S}^d$. O produto interno de Aitchison é definido como:*

$$\langle\boldsymbol{x},\boldsymbol{y}\rangle_A=\frac{1}{2d}\sum_{i=1}^{d}\sum_{j=1}^{d}\ln\frac{x_i}{x_j}\ln\frac{y_i}{y_j} \tag{2.7}$$

Exemplo 2.3.3. *Sejam $\boldsymbol{x}=(0.775,0.195,0.03)$ e $\boldsymbol{y}=(0.095,0.535,0.37)$ duas composições em $\mathcal{S}^3$. O produto interno de Aitchison de $\boldsymbol{x}$ por $\boldsymbol{y}$ é dado por:*

$$\begin{aligned}\langle\boldsymbol{x},\boldsymbol{y}\rangle_A&=\frac{1}{2\times3}\sum_{i=1}^{3}\sum_{j=1}^{3}\ln\frac{x_i}{x_j}\ln\frac{y_i}{y_j}\\&=\frac{1}{6}\times(-12.231)\\&=-2.039\end{aligned}$$

Definição 2.3.4 (Norma de Aitchison)**.** *Seja $\boldsymbol{x}=(x_1,\ldots,x_d)$ uma composição em $\mathcal{S}^d$. A norma de Aitchison é definida como a raiz quadrada do produto interno de $\boldsymbol{x}$ por si mesmo, isto é:*

$$\|\boldsymbol{x}\|_A=\sqrt{\langle\boldsymbol{x},\boldsymbol{x}\rangle_A}=\sqrt{\frac{1}{2d}\sum_{i=1}^{d}\sum_{j=1}^{d}\Big(\ln\frac{x_i}{x_j}\Big)^2} \tag{2.8}$$

Exemplo 2.3.4. *Sejam* $\boldsymbol{x} = (0.775, 0.195, 0.03)$ *e* $\boldsymbol{y} = (0.095, 0.535, 0.37)$ *duas composições em* $\mathcal{S}^3$*. A norma de Aitchison de* $\boldsymbol{x}$ *e de* $\boldsymbol{y}$ *é dada por:*

$$\begin{aligned}\|\boldsymbol{x}\|_A &= \sqrt{\frac{1}{2\times 3}\sum_{i=1}^{3}\sum_{j=1}^{3}\Big(\ln\frac{x_i}{x_j}\Big)^2}\\ &= \sqrt{\frac{1}{6}\times 31.962}\\ &= 2.308\end{aligned}$$

$$\begin{aligned}\|\boldsymbol{y}\|_A &= \sqrt{\frac{1}{2\times 3}\sum_{i=1}^{3}\sum_{j=1}^{3}\Big(\ln\frac{y_i}{y_j}\Big)^2}\\ &= \sqrt{\frac{1}{6}\times 9.944}\\ &= 1.287\end{aligned}$$

Analogamente às operações usuais no espaço real, é também possível realizar a operação de perturbação com o inverso de uma composição resulta numa nova operação, designada por diferença da perturbação e denotada com o sinal $\ominus$. Concretamente:

$$\begin{aligned}\mathbf{x}\oplus\mathbf{y}^{-1} &= \mathbf{x}\oplus\Big((-1)\odot\mathbf{y}\Big)\\ &= (x_1,\ldots,x_d)\oplus\mathcal{C}(y_1^{-1},\ldots,y_d^{-1})\\ &= (x_1,\ldots,x_d)\oplus\mathcal{C}\Big(\frac{1}{y_1},\ldots,\frac{1}{y_d}\Big)\\ &= (x_1,\ldots,x_d)\oplus\left(\frac{\frac{1}{y_1}}{\sum_{i=1}^{d}\frac{1}{y_i}},\ldots,\frac{\frac{1}{y_d}}{\sum_{i=1}^{d}\frac{1}{y_i}}\right)\\ &= \left(Big(\frac{x_1\frac{1}{y_1}}{\sum\frac{1}{y_i}\frac{\sum\frac{x_i}{y_i}}{\sum\frac{1}{y_i}}},\ldots,\frac{x_d\frac{1}{y_d}}{\sum\frac{1}{y_i}\frac{\sum\frac{x_i}{y_i}}{\sum\frac{1}{y_i}}}\right)\\ &= \Big(\frac{x_1\frac{1}{y_1}}{\sum\frac{x_i}{y_i}},\ldots,\frac{x_d\frac{1}{y_d}}{\sum\frac{x_i}{y_i}}\Big)\\ &= \mathcal{C}\Big(\frac{x_1}{y_1},\ldots,\frac{x_d}{y_d}\Big)\\ &= (x_1,\ldots,x_d)\ominus(y_1,\ldots,y_d)\\ &= \mathbf{x}\ominus\mathbf{y}\end{aligned}$$

No caso particular em que $\mathbf{x} = \mathbf{y}$ e com a diferença de perturbação, prova-se que o

elemento neutro corresponde à perturbação de uma composição com o seu inverso:

$$\begin{aligned}\mathbf{x} \oplus \mathbf{x}^{-1} &= \mathbf{x} \oplus \Big((-1) \odot \mathbf{x}\Big) \\ &= (x_1, \ldots, x_d) \oplus (x_1^{-1}, \ldots, x_d^{-1}) \\ &= \mathcal{C}\Big(\frac{x_1}{x_1}, \ldots, \frac{x_d}{x_d}\Big) \\ &= \mathcal{C}(1, \ldots, 1) \\ &= \mathbf{n}\end{aligned}$$

Definição 2.3.5 (Distância de Aitchison)**.** *Sejam $\boldsymbol{x} = (x_1, \ldots, x_d)$ e $\boldsymbol{y} = (y_1, \ldots, y_d)$ duas composições em $\mathcal{S}^d$. A distância entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é definida como:*

$$d(\boldsymbol{x}, \boldsymbol{y})_A = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_A = \sqrt{\frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \Big(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\Big)^2} \tag{2.9}$$

Exemplo 2.3.5. *Sejam $\boldsymbol{x} = (0.775, 0.195, 0.03)$ e $\boldsymbol{y} = (0.095, 0.535, 0.37)$ duas composições em $\mathcal{S}^3$. A distância entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é dada por:*

$$\begin{aligned}d(\boldsymbol{x}, \boldsymbol{y})_A &= \sqrt{\frac{1}{2 \times 3} \sum_{i=1}^{3} \sum_{j=1}^{3} \Big(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\Big)^2} \\ &= \sqrt{\frac{1}{6} \times 66.369} \\ &= 3.326\end{aligned}$$

Estes conceitos permitem estabelecer uma estrutura de espaço vetorial linear Euclidiano que é designada por Geometria de Aitchison. Estas definições são baseadas em logaritmos dos rácios (em inglês *logratios*) entre as partes dos dados composicionais [1].

Uma das consequências do uso de *logratios* é que, como a soma das partes composicionais é irrelevante, qualquer composição na mesma classe de equivalência origina o mesmo *logratio*, isto é, o *logratio* da composição $\mathbf{x} = (x_1, \ldots, x_d)$ e de todas as que se encontram na mesma classe, $\mathbf{x}_\lambda = \lambda\mathbf{x} = (\lambda x_1, \ldots, \lambda x_d), \forall \lambda > 0$, é igual, uma vez que:

$$\begin{aligned}\ln \frac{\lambda x_i}{\lambda x_j} &= (\ln \lambda + \ln x_i) - (\ln \lambda + \ln x_j) \\ &= \ln \lambda + \ln x_i - \ln \lambda - \ln x_j \\ &= \ln \frac{x_i}{x_j}\end{aligned} \tag{2.10}$$

De modo análogo, e uma vez que se usa as log-razões no cálculo do produto interno, obtém-se:

$$\langle \mathbf{x}, \mathbf{x}_\lambda \rangle_A = \langle \mathbf{x}, \mathbf{x} \rangle_A = \|\mathbf{x}\|_A^2 \tag{2.11}$$

$$d(\mathbf{x}, \mathbf{x}_\lambda)_A = d(\mathbf{x}, \mathbf{x})_A \tag{2.12}$$

Para a equação (2.11) verifica-se que:

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{x}_\lambda \rangle_A &= \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \ln \frac{x_i}{x_j} \ln \frac{\lambda x_i}{\lambda x_j} \\
&= \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \ln \frac{x_i}{x_j} \ln \frac{x_i}{x_j}, \quad \text{pela equação (2.10)} \\
&= \langle \mathbf{x}, \mathbf{x} \rangle_A \\
&= \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \ln \frac{x_i}{x_j} \ln \frac{x_i}{x_j} \\
&= \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\ln \frac{x_i}{x_j} \right)^2 \\
&= \left(\sqrt{\tfrac{1}{2d} \textstyle\sum_{i=1}^{d} \sum_{j=1}^{d} \left(\ln \frac{x_i}{x_j} \right)^2} \right)^2 \\
&= \|\mathbf{x}\|_A^2
\end{aligned}$$

Para a equação (2.12) tem-se:

$$\begin{aligned}
d(\mathbf{x}, \mathbf{x}_\lambda)_A &= \sqrt{\frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\ln \frac{x_i}{x_j} - \ln \frac{\lambda x_i}{\lambda x_j} \right)^2} \\
&= \sqrt{\frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\ln \frac{x_i}{x_j} - \ln \frac{x_i}{x_j} \right)^2}, \quad \text{pela equação (2.10)} \\
&= d(\mathbf{x}, \mathbf{x})_A = 0
\end{aligned}$$

Tal como acontece com os conceitos métricos, os resultados das operações de perturbação e potenciação também não dependem da constante das composições [1]. Além disso, a estrutura geométrica de $\mathcal{S}^d$ satisfaz as propriedades usuais, tal como a compatibilidade da distância com a perturbação e potenciação, isto é, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^d$ e $\lambda \in \mathbb{R}$, tem-se:

$$d(\mathbf{x} \oplus \mathbf{z}, \mathbf{y} \oplus \mathbf{z})_A = d(\mathbf{x}, \mathbf{y})_A \tag{2.13}$$

$$d(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y})_A = |\alpha| \; d(\mathbf{x}, \mathbf{y})_A \tag{2.14}$$

Além da compatibilidade da distância, outras propriedades típicas de um espaço Euclidiano são igualmente válidas no simplex, $\mathcal{S}^d$ [25].

1. Desigualdade de Cauchy-Schwartz:

$$|\langle \mathbf{x}, \mathbf{y} \rangle_A| \leq \|\mathbf{x}\|_A \cdot \|\mathbf{y}\|_A \tag{2.15}$$

2. Teorema de Pitágoras, se $\mathbf{x}, \mathbf{y}$ são ortogonais, isto é, $\langle \mathbf{x}, \mathbf{y} \rangle_A = 0$, então:

$$\|\mathbf{x} \ominus \mathbf{y}\|_A^2 = \|\mathbf{x}\|_A^2 + \|\mathbf{y}\|_A^2 \tag{2.16}$$

3. Desigualdade Triangular:

$$d(\mathbf{x}, \mathbf{y})_A \leq d(\mathbf{x}, \mathbf{z})_A + d(\mathbf{z}, \mathbf{x})_A \tag{2.17}$$

Tal como ocorre na geometria Euclidiana, usando a norma e o produto interno de Aitchison, é possível estabelecer o ângulo α entre dois vetores composicionais. Assim, o ângulo no simplex entre as composições, $\mathbf{x}$ e $\mathbf{y}$, é dado à custa da seguinte relação [5]:

$$\cos \alpha = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_A}{\|\mathbf{x}\|_A \cdot \|\mathbf{y}\|_A} \tag{2.18}$$

Exemplo 2.3.6. *Usando os valores do produto interno, calculado no exemplo 2.3.3, e das normas, calculadas em 2.3.4, para as composições $\boldsymbol{x}$ e $\boldsymbol{y}$, tem-se:*

$$\cos \alpha = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A}{\|\boldsymbol{x}\|_A \cdot \|\boldsymbol{y}\|_A} = \frac{-2.039}{2.308 \times 1.287} = -0.686$$

$$\textit{Logo}, \alpha = \arccos(-0.686) = 133.31^\circ$$

2.4 Transformações de Dados Composicionais

Frequentemente, a análise de dados composicionais é associada com a aplicação de uma transformação apropriada e, de seguida, a utilização dos métodos estatísticos convencionais como se dados multivariados comuns se tratassem. Apesar de, numa perspetiva prática, ser verdade muitas vezes, a dificuldade com este tipo de pensamento é a interpretação de resultados. Após a aplicação de uma transformação, não se está a trabalhar mais com as composições originais, mas sim com as suas transformações e, por isso, a interpretação dos resultados deve levar isso em consideração [5], levando em conta o significado e o propósito da transformação [1].

Uma transformação pode também ser vista como a representação de uma composição num sistema de coordenadas, com respeito à geometria de Aitchison. Assim, resta estabelecer o que significa uma transformação em termos de representação em coordenadas [1]. A transformação a ser aplicada aos dados composicionais tem por base a análise estatística das log-razões (do inglês *Logratio Analysis*). Esta abordagem surgiu da importância reconhecida do princípio de invariância de escala, cuja aplicabilidade exige que se trabalhe com razões entre as componentes, que anula a constante de escala [4], demonstrada na equação 2.10. A transformação log-razão é uma correspondência biunívoca em $\mathbb{R}$ e o tratamento matemático de um quociente é mais fácil em termos do seu logaritmo e, por essa razão, Aitchison propôs a adoção de uma técnica de transformação envolvendo logaritmos de razões das componentes [4].

Uma dada composição $\mathbf{x} = (x_1, \ldots, x_d)$ pode originar diversas transformações das suas componentes. Assim, coloca-se a questão de qual a transformação log-razão que

se deve escolher. Posto isto, há a necessidade de introduzir o conceito de log-contraste (do inglês *logcontrast*), que pode ser considerado como uma combinação linear no simplex [5].

Definição 2.4.1 (Log-contraste). *Sejam $\boldsymbol{x} = (x_1, \ldots, x_d)$ uma composição em $\mathcal{S}^d$ e $\boldsymbol{a} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$, com $\alpha_1 + \cdots + \alpha_d = 0$, o vetor dos coeficientes, $\alpha_i \in \mathbb{R}$, $\forall i \in \{1, \ldots, d\}$. Um log-contraste de $\boldsymbol{x}$ é uma combinação linear definida como:*

$$\boldsymbol{a}^T \ln \boldsymbol{x} = \sum_{i=1}^{d} \alpha_i \ln x_i \tag{2.19}$$

O log-contraste é invariante quanto à escala, um dos princípios da análise composicional. Assim, $\forall k > 0$, tem-se:

$$\begin{aligned} \text{Log-razão:} \mathbb{R}^2 &\to \mathbb{R} \\ (x, y) &\to \ln \frac{x}{y}, y \neq 0, x, y > 0 \end{aligned}$$

$$\begin{aligned} \mathbf{a}^T \ln(k\mathbf{x}) &= \sum_{i=1}^{d} \alpha_i \ln(kx_i) \\ &= \sum_{i=1}^{d} \alpha_i \Big(\ln k + \ln x_i \Big) \\ &= \sum_{i=1}^{d} \alpha_i \underbrace{\ln k}_{\text{const.}} + \sum_{i=1}^{d} \alpha_i \ln x_i \\ &= \ln k \times \underbrace{\sum_{i=1}^{d} \alpha_i}_{=0} + \sum_{i=1}^{d} \alpha_i \ln x_i \\ &= \sum_{i=1}^{d} \alpha_i \ln x_i \\ &= \mathbf{a}^T \ln \mathbf{x} \end{aligned}$$

Com o intuito de evitar a ambiguidade dos log-contrastes, que pode acontecer, por exemplo, na comparação de log-contrastes, surge o conceito de log-contrastes ortogonais.

Definição 2.4.2 (Log-contrastes ortogonais). *Seja $\boldsymbol{x}$ uma composição em $\mathcal{S}^d$ e sejam $\boldsymbol{a} = (\alpha_1, \ldots, \alpha_d)$, $\boldsymbol{b} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ dois vetores de coeficientes, com $\sum_{i=1}^{d} \alpha_i = 0$ e $\sum_{i=1}^{d} \beta_i = 0$ com $\alpha_i, \beta_i \in \mathbb{R}$, $\forall i \in \{1, \ldots, d\}$.*
Dois log-contrastes, $\boldsymbol{a}^T \ln \boldsymbol{x}$ e $\boldsymbol{b}^T \ln \boldsymbol{x}$, são ortogonais se e só se $\boldsymbol{a}^T \boldsymbol{b} = 0$.

Geralmente, muitas das dificuldades relacionadas com a análise de dados composicionais podem ser ultrapassadas escolhendo o log-contraste apropriado. Essa escolha depende do problema e da interpretação da composição [4], pois, não há nenhuma regra que estabeleça, à partida, que uma transformação seja melhor que as outras. Todavia, é essencial que a transformação selecionada analise, de forma adequada, os dados de modo a que essa escolha apenas dependa do problema, da interpretação da composição e dos objetivos da análise [5].

Os dados composicionais induzem o seu próprio espaço amostral: eles são representados no simplex, com a geometria de Aitchison, que apresenta diferenças substanciais em relação à geometria Euclidiana usual. Por esse motivo, os métodos estatísticos *standard* utilizados na geometria Euclidiana não podem ser diretamente aplicados às composições [26]. Como solução para esse problema, existe uma família de transformações log-razões que permitem expressar os dados composicionais do simplex no espaço real Euclidiano [26], [5].

Em 1986, John Aitchison introduziu a primeira transformação de log-razão designada por log-razões aditiva, (do inglês *additive log-ratio*), abreviada de alr. Todavia, esta transformação não era isométrica, isto é, as distâncias entre as composições em coordenadas transformadas não são iguais às distâncias entre as composições originais. Assim, numa tentativa de ultrapassar esse inconveniente, foi introduzida uma nova transformação, baseada na média geométrica das partes das composições designada de transformação de log-razões centradas (do inglês *centered log-ratio*), abreviada de clr. Contudo, nem a transformação alr nem a clr podem ser diretamente associadas a um sistema de coordenadas ortogonal no simplex, o que levou a que, em 2003, a transformação de log-razões isométrica (do inglês *isometric log-ratio*), abreviada de ilr, fosse proposta por Egozcue e colaboradores [27]. A transformação ilr que é uma isometria entre $\mathcal{S}^d$ e $\mathbb{R}^{d-1}$, ultrapassando, desse modo, os problemas das outras transformações [26].

2.4.1 Transformação alr

Definição 2.4.3. *Seja $\boldsymbol{x}$ uma composição de d partes no simplex $\mathcal{S}^d$.*
Designa-se por transformação de log-razões aditivas de $\boldsymbol{x}$, relativamente à k-ésima componente e denota-se por $\text{alr}_k(\boldsymbol{x})$*, à transformação: $\boldsymbol{x} \in \mathcal{S}^d \to \boldsymbol{x}^{(k)} \in \mathbb{R}^{d-1}$, para algum $k \in \{1, \ldots, d\}$, definida por:*

$$\begin{aligned} \boldsymbol{x}^{(k)} &= \text{alr}_k(\boldsymbol{x}) \\ &= (x_1^{(k)}, \ldots, x_{d-1}^{(k)}) \\ &= \left(\ln \frac{x_1}{x_k}, \ldots, \ln \frac{x_{k-1}}{x_k}, \ln \frac{x_{k+1}}{x_k}, \ldots, \ln \frac{x_d}{x_k} \right) \end{aligned} \tag{2.20}$$

Exemplo 2.4.1. *Seja $\boldsymbol{x} = (11.4, 52.7, 35.9)$ uma composição. A transformação de log-razões*

aditivas de $\boldsymbol{x}$, $\text{alr}(\boldsymbol{x})$, *relativamente à 2ª componente é dada por:*

$$\begin{aligned}\boldsymbol{x}^{(2)} &= \text{alr}_2(\boldsymbol{x}) \\ &= (x_1^{(2)}, x_3^{(2)}) \\ &= \Big(\ln\frac{11.4}{52.7}, \ln\frac{35.9}{52.7}\Big) \\ &= (-1.53, -0.38)\end{aligned}$$

Na transformação $\mathbf{x}_k$, fixa-se uma das componentes do vetor como denominador da log-razão, neste caso a componente x_k. Contudo, qualquer uma das partes da composição pode ser a escolhida como a variável razão nas coordenadas. Assim, selecionando qualquer outra parte para denominador, obter-se-ia uma nova transformação alr. Neste caso em particular, uma vez que a composição $\mathbf{x}$ tem d partes, seria possível obter d diferentes transformações alr. Deste modo, a escolha da componente a utilizar depende do contexto, mas também da adequação dos resultados para a visualização e exploração de todos. Por exemplo, ao realizar um estudo sobre a diversidade dos solos em diversos pontos do país, o investigador pode ter interesse em conhecer as log-razões dos constituintes do solo em relação ao chumbo e, por isso, o chumbo ser a parte da composição do solo selecionada como razão.

No caso de uma matriz $\mathbf{X}_{(n\times d)}$ de dados composicionais com composições definidas por $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ nas linhas da matriz, para $i \in \{1, \ldots, n\}$, a matriz de coordenadas alr, relativamente à k-ésima componente dos dados é formada pelas linhas dadas por:

$$\begin{aligned}\mathbf{x}_i^{(k)} &= \text{alr}_k(\mathbf{x}_i) \\ &= \Big(\ln\frac{x_{i1}}{x_{ik}}, \ldots, \ln\frac{x_{i,k-1}}{x_{ik}}, \ln\frac{x_{i,k+1}}{x_{ik}}, \ldots, \ln\frac{x_{id}}{x_{ik}}\Big)\end{aligned} \tag{2.21}$$

De notar que a matriz resultante da transformação alr tem dimensão $n \times (d-1)$. Uma vez que $\ln(\frac{x_{ik}}{x_{ik}}) = 0$ e, pelo princípio da coerência subcomposicional, que diz que qualquer parte não informativa pode ser retirada sem que isso afete os resultados, uma vez que a coluna de zeros que seria obtida não traduzia, na prática, qualquer informação, ela é eliminada da matriz da transformação, tal como acontece com os vetores composicionais alr-transformados.

Nas coordenadas alr, as operações de perturbação e potenciação continuam a ser válidas, isto é, para $\mathbf{x}, \mathbf{y} \in \mathcal{S}^d$, $c \in \mathbb{R}$ e $\forall k \in \{1, \ldots, d\}$, tem-se:

$$\text{alr}_k(\mathbf{x} \oplus \mathbf{y}) = \text{alr}_k(\mathbf{x}) + \text{alr}_k(\mathbf{y}) \tag{2.22}$$

$$\text{alr}_k(c \odot \mathbf{x}) = c \cdot \text{alr}_k(\mathbf{x}) \tag{2.23}$$

Verifique-se para o caso $k = 3$.

Sejam $\mathbf{x}, \mathbf{y} \in \mathcal{S}^3$, $c \in \mathbb{R}$. Tome-se, sem perda de generalidade, como denominador a segunda componente.

- Perturbação:

$$\begin{aligned}
\text{alr}_2(\mathbf{x} \oplus \mathbf{y}) &= \text{alr}_2\Big((x_1, x_2, x_3) \oplus (y_1, y_2, y_3)\Big) \\
&= \text{alr}_2\, \mathcal{C}(x_1y_1, x_2y_2, x_3y_3) \\
&= \text{alr}_2\, \Big(\frac{x_1y_1}{x_1y_1 + x_2y_2 + x_3y_3}, \frac{x_2y_2}{x_1y_1 + x_2y_2 + x_3y_3}, \frac{x_3y_3}{x_1y_1 + x_2y_2 + x_3y_3}\Big) \\
&= \Big(\ln \frac{\frac{x_1y_1}{x_1y_1+x_2y_2+x_3y_3}}{\frac{x_2y_2}{x_1y_1+x_2y_2+x_3y_3}}, \ln \frac{\frac{x_3y_3}{x_1y_1+x_2y_2+x_3y_3}}{\frac{x_2y_2}{x_1y_1+x_2y_2+x_3y_3}}\Big) \\
&= \Big(\ln \frac{x_1y_1}{x_2y_2}, \ln \frac{x_3y_3}{x_2y_2}\Big) \\
&= \Big(\ln \frac{x_1}{x_2} + \ln \frac{y_1}{y_2}, \ln \frac{x_3}{x_2} + \ln \frac{y_3}{y_2}\Big) \\
&= \Big(\ln \frac{x_1}{x_2}, \ln \frac{x_3}{x_2}\Big) + \Big(\ln \frac{y_1}{y_2}, \ln \frac{y_3}{y_2}\Big) \\
&= \text{alr}_2(\mathbf{x}) + \text{alr}_2(\mathbf{y})
\end{aligned}$$

- Potenciação:

$$\begin{aligned}
\text{alr}_2(c \odot \mathbf{x}) &= \text{alr}_2\, \mathcal{C}(x_1^c, x_2^c, x_3^c) \\
&= \text{alr}_2\Big(\frac{x_1^c}{x_1^c + x_2^c + x_3^c}, \frac{x_2^c}{x_1^c + x_2^c + x_3^c}, \frac{x_3^c}{x_1^c + x_2^c + x_3^c}\Big) \\
&= \Big(\ln \frac{\frac{x_1^c}{x_1^c+x_2^c+x_3^c}}{\frac{x_2^c}{x_1^c+x_2^c+x_3^c}}, \ln \frac{\frac{x_3^c}{x_1^c+x_2^c+x_3^c}}{\frac{x_2^c}{x_1^c+x_2^c+x_3^c}}\Big) \\
&= \Big(\ln \frac{x_1^c}{x_2^c}, \ln \frac{x_3^c}{x_2^c},\Big) \\
&= \Big(\ln \Big(\frac{x_1}{x_2}\Big)^c, \ln \Big(\frac{x_3}{x_2}\Big)^c\Big) \\
&= \Big(c \ln \frac{x_1}{x_2}, c \ln \frac{x_3}{x_2},\Big) \\
&= c \cdot \Big(\ln \frac{x_1}{x_2}, \ln \frac{x_3}{x_2}\Big) \\
&= c \cdot \text{alr}_2(\mathbf{x})
\end{aligned}$$

Contudo, o produto interno, a norma e a distância de Aitchison não preservam as suas propriedades com esta transformação, por exemplo, $\langle \mathbf{x}, \mathbf{y} \rangle_A \neq \langle \text{alr}_k(\mathbf{x}), \text{alr}_k(\mathbf{y}) \rangle_A$. Além disso, a interpretação das coordenadas alr conduz a interpretações que podem ser imprecisas em termos das suas coordenadas originais, uma vez que a interpretação de uma coordenada não pode ser feita em termos de uma parte, pois, para cada componente de $x^{(k)}$, a informação que se obtém é apenas em relação à parte x_k, mas desconhece-se a log-razão entre as demais partes [1].

Uma vez que as coordenadas alr são obtidas uma a uma, é possível regressar à composição original. Para tal, define-se a inversa da transformação alr, alr^{-1}. Fixe-se

x_d como denominador da log-razão. Assim, a inversa da transformação alr_d, denotada por alr_d^{-1}, é definida por $\text{alr}^{-1} : \mathbf{x}^{(d)} \in \mathbb{R}^{d-1} \to \mathbf{x} \in \mathcal{S}^d$.

$$\text{alr}_2(\mathbf{x}) = \Big(\underbrace{\ln \frac{x_1}{x_2}}_{y_1}, \underbrace{\ln \frac{x_3}{x_2}}_{y_3}\Big)$$

Então:

$$y_1 = \ln \frac{x_1}{x_2} \Leftrightarrow \exp y_1 = \frac{x_1}{x_2} \Leftrightarrow x_1 = x_2 \exp y_1$$
$$y_3 = \ln \frac{x_3}{x_2} \Leftrightarrow \exp y_3 = \frac{x_3}{x_2} \Leftrightarrow x_3 = x_2 \exp y_3$$

$$x_1 + x_2 + x_3 = 1 \Leftrightarrow x_2(\exp y_1 + 1 + \exp y_3) = 1 \Leftrightarrow x_2 = \frac{1}{\exp y_1 + 1 + \exp y_3}$$

Deste modo, generalizando para todas as partes e simplificando a notação, tem-se:

- para $j \in \{1, \ldots, d\}$, $j \neq k$:

$$x_j = \frac{\exp(x_j^{(k)})}{\exp(x_1^{(k)}) + \cdots + \exp(x_{k-1}^{(k)}) + 1 + \exp(x_{k+1}^{(k)}), \cdots + \exp(x_d^{(k)})}$$

- para $j = k$:

$$x_j = \frac{1}{\exp(x_1^{(k)}) + \cdots + \exp(x_{k-1}^{(k)}) + 1 + \exp(x_{k+1}^{(k)}), \cdots + \exp(x_d^{(k)})}$$

Apesar da simplicidade da transformação alr, as suas principais desvantagens são a sua subjetividade na escolha da variável razão, pois cada escolha dá origem a diferentes composições para a mesma composição original, e o facto desta transformação originar um sistema de coordenadas não ortogonal. Por este motivo, esta transformação não satisfaz o princípio de invariância de permutação e, por isso, a análise de dados composicionais, através deste tipo de transformação, pode levar a conclusões pouco adequadas [5].

2.4.2 Transformação clr

Como tentativa de ultrapassar as limitações que a transformação alr apresenta, Aitchison propôs a transformação de log-razões centradas, onde cada composição de d partes dá origem a um vetor de d coordenadas clr.

Definição 2.4.4. *Seja $\boldsymbol{x}$ uma composição de d partes no simplex $\mathcal{S}^d$.*
Designa-se por transformação de log-razões centradas de $\boldsymbol{x}$, $\text{clr}(\boldsymbol{x})$, a transformação clr*: $\boldsymbol{x} \in \mathcal{S}^d \to \boldsymbol{y} \in \mathbb{R}^d$, definida por:*

$$\begin{aligned} \boldsymbol{y} &= \text{clr}(\boldsymbol{x}) \\ &= \Big(\ln \frac{x_1}{\sqrt[d]{\prod_{j=1}^d x_j}}, \ldots, \ln \frac{x_d}{\sqrt[d]{\prod_{j=1}^d x_j}}\Big) \end{aligned} \tag{2.24}$$

Exemplo 2.4.2. *Seja* $\boldsymbol{x} = (11.4, 52.7, 35.9)$ *uma composição que representa a constituição da amostra a* 60.1 *m de profundidade do conjunto de dados 'Arctic Lake'.*
Uma vez que $\sqrt[3]{\prod_{j=1}^{3} x_j} = \sqrt[3]{11.4 \times 52.4 \times 35.9} = 24.84$*, então a transformação de log-razões centradas de* $\boldsymbol{x}$*,* $\text{clr}(\boldsymbol{x})$*, é dada por:*

$$\begin{aligned}\boldsymbol{y} &= \text{clr}(\boldsymbol{x}) \\ &= \text{clr}(11.4, 52.7, 35.9) \\ &= \Big(\ln \frac{11.4}{27.84}, \ln \frac{52.7}{27.84}, \ln \frac{35.9}{27.84}\Big) \\ &= (\ln 0.41, \ln 1.89, \ln 1.29) \\ &= (-0.89, 0.64, 0.25)\end{aligned}$$

Também é possível estabelecer a transformação inversa, clr^{-1}: $\mathbf{y} \in \mathbb{R}^d \to \mathbf{x} \in \mathcal{S}^d$ definida por:

$$\begin{aligned}\text{clr}^{-1}(\mathbf{y}) &= \mathcal{C}(\exp(y_1), \ldots, \exp(y_d)) \\ &= \Big(\frac{\exp y_1}{\exp y_1 + \cdots + \exp y_d}, \ldots, \frac{\exp y_d}{\exp y_1 + \cdots + \exp y_d}\Big)\end{aligned} \tag{2.25}$$

No caso de uma matriz $\mathbf{X}_{(n\times d)}$ de dados composicionais com as composições definidas por $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ nas linhas da matriz, para $i \in \{1, \ldots, n\}$, a matriz de coordenadas clr é formada pelas linhas dadas por:

$$\begin{aligned}\mathbf{y}_i &= \text{clr}(\mathbf{x}_i) \\ &= \Big(\ln \frac{x_{i1}}{\text{gm}(\mathbf{x}_i)}, \ldots, \ln \frac{x_{id}}{\text{gm}(\mathbf{x}_i)}\Big)\end{aligned} \tag{2.26}$$

onde $\text{gm}(\mathbf{x}_i)$ designada a média geométrica da composição $\mathbf{x}_i$, isto é:

$$\text{gm}(\mathbf{x}_i) = \sqrt[d]{\prod_{j=1}^{d} x_{ij}} \tag{2.27}$$

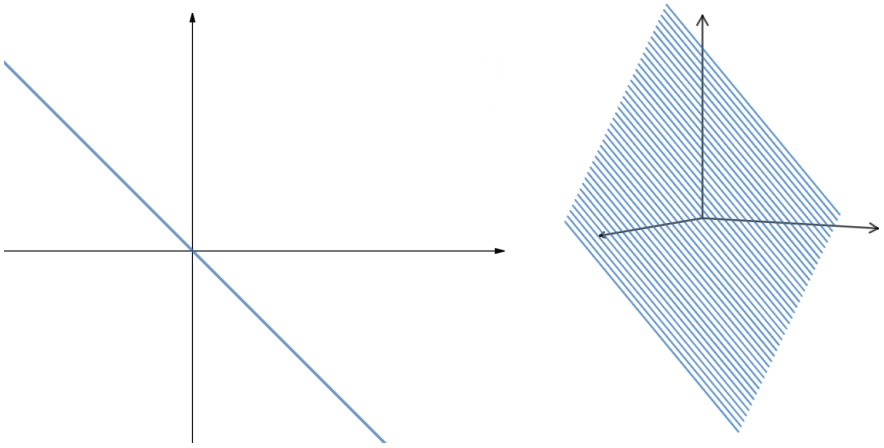
Numa matriz, cada linha, que corresponde a uma composição, é dividida pela média geométrica da própria composição. Além disso, quer a composição original, quer a composição resultante da aplicação da transformação clr, têm o mesmo número de partes e, por isso, a matriz resultante tem a mesma dimensão que a matriz original.

Assim, em comparação com a transformação alr, a clr não apresenta subjetividade na escolha da razão, uma vez que é sempre a média geométrica da composição, que trata cada componente simetricamente [1], uma vez que compara com uma quantidade global. Todavia, seria possível substituir o denominador por qualquer constante positiva, pois esta não unicidade é consistente com o conceito de composições como classes de equivalência [28]. Além disso, apresenta também vantagens em relação à transformação alr no que diz respeito à interpretabilidade, uma vez que se interpreta a log-razão de cada parte em relação a um todo.

Contudo, a transformação também apresenta algumas desvantagens. A principal é que a soma das partes da composição clr transformada é zero, isto é:

$$\begin{aligned}
\sum_{j=1}^{d} y_j &= \sum_{j=1}^{d} \ln \frac{x_j}{\mathrm{gm}(\mathbf{x})} \\
&= \sum_{j=1}^{d} \ln \frac{x_j}{\sqrt[d]{\prod_{j=1}^{d} x_j}} \\
&= \sum_{j=1}^{d} \ln x_j - \sum_{j=1}^{d} \ln \sqrt[d]{\prod_{j=1}^{d} x_j} \\
&= \sum_{j=1}^{d} \ln x_j - \sum_{j=1}^{d} \frac{1}{d} \ln \Big(\prod_{j=1}^{d} x_j \Big) \\
&= \sum_{j=1}^{d} \ln x_j - \frac{1}{d} \times d \times \ln \Big(\prod_{j=1}^{d} x_j \Big) \\
&= \sum_{j=1}^{d} \ln x_j - \ln(x_1 \times \cdots \times x_d) \\
&= \sum_{j=1}^{d} \ln x_j - (\ln x_1 + \cdots + \ln x_d) = 0
\end{aligned} \tag{2.28}$$

Esta particularidade da transformação clr é que os dados nesta representação são colineares e, por isso, a composição transformada estará contida no hiperplano $\{(y_1, \ldots, y_d) \in \mathbb{R}^d : y_1 + \cdots + y_d = 0\}$, o qual passa pela origem de $\mathbb{R}^d$ e é ortogonal ao vetor $(1, \ldots, 1)$.



(a) $d = 2$ (b) $d = 3$

Figura 2.3: Espaço das transformações clr numa composição com (a) 2 e com (b) 3 elementos.

Deste modo, se se pensar numa matriz de composições, ao expressar cada observa-

ção como dados clr-transformados, a matriz resultante não tem caraterística completa e a respetiva matriz de covariância será singular [1]. Esta propriedade da transformação clr é uma desvantagem no que a deteção de observações atípicas diz respeito, uma vez que impossibilita a obtenção da inversa da matriz de covariâncias e, consequentemente, o cálculo da distância de Mahalanobis, habitualmente usada.

Outra desvantagem da transformação clr é que o facto de se usar a média geométrica da composição como denominador da log-razão não garante que a média geométrica da composição coincida com a média geométrica das eventuais subcomposições e, por isso, não há garantias que satisfaça o princípio da coerência subcomposicional.

Exemplo 2.4.3. *Seja* $\boldsymbol{x} = (11.4, 52.7, 35.9)$ *uma composição. Tome-se a subcomposição constituída pela retirada da primeira parte da composição* $\boldsymbol{x}$, $\boldsymbol{x}_s = (52.7, 35.9)$.

- *média geométrica de* $\boldsymbol{x}$*:* $gm(\boldsymbol{x}) = (11.4 \times 52.7 \times 35.9)^{\frac{1}{3}} = 27.84$
- *média geométrica de* $\boldsymbol{x}_s$*:* $gm(\boldsymbol{x}_s) = (52.7 \times 35.9)^{\frac{1}{2}} = 43.50$

Tal como acontece na transformação alr, a transformação clr também permite reduzir as operações de perturbação e potenciação à soma e ao produto usuais em $\mathbb{R}^d$. Deste modo, se $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}^d$, $c \in \mathbb{R}$, tem-se:

$$\text{clr}(\mathbf{x}_1 \oplus \mathbf{x}_2) = \text{clr}(\mathbf{x}_1) + \text{clr}(\mathbf{x}_1) \tag{2.29}$$
$$\text{clr}(c \odot \mathbf{x}_1) = c \cdot \text{clr}(\mathbf{x}_1) \tag{2.30}$$

Com as coordenadas clr, é possível estabelecer uma estrutura de espaço métrico no simplex. Define-se:

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \langle \text{clr}(\mathbf{x}_1), \text{clr}(\mathbf{x}_2) \rangle \tag{2.31}$$
$$\|\mathbf{x}_1\|_A = \|\text{clr}(\mathbf{x}_1)\| \tag{2.32}$$
$$d(\mathbf{x}_1, \mathbf{x}_2)_A = d(\text{clr}(\mathbf{x}_1), \text{clr}(\mathbf{x}_2)) \tag{2.33}$$
$$\tag{2.34}$$

Uma vez que a definição de produto interno de Aitchison e distância de Aitchison se verificam para a transformação, os coeficientes representam uma isometria: todos os conceitos métricos do simplex mantém-se após aplicar a transformação clr.

2.4.3 Transformação ilr e Coordenadas Pivô

A terceira transformação é a ilr, proposta em [27]. A classe das coordenadas das log-razões isométricas forma uma base ortogonal no simplex, que é um dos pontos fulcrais na Geometria de Aitchison.

Definição 2.4.5 (Base Ortonormal no Simplex)**.** *Seja* $\mathcal{S}^d$ *um simplex de* d *partes. O conjunto de vetores* $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d-1}\}$, $\boldsymbol{e}_j \in \mathcal{S}^d$ *e* $j \in \{1, \ldots, d-1\}$, *é uma base ortonormal de* $\mathcal{S}^d$ *se:*

i. $\langle \boldsymbol{e}_j, \boldsymbol{e}_k \rangle_A = 0$, *para* $j \neq k$;

ii. $\|\boldsymbol{e}_j\|_A = 1$, $j \in \{1, \ldots, d-1\}$.

Assim, o vetor resultante da transformação ilr estará em $\mathbb{R}^{d-1}$ e uma implicação prática é que, deste modo, evita a singularidade da matriz de covariâncias que ocorre com as coordenadas clr, mas existe uma infinidade de possibilidades para definir um sistema de bases ortonormal.

Exemplo 2.4.4. *Sejam* $\boldsymbol{b}$ *e* $\boldsymbol{b}'$ *dois conjuntos de vetores pertencentes ao simplex* $\mathcal{S}^3$. *Seja* $\boldsymbol{b} = \{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ *uma base ortonormal de* $\mathcal{S}^3$ *com:*

- $\boldsymbol{e}_1 = \mathcal{C}\left(e^{\frac{2\sqrt{6}}{6}}, e^{-\frac{\sqrt{6}}{6}}, e^{-\frac{\sqrt{6}}{6}} \right) = \left(\frac{e^{\frac{2\sqrt{6}}{6}}}{a_1}, \frac{e^{-\frac{\sqrt{6}}{6}}}{a_1}, \frac{e^{-\frac{\sqrt{6}}{6}}}{a_1} \right)$
- $\boldsymbol{e}_2 = \mathcal{C}\left(e^{0}, e^{\frac{\sqrt{2}}{2}}, e^{-\frac{\sqrt{2}}{2}} \right) = \left(\frac{e^{0}}{a_2}, \frac{e^{\frac{\sqrt{2}}{2}}}{a_2}, \frac{e^{-\frac{\sqrt{2}}{2}}}{a_2} \right)$

onde $a_1 = e^{\frac{2\sqrt{6}}{6}} + e^{\frac{-\sqrt{6}}{6}} + e^{\frac{-\sqrt{6}}{6}}$ *e* $a_2 = 1 + e^{\frac{\sqrt{2}}{2}} + e^{\frac{-\sqrt{2}}{2}}$.

A média geométrica de $\boldsymbol{e}_1$ *e* $\boldsymbol{e}_2$ *são dadas, respetivamente, por:*

- $\mathrm{gm}(\boldsymbol{e}_1) = \sqrt[3]{\frac{e^{\frac{2\sqrt{6}}{6}}}{a_1} \times \frac{e^{-\frac{\sqrt{6}}{6}}}{a_1} \times \frac{e^{-\frac{\sqrt{6}}{6}}}{a_1}} = \frac{1}{a_1} \sqrt[3]{e^{\frac{2\sqrt{6}}{6}} \times e^{-\frac{\sqrt{6}}{6}} \times e^{-\frac{\sqrt{6}}{6}}} = \frac{1}{a_1}\sqrt{e^0} = \frac{1}{a_1}$
- $\mathrm{gm}(\boldsymbol{e}_2) = \sqrt[3]{\frac{e^{0}}{a_2} \times \frac{e^{\frac{\sqrt{2}}{2}}}{a_2} \times \frac{e^{-\frac{\sqrt{2}}{2}}}{a_2}} = \frac{1}{a_2} \sqrt[3]{e^{0} \times e^{\frac{\sqrt{2}}{2}} \times e^{-\frac{\sqrt{2}}{2}}} = \frac{1}{a_2}\sqrt{e^0} = \frac{1}{a_2}$

Seja $\boldsymbol{b}' = \{\boldsymbol{e}'_1, \boldsymbol{e}'_2\}$ *uma outra base ortonormal de* $\mathcal{S}^3$ *com:*

- $\boldsymbol{e}'_1 = \mathcal{C}\left(e^{-\frac{\sqrt{2}}{2}}, e^{0}, e^{\frac{\sqrt{2}}{2}} \right) = \left(\frac{e^{-\frac{\sqrt{2}}{2}}}{a'_1}, \frac{e^{0}}{a'_1}, \frac{e^{\frac{\sqrt{2}}{2}}}{a'_1} \right)$
- $\boldsymbol{e}'_2 = \mathcal{C}\left(e^{\frac{3}{6}}, e^{\frac{-\sqrt{18}}{6}}, e^{-\frac{3}{6}} \right) = \left(\frac{e^{\frac{3}{6}}}{a'_2}, \frac{e^{\frac{-\sqrt{18}}{6}}}{a'_2}, \frac{e^{\frac{3}{6}}}{a'_2} \right)$

A média geométrica de $\boldsymbol{e}'_1$ *e* $\boldsymbol{e}'_2$ *são dadas, respetivamente, por:*

- $\mathrm{gm}(\boldsymbol{e}'_1) = \sqrt[3]{\frac{e^{-\frac{\sqrt{2}}{2}}}{a'_1} \times \frac{e^{0}}{a'_1} \times \frac{e^{\frac{\sqrt{2}}{2}}}{a'_1}} = \frac{1}{a'_1} \sqrt[3]{e^{\frac{-\sqrt{2}}{2}} \times e^{0} \times e^{\frac{\sqrt{2}}{2}}} = \frac{1}{a'_1}\sqrt{e^0} = \frac{1}{a'_1}$
- $\mathrm{gm}(\boldsymbol{e}'_2) = \sqrt[3]{\frac{e^{\frac{3}{6}}}{a'_2} \times \frac{e^{\frac{-\sqrt{18}}{6}}}{a'_2} \times \frac{e^{\frac{3}{6}}}{a'_2}} = \frac{1}{a'_2} \sqrt[3]{e^{\frac{3}{6}} \times e^{\frac{-\sqrt{18}}{6}} \times e^{\frac{3}{6}}} = \frac{1}{a'_2}\sqrt{e^0} = \frac{1}{a'_2}$

onde $a'_1 = e^{\frac{-\sqrt{2}}{2}} + 1 + e^{\frac{\sqrt{2}}{2}}$ *e* $a'_2 = e^{\frac{3}{6}} + e^{\frac{-\sqrt{18}}{6}} + e^{\frac{3}{6}}$.

Em primeiro lugar, para a base $\mathbf{b}$*, pode-se notar que:*

$$\begin{aligned}
\operatorname{clr}(\boldsymbol{e}_1) &= \left(\ln \frac{\frac{e^{\frac{2\sqrt{6}}{6}}}{a_1}}{\operatorname{gm}(\boldsymbol{e}_1)}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a_1}}{\operatorname{gm}(\boldsymbol{e}_1)}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a_1}}{\operatorname{gm}(\boldsymbol{e}_1)} \right) \\
&= \left(\ln \frac{\frac{e^{\frac{2\sqrt{6}}{6}}}{a_1}}{\frac{1}{a_1}}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a_1}}{\frac{1}{a_1}}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a_1}}{\frac{1}{a_1}} \right) \\
&= \left(\ln e^{\frac{2\sqrt{6}}{6}}, \ln e^{-\frac{\sqrt{6}}{6}}, \ln e^{-\frac{\sqrt{6}}{6}} \right) \\
&= \left(\frac{2\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6} \right)
\end{aligned}$$

$$\begin{aligned}
\operatorname{clr}(\boldsymbol{e}_2) &= \left(\ln \frac{\frac{e^{0}}{a_2}}{\operatorname{gm}(\boldsymbol{e}_2)}, \ln \frac{\frac{e^{\frac{\sqrt{2}}{2}}}{a_2}}{\operatorname{gm}(\boldsymbol{e}_2)}, \ln \frac{\frac{e^{-\frac{\sqrt{2}}{2}}}{a_2}}{\operatorname{gm}(\boldsymbol{e}_2)} \right) \\
&= \left(\ln \frac{\frac{e^{0}}{a_2}}{\frac{1}{a_2}}, \ln \frac{\frac{e^{\frac{\sqrt{2}}{2}}}{a_2}}{\frac{1}{a_2}}, \ln \frac{-\frac{e^{-\frac{\sqrt{2}}{2}}}{a_2}}{\frac{1}{a_2}} \right) \\
&= \left(\ln e^{0}, \ln e^{\frac{\sqrt{2}}{2}}, \ln e^{-\frac{\sqrt{2}}{2}} \right) \\
&= \left(0, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right)
\end{aligned}$$

De modo análogo, para a base $\mathbf{b}'$*, pode-se notar que:*

$$\begin{aligned}
\operatorname{clr}(\boldsymbol{e}'_1) &= \left(\ln \frac{\frac{e^{-\frac{\sqrt{2}}{2}}}{a'_1}}{\operatorname{gm}(\boldsymbol{e}'_1)}, \ln \frac{\frac{e^{0}}{a'_1}}{\operatorname{gm}(\boldsymbol{e}'_1)}, \ln \frac{\frac{e^{\frac{\sqrt{2}}{2}}}{a'_1}}{\operatorname{gm}(\boldsymbol{e}'_1)} \right) \\
&= \left(\ln \frac{\frac{e^{-\frac{\sqrt{2}}{2}}}{a'_1}}{\frac{1}{a_1}}, \ln \frac{\frac{e^{0}}{a'_1}}{\frac{1}{a'_1}}, \ln \frac{\frac{e^{-\frac{\sqrt{2}}{2}}}{a_1}}{\frac{1}{a'_1}} \right) \\
&= \left(\ln e^{-\frac{\sqrt{2}}{2}}, \ln e^{0}, \ln e^{\frac{\sqrt{2}}{2}} \right) \\
&= \left(-\frac{\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2} \right)
\end{aligned}$$

$$\begin{aligned}
\operatorname{clr}(\boldsymbol{e}'_2) &= \left(\ln \frac{\frac{e^{\frac{3}{6}}}{a'_2}}{\operatorname{gm}(\boldsymbol{e}'_2)}, \ln \frac{\frac{e^{\frac{-\sqrt{18}}{6}}}{a'_2}}{\operatorname{gm}(\boldsymbol{e}'_2)}, \ln \frac{\frac{e^{\frac{3}{6}}}{a'_2}}{\operatorname{gm}(\boldsymbol{e}'_2)} \right) \\
&= \left(\ln \frac{\frac{e^{\frac{3}{6}}}{a_2}}{\frac{1}{a'_2}}, \ln \frac{\frac{e^{\frac{-\sqrt{18}}{6}}}{a_2}}{\frac{1}{a'_2}}, \ln \frac{\frac{e^{\frac{3}{6}}}{a_2}}{\frac{1}{a'_2}} \right) \\
&= \left(\ln e^{\frac{3}{6}}, \ln e^{\frac{-\sqrt{18}}{6}}, \ln e^{\frac{3}{6}} \right) \\
&= \left(\frac{3}{6}, \frac{-\sqrt{18}}{6}, \frac{3}{6} \right)
\end{aligned}$$

Assim, para a base $\boldsymbol{b}$*, tem-se:*

$$\begin{aligned}\langle \boldsymbol{e}_1, \boldsymbol{e}_2\rangle_A &= \langle clr(\boldsymbol{e}_1), clr(\boldsymbol{e}_2)\rangle \\ &= \langle \Big(\frac{2\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}\Big), \Big(0, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\Big)\rangle_A \\ &= \frac{2\sqrt{6}}{6} \times 0 + \frac{-\sqrt{6}}{6} \times \frac{\sqrt{2}}{2} + \frac{-\sqrt{6}}{6} \times \frac{-\sqrt{2}}{2} = 0\end{aligned}$$

$$\|\boldsymbol{e}_1\|_A = \|clr(\boldsymbol{e}_1)\| = \sqrt{\Big(\frac{2\sqrt{6}}{6}\Big)^2 + \Big(\frac{-\sqrt{6}}{6}\Big)^2 + \Big(\frac{-\sqrt{6}}{6}\Big)^2} = 1$$

$$\|\boldsymbol{e}_2\|_A = \|clr(\boldsymbol{e}_2)\| = \sqrt{0^2 + \Big(\frac{\sqrt{2}}{2}\Big)^2 + \Big(\frac{-\sqrt{2}}{2}\Big)^2} = 1$$

Para a base $\boldsymbol{b}'$*, tem-se:*

$$\begin{aligned}\langle \boldsymbol{e}'_1, \boldsymbol{e}'_2\rangle_A &= \langle clr(\boldsymbol{e}'_1), clr(\boldsymbol{e}'_2)\rangle \\ &= \langle \Big(\frac{-\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2}\Big), \Big(\frac{1}{2}, \frac{-\sqrt{18}}{6}, \frac{1}{2}\Big)\rangle_A \\ &= \frac{-\sqrt{2}}{2} \times \frac{1}{2} + 0 \times \frac{-\sqrt{18}}{6} + \frac{-\sqrt{2}}{2} \times \Big(-\frac{1}{2}\Big) = 0\end{aligned}$$

$$\|\boldsymbol{e}'_1\|_A = \|clr(\boldsymbol{e}'_2)\| = \|clr(\boldsymbol{e'}_1)\| = \sqrt{\Big(\frac{-\sqrt{2}}{2}\Big)^2 + 0^2 + \Big(\frac{\sqrt{2}}{2}\Big)^2} = 1$$

$$\|\boldsymbol{e}'_2\|_A = \sqrt{\Big(\frac{1}{2}\Big)^2 + \Big(\frac{-\sqrt{18}}{6}\Big)^2 + \Big(-\frac{1}{2}\Big)^2} = 1$$

Pode-se agora definir as coordenadas ilr-transformadas.

Definição 2.4.6. *Sejam* $\boldsymbol{x}$ *uma composição em* $\mathcal{S}^d$ *e o conjunto de vetores* $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d-1}\} \in \mathcal{S}^d$ *uma base ortonormal desse espaço. Designa-se por transformação de log-razões isométricas de* $\boldsymbol{x}$ *em relação à base* $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d-1}\}$ *e denota-se* $\text{ilr}(\boldsymbol{x})$ *, a transformação ilr:* $\boldsymbol{x} \in \mathcal{S}^d \to \boldsymbol{z} \in \mathbb{R}^{-d}$ *definida por:*

$$\boldsymbol{z} = ilr(\boldsymbol{x}) \tag{2.35}$$
$$= (z_1, \ldots, z_{d-1}) \tag{2.36}$$

com $z_j = \langle \boldsymbol{x}, \boldsymbol{e}_j\rangle_A$ *ou, de modo equivalente,* $z_j = \langle clr(\boldsymbol{x}), clr(\boldsymbol{e}_j)\rangle$*,* $j \in \{1, \ldots, d-1\}$*.*

Qualquer transformação do tipo ilr baseia-se na escolha de uma base ortonormal do hiperplano $\mathcal{H}$: $\{y = (y_1, \ldots, y_d) \in \mathbb{R}^d : y_1 + \cdots + y_d = 0\}$ (exemplificado na figura 2.3). Desse modo, qualquer transformação do tipo ilr baseia-se na escolha de uma base ortogonal do hiperplano $\mathcal{H}$, que é formado pela transformação clr. Nesta base, a composição $\mathbf{x} \in \mathcal{S}^d$ passa a ser descrita por vetores não colineares transformados por ilr [5], ultrapassando, assim, um dos principais inconvenientes da transformação clr.

Uma matriz de dados composicionais $\mathbf{X}_{(n\times d)}$ também pode ser representada pela sua transformação ilr. Deste modo, tem-se a chamada matriz das coordenadas pivô $\mathbf{Z}_{(n\times(d-1))}$, com $i = \{1, \ldots, n\}$, a qual é formada pelos elementos da forma:

$$z_{ij} = \sqrt{\frac{d-j}{d-j+1}} \ln \frac{x_{ij}}{\sqrt[d-j]{\prod_{k=j+1}^{d} x_{ik}}}, \quad j = \{1, \ldots, d-1\} \tag{2.37}$$

As coordenadas ilr obtidas da forma expressa em (2.37), que resulta de uma escolha particular para uma base, são geralmente designadas de coordenadas pivô, o que se torna intuitivo uma vez que uma das partes é definida como sendo o pivô e aparece somente na primeira coordenada. Assim, a escolha da parte que irá ter o papel de pivô é de extrema importância, uma vez que define o sistema de coordenadas como um todo [1]. Se se tomar, por exemplo, x_1 como pivô, as coordenadas ilr serão da forma:

$$\begin{aligned}
z_1 &= \sqrt{\frac{d-1}{d}} \ln \frac{x_1}{\sqrt[d-1]{\prod_{k=2}^{d} x_k}} \\
z_2 &= \sqrt{\frac{d-2}{d-1}} \ln \frac{x_2}{\sqrt[d-2]{\prod_{k=3}^{d} x_k}} \\
&\ldots \\
z_{d-2} &= \sqrt{\frac{2}{3}} \ln \frac{x_{d-2}}{\sqrt{x_{d-1} x_d}} \\
z_{d-1} &= \sqrt{\frac{1}{2}} \ln \frac{x_{d-1}}{x_d}
\end{aligned}$$

Exemplo 2.4.5. *Utilizando a mesma composição que no exemplo 2.4.2,* $\boldsymbol{x} = (11.4, 52.7, 35.9)$*, a sua transformação ilr com pivô* x_1 *é dada por:*

$$\begin{aligned}
\boldsymbol{z} &= \text{ilr}(\boldsymbol{x}) \\
&= \text{ilr}(11.4, 52.7, 35.9) \\
&= \left(\sqrt{\frac{3-1}{3}} \ln \frac{x_1}{\sqrt[3-1]{x_2 \times x_3}}, \sqrt{\frac{3-2}{3-1}} \ln \frac{x_2}{\sqrt[3-2]{x_3}}\right) \\
&= \left(\sqrt{\frac{2}{3}} \ln \frac{11.4}{\sqrt{52.7 \times 35.9}}, \sqrt{\frac{1}{2}} \ln \frac{52.7}{35.9}\right) \\
&= (-1.09, 0.27)
\end{aligned}$$

Como já foi referido, as coordenadas pivô têm a particularidade de a parte x_1, selecionada para pivô, apenas aparecer na primeira coordenada, z_1 [1]. Todavia, o mesmo

não acontece para as restantes partes: x_2 aparece nas coordenadas z_1 e z_2, x_3 aparece nas coordenadas z_1, z_2 e z_3 e por assim adiante. Isolando a primeira parte de uma coordenada, i.e., z_1, é de interesse, uma vez que ela sumariza toda a informação relativa, em termos de logaritmos, acerca de x_1. Note-se que:

$$
\begin{aligned}
z_1 &= \sqrt{\frac{d-1}{d}} \ln \frac{x_1}{\sqrt[d-1]{\prod_{k=2}^{d} x_k}} \quad \textbf{(*)} \\
&= \sqrt{\frac{d-1}{d}} \ln x_1 - \sqrt{\frac{d-1}{d}} \ln \Big(\prod_{k=2}^{d} x_k \Big)^{\frac{1}{d-1}} \\
&= \sqrt{\frac{d-1}{d}} \ln x_1 - \sqrt{\frac{d-1}{d}} \sqrt{\Big(\frac{1}{d-1}\Big)^2} \ln(x_2 \times \cdots \times x_d) \\
&= \sqrt{\frac{d-1}{d}} \ln x_1 - \sqrt{\frac{1}{(d-1)d}} \ln(x_2 \times \cdots \times x_d) \\
&= \sqrt{\frac{d-1}{d-1}} \sqrt{\frac{d-1}{d}} \ln x_1 - \sqrt{\frac{1}{(d-1)d}} \ln(x_2 \times \cdots \times x_d) \\
&= (d-1) \sqrt{\frac{1}{(d-1)d}} \ln x_1 - \sqrt{\frac{1}{(d-1)d}} \Big(\ln x_2 + \cdots + \ln x_d \Big) \\
&= \sqrt{\frac{1}{(d-1)d}} \Big((\ln x_1 - \ln x_2) + (\ln x_1 - \ln x_3) + \cdots + (\ln x_1 - \ln x_d) \Big) \\
&= \sqrt{\frac{1}{(d-1)d}} \Big(\ln \frac{x_1}{x_2} + \ln \frac{x_1}{x_3} + \cdots + \ln \frac{x_1}{x_d} \Big) \quad \textbf{(**)}
\end{aligned}
$$

Assim, z_1 pode ser interpretada como a dominância relativa de x_1 em relação às outras partes, "em média". Se:

- $z_1 > 0 \rightarrow x_1$ é dominante em relação à média geométrica das outras partes;
- $z_1 < 0 \rightarrow x_1$ não é dominante em relação à média geométrica das outras partes;
- $z_1 = 0 \rightarrow x_1$ está balanceada em relação à média geométrica das outras partes.

Esta interpretação, contudo, apenas pode ser feita para a primeira coordenada e é sobretudo nela que o interesse se deve focar, pela quantidade de informação que traduz.

É possível obter a primeira coordenada dos coeficientes clr-transformados, $y_1 = \ln \frac{x_1}{\sqrt[d]{\prod x_j}}$, a menos de um fator de escala a partir da primeira coordenada da transformação ilr z_1 dada pela equação (2.37), com $j = 1$, ou seja, **(**)**. Observando **(**)**, conclui-se que:

$$z_1 = \sqrt{\frac{d-1}{d}} y_1 \tag{2.38}$$

pelo que

$$y_1 = \sqrt{\frac{d}{d-1}} z_1 \tag{2.39}$$

Exemplo 2.4.6. *Do exemplo 2.4.2, para a composição* $\boldsymbol{x} = (11.4, 52.7, 35.9)$ *tem-se que* $y_1 = -0.89$ *e do exemplo 2.4.5, para a mesma composição, tem-se que* $z_1 = -1.09$*. Assim, verifica-se que:*

$$z_1 = -1.09 = \sqrt{\frac{3}{2}} \times -0.89 = \sqrt{\frac{3}{2}}\, y_1$$

A relação 2.39 pode fazer crer que se trata apenas de um fator entre as duas transformações. Tal só se verifica entre a primeira componente de uma transformação com a primeira componente da outra transformação. Enquanto que a parte x_1 da composição está, exclusivamente, na primeira coordenada de $\mathbf{z}$, x_1 aparece em todas as coordenadas de $\mathbf{y}$, na média geométrica, que é o denominador da log-razão.

Tal como acontece nas outras transformações de dados composicionais, é possível estabelecer a inversa da transformação ilr e, desse modo, regressar às partes originais da composição. A inversa da transformação ilr, denotada por ilr^{-1}: $\mathbf{z} \in \mathbb{R}^{d-1} \to \mathbf{x} \in \mathcal{S}^d$ é dada por:

- para $j = 1$:
  $x_1 = \exp\left(\frac{\sqrt{d-1}}{\sqrt{d}} z_1\right)$
- para $j = \{2, \ldots, d-1\}$:
  $x_j = \exp\left(-\sum_{k=1}^{j-1} \frac{1}{\sqrt{(d-k+1)(d-k)}} z_k + \frac{\sqrt{d-j}}{\sqrt{d-j+1}} z_j\right)$
- para $j = d$:
  $x_d = \exp\left(-\sum_{k=1}^{d-1} \frac{1}{\sqrt{(d-k+1)(d-k)}} z_d\right)$

De realçar que, em $\mathcal{S}^d$, podem ser definidas inúmeras bases ortonormais. Devido à diversidade de bases possíveis, cada uma delas dará origem a diferentes coordenadas ilr-transformadas [1].

A base ortonormal de vetores correspondente às coordenadas pivô definidas como na equação 2.37 é dada por:

$$\mathbf{v}_j = \sqrt{\frac{d-j}{d-j+1}}\left(0, 0, \ldots, 0, 1, -\frac{1}{d-j}, \ldots, -\frac{1}{d-j}\right) \tag{2.40}$$

para $j \in \{1, \ldots, d-1\}$, onde as primeiras $j-1$ entradas são zero. Esses vetores, vistos como colunas numa matriz $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_{d-1}]$ de dimensão $d \times (d-1)$, são formados pelos coeficientes clr da base composicional original (definição 2.4.5). Daí pode-se tirar imediatamente uma relação entre os coeficientes clr e as coordenadas pivô [5]:

$$\mathbf{y} = \mathbf{V}\mathbf{z} \;\text{ e }\; \mathbf{z} = \mathbf{V}^T\mathbf{y} \tag{2.41}$$

Esta relação linear é verdadeira para quaisquer coordenadas ilr, e não apenas para as coordenadas pivô [5].

Tal como ocorre com os coeficientes clr, também as coordenadas ilr representam uma isometria. Deste modo, se $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}^d$, $c \in \mathbb{R}$, tem-se:

$$\text{ilr}(\mathbf{x}_1 \oplus \mathbf{x}_2) = \text{ilr}(\mathbf{x}_1) + \text{ilr}(\mathbf{x}_1) \tag{2.42}$$

$$\text{ilr}(c \odot \mathbf{x}_1) = c \cdot \text{ilr}(\mathbf{x}_1) \tag{2.43}$$

Todos os conceitos métricos também se mantém se se tomar as coordenadas ilr definindo:

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \langle \text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2) \rangle \tag{2.44}$$

$$\|\mathbf{x}_1\|_A = \|\text{ilr}(\mathbf{x}_1)\| \tag{2.45}$$

$$d(\mathbf{x}_1, \mathbf{x}_2)_A = d(\text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2)) \tag{2.46}$$

$$\tag{2.47}$$

Até ao momento, o pivô considerado foi sempre a primeira parte da composição $\mathbf{x}$, x_1. Contudo, de um ponto de vista prático, nem sempre é a primeira parte da composição que tem interesse de ser comparada com as demais - uma outra parte pode ser a que suscita o interesse na análise. Nesse caso, basta realizar uma permutação e colocar, na primeira posição, a parte da composição que, efetivamente, tem a primazia do interesse no estudo. Assim, se a parte x_l é que, efetivamente, faz mais sentido estabelecer uma comparação com as demais partes, a composição pode ser reescrita como:

$$\mathbf{x}^{(l)} = (x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_d) \tag{2.48}$$

De realçar que as demais partes da composição não têm a sua ordem alterada. Partindo da permutação da composição estabelecida na equação (2.48), é possível estabelecer as coordenadas pivô:

$$z_j^{(l)} = \sqrt{\frac{d-j}{d-j+1}} \ln \frac{x_j^{(l)}}{\sqrt[d-j]{\prod_{k=j+1}^{d} x_k^{(l)}}}, \quad j = \{1, \ldots, d-1\} \tag{2.49}$$

Capítulo 3

Métodos de Deteção de *Outliers* Multivariados

3.1 Abordagem Comediana

A Abordagem Comediana (do inglês *Comedian Approach*) é um método de deteção de observações atípicas multivariadas que usa a comediana (do inglês *comedian*) como medida alternativa de dependência entre duas variáveis aleatórias.

Definição 3.1.1 (Comediana)**.** *Sejam X_1 e X_2 duas variáveis aleatórias. A comediana de X_1 e X_2 é definida como:*

$$COM(X_1, X_2) = med[(X_1 - med(X_1))(X_2 - med(X_2))] \tag{3.1}$$

onde med *denota a mediana. Para obter a comediana amostral seja $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})$, $j \in 1, 2$, o vetor de n observações amostrais da população X_i. Define-se comediana amostral como:*

$$COM(\boldsymbol{x}_1, \boldsymbol{x}_2) = med_i[(\boldsymbol{x}_{i1} - med(\boldsymbol{x}_1)) * (\boldsymbol{x}_{i2} - med(\boldsymbol{x}_2))] \tag{3.2}$$

onde $*$ denota o produto elemento a elemento.

Exemplo 3.1.1. *Tome-se, novamente, o conjunto de dados 'Arctic Lake' como referência para o exemplo.*
Sejam $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$ os vetores amostrais das variáveis Lodo e Argila, respetivamente. O valor da comediana amostral para as variáveis em questão é:

$$\begin{aligned}
COM(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \text{med}[(\boldsymbol{x}_{i1} - \text{med}(\boldsymbol{x}_1)) * (\boldsymbol{x}_{i2} - \text{med}(\boldsymbol{x}_2))] \\
&= \text{med}\Big([19.5 - 48 \quad \ldots \quad 47.8 - 48][3 - 34.9 \quad \ldots \quad 52.2 - 34.9]\Big) \\
&= \text{med}\Big([-28.5 \quad \ldots \quad -0.2][-31.9 \quad \ldots \quad 15.3]\Big) \\
&= \text{med}\ (909.15, \ldots, -3.06) \\
&= 9.18
\end{aligned}$$

Este conceito generaliza a definição de Desvio Absoluto Mediano (em inglês *Median Absolute Deviation* e abreviado como MAD), para duas variáveis.

Definição 3.1.2 (Desvio Absoluto Mediano)**.** *Seja X uma variável aleatória. O Desvio Absoluto Mediano, conhecido simplesmente como MAD de X, é dado por:*

$$MAD(X) = med(|X - med(X)|) \tag{3.3}$$

onde med *denota a mediana. Relativamente ao conceito amostral, dada uma amostra $\boldsymbol{x} = (x_1, \ldots, x_n)$ de uma população X, define-se MAD amostral como:*

$$MAD(\boldsymbol{x}) = med(|x_1 - \text{med}(\boldsymbol{x})|, \ldots, |x_n - \text{med}(\boldsymbol{x})|) \tag{3.4}$$

Exemplo 3.1.2. *Sejam $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$ os vetores com os valores amostrais das variáveis Lodo e Argila do conjunto de dados 'Arctic Lake', respetivamente. O valor amostral do Desvio Absoluto Mediano para cada uma das variáveis é:*

$$\begin{aligned} MAD(\boldsymbol{x}_1) &= med|(\boldsymbol{x}_1 - med(\boldsymbol{x}_1)| \\ &= med|19.5 - 48|, |24.9 - 48|, \ldots, |48 - 48|, |47.8 - 48)|) \\ &= med(28.5, 23.1, \ldots, 0, 0.2) \\ &= 48 \end{aligned}$$

$$\begin{aligned} MAD(\boldsymbol{x}_2) &= med|(\boldsymbol{x}_2 - med(\boldsymbol{x}_2))| \\ &= med(|3 - 34.9|, |3.2 - 34.9|, \ldots, |49.5 - 34.9|, |50.2 - 34.9|) \\ &= med(31.9, 31.7, \ldots, 14.6, 15.3) \\ &= 11.9 \end{aligned}$$

A MAD é uma medida robusta de variabilidade de uma amostra univariada que traduz a mediana dos resíduos dos dados, sendo os resíduos medidos pela diferença de cada observação à mediana das observações. O MAD, como é uma medida robusta, é mais resistente a *outliers* (observações atípicas) que o desvio padrão. Além disso, esta medida tem a propriedade desejável do seu ponto de rutura (*breakdown point*) ser de 50%, que tem o valor máximo possível, uma vez que seria necessário que sensivelmente metade dos dados fossem inflacionados, isto é, que tendessem para o infinito para que a mediana fosse afetada [29] e, desse modo, alterar o valor do MAD.

É possível generalizar o MAD ao estabelecer uma relação entre este e a comediana.

Teorema 3.1.1. *A comediana é igual ao quadrado de MAD quando se considera $X = Y$.*

Demonstração.

$$\begin{aligned} \text{COM}(X, X) &= \text{med}\Big((X - \text{med}(X))(X - \text{med}(X))\Big) \\ &= \text{med}((X - \text{med}(X))^2) \\ &= \text{med}(|X - \text{med}(X)||X - \text{med}(X)|) \\ &= \text{med}(|X - \text{med}(X)|)\text{med}(|X - \text{med}(X)|) \\ &= \text{MAD}(X)\ \ \text{MAD}(X) \\ &= \text{MAD}^2(X) \end{aligned}$$

■

Neste sentido, é possível estabelecer um paralelismo entre MAD e comediana com o desvio padrão, σ, e a covariância a qual satisfaz a condição $\sigma^2 = \text{COV(X,Y)}$, para X=Y. A covariância requer a existência dos dois primeiros momentos das variáveis aleatórias X e Y, isto é, da média (primeiro momento) e da variância (segundo momento centrado), enquanto que a $\text{COM}(X, Y)$ existe sempre [6]. Como a comediana é, efetivamente, uma mediana, herda o ponto de rutura dela, tendo, deste modo, também o maior ponto de rutura possível.

A comediana obedece às seguintes propriedades:

- simetria $\rightarrow \text{COM}(X, Y) = \text{COM}(Y, X)$;
- invariante quanto à localização $\rightarrow \text{COM}(X, Y + b) = \text{COM}(X, Y), \forall b \in \mathbb{R}$;
- invariante quanto à escala $\rightarrow \text{COM}(X, aY) = a\text{COM}(Y, X), \forall a \in \mathbb{R}$.

Hall e Welsh, no seu artigo *Limit Theorems for the Median Deviation* [30], fazem a demonstração, sob certas condições (nomeadamente, a função de distribuição ser suave e sem ser estabelecida a hipótese de simetria) de duas caraterísticas importantes do estimador do MAD: a sua consistência e a sua normalidade assintótica. Concretamente, dada uma amostra aleatória, $X_1, \ldots, X_n$, de uma população X, o MAD amostral, denotado por $\widehat{MAD}(.)$, é um estimador consistente do parâmetro populacional, MAD(.). O estimador é consistente pois aproxima-se do parâmetro à medida que o tamanho amostral aumenta, isto é, o estimador, converge em probabilidade para o parâmetro. Formalmente, tem-se:

$$\lim_{n \to +\infty} \Pr(|\widehat{\text{MAD}} - \text{MAD}| > \epsilon) = 0, \forall \epsilon > 0 \tag{3.5}$$

Relativamente à normalidade assintótica significa que, com o aumento do tamanho da amostra, o estimador, convenientemente normalizado, tende, em distribuição, para uma lei gaussiana, nomeadamente [29]:

$$\sqrt{n}(\widehat{\text{MAD}}(X) - \text{MAD}(X)) \overset{d}{\sim} \text{N}\Big(0, \frac{1}{4A^2}\Big(1 + \frac{B}{f(F^{-1}(\frac{1}{2}))^2}\Big)\Big) \tag{3.6}$$

onde f e F referem-se, respetivamente, à função densidade de probabilidade e à função distribuição de probabilidade, com:

- $A = f(F^{-1}(\frac{1}{2}) + \text{MAD}(X)) + f(F^{-1}(\frac{1}{2}) - \text{MAD}(X)) > 0$,
- $C = f(F^{-1}(\frac{1}{2}) - \text{MAD}(X)) - f(F^{-1}(\frac{1}{2}) + \text{MAD}(X))$,
- $B = C^2 + 4Cf(F^{-1}(\frac{1}{2}))\Big(1 - F(F^{-1}(\frac{1}{2}) + \text{MAD}(X)) - F(F^{-1}(\frac{1}{2}) - \text{MAD}(X))\Big)$

Para a convergência acontecer, é necessário que $\widehat{\text{MAD}} - \text{MAD}$ decresça mais rapidamente do que cresce $\sqrt{n}$, ou seja, tem de ir para 0 mais depressa do que cresce $\sqrt{n}$ [14].

De modo semelhante, resultados análogos foram definidos para a comediana.

Com base na definição de comediana, surge, como medida alternativa de coeficiente de correlação, baseando-se no uso da mediana e não do valor esperado, a correlação mediana (em inglês, *median correlation*):

Definição 3.1.3 (Correlação Mediana)**.** *Sejam X e Y duas variáveis aleatórias. A correlação mediana entre X e Y é definida como:*

$$\delta(X,Y) = \frac{COM(X,Y)}{MAD(X)MAD(Y)} \tag{3.7}$$

a qual toma valores no intervalo $\delta \in [-1,1]$. Similarmente às definições anteriores, a correlação mediana amostral é dada por:

$$\delta(\boldsymbol{x},\boldsymbol{y}) = \frac{COM(\boldsymbol{x},\boldsymbol{y})}{MAD(\boldsymbol{x})MAD(\boldsymbol{y})} \tag{3.8}$$

para quaisquer vetores amostrais $\boldsymbol{x}$ e $\boldsymbol{y} \in \mathbb{R}^d$.

Exemplo 3.1.3. *Tome-se os valores obtidos nos exemplos* **??** *e 3.1.2. Dado que $COM(\boldsymbol{x}_1,\boldsymbol{x}_2) = 9.18$, $MAD(\boldsymbol{x}_1) = 48$ e $MAD(\boldsymbol{x}_2) = 11.9$, obtém-se:*

$$\begin{aligned}\delta(\boldsymbol{x}_1,\boldsymbol{x}_2) &= \frac{COM(\boldsymbol{x}_1,\boldsymbol{x}_2)}{MAD(\boldsymbol{x}_1)MAD(\boldsymbol{x}_2)}\\ &= \frac{9.18}{48 \times 11.9}\\ &\approx 0.016\end{aligned}$$

A correlação mediana toma valor 1 ou -1 quando as variáveis são completamente dependentes e, opostamente, toma valor 0 quando elas são independentes.

Generalizando o conceito de comediana para dados multivariados, é possível obter a matriz das comedianas.

Definição 3.1.4 (Matriz de Comedianas)**.** *Seja $\boldsymbol{X}_{(n\times d)}$ a matriz de dados relativa às variáveis X_j, com $j \in \{1,\ldots,d\}$. Então, a matriz de comedianas, de dimensão $d \times d$, é definida como:*

$$\boldsymbol{COM}(\boldsymbol{X}) = [COM(X_{j_1}, X_{j_2})]_{j_1,j_2 \in \{1,\ldots,d\}} \tag{3.9}$$

Do mesmo modo, é possível definir a matriz de correlações comedianas $\boldsymbol{\delta}$.

Definição 3.1.5 (Matriz de Correlações Comedianas)**.** *Seja $\boldsymbol{X}_{(n\times d)}$ a matriz de dados relativa às variáveis X_j, com $j \in \{1,\ldots,d\}$. Então, a matriz de correlação comediana, de dimensão $d \times d$, é definida como:*

$$\boldsymbol{\delta} = \boldsymbol{D}\ \boldsymbol{COM}(\boldsymbol{X})\ \boldsymbol{D} \tag{3.10}$$

onde $\boldsymbol{D}$ é uma matriz diagonal $diag\Big(\frac{1}{MAD(X_1)},\ldots,\frac{1}{MAD(X_d)}\Big)$.

Apesar destas estimativas terem um elevado ponto de rutura, surge um problema no que à matriz de comedianas diz respeito, uma vez que, embora seja uma alternativa robusta à matriz de covariância, não é, em geral, (semi)definida positiva. Para ultrapassar esse obstáculo, Sajesh e Srinivasan em [31], sugerem um procedimento de três etapas para ultrapassar essa inconveniência e, desse modo, obter estimativas robustas para a localização e dispersão.

i. Calcular os valores próprios, λ_j, e os vetores próprios, $\mathbf{e}_j$, com $j \in \{1, \ldots, d\}$, de $\boldsymbol{\delta}(\mathbf{X})$ e designar por $\mathbf{E}$ a matriz cuja colunas são os vetores próprios e por $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$, com $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, tal que $\boldsymbol{\delta}(\mathbf{X}) = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T$;

ii. Seja $\mathbf{D}$ a matriz diagonal $\text{diag}\Big(\frac{1}{\text{MAD}(X_1)}, \ldots, \frac{1}{\text{MAD}(X_d)}\Big)$. Calcula-se então:

$$\mathbf{Q} = \mathbf{D}^{-1}\mathbf{E}$$

$$\mathbf{z}_i = \mathbf{Q}^{-1}\mathbf{x}_i$$

com $i \in \{1, \ldots, n\}$;

iii. As estimativas robustas para a localização, $\mathbf{m}(\mathbf{X})$, e para a dispersão, $\mathbf{S}(\mathbf{X})$, são dadas por:

$$\mathbf{m}(\mathbf{X}) = \mathbf{Q}\mathbf{l} \tag{3.11}$$

$$\mathbf{S}(\mathbf{X}) = \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^T \tag{3.12}$$

onde:

* $\boldsymbol{\Gamma}$ é uma matriz diagonal cujas entradas são dadas por $(\text{MAD}(Z_j))^2$;
* $\mathbf{l} = (\text{med}(Z_1), \ldots, \text{med}(Z_d))$;
* Z_j é a $j-$ésima coluna $\mathbf{Q}$.

É possível fazer uma melhoria nestas estimativas, utilizando um processo iterativo, substituindo $\mathbf{S}$ por δ e repetir as etapas enumeradas anteriormente.

Conhecidos os estimadores robustos para a localização e dispersão, o interesse é verificar se, de facto, uma observação é ou não atípica (*outlier*). Um dos procedimentos mais usados é recorrer à distância de Mahalanobis que, sob a normalidade dos dados $d-$ dimensionais, segue uma distribuição de qui-quadrado com d graus de liberdade. Em contexto robusto, recorre-se à distância de Mahalanobis robusta, definida como:

$$\text{RD}(\mathbf{x}_i, \mathbf{m}) = rd_i = (\mathbf{x}_i - \mathbf{m})^T\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{m}), \quad i = \{1, \ldots, n\} \tag{3.13}$$

com $\mathbf{m}$ e $\mathbf{S}$ definidos como nas equações 3.11 e 3.12, respetivamente.

Tendo sido calculada a distância robusta das observações, é necessário estabelecer como utilizá-la para definir uma observação como discrepante. Para tal, é necessário

definir um valor de corte (do inglês *cut off value*) de modo a determinar potenciais *outliers*. O valor de corte é dado por:

$$cv = 1.4826 \times \frac{\chi^2_{d;1-\alpha} \times \text{med}(rd_1, \ldots, rd_n)}{\chi^2_{d;0.5}} \tag{3.14}$$

onde α refere-se à probabilidade de erro do tipo I que se está disposto a cometer, $\chi^2_{d;1-\alpha}$ e $\chi^2_{d;0.50}$ referem-se, respetivamente, aos quantis $1-\alpha$ e 0.5 da distribuição qui-quadrado com d graus de liberdade e 1.4826 é um fator de correção, que corresponde ao inverso do valor do quantil 0.75 da distribuição Normal *standarizada*, de modo a que a MAD não seja enviesada [32].

Assim, uma observação é declarada como *outlier* se $\text{RD}(\mathbf{x}_i, \mathbf{m}) > cv$.

Ao usar o valor de corte definido em 3.14 para a distância robusta de Mahalanobis (equação 3.13) pode ser definida uma função de peso e as estimativas robustas para a localização e dispersão obtidas. Estas estimativas são definidas positivas e aproximadamente invariantes afins. Além disso, essas estimativas para a localização e dispersão, obtidas pela comediana, têm um ponto de rutura elevado, que ajuda na deteção de um grupo vasto de observações atípicas (*outliers*). Por estudos prévios realizados, pode-se afirmar que a eficiência deste método usando a distância de Mahalanobis robusta aumenta com o aumento da dimensão do conjunto de dados [19].

3.2 Estimador de Stahel-Donoho Ajustado

3.2.1 Estimador de Stahel-Donoho

O estimador de Stahel-Donoho proposto, de forma independente, por Stahel ([33]) e Donoho ([34]), é usado no cálculo de distâncias robustas e, desse modo, verificar se uma dada amostra tem ou não *outliers*.

Esses estimadores, para a localização e dispersão, são calculados atribuindo um peso decrescente às observações que estão distantes em relação a alguma projeção dos dados no espaço univariado.

Definição 3.2.1 (Atipicidade de Stahel-Donoho)**.** *A atipicidade segundo Stahel-Donoho de uma observação $\boldsymbol{x}_i$ (do inglês Stahel-Donoho Outlyingness), pertencente a uma matriz de dados $\boldsymbol{X}_{(n\times d)}$, é dada pelo valor numérico:*

$$SDO_i = \sup_{\boldsymbol{a}\in\mathcal{S}_d} \frac{|\boldsymbol{a}^T\boldsymbol{x}_i - t(\boldsymbol{a}^T\boldsymbol{X})|}{V(\boldsymbol{a}^T\boldsymbol{X})}, \;\; i \in \{1, \ldots, n\} \tag{3.15}$$

onde t e V são, respetivamente, os estimadores para a localização e dispersão e $\mathcal{S}_d = \{\boldsymbol{a} \in \mathbb{R}^d : \|\boldsymbol{a}\| = 1\}$, o conjunto das direções univariadas.

Se, como alternativa às medidas de localização e dispersão usuais (vetor de médias e matriz de covariância), se optar por utilizar medidas robusta, como a mediana e o MAD, ter-se-á [35]:

$$\text{SDO}_i = \sup_{\mathbf{a}\in\mathcal{S}_d} \frac{|\mathbf{a}^T\mathbf{x}_i - \text{med}(\mathbf{a}^T\mathbf{X})|}{\text{MAD}(\mathbf{a}^T\mathbf{X})} \tag{3.16}$$

Após determinar a atipicidade, é possível calcular os estimadores ponderados para a localização, $t(\mathbf{X})$, e dispersão, $V(\mathbf{X})$ [19], [36]:

$$t(\mathbf{X}) = \frac{\sum_{i=1}^n w(r_i)\mathbf{x}_i}{\sum_{i=1}^n w(r_i)} \tag{3.17}$$

$$V(\mathbf{X}) = \frac{\sum_{i=1}^n w(r_i)(\mathbf{x}_i - t(\mathbf{X}))(\mathbf{x}_i - t(\mathbf{X}))^T}{\sum_{i=1}^n w(r_i)} \tag{3.18}$$

onde $w : \mathbb{R}_0^+ \to \mathbb{R}^+$ é uma função de pesos, positiva e decrescente. Há várias funções que podem ser utilizadas com este intuito: funções de peso de Huber, funções de perda logarítmica (do inglês *log loss functions*), funções quadráticas de perda, funções logísticas de perda, funções de perda exponencial, entre outras [37].

Quanto maior a atipicidade de uma observação, menor deverá ser o peso a atribuir a essa observação, isto é, uma observação tida como *outlier* terá um peso inferior que uma observação que esteja no conjunto predominante (não atípicas).

Se os dados são provenientes de uma distribuição Normal $d-$dimensional, prova-se que os SDO_i seguem assintoticamente uma distribuição χ^2_d [35]. Deste modo, pode-se classificar-se como *outlier* a observação x_i se o seu respetivo SDO_i ultrapassa o valor $\chi^2_{d;1-\alpha}$, com α sendo a probabilidade de cometer um erro de tipo I previamente fixada. Todavia, a identificação de *outliers* pela medida SDO é afetada dos dois mesmos problemas que os métodos baseados na distância de Mahalanobis:

- Geralmente assume-se que as observações $\mathbf{x}_i$ provêm de uma distribuição Normal $d-$dimensional. Quando não se tem conhecimento sobre a forma das observações, a distribuição que os SDO_i seguem é, em geral, desconhecida (apesar de ter que ser limitada à esquerda por zero e enviesada à direita) pelo que o uso do quantil de ordem $1-\alpha$ da distribuição χ^2_d para a identificação de *outliers* pode não ser válida;
- A medida SDO apenas é adequada para dados elípticos simétricos, pois depende do MAD que, como medida de dispersão, assume que os dados estão dos dois lados da mediana.

Apesar destes problemas, esta abordagem, se as condições se verificarem (por exemplo, a população ser proveniente de uma distribuição Normal $p-$dimensional), é bastante robusta uma vez que o estimador tem um ponto de rutura que tende assintoticamente para 50% [35].

A maior dificuldade na utilização desde método recai sobre o cálculo dos valores de atipicidade (SDO). Este cálculo depende das projeções, **a**, que são uma infinidade. Se

pensarmos que os dados têm dimensão $d = 2$, os vetores de projeções são aqueles que têm origem no centro do referencial e extremidade um dos pontos da circunferência unitária, logo há uma imensidade de possíveis vetores que se podem selecionar (figura 3.1).

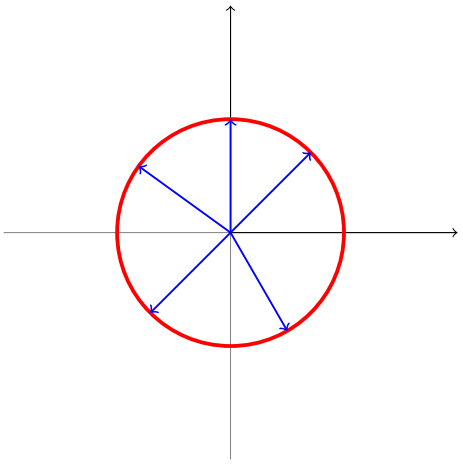


Figura 3.1: Representação, em $\mathbb{R}^2$, de cinco direções **a** selecionadas (a azul) do conjunto $\mathcal{S}_2$.

Na prática, não se consegue trabalhar com uma infinidade de projeções. Por norma, estipula-se como número de projeções a utilizar $m = 250d$, numa tentativa de equilíbrio entre eficácia do método e tempo computacional [32].

Inicialmente, não havia nenhum método satisfatório para calcular as medidas SDO e, desse modo, este método não possuía aplicação prática para detetar observações atípicas. Contudo, mais tarde, Gasko e Donoho propuserem um método que usa este tipo de estimador para identificar pontos de influência em regressão múltipla [19].

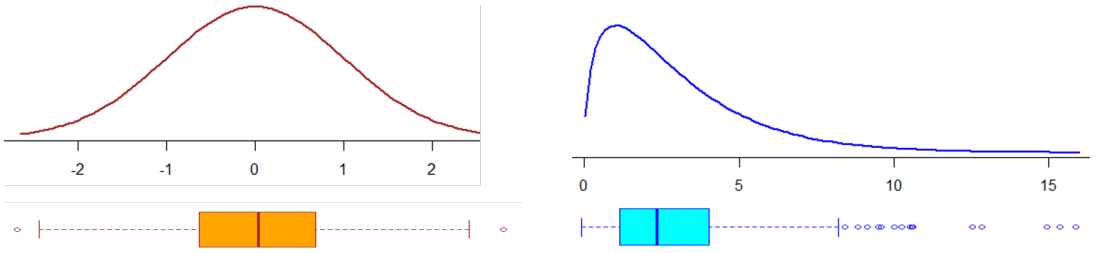
3.2.2 Atipicidade Ajustada

De modo a ultrapassar um dos maiores obstáculos deste método, que é partir do pressuposto que os dados são provenientes de uma distribuição elíptica e simétrica, por vezes usam-se transformações de simetria numa ou, eventualmente, em todas as variáveis. Algumas das transformações mais comuns são a logarítmica ou a Box-Cox. Esta abordagem é usada sobretudo quando a transformação destas variáveis tem significado físico. Todavia, estas transformações necessitam de um maior pré-processamento, não são invariantes afins e, nem sempre, se consegue interpretar o seu significado. Além disso, a transformação Box-Cox é baseada no Método de Máxima Verosimilhança e, consequentemente, não é robusta a observações atípicas [32].

Para levar em consideração um possível enviesamento dos dados, que pode ser observado por gráficos de dispersão univariados, para se ter uma ideia de como os dados se comportam, Hubert e Van der Veeken, no seu artigo *Outlier Detection for Skewed Data* [32], propõem um ajuste ao estimador SDO, de modo a permitir a assimetria nos

dados. Esta nova medida é designada por Atipicidade Ajustada (em inglês *Adjusted Outlyingness*, abreviada como AO). Este método baseia-se no *box plot* ajustado para dados enviesados e, essencialmente, define, para cada variável, escalas diferentes para as observações que estão à esquerda e à direita da mediana.

O *box plot* clássico, proposto por Tukey, tem por base a distribuição Normal e declara como *outliers* moderados as observações que distam entre 1.5 e 3 vezes da distância interquartil, AIQ, (isto é, a diferença entre os quantis de ordem 0.25 e 0.75) - observações fora do intervalo $[Q_{0.25} - 1.5 \times \text{AIQ}; Q_{0.75} + 1.5 \times \text{AIQ}]$ - e como *outliers* severos as observações que distam 3 ou mais vezes da distância interquartil - observações fora do intervalo $[Q_{0.25} - 3 \times \text{AIQ}; Q_{0.75} + 3 \times \text{AIQ}]$. De facto, o *box plot* é uma ferramenta de visualização bastante útil, mas que pode não dar uma ideia concreta de quais serão as potenciais observações atípicas no caso de distribuições com caudas pesadas ou enviesadas.



(a) Distribuição Normal (b) Distribuição Qui-Quadrado

Figura 3.2: Gráficos das funções de densidade provenientes de uma amostra aleatória ($n = 500$) e respetivos *box plots* de (a) uma distribuição Normal $\mathcal{N}(0, 1)$ e de (b) uma distribuição qui-quadrado χ^2_3.

Como visto na figura 3.2, o *box plot* não é uma boa opção para dados provenientes de distribuições assimétricas, como o caso da distribuição Qui-quadrado. Desse modo e, após diversas tentativas que se revelaram pouco adequadas (por exemplo, o *SIQR box plot*, por Kimber, em 1990, e Aucremanne, em 2004 [39]) surgiu o *box plot* ajustado. O *box plot* ajustado difere do *box plot* clássico por definir as barreiras de modo diferente utilizando, para tal, *medcouple*.

Para medir o enviesamento de uma amostra univariada $\mathbf{x} = (x_1, \ldots, x_n)$, proveniente de uma distribuição contínua e unimodal, é utilizada a estatística robusta *medcouple* (MC) [39]. A *medcouple* é uma medida de enviesamento definida como a mediana de diferenças (normalizadas) de desvios das observações à esquerda e à direita da mediana (de notar que, para o cálculo da mediana, as observações precisam de estar ordenadas). De modo formal, pode-se definir *medcouple* como:

$$\text{MC} = \underset{x_i \leq Q_{0.50} \leq x_j}{\text{med}} h(x_i, x_j) \tag{3.19}$$

onde $Q_{0.50}$ corresponde à mediana amostral e, para $x_i \neq x_j$, a função *kernel* h é dada

por:

$$h(x_i, x_j) = \frac{(x_j - Q_{0.50}) - (Q_{0.50} - x_i)}{x_j - x_i}, \quad x_i, x_j \in \mathbf{x} \tag{3.20}$$

A *medcouple* é uma estatística não paramétrica (i.e., estatística sobre a qual não se tem conhecimento sobre a distribuição da população) que pertence à classe das estatísticas-L: estatísticas que são uma combinação linear das estatísticas de ordem, isto é, que são extraídas dos dados ordenados. Por exemplo, a estatística de primeira ordem corresponde ao mínimo e a estatística de $n-$ésima ordem corresponde ao máximo.

Esta estatística, sendo não paramétrica, tem a vantagem de ser calculada para qualquer distribuição. Além disso, tem um ponto de rutura de 25%, isto é, cerca de um quarto da amostra pode ser contaminada sem que o estimador entre em rutura [38].

A *medcouple* toma valores compreendidos entre -1 e 1. A distribuição é tanto mais simétrica quanto mais próximo o valor MC estiver de 0; pelo contrário, quanto mais assimétrica, mais perto em valor absoluto a MC está de 1. Deste modo:

$$\begin{cases} -1 \leq \text{MC} < 0 & \rightarrow \text{distribuição enviesada à esquerda} \\ \text{MC} = 0 & \rightarrow \text{distribuição simétrica} \\ 0 < \text{MC} \leq 1 & \rightarrow \text{distribuição enviesada à direita} \end{cases} \tag{3.21}$$

Contudo, o cálculo desta estatística não é trivial. Os algoritmos mais comuns para cálculo da *medcouple* são de dois tipos:

- algoritmo *naïve*;
- algoritmos rápidos, como por exemplo o comparar um valor com a matriz kernel da função h definida em 3.20.

O *box plot standard* tem a sua cerca (*fence*) definida pelo intervalo:

$$[Q_{0.25} - 1.5 \text{ AIQ}; Q_{0.75} + 1.5 \text{ AIQ}] \tag{3.22}$$

Utilizando a função h definida em 3.20, os novos limites da cerca podem ser definidos como:

$$[Q_{0.25} - h_i(\text{MC}) \text{ AIQ}; Q_{0.75} + h_s(\text{MC}) \text{ AIQ}] \tag{3.23}$$

onde h_i e h_s referem-se aos valores da função para o limite inferior e limite superior, respetivamente. Além disso, é necessário estabelecer que, no caso particular de $\text{MC} = 0$, isto é, de não haver enviesamento e, por isso, a distribuição ser simétrica, tem-se $h_i(\text{MC}) = 1.5 = h_s(\text{MC})$. Estes valores permitem que a cerca seja assimétrica e, desse modo, é possível obter o *box plot* ajustado.

Neste sentido, foram estudados três modelos diferentes de cerca:

1. <u>modelo linear</u>, com limites:

- $h_i(\text{MC}) = 1.5 + a_1\ \text{MC}$
- $h_s(\text{MC}) = 1.5 + a_2\ \text{MC}$

2. modelo quadrático, com limites:
   - $h_i(\text{MC}) = 1.5 + a_1\ \text{MC} + b_1\ \text{MC}^2$
   - $h_s(\text{MC}) = 1.5 + a_2\ \text{MC} + b_2\ \text{MC}^2$
3. modelo exponencial, com limites:
   - $h_i(\text{MC}) = 1.5\ e^{c_1 \text{MC}}$
   - $h_s(\text{MC}) = 1.5\ e^{c_2 \text{MC}}$

com $a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{R}$. Estas constantes são calculadas de modo a que a percentagem de observações atípicas declaradas seja aproximadamente de 0.7% [39] e, para tal, são testadas diversas distribuições de modo a que a escolha seja o mais adequada possível, tendo em conta esse critério.

Para descobrir as constante acima, é necessário estabelecer os quantis da cerca. Uma vez que a regra é declarar 0.7% das observações como possíveis *outliers*, pode-se então definir $\alpha = 0.0035$ e $\beta = 0.9965$ como os valores para os quantis que definem a cerca. Deste modo, as constantes para o modelo linear podem ser calculadas como:

$$\begin{cases} Q_{0.25} - (1.5 + a_1\ \text{MC})\text{AIQ} \approx Q_\alpha \\ Q_{0.75} - (1.5 + a_2\ \text{MC})\text{AIQ} \approx Q_\beta \end{cases} \Leftrightarrow \begin{cases} a_1 \approx \frac{1}{\text{MC}} \left(\frac{Q_{0.25} - Q_\alpha}{\text{AIQ}} - 1.5 \right) \\ a_2 \approx \frac{1}{\text{MC}} \left(\frac{Q_\beta - Q_{0.75}}{\text{AIQ}} - 1.5 \right) \end{cases} \tag{3.24}$$

De modo análogo pode-se obter as constantes para o modelo quadrático e para o modelo exponencial.

$$\begin{cases} Q_{0.25} - (1.5 + a_1\ \text{MC} + b_1\ \text{MC}^2)\text{AIQ} & \approx Q_\alpha \\ Q_{0.75} - (1.5 + a_2\ \text{MC} + b_2\ \text{MC}^2)\text{AIQ} & \approx Q_\beta \\ b_1, b_2 \neq 0 \end{cases} \Leftrightarrow \begin{cases} a_1 + b_1 \text{MC} & \approx \frac{1}{\text{MC}} \left(\frac{Q_{0.25} - Q_\alpha}{\text{AIQ}} - 1.5 \right) \\ a_2 + b_2 \text{MC} & \approx \frac{1}{\text{MC}} \left(\frac{Q_\beta - Q_{0.75}}{\text{AIQ}} - 1.5 \right) \\ b_1, b_2 \neq 0 \end{cases} \tag{3.25}$$

No caso particular de $a_1, a_2 = 0$, o modelo quadrático reduz-se a:

$$\begin{cases} b_1 \approx \frac{1}{\text{MC}^2} \left(\frac{Q_{0.25} - Q_\alpha}{\text{AIQ}} - 1.5 \right) \\ b_2 \approx \frac{1}{\text{MC}^2} \left(\frac{Q_\beta - Q_{0.75}}{\text{AIQ}} - 1.5 \right) \\ b_1, b_2 \neq 0 \end{cases} \tag{3.26}$$

Para o modelo de cerca exponencial, tem-se:

$$\begin{cases} Q_{0.25} - (1.5\ e^{c_1 \text{MC}})\text{AIQ} & \approx Q_\alpha \\ Q_{0.75} - (1.5\ e^{c_2 \text{MC}})\text{AIQ} & \approx Q_\beta \end{cases} \Leftrightarrow \begin{cases} e^{c_1 \text{MC}} & \approx \frac{Q_{0.25} - Q_\alpha}{1.5\text{AIQ}} \\ e^{c_2 \text{MC}} & \approx \frac{Q_\beta - Q_{0.75}}{1.5\text{AIQ}} \end{cases} \Leftrightarrow \begin{cases} c_1 & \approx \frac{1}{\text{MC}} \ln\left(\frac{Q_{0.25} - Q_\alpha}{1.5\text{AIQ}}\right) \\ c_2 & \approx \frac{1}{\text{MC}} \ln\left(\frac{Q_\beta - Q_{0.75}}{1.5\text{AIQ}}\right) \end{cases} \tag{3.27}$$

Dos três modelos, o mais utilizado, por apenas conter um parâmetro para estimar (para cada lado da cerca) e ser aquele que apresenta um melhor desempenho nas várias distribuições testadas em [39], é o exponencial.

Por motivos de robustez e para facilidade do modelo, os valores obtidos por estimativa para $c_1 = -3.79$ e $c_2 = 3.87$ foram arredondados, por defeito, para $c_1 = -4$ e $c_2 = 3$, de modo a que a cerca seja menor e, desse modo, mais robusto o modelo. Estes valores não são iguais, uma vez que, no estudo para os obter, apenas foram consideradas distribuições simétricas ou enviesadas à direita. De modo equivalente, para distribuições enviesadas à esquerda, as constantes são definidas por $c_1 = -3$ e $c_2 = 4$.

Assim, as cercas para os *box plots* ajustados são dadas, para MC > 0 e para MC < 0, respetivamente, por:

$$[Q_{0.25} - 1.5e^{-4\text{MC}}; Q_{0.75} + 1.5e^{3\text{MC}}] \tag{3.28}$$

$$[Q_{0.25} - 1.5e^{-3\text{MC}}; Q_{0.75} + 1.5e^{4\text{MC}}] \tag{3.29}$$

Retomando a definição de atipicidade de Stahel-Donoho robusta de uma observação x_i, definida em 3.16 e não levando em consideração as direções, tem-se:

$$\text{SDO}_i = \frac{|x_i - \text{med}(\mathbf{X})|}{\text{MAD}(\mathbf{X})} \tag{3.30}$$

Partindo deste conceito, pode-se introduzir então a definição de atipicidade ajustada (AO).

Definição 3.2.2 (Atipicidade Ajustada)**.** *Seja $\boldsymbol{x} = (x_1, \ldots, x_n)$ uma amostra univariada de tamanho n. A atipicidade ajustada (em inglês Adjusted Outlyingness) de x_i ao valor numérico dado por:*

$$AO(\boldsymbol{x}_i) = AO_i = \begin{cases} \frac{x_i - med(X)}{w_2 - med}, & x_i > med(\boldsymbol{x}) \\ \\ \frac{med(X) - x_i}{med - w_1}, & x_i < med(\boldsymbol{x}) \end{cases} \tag{3.31}$$

onde w_1 e w_2 referem-se, respetivamente, ao limite inferior e ao limite superior da cerca do box plot ajustado.

Esta definição permite que, por exemplo, duas observações que estejam à mesma distância da mediana, uma à esquerda e outra à direita, tenham valor AO diferente. Isto deve-se ao facto de, com escalas diferentes, o mesmo valor ter significados distintos e uma das observações ser declarada como atípica e a outra, com a mesma distância

da mediana, não ser considerada como *outlier*.

Tal como a *Stahel-Donoho Outlyingness*, a AO é invariante quanto à localização e quanto à escala, pois não é afetada pela mudança do centróide dos dados ou escala dos mesmos [39].

Generalizando para observações multivariadas o conceito de *Adjusted Outlyingness* para uma matriz de dados $d-$dimensionais $\mathbf{X}_{(n\times d)}$, onde $\mathbf{x}_i$ corresponde a uma observação, $i \in \{1, \ldots, n\}$, tem-se:

$$\mathrm{AO}_i = \mathrm{AO}(\mathbf{x}_i) = \sup_{\mathbf{a}\in\mathcal{S}_d} (\mathbf{a}^T \mathbf{x}_i) \tag{3.32}$$

Tal como acontece no estimador SDO, não é possível calcular as projeções sobre todas as direções na AO. Depois de calculado o valor de atipicidade ajustado para cada observação, pode-se então decidir se ela é ou não um *outlier*. No caso da AO, em que se está a assumir que os dados podem não ser provenientes de uma distribuição simétrica (por exemplo, a distribuição Normal), ou seja, é uma abordagem livre de distribuição, logo a distribuição das AO não tem de seguir uma distribuição χ^2_d, como acontece quando se assume que eles são provenientes de uma distribuição Normal $d-$ dimensional; contudo, esta será enviesada à direita e limitada à esquerda em zero.

Depois de calculados as AO para o conjunto de dados em $\mathbf{X}$, pode ser construído o *box plot* ajustado com esses valores e declarar uma observação multivariada como atípica se ultrapassar o limite superior da cerca do *outlier* ajustado, sendo o valor de corte (*cut off value*) definido como:

$$cv = Q_{0.75} + 1.5e^{3\mathrm{MC}}\mathrm{AIQ} \tag{3.33}$$

Assim, as observações x_i que tenham valor de AO maior ao cv são declarados como *outliers*.

3.3 Aplicabilidade de Métodos Baseados em Distância Robusta a Dados Composicionais

3.3.1 Aplicabilidade à Abordagem Comediana

Para aplicar a Abordagem Comediana (do inglês *Comedian Approach*) aos dados composicionais, é necessário verificar que todas as condições se verificam para as transformações composicionais. Neste caso em particular, a transformação utilizada será a clr.

Em primeiro lugar irá demonstrar-se que as medidas utilizadas na Abordagem Comediana podem ser calculadas com os dados clr-transformados.

Seja $\mathbf{X}_{(n\times d)}$ uma matriz de dados composicionais. Para qualquer composição $\mathbf{x}_i = [x_{i1}x_{i2}\ldots x_{id}]$ (linha da matriz $\mathbf{X}$), a transformação clr conduz a ter:

$$\begin{aligned}\text{clr}(\mathbf{x}_i) &= \text{clr}\ (x_{i1},\ldots,x_{id}) \\ &= \left(\ \ln\frac{x_{i1}}{\sqrt[d]{\prod_{k=1}^{d}x_k}},\ldots,\ln\frac{x_{id}}{\sqrt[d]{\prod_{k=1}^{d}x_k}}\ \right) \\ &= \left(\ \ln\frac{x_{i1}}{\text{gm}(\mathbf{x}_i)},\ldots,\ln\frac{x_{id}}{\text{gm}(\mathbf{x}_i)}\ \right)\end{aligned}$$

A mediana da $j-$ésima componente, X_j, dos dados clr-transformados é dada por:

$$\text{med}(X_j) = \text{med}\left(\ \ln\frac{x_{1j}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nj}}{\text{gm}(\mathbf{x}_n)}\ \right)$$

Deste modo, pode ser definida a comediana das variáveis clr-transformadas como:

$$\begin{aligned}\text{COM}(X_j,X_k) &= \text{med}\Big((X_j-\text{med}(X_j))(X_k-\text{med}(X_k))\Big) \\ &= \text{med}\Big(\ \Big(\ \Big(\ln\frac{x_{1j}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nj}}{\text{gm}(\mathbf{x}_n)}\Big)-\text{med}\Big(\ln\frac{x_{1j}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nj}}{\text{gm}(\mathbf{x}_n)}\Big)\ \Big) \\ &\quad \Big(\ \Big(\ln\frac{x_{1k}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nk}}{\text{gm}(\mathbf{x}_n)}\Big)-\text{med}\Big(\ln\frac{x_{1k}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nk}}{\text{gm}(\mathbf{x}_n)}\Big)\ \Big)\Big) \\ &= \text{med}\Big(\ (\ x'_{1j},\ldots,x'_{dj}\ )*(\ x'_{1k},\ldots,x'_{dk}\ )\ \Big)\end{aligned}$$

onde $x'_{ij} = \ln\frac{x_{ij}}{\text{gm}(\mathbf{x}_i)} - \text{med}\Big(\ln\frac{x_{1j}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nj}}{\text{gm}(\mathbf{x}_n)}\Big)$.

De modo análogo pode verificar-se que a transformação clr pode ser aplicada na MAD:

$$\begin{aligned}\text{MAD}(X_j) &= \text{med}\Big(|X_j-\text{med}(X_j)|\Big) \\ &= \text{med}\Big(\ \Big|\ \Big(\ln\frac{x_{1j}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nj}}{\text{gm}(\mathbf{x}_n)}\Big)-\text{med}\Big(\ln\frac{x_{1j}}{\text{gm}(\mathbf{x}_1)},\ldots,\ln\frac{x_{nj}}{\text{gm}(\mathbf{x}_n)}\Big)\ \Big|\ \Big) \\ &= \text{med}\ (\ |x'_{1j}|,\ldots,|x'_{dj}|\ )\end{aligned}$$

Após provar que a comediana e a MAD podem ser calculadas em dados clr-transformados, irá estabelecer-se a correlação mediana.

$$\begin{aligned}\delta(X_j,X_k) &= \frac{\text{COM}(X_j,X_k)}{\text{MAD}(X_j)\ \ \text{MAD}(X_k)} \\ &= \frac{\text{med}\Big(\ (\ x'_{1j},\ldots,x'_{dj}\ )*(\ x'_{1k},\ldots,x'_{dk}\ )\ \Big)}{\text{med}\ (\ |x'_{1j}|,\ldots,|x'_{dj}|\ )\ \text{med}\ (\ |x'_{1k}|,\ldots,|x'_{dk}|\ )} \\ &= \frac{\text{med}\ (\ x'_{1j}x'_{1k},\ldots,x'_{dj}x'_{dk}\ )}{\text{med}\ (\ |x'_{1j}|,\ldots,|x'_{dj}|\ )\ \text{med}\ (\ |x'_{1k}|,\ldots,|x'_{dk}|\ )}\end{aligned}$$

Uma vez que existe a comediana entre duas variáveis numa matriz clr-transformada, pode-se estabelecer, de modo análogo, a matriz comediana, $\mathbf{COM}(\mathbf{X})$, em que cada entrada a_{jk} corresponde ao valor de $\text{COM}(X_j, X_k)$, com $j, k \in \{1, \ldots, d\}$. Recorde que $a_{jj} = \text{COM}(X_j, X_j) = \text{MAD}^2(X_j)$.

De modo semelhante, pode-se estabelecer a matriz de correlação comediana para a transformação clr.

$$
\begin{aligned}
\boldsymbol{\delta} &= \mathbf{D}\,\mathbf{COM}(\mathbf{clr}(\mathrm{X}))\,\mathbf{D} \\
&= \text{diag}\Big(\text{MAD}(X_j)^{-1}\Big) \begin{pmatrix} a_{11} & \ldots & a_{1d} \\ \vdots & \vdots & \vdots \\ a_{d1} & \ldots & a_{dd} \end{pmatrix} \text{diag}\Big(\text{MAD}(X_j)^{-1}\Big) \\
&= \text{diag}\Big(\text{MAD}(X_j)^{-1}\Big) \begin{pmatrix} \frac{a_{11}}{\text{MAD}(X_1)} & \cdots & \frac{a_{1d}}{\text{MAD}(X_d)} \\ \vdots & \vdots & \vdots \\ \frac{a_{d1}}{\text{MAD}(X_1)} & \cdots & \frac{a_{dd}}{\text{MAD}(X_d)} \end{pmatrix} \\
&= \begin{pmatrix} \frac{a_{11}}{\text{MAD}^2(X_1)} & \cdots & \frac{a_{1d}}{\text{MAD}(X_1)\ \text{MAD}(X_d)} \\ \vdots & \vdots & \vdots \\ \frac{a_{d1}}{\text{MAD}(X_d)\ \text{MAD}(X_1)} & \cdots & \frac{a_{dd}}{\text{MAD}^2(X_d)} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \cdots & \frac{a_{1d}}{\text{MAD}(X_1)\ \text{MAD}(X_d)} \\ \vdots & \vdots & \vdots \\ \frac{a_{d1}}{\text{MAD}(X_d)\ \text{MAD}(X_1)} & \cdots & 1 \end{pmatrix}
\end{aligned}
$$

onde $\mathbf{D}$ é uma matriz $\text{diag}\Big(\text{MAD}(X_1)^{-1}, \ldots, \text{MAD}(X_d)^{-1}\Big) = \text{diag}\Big(\text{MAD}(X_j)^{-1}\Big)$.

Uma vez que todas as medidas e matrizes da Abordagem Comediana são possíveis de calcular com dados clr-transformados, as estimativas robustas para a localização, $\mathbf{m}$, e para a dispersão, $\mathbf{S}$, podem ser calculadas e, desse modo, utilizando a distância robusta de Mahalanobis, os eventuais *outliers* do conjunto de dados clr-transformados detetados.

3.3.2 Aplicabilidade à Atipicidade Ajustada

Ao contrário da Abordagem Comediana, na qual foi utilizada a composição clr, pela sua vantagem no que à interpretabilidade dos resultados diz respeito, na Atipicidade Ajustada, essa transformação não pode ser utilizada neste método. A particularidade de a soma de uma observação, com a transformação clr, somar zero impede a utilização deste método.

Seja **x** uma observação de tamanho d. A sua transformação clr corresponde a:

$$\mathbf{y} = \text{clr}(\mathbf{x}) = (y_1, \ldots, y_d), \text{ com } y_j = \ln \frac{x_j}{\sqrt[d]{\prod_{k=1}^{d} x_k}} \tag{3.34}$$

Então, obtém-se:

$$\sum_{j=1}^{d} y_j = y_1 + \cdots + y_d = 0 \tag{3.35}$$

como demonstrado na equação (2.28).

Desde modo, as observações de uma composição clr têm uma grande consequência: não é possível considerar somente um dos coeficientes clr para análise sem levar os demais em consideração. Isto traduz-se numa limitação numa análise univariada, por exemplo [1].

Uma implicação prática da soma coeficientes desta transformação ser zero é que, numa matriz com n observações onde todas elas somam zero, não se terá independência linear.

Definição 3.3.1 (Dependência Linear)**.** *Seja* $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}, \boldsymbol{u}_j \in \mathbb{R}^d$.

Diz-se que o conjunto de vetores é linearmente dependente se existem escalares $\alpha_1, \ldots, \alpha_n$*, não todos nulos, tais que:*

$$\sum_{i=1}^{n} \alpha_i \boldsymbol{u}_i = \alpha_1 \boldsymbol{u}_1 + \cdots + \alpha_n \boldsymbol{u}_n = \boldsymbol{0} \tag{3.36}$$

Assim, se existirem pelo menos dois vetores que sejam combinação linear de um outro, eles não serão linearmente independentes.

Proposição 3.3.1. *Seja* $M \in M_{(n\times n)}(\mathbb{R})$.
Então $det(M) = 0$ *se e só se as colunas de* M *são linearmente dependentes em* $\mathbb{R}^n$ *(com a estrutura de espaço vetorial sobre* $\mathbb{R}$*).*

Assim, se as colunas são linearmente independentes, pelo menos alguma é combinação linear de outras. Desse modo, pode-se concluir que $\det(M) = 0$, usando as propriedades dos determinantes relativamente às colunas [40] e, desse modo, a caraterística será menor que o número de colunas da mesma (relação análoga para as linhas, uma vez que $\det(M) = \det(M^T)$).

No caso em que a matriz não é quadrada, isto é $n \times n$, a caraterística da matriz é dada por:

$$\text{car}M \leq \min\{n, d\} \tag{3.37}$$

onde n corresponde ao número de linhas e d ao número de colunas da matriz. No caso de haver linhas ou colunas linearmente dependentes ter-se-á:

$$\text{car}M < \min\{n, d\} \tag{3.38}$$

Assim, a matriz resultante dos coeficientes clr não é *full rank*, isto é, a sua caraterística não corresponde a car$M = \min\{n, d\}$, e, por isso, a correspondente matriz de covariância é singular (o determinante é zero e, por isso, não invertível). Neste caso em particular, a caraterística da matriz das observações clr transformadas é car$M = \min\{n, d\} - 1$. Habitualmente, o número de componentes é inferior ao número de observações e, nesse caso, pode ser dado por car$M = d - 1$.

De uma forma gráfica, é possível visualizar o problema desta composição. Como é possível visualizar na figura 2.3, o espaço de uma composição com $d = 2$ elementos é um espaço de dimensão 1 e o espaço de uma composição com $d = 3$ elementos é um espaço de dimensão 2 e, desse modo, não é possível gerar todo o espaço com os coeficientes clr.

Pelo facto de as composições não gerarem o espaço completo e formarem um hiperplano, a opção recai sobre as coordenadas ilr para o método da Atipicidade Ajustada. Estas têm a vantagem de formar uma base ortonormal no hiperplano formado pelos coeficientes clr e existe uma infinidade de possibilidades de definir tal sistema ortogonal [1].

Os coeficientes clr são centrados, mas o mesmo não se passa com as coordenadas ilr. Por esse motivo, o uso da Atipicidade Ajustada (AO), inspirada no *box plot* ajustado, é uma escolha que não pressupõe qualquer comportamento univariado sobre os dados.

No cálculo da Atipicidade Ajustada é utilizada a *medcouple* (MC) - medida de enviesamento de uma amostra univariada. Uma vez que, para utilizar a MC, não é necessário ter conhecimento prévio sobre a distribuição e é sempre possível de a calcular e, desse modo, o uso de dados composicionais não constituí um obstáculo ao uso desta metodologia.

Uma vez que a escolha do *pivot* implica que a interpretação feita dos resultados seja diferente e de modo a que nenhuma escolha seja preterida em relação às demais, optou-se por calcular a AO em relação a todas as transformações ilr possíveis, isto é, irá realizar-se as d pivotagens e, daí, tirar as devidas conclusões. Além disso, uma vez que este método escolhe as direções segundo as quais será realizada a procura de *outliers* de forma aleatória, de modo a tentar evitar que o acaso de uma direção classifique uma observação como atípica, irá repetir-se este processo de procura diversas vezes e, assim, ter resultados mais robustos.

Capítulo 4

Metodologias Gráficas

Quando se tenta descrever um conjunto de dados multivariados, em particular, um conjunto de dados composicionais, as medidas descritivas, como o vetor de médias ou a matriz de covariâncias, não são por si por si só informativas, uma vez que, por exemplo, um vetor de médias ou qualquer outro na sua classe de equivalência traduzem exatamente a mesma informação, devido às propriedades do simplex. Contudo, com os métodos usuais, eles são vistos como vetores diferentes, uma vez que é analisado o seu valor absoluto e, por isso, um vetor e o mesmo vetor a menos de uma constante $c \in \mathbb{R}^+$, correspondem a vetores diferentes, mas na perspetiva composicional, eles pertencem à mesma classe de equivalência.

A representação gráfica de um conjunto de dados é, sem dúvida, um método muito vantajoso em qualquer análise exploratória de dados, visto que permite a quem está a realizar a análise visualizar tendências nos dados [5]. No âmbito dos dados composicionais, as metodologias gráficas usualmente utilizadas são os diagramas ternários, os gráficos de dispersão univariados e os biplots. Além disso, pode-se também utilizar o *bagplot* para visualizar a dispersão dos dados.

4.1 Diagrama Ternário

Os diagramas ternários são uma representação de dados composicionais, com a restrição das observações representadas terem soma constante. Com a operação de fecho, é possível que todas as observações de um conjunto de dados tenham a mesma soma das partes. Para representação nos diagramas ternários, as constantes k geralmente utilizadas são 1 e 100, correspondendo respetivamente a proporções e percentagens.

Um diagrama ternário corresponde a uma representação no plano bidimensional do simplex de dimensão 3, $\mathcal{S}^3$ (representado na figura 2.3b), e que é, frequentemente utilizado na área das geociências. Quando os dados são caraterizados por mais de três elementos, estes são amalgamados em três componentes [41] (por exemplo, agrupar elementos presentes nos solos pelos suas caraterísticas: semimetais, metais alcalinoterrosos, metais alcalinos, etc).

Definição 4.1.1 (Diagrama Ternário). *Um diagrama ternário é uma representação dos pontos em $\mathcal{S}^3$ num triângulo equilátero de vértices X_1, X_2, X_3. Dada uma composição $\boldsymbol{x} = (x_1, x_2, x_3)$, esta é representada a uma distância x_1 do lado oposto do 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.*

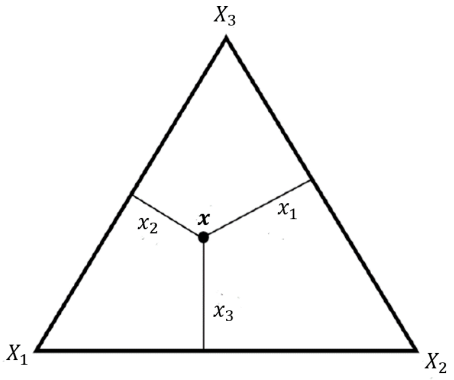


Figura 4.1: Diagrama ternário.

Na figura 4.1, as bordas do diagrama ternário correspondem às marginais (relativamente aos vértices opostos às bordas) nulas na representação em $\mathbb{R}^3$ e os vértices representam composições com uma parte igual a k e as restantes iguais a zero. No que diz respeito ao centro do conjunto de dados ele é representado no baricentro do diagrama (ponto de intersecção das medianas de um triângulo).

Para construir um diagrama ternário em coordenadas cartesianas, é necessário fixar os vértices do triângulo. A sua ordenação deverá ser feita no sentido contrário ao dos ponteiros do relógio. Assim, $X_1 = (u_0, v_0)$ que corresponde à origem, $X_2 = (u_0 + 1, v_0)$ e $X_3 = (u_0 + \frac{1}{2}, v_0 + \frac{\sqrt{3}}{2})$, que na primeira coordenada corresponde ao ponto médio de u_0 e $u_0 + 1$ e a segunda coordenada a $v_0 + \tan 60° = v_0 + \frac{\sqrt{3}}{2}$. Assim, para fazer a representação de uma composição de três partes, $\mathbf{x} = (x_1, x_2, x_3)$ com $x_1 + x_2 + x_3 = k$, é necessário conhecer as suas coordenadas cartesianas, que são obtidas pela combinação linear convexa das coordenadas dos vértices, isto é [5]:

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

Exemplo 4.1.1. *Tome-se, uma vez mais, como exemplo, o conjunto de dados 'Arctic Lake'. Seja $\boldsymbol{x} = (53.4, 36.8, 9.8)$ uma composição desse conjunto de dados constituída por areia, lodo e argila respetivamente. Neste caso em particular $k = 53.4 + 36.8 + 9.8 = 100$ e , por isso, cada lado do diagrama pode ser dividido em 100 partes iguais e, com as linhas paralelas respetivamente ao lado oposto a cada vértice, assinalar o valor correspondente a cada coordenada. Deste modo, tem-se:*

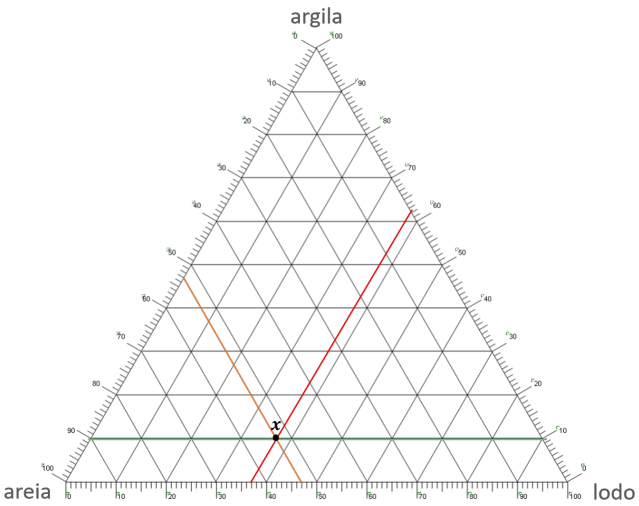


Figura 4.2: Representação de uma composição num diagrama ternário.

Na figura 4.2, marcaram-se as linhas auxiliares correspondentes às constantes para cada uma das partes da composição. Assim, a laranja está assinalada a linha da constante 53.4% *para a areia, a vermelho está assinalada a linha da constante* 36.8% *para a lodo e a verde está assinalada a linha da constante* 9.8% *para a argila. Na prática, apenas seriam necessárias duas dessas linhas, pois, após definida a constante de fecho, a representação da terceira parte da componente fica completamente estabelecida.*

Para interpretar corretamente um diagrama ternário, pode-se recorrer da propriedade de que os segmentos ortogonais que ligam um determinado ponto do diagrama ternário aos seus lados - as alturas desse ponto - têm soma constante igual a k [4]. Por consequência, se uma dada composição tiver a sua representação muito próxima de uma aresta do triângulo isso implica a dominância das partes que formam essa aresta e, de modo equivalente, se uma certa composição estiver próxima de um vértice, isso significa a dominância da parte associada a esse vértice. Deste modo, para analisar dados composicionais representados num diagrama ternário, deverá estar-se atento aos seguintes padrões [4]:

i. se as composições concentram-se num vértice isso indica a dominância da parte associada a esse vértice;

ii. se as composições distribuem-se ao longo de uma aresta isso indica a dominância das partes associadas a essa aresta;

iii. se as composições concentram-se em torno do baricentro do simplex isso implica que as partes representadas têm proporções aproximadamente iguais;

iv. se as composições formam um padrão linear paralelo a um dos lados isso significa que as proporções da parte associada ao vértice oposto nas composições é (aproximadamente) constante;

v. se as composições formam um padrão linear (aproximadamente) perpendicular a um dos lados isso significa que as partes associadas a esse lado são (aproximadamente) proporcionais (reduzida variabilidade relativa);

vi. se as composições estiverem dispersas no diagrama, indica que as partes apresentam elevada variabilidade relativa entre si.

Exemplo 4.1.2. *De modo análogo, é possível representar todo o conjunto de dados 'Arctic Lake' num diagrama ternário. Neste caso em particular, opta-se por* $k = 1$*, mas qualquer outra constante daria origem a um diagrama ternário equivalente.*

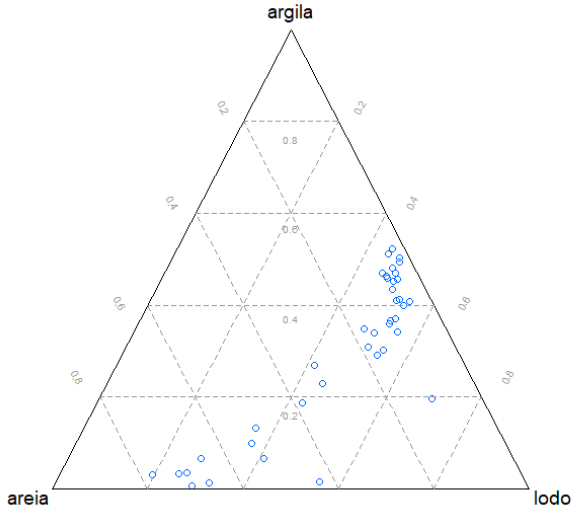


Figura 4.3: Representação do conjunto de dados 'Arctic Lake' num diagrama ternário.

Assim, por observação do diagrama ternário da figura 4.3, pode-se concluir que cerca de metade das composições estão muito do lado que tem como extremidades a argila e o lodo e, por isso, uma maior proporção dessas partes e uma pequena proporção de areia; as restantes observações, ligeiramente mais dispersas, estão próximas do lado cujas extremidades são a areia e lodo e, por isso, uma maior proporção dessas partes em oposição à argila, que aparece em proporções muito pequenas nessas composições.

4.2 Gráfico de Dispersão Univariado

Os gráficos de dispersão univariados consideram o conjunto de dados composicional ilr-transformados no espaço Euclediano. Assim, cada parte da composição é representada de forma univariada paralelamente com a primeira coordenada pivô para cada uma das observações. Há duas possibilidades de representação deste tipo de gráficos [1]:

- gráficos de dispersão de cada uma das variáveis ilr-transformadas ($z_1^{(l)}, l = 1, \ldots, d$), um gráfico para cada pivô. No eixo dos xx está o índice das composições colocadas aleatoriamente (figura 4.4);

- gráficos de dispersão das coordenadas paralelas, onde as partes em cada observação são unidas por segmentos de reta entre as diversas variáveis (figura 4.5).

Nas figuras 4.4 e 4.5 ilustram-se os gráficos supra referidos com os dados do conjunto 'Arctic Lake'.

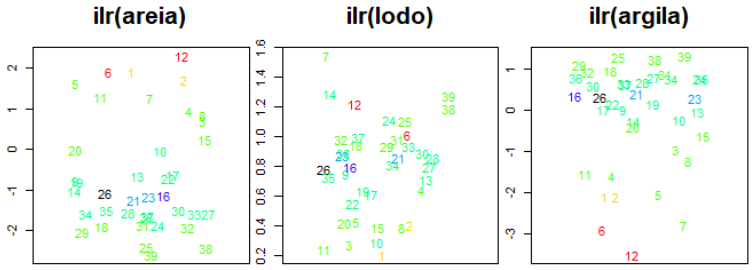


Figura 4.4: Gráficos de dispersão univariados para os elementos constituintes do solo.

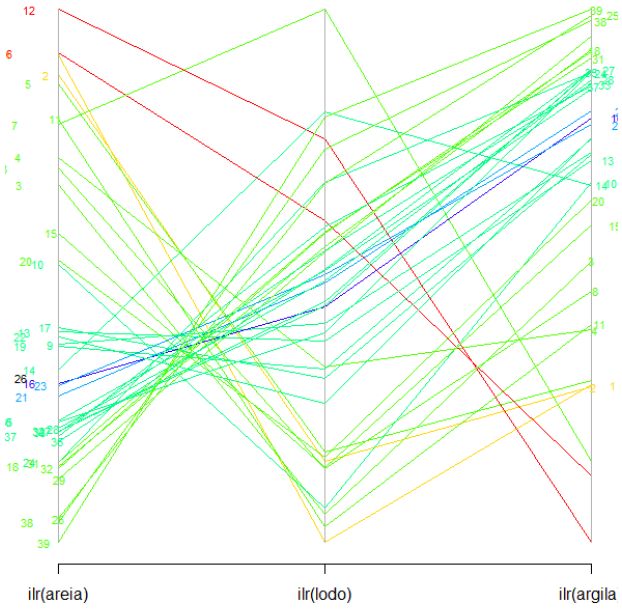


Figura 4.5: Gráficos de dispersão das coordenadas paralelas para os elementos constituintes do solo.

No gráfico da figura 4.4, a disposição das composições no eixo horizontal é arbitrária, de modo a não impor nenhum enviesamento, não tendo qualquer tipo de interpretação [1].

Quer no gráfico da figura 4.4, quer o da figura 4.5, as observações estão representadas numa gradação de cores. Isso acontece porque estes gráficos são feitos com recurso à *package* `mvoutlier` que, para a matriz das coordenadas pivô, $\mathbf{X}_j$ (matriz que considera a primeira coordenada da respetiva transformação ilr), calcula para cada linha

a mediana e, de seguida, calcula a distância de cada coordenada pivô à sua mediana. Além da mediana, calcula também o valor da média e se o valor da mediana for superior ao da média, significa que mais de 50% das distâncias calculadas têm valor superior à média e são assinaladas a vermelho. Isto significa que as proporções das observações em relação às restantes destaca-se. É também possível usar símbolos de acordo com os pontos de corte para os quantis $Q_{0.25}, Q_{0.5}$ e $Q_{0.75}$ e o valor de corte para o quadrado da distância de Mahalanobis. Observações que ultrapassam o valor de corte são, usualmente, assinaladas com uma cruz. Neste caso em particular, optou-se não utilizar a opção dos símbolos para a distância de Mahalanobis.

4.3 Bagplot

O *bagplot*, também designado de *box plot* bivariado, é uma generalização do *box plot* para dados bivariados. Este gráfico baseia-se também no conceito de contornos de profundidade. Assim, o contorno de profundidade k contém as observações que têm profundidade de localização maior ou igual a k.

O *bagplot* consiste em três polígonos concêntricos, chamados de bolsa (do inglês *bag*), cerca (do inglês *fence*) e ciclo (do inglês *loop*). O saco é o polígono mais interior e contém 50% dos dados, a cerca é a linha poligonal do polígono exterior e que separa os *inliers* (observações não atípicas) dos *outliers* e o ciclo é a região que contém os pontos exteriores ao saco mas dentro da cerca [43].

Exemplo 4.3.1. *Tome-se, uma vez mais, o conjunto de dados 'Arctic Lake'. Utilizando as variáveis 'Areia' e 'Lodo', é possível construir o respetivo bagplot.*

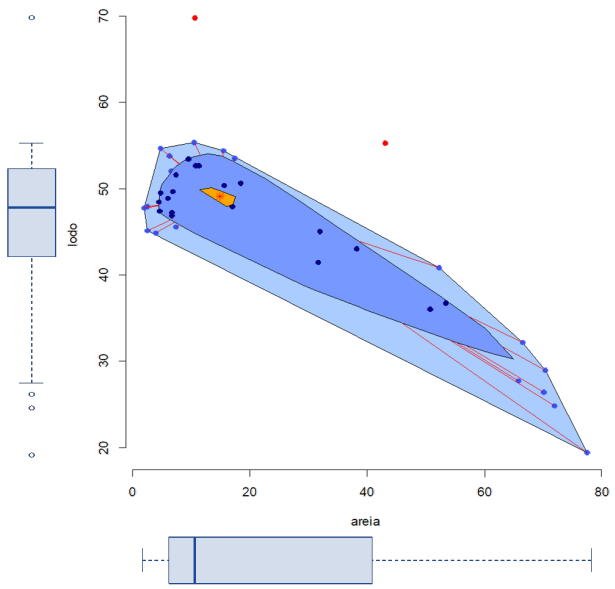


Figura 4.6: *Bagplot* e respetivos *box plots* para as variáveis Areia e Lodo do conjunto de dados 'Arctic Lake'.

Pela observação da figura 4.6, é possível observar o saco, região representada a azul escuro e que contém 50% das observações, o ciclo, representado a azul claro, e a cerca, que corresponde às linhas exteriores do polígono do ciclo. Neste *bagplot* são considerados dois *outliers* (pontos representados a vermelho), que são os pontos que estão no exterior da cerca.

Dentro do saco, representado por um asterisco vermelho está a mediana de Tukey - ponto com a maior profundidade do semiespaço, que está contida numa região laranja, definida pelos vértices do polígono da região central. Pelo tamanho do saco, pode-se concluir que existe uma grande variabilidade dos dados, mais da areia do que do lodo, e que a correlação entre a areia e o lodo é negativa, o que faz sentido, uma vez que ao aumentar-se a quantidade de areia presente numa amostra de solo, a quantidade de lodo será menor.

Tal como acontece no *box plot* univariado, o *bagplot* também permite visualizar diversas caraterística dos dados: a sua localização (a mediana de profundidade), dispersão (o tamanho do saco), correlação (a orientação do saco), enviesamento (a forma do saco e do ciclo) e as caudas (os pontos perto da fronteira do ciclo) e os *outliers* [43].

Capítulo 5

Aplicação de Metodologias de Deteção de *Outliers*

Os *outliers* são dados que se diferenciam drasticamente dos demais, por algum motivo, e que podem, eventualmente, causar anomalias nos resultados obtidos numa análise estatística que não os contemple como observações atípicas [47]. Também, por si só, despertam interesse uma vez que, sabendo que, alguma observação é tida como atípica em relação às restantes, irá procurar-se motivos para que justifiquem essas circunstâncias.

Neste Capítulo, o objetivo é, utilizando as metodologias numéricas descritas no Capítulo 3, aliadas a metodologias gráficas tipicamente usadas no âmbito dos dados composicionais, detetar potenciais *outliers* em dois conjuntos de dados reais. Serão considerados:

* Dados epidemiológicos: este primeiro conjunto de dados refere-se ao problema epidémico do HIV/SIDA numa área geográfica heterogénea - o arquipélago de Cabo Verde [11].

* Dados da qualidade dos solos: este segundo conjunto de dados é referente à qualidade do solo e a sua eventual contaminação por hidrocarbonetos aromáticos policíclicos, HAP, em Lisboa, Portugal [48].

5.1 Dados Epidemiológicos

5.1.1 Contextualização do Problema

Acredita-se que o Vírus da Imunodefiência Humana (VIH/HIV) tenha surgido no centro-oeste da África durante o início do século XX e oficialmente reconhecido pelo Centro de Controle de Prevenção de Doenças, nos Estados Unidos da América nos anos 80 [49]. O HIV é um vírus que invade o sistema imunitário e destrói as suas defesas, enfraquecendo, desse modo, a capacidade do organismo combater doenças e infeções e, em casos extremos, provocar a morte [50].

Uma confusão que é frequentemente associada a esta doença é entre HIV (Vírus da Imunodeficiência Humana) e SIDA (Síndrome de Imunodeficiência Adquirida). Na verdade, são diagnósticos diferentes – a infeção por VIH pode conduzir a uma doença ou síndrome, condição conhecida como SIDA. Deste modo, pode ter-se uma infeção por VIH sem adquirir SIDA, sendo muitas as pessoas com infeção por VIH que vivem durante anos sem desenvolver SIDA. A SIDA ocorre quando o VIH condiciona danos importantes ou severos no sistema imunitário e é uma condição complexa, com sintomas que variam de doente para doente, dentro dos quais se destacam a tuberculose, a pneumonia, outras infeções e, ainda, certos tipos de neoplasias (cancro) [51].

O HIV transmite-se, na maioria das vezes através de relações sexuais (quer por sexo vaginal, anal ou oral), mas também por partilha de seringas ou outro tipo de material perfurante e ainda de mãe para filho, durante a gravidez, nascimento ou amentação, uma vez que este vírus encontra-se nos fluídos corporais (sémen, fluídos vaginais e anais, sangue e leite materno). Assim, apesar do que muitos ainda erroneamente julgam o HIV não se transmite por beijos, saliva, contato com pele intata e saudável, partilha de itens pessoais (como talheres, sanitas, toalhas ou piscinas) ou contato com animais ou insetos [50].

Até ao momento, não existe uma cura ou vacina para a SIDA. Contudo, os tratamentos realizados com antirretrovirais (ARV) têm apresentado melhorias na qualidade da saúde dos doentes, prolongam a vida e reduzem o risco da transmissão do HIV. Apesar das vantagens que este tipo de tratamento tem nos pacientes, como não são uma cura, eles não voltarão a ser completamente saudáveis e têm efeitos secundários, além de serem tratamentos dispendiosos [52] e, por esse motivo, criam uma desigualdade no que diz respeito ao acesso ao tratamento nos diversos pontos do globo. Os países menos desenvolvidos, como os que se situam no continente Africano, quer pela falta de cuidados de saúde, quer pela falta de informações e conhecimento sobre como prevenir o HIV, bem como a carência económica generalizada, são os possuem maiores número de novos casos. Segundo dados do UNAIDS [1], Programa Conjunto da ONU para o HIV/SIDA, em 2018 houve 1.7 milhões de novos casos no mundo deste vírus, dos quais 1.08 milhões foram no continente africano, o que corresponde a 63.5% do total de novos casos.

Em relação à SIDA, e segundo o modelo epidemiológico SICA [54], a população é dividida em quatro grupos de indivíduos mutuamente exclusivos:

- indivíduos suscetíveis (S);
- indivíduos infetados pelo HIV sem sintomas clínicos de SIDA, mas capazes de transmitir o HIV a outros (I);
- indivíduos infetados por HIV em tratamento com carga viral remanescente baixa (estado crónico) (C);
- indivíduos infetados com HIV com sintomas clínicos de SIDA (A).

[1] dados disponíveis em `https://www.unaids.org/en`)

Neste estudo, pretende-se analisar, na perspetiva composicional, dados resultantes deste modelo SICA à realidade do arquipélago de Cabo Verde. Este país, de língua oficial portuguesa e com $4033\,\text{km}^2$ de área, situa-se a cerca de 570 km da costa da África Ocidental e é constituído por 10 ilhas, das quais apenas 9 são habitadas.

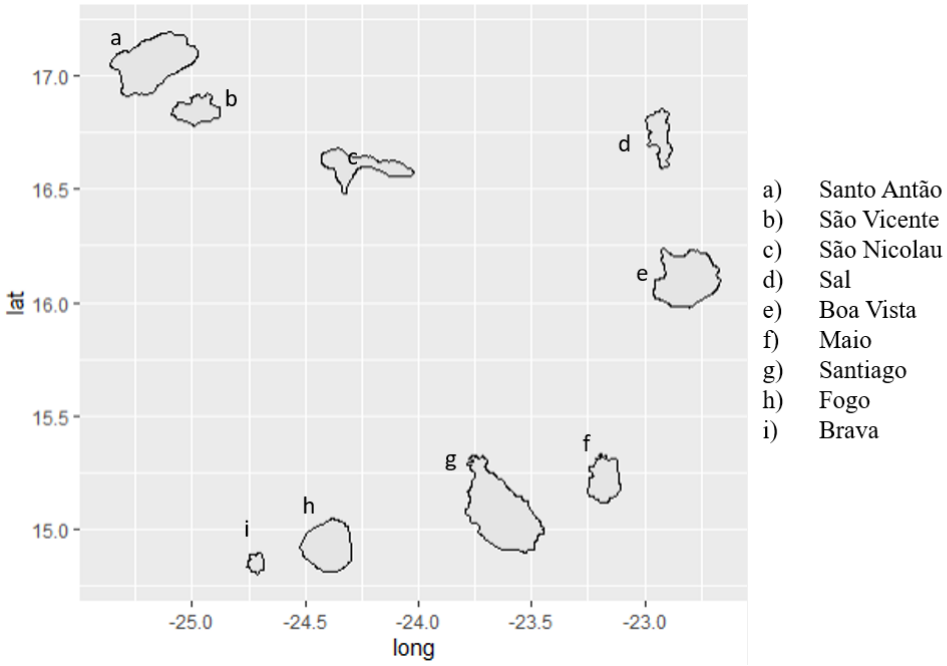


Figura 5.1: Ilhas habitáveis do Arquipélago de Cabo Verde.

Em 2015, Cabo Verde ocupava o 122º lugar no ranking de 187 países no Índice de Desenvolvimento Humano (IDH). A esperança média de vida, estimada em 71 anos, é a mais elevada da África subsaariana. Em 2011, 94% das crianças com menos um ano de idade foram vacinadas. Da mesma forma, o resultado da educação situa Cabo Verde no nível mais alto da África subsaariana: a taxa de alfabetização de adultos é estimada em 87%, embora ainda existam diferenças entre homens e mulheres [53].

De acordo com o Ministério da Saúde de Cabo Verde, em 2007, na subregião do Oeste Africano, Cabo Verde foi um dos países com os melhores indicadores de saúde da população, graças a um esforço persistente que consistiu na criação de infraestruturas, desenvolvimento da força de profissionais de saúde, organização dos serviços, alocação cuidadosa de recursos e leis que suportam a institucionalização do sistema de saúde, o que, para uma nação com poucos recursos financeiros, traduz-se num grande esforço. Contudo, este arquipélago ainda apresenta desigualdades no acesso às especialidades de saúde, particularmente nas suas ilhas e mais periféricas (Fogo, Brava e Santo Antão), que enfrentam maiores dificuldades [53].

Mesmo assim, Cabo Verde é tido como um exemplo a seguir no que diz respeito ao continente africano, pela medidas implementadas nos últimos anos de modo a evitar a propagação do HIV na população. Algumas dessas medidas foram a integração de conteúdos sobre o VIH nos programas curriculares de todas as escolas primárias e se-

cundárias, a solidariedade e cooperação com o Governo do Brasil através da doação de medicamentos antirretrovirais, a possibilidade de as mulheres gestantes poderem realizar o teste nas consultas de pré-natal e a ampliação dos testes anónimos e voluntários na população, entre outras [55]. Com uma população mais instruída e sensibilizada para o problema, a taxa de contágio tem vindo a decrescer nos últimos anos.

5.1.2 Análise e Discussão de Resultados

Relativamente às quantidades S, I, C, e A referidas anteriormente, a população das ilhas de Cabo Verde segundo dados de 2015, é composta de acordo com a tabela 5.1 ([54; 55]):

Ilhas	**S**	**I**	**C**	**A**	Total
Santo Antão	40388	10	93	9	40500
São Vicente	80763	32	186	19	81000
São Nicolau	12381	7	29	3	12420
Sal	33642	22	78	8	33750
Boa Vista	14404	10	33	3	14450
Maio	9657	5	16	2	9680
Fogo	35735	15	82	8	5840
Brava	5681	5	13	1	5700
Santiago	293084	303	676	87	294150

Tabela 5.1: Distribuição da população de Cabo Verde por ilhas e grupos em relação ao HIV, nas variáveis SICA

Como é possível observar na Tabela 5.1, as ilhas têm um número de habitantes muito diverso: enquanto que a ilha Brava tem apenas 5700 habitantes, Santiago, a maior ilha do arquipélago tem 294150 habitantes. Contudo, uma vez que a análise realizada será na perspetiva composicional, os valores absolutos, por si só, não são o foco de interesse.

Começa-se pela representação dos dados usando o diagrama ternário. Uma vez que, claramente, a proporção de susceptíveis é muito superior à dos demais grupos, não terá interesse um diagrama com a variável S. Na figura 5.2, representa-se o diagrama ternário com as variáveis I, C, A.

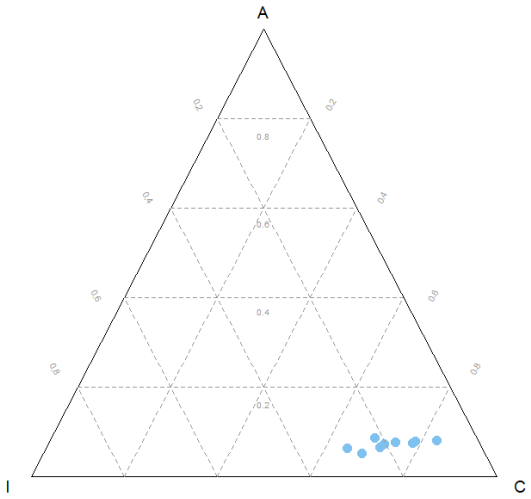


Figura 5.2: Diagrama ternário para as variáveis I, C, A.

Pela observação deste diagrama, conclui-se que existe uma maior concentração de indivíduos em estado crónico (C) nas 9 ilhas e que a proporção dos infetados com sintomas clínicos (A) é aproximadamente constante relativamente às outras duas.

Uma vez que os dados têm dimensão 3 quando utilizada a transformação ilr e dimensão 4 quando utilizada a transformação clr, não é possível uma representação visual como um todo num espaço de dimensão 2, isto é, no plano. Contudo, é possível recorrer a algumas metodologias gráficas (por exemplo, *box plot*, *bagplot*) que permitem ter uma ideia de como os dados se comportam, mesmo que não como um todo.

Para visualizar a distribuição dos dados numa perspetiva composicional comparando cada parte com o todo, tome-se os dados clr-transformados e a representação destes em *box plots* (figura 5.3) e *bagplots* (figura 5.4).

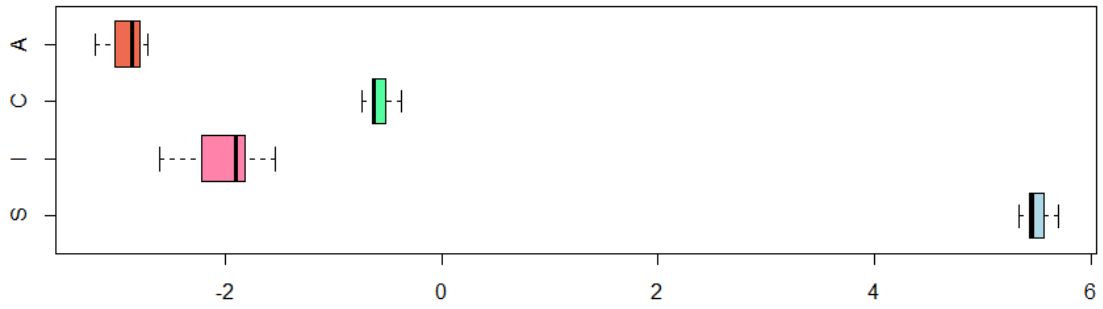


Figura 5.3: *Box plots* para os grupos de infeção, segundo o modelo SICA.

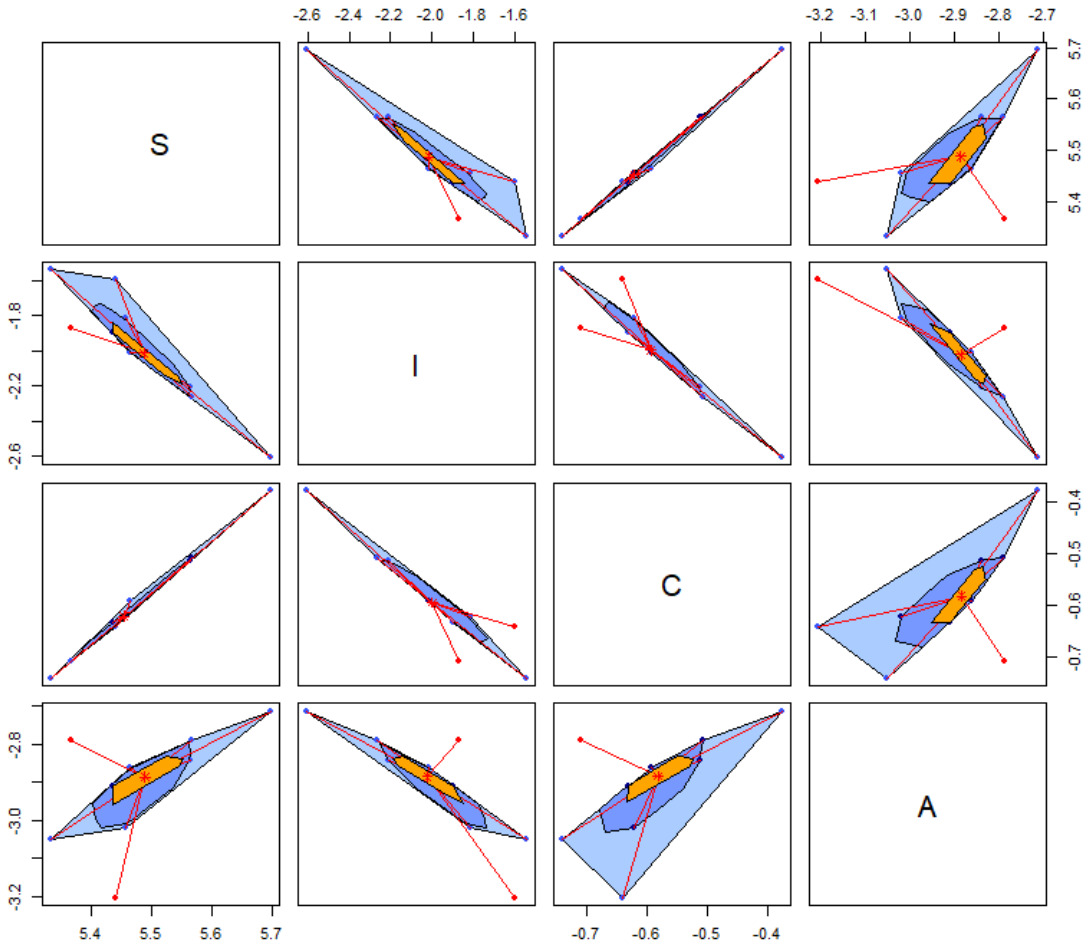


Figura 5.4: Bagplot bidimensional dos dados clr-transformados dos grupos do modelo epidimológico SICA.

Numa análise univariada, pela observação da figura 5.3 não se destacam log-razões clr atípicas. No entanto, numa análise bivariada, pela observação dos *bagplots* da figura 5.4 e após identificação dos pontos fora da cerca, é possível constatar que a ilha Maio é considerada como *outliers* em quatros dos seis *bagplots* (clr(S) vs. clr(I), clr(S) vs. clr(A), clr(I) vs. clr(C) e clr(C) vs. clr(A)), a ilha Brava duas vezes (clr(S) vs. clr(A) e clr(I) vs. clr(A)) e a ilha Sal uma vez (clr(I) vs. clr(A)). Assim, numa análise bivariada foi possível detetar observações atípicas não detetadas na análise univariada. Por exemplo, para o *bagplot* clr(C) vs. clr(A), existe uma observação atípica (ilha de Maio), significando que o par clr-transformado de infetados crónicos e infetados com sintomas foge ao padrão do restantes pares, que obedecem a um padrão de comportamento com correlação positiva.

As metodologias gráficas utilizadas até ao momento não permitem um consenso de que ilha ou ilhas podem ter um comportamento fora do padrão. Isto deve-se, em parte, ao facto que, dependendo da metodologia utilizada, o que é visto como atípico, depende em cada caso. Além disso, nenhuma das metodologias trata os dados como

um todo.

Deste modo, surge a necessidade de introduzir metodologias numéricas, que tratem os dados como um todo. As metodologias selecionadas, neste caso, são a Abordagem Comediana e o Estimador de Stahel-Donoho Ajustado, introduzidas no Capítulo 3.

A Abordagem Comediana baseia-se na distância robusta de Mahalanobis, após serem calculadas estimativas robustas para o vetor de médias e matriz de covariância, considerando a comediana como medida alternativa de dependência entre duas variáveis aleatórias. Neste caso, a transformação composicional escolhida é a clr, pelas vantagens que apresenta na sua interpretabilidade dos resultados. Contudo, de modo análogo, podia-se utilizar a transformação ilr.

Assim, após a implementação das diversas etapas desta metodologia, obteve-se, para estimativas robustas para a localização, $\mathbf{m}(\text{clr}(\mathbf{X}))$, e para a dispersão, $\mathbf{S}(\text{clr}(\mathbf{X}))$, são dadas por:

$$\mathbf{m}(\text{clr}(\mathbf{X})) = \begin{pmatrix} 5.437 \\ -1.915 \\ -0.636 \\ -2.907 \end{pmatrix}$$

$$\mathbf{S}(\text{clr}(\mathbf{X})) = \begin{pmatrix} 0.000164 & 0.000592 & 0.000111 & -0.000113 \\ 0.000592 & 0.003899 & 0.000538 & -0.000615 \\ 0.000111 & 0.000538 & 0.000141 & -0.000102 \\ -0.000113 & -0.000615 & -0.000102 & 0.000157 \end{pmatrix}$$

Partindo destas estimativas, a distância de Mahalanobis robusta, para um nível de confiança $\alpha = 0.025$, obtida para cada uma das ilhas do arquipélago de Cabo Verde está disposta na Tabela 5.2.

Ilhas	**Distância Robusta**
Santo Antão	3619.52
São Vicente	988.74
São Nicolau	104.83
Sal	0.67
Boa Vista	174.36
Maio	303.07
Fogo	720.17
Brava	1378.82
Santiago	921.463

Tabela 5.2: Distância robusta de Mahalanobis obtida para cada uma das ilhas, pela Abordagem Comediana.

Para estabelecer se alguma das observações é atípica, é necessário definir o valor

de corte (*cut off value*) que estabelece a partir de que distância uma ilha é considerada *outlier*. Neste caso em particular, o valor de corte é dado por:

$$\begin{aligned} cv &= 1.4826 \times \frac{\chi^2_{4;1-0.025} \times \text{med}(rd_1, \ldots, rd_n)}{\chi^2_{4;0.5}} \\ &= 1.4826 \times \frac{\chi^2_{4;0.975} \times \text{med}(3619.52, \ldots, 921.46)}{\chi^2_{4;0.5}} \\ &= 1.4826 \times \frac{11.14 \times 720.17}{3.36} \\ &= 3544.56 \end{aligned}$$

Comparando com os valores das distâncias robustas da tabela 5.2, pode-se proceder à classificação das observações como indicado na tabela 5.3.

Ilhas	**Distância Robusta**	*Outlier*?
Santo Antão	3619.52	Sim
São Vicente	988.74	Não
São Nicolau	104.83	Não
Sal	0.67	Não
Boa Vista	174.36	Não
Maio	303.07	Não
Fogo	720.17	Não
Brava	1378.82	Não
Santiago	921.463	Não

Tabela 5.3: Distância robusta de Mahalanobis e classificação obtida para cada uma das ilhas, pela Abordagem Comediana.

Por este método, apenas a ilhas de Santo Antão é considerada como atípica, com uma distância muito superior às restantes. Na prática, se outros valores superiores de α fossem tomados, haveria a possibilidade de não classificar como *outlier* a ilha de Santo Antão. Contudo, o valor de $\alpha = 0.025$ é o que se toma habitualmente na literatura para definir o quantil de corte (ver [56], [1]).

A outra metodologia numérica explorada foi Atipicidade Ajustada, que tem por base o Estimador de Stehel-Donoho, com a atribuição de pesos às diferentes observações e esse peso deverá ser tanto menor, quanto maior foi a sua atipicidade em relação a alguma direção, utilizando a *medcouple* como medida de assimetria e, desse modo, não pressupondo qualquer tipo de distribuição dos dados.

Uma vez que não é possível estabelecer um base ortonormada de dimensão 4 com a transformação clr, há a impossibilidade de usar essa transformação, pelo que será tomada, nesta metodologia, a transformação ilr. Assim, de modo a não desprezar qualquer das variáveis, realiza-se a pivotagem em relação a todas as partes possíveis e,

então, é realizada a deteção de *outliers* por via da Atipicidade Ajustada (do inglês *Adjusted Outlyingness*), para cada uma das coordenadas pivô. Além disso, uma vez que esta abordagem usa um conjunto de direções aleatórias (usualmente $d = 250p$, onde p designa o número de variáveis), optou-se por repetir o processo um determinado número de vezes, n_{iter}, de modo a que o acaso de uma direção tomada não influencie de forma demasiado pronunciada os resultados. Optou-se, também, por considerar o *box plot* enviesado à direita, $[Q_{0.25} - 1.5e^{-3\text{MC}}; Q_{0.75} + 1.5e^{4\text{MC}}]$, uma vez que o tipo de assimetria para a distribuição de distâncias são, em geral, desse género [32].

Ilhas	ilr(S)	ilr(I)	ilr(C)	ilr(A)
Santo Antão	0	0	0	0
São Vicente	0	0	0	0
São Nicolau	0	0	0	0
Sal	0	0	0	0
Boa Vista	0	1	0	0
Maio	0	0	0	1
Fogo	0	0	0	0
Brava	0	1	8	2
Santiago	0	0	0	0

Tabela 5.4: Número de vezes que cada ilha é declarada como *outlier*, para $n_{\text{iter}} = 10$.

Ilhas	ilr(S)	ilr(I)	ilr(C)	ilr(A)
Santo Antão	0	0	0	0
São Vicente	0	0	0	0
São Nicolau	0	0	0	0
Sal	0	0	0	0
Boa Vista	0	15	0	0
Maio	0	0	0	1
Fogo	0	0	0	0
Brava	0	15	72	20
Santiago	0	0	0	0

Tabela 5.5: Número de vezes que cada ilha é declarada como *outlier*, para $n_{\text{iter}} = 100$.

Ilhas	ilr(S)	ilr(I)	ilr(C)	ilr(A)
Santo Antão	0	0	0	0
São Vicente	0	0	0	0
São Nicolau	0	0	0	0
Sal	0	0	0	0
Boa Vista	0	163	0	0
Maio	0	0	0	43
Fogo	0	0	0	0
Brava	5	174	746	229
Santiago	0	0	0	0

Tabela 5.6: Número de vezes que cada ilha é declarada como *outlier*, para $n_{\text{iter}} = 1000$.

Pela observação das tabelas 5.4, 5.5 e 5.6, é possível perceber que a ilha Brava é declarada um número maior de vezes como atípica, sobretudo quando a análise resulta tomando os dados ilr transformados com pivotagem na variável C, mas quando se executa o processo 1000 vezes, é considerada também como tal um número considerável de vezes para os pivôs I e A, 17% e 23% respetivamente.

Pelas metodologias utilizadas, quer a Abordagem Comediana, quer a Atipicidade Ajustada, não foi possível identificar *outliers* comuns. De realçar que as duas abordagens utilizaram diferentes transformações e, visto que a interpretação dessas mesmas transformações é diferente, o que cada uma delas classifica como *outlier* tem um significado diferente. Por outro lado, convém relembrar que, de acordo com os autores, a Abordagem Comediana apresenta um melhor desempenho quando a dimensão é mais elevada.

Tomando a transformação clr e com a Abordagem Comediana, foi identificada a ilha de Santa Antão como *outlier*. Tomando a transformação ilr e com a Atipicidade Ajustada, foi identificada como atípica a ilha Brava.

Conhecendo a realidade do arquipélago de Cabo Verde, consegue-se compreender a dualidade de resultados:

- A ilha de Santo Antão é uma das mais ilhas periféricas em relação à ilha de Santigo, onde se localiza a capital do país, Praia, e que concentra a maior quantidade de serviços, incluindo os de saúde e, por isso, os acessos não são tão facilitados;
- A ilha Brava, em termos absolutos, é a que tem uma menor densidade populacional e, portanto, um caso num dos grupos I, C ou A, ao comparar um dos grupos com os demais grupos (que é o objetivo da transformação ilr), terá um peso maior do que em ilhas com mais população.

Utilizando as representações gráficas da primeira coordenada ilr, isto é, recorrendo aos gráficos de dispersão univariada (figura 5.5), poder-se-á, de algum modo, compreender o que tipifica estas duas ilhas.

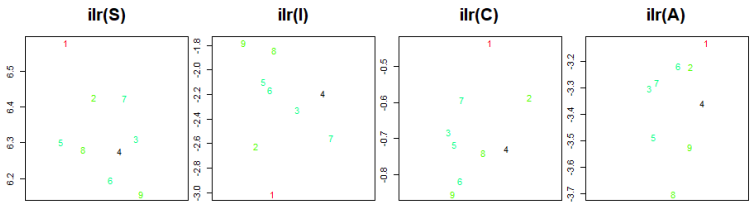


Figura 5.5: Gráficos de dispersão univariados para os grupos do modelo SICA.

Os gráficos de dispersão univariados representados na figura 5.5 salientam, em termos relativos:

- para a observação 1, que corresponde à ilha de Santo Antão, a dominância positiva das variáveis S, C e A e a dominância negativa da variável I;
- para a observação 8, que corresponde à ilha Brava, a variável A, comparativamente com as restantes, é negativamente dominante, uma vez que esta observação é a que apresenta um menor valor na coordenada ilr(A).

Em resumo, na composição epidemiológica das 9 ilhas, a ilha de Santo Antão destaca-se por exibir um elevado contraste entre o grupo de paciente infetados sem sintomas (muito baixo) e os demais grupos (muito elevado), enquanto que a ilha Brava destaca-se por uma baixa predominância de pacientes infetados com sintomas clínicos.

5.2 Dados da Qualidade dos Solos

5.2.1 Contextualização do Problema

O solo corresponde à camada superficial da crosta terrestre, constituída por partículas minerais, matéria orgânica, água, ar e organismos vivos [57] e é um elemento fundamental no ecossistema uma vez que permite a produção de alimentos, é o habitat para uma grande variedade de organismos, permite o armazenamento e filtragem de água, o ciclo dos nutrientes, o armazenamento de carbono e suporte da vida humana [48]. Nas áreas urbanas, os solos também desempenham um papel de extrema importância, uma vez que servem de suporte ao seu desenvolvimento (suporte e fonte de material para obras, agricultura urbana, áreas verdes, meio de descarte de resíduos, entre outros) e, por isso, é considerado um recurso vital, embora seja escasso e perecível. Por esse motivo, assegurar a qualidade do solo é, sem dúvida, uma questão importante e que pode ser desafiante devido à heterogeneidade espacial natural [48].

A qualidade do solo não está somente relacionada com o grau de poluição do solo; é também definida como "a capacidade de um solo funcionar dentro dos limites do ecossistema e do uso da terra para sustentar a produtividade biológica, manter a qualidade ambiental e promover a saúde de plantas e animais"[58] e do ser humano. A qualidade dos solos pode ser afetada por um conjunto de fatores. No caso das áreas urbanas, que estão sujeitas a um crescimento, quer a nível de atividade, quer a nível de população,

há um conjunto de razões que afetam a qualidade dos solos urbanos tais como contaminação devido ao uso de transportes, volatilização de contaminantes, lixiviação para águas subterrâneas ou escoamento para ecossistemas aquáticos próximos, deposição de partículas no ar e outras partículas de outras origens, como materiais de construção, lixo, materiais desgastados de pavimentos ou automóveis [48].

Um solo contaminado pode ter um forte impacto na saúde das pessoas e, em casos mais graves, ser o motivo de doenças cancerígenas ou defeitos de nascimento. Os contaminantes orgânicos hidrofóbicos (HOCs), tais como os hidrocarbonetos aromáticos policíclicos (HAPs/PAHs) ou os policlorobifenilos (PCBs), são o resultado da acumulação, a longo prazo, de partículas provenientes dos diversos tipos de contaminação supra citados [48].

A avaliação da qualidade do solo consiste na determinação da presença de contaminantes no solo e da sua concentração, num dado momento. Esses elementos são designados de elementos potencialmente tóxicos (PTE). Para uniformizar a quantidade de contaminantes nos determinados tipos de solo, tais como os PAHs, PCBs e elementos químicos, entre outros contaminantes, que um solo pode ter, para ser considerado contaminado ou não, a Agência Portuguesa do Ambiente possui tabelas de referência, disponíveis no seu guia técnico [59].

A área urbana de Lisboa é a maior de Portugal, concentrando um elevado número de empresas, serviços e pessoas. Como a maioria das grandes cidades a nível mundial, uma das consequências da sua grande dinâmica é a pegada antropológica, que afeta também os solos. Assim, é uma área que tem todo o interesse de ser estudada para perceber até que ponto os diferentes tipos de solo, dispersos pela área urbana de Lisboa, são afetados.

5.2.2 Análise e Discussão de Resultados

A presente análise é baseada em dados discutidos em [48] relativos à área urbana de Lisboa. O conjunto de dados é constituído por 49 amostras de solo, de diversos locais, entre os quais estradas, aeroportos, jardins ornamentais, parques e espaços abertos, pátios de escolas e parques de diversões, de acordo com a figura 5.6.

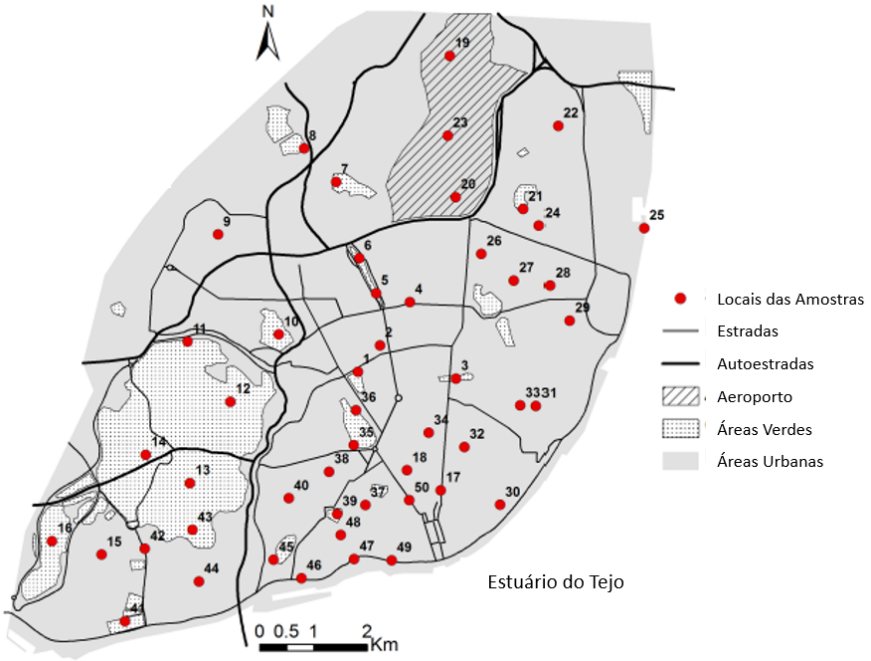


Figura 5.6: Local das amostras na área urbana de Lisboa [imagem retirada da figura 5.1 de [48]].

O conjunto de dados sobre o qual incide este estudo é composto por 11 variáveis, que correspondem aos seguintes 11 elementos químicos:

- **Al** (Alumínio);
- **Fe** (Ferro);
- **Ca** (Cálcio);
- **Co** (Cobalto);
- **Cr** (Cromo);
- **Ni** (Níquel);
- **As** (Arsénio);
- **Cu** (Cobre);
- **Pb** (Chumbo);
- **Zn** (Zinco);
- **Hg** (Mercúrio).

Foi analisado, no conjunto de dados, se havia valores omissos. Para as observações 1 e 14, na variável Hg, não havia valores introduzidos. Optou-se por manter estas observações no conjunto e, para se poder aplicar de forma direta as técnicas estatísticas usuais, imputar esses valores. Os dois valores omissos foram imputados utilizando a técnica do k-vizinho mais próximo (também designado em inglês de k-NN), com $k = 3$ e com a métrica de Aitchison.

Uma vez que a aplicação das transformações composicionais descritas no Capítulo 2 obrigam a que os dados sejam estritamente maiores que zero (pois o logaritmo de zero não está definido no espaço real), foi também testada a existência de zeros. Apenas a observação 43 da variável Hg tem valor zero. Uma vez que se desconhece o motivo desse zero (se foi originado por arrendondamento, se a quantidade de Hg detetada é abaixo daquela que o método/processo permite ou se, de facto, não existe o elemento na amostra de solo) [1], optou-se por retirar essa observação do conjunto de dados.

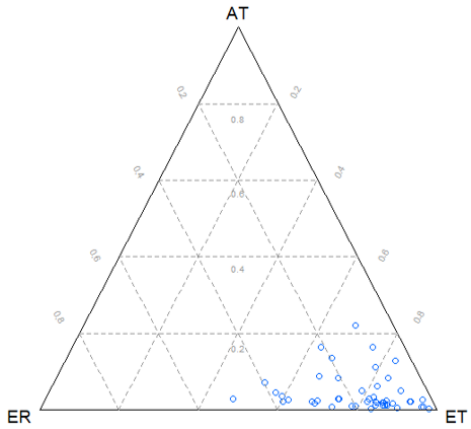
Devido à remoção da observação 43, em algumas metodologias gráficas irá aparecer rotulada uma observação 43, mas que, na prática corresponde à observação 44 do mapa da figura 5.6. Assim, todas as observações subsequentes tem que se interpretar à luz deste facto: as observações 1 a 42 correspondem exatamente às do mapa, quando estiver rotulado 43 deverá ler-se 44, quando estiver rotulado 44 deverá interpretar-se como a observação 45 do mapa e assim consecutivamente.

Após este pré-processamento dos dados, as metodologias estatísticas usuais podem ser implementadas. O conjunto de dados que foi utilizado encontra-se no Anexo A].

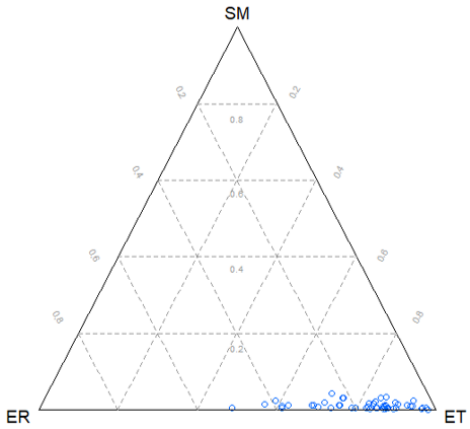
Começa-se pela representação dos dados usando o diagrama ternário. Neste caso, uma vez que existem 11 variáveis, realizar ${}^{11}C_3 = 165$ diagramas, além de moroso, poderá ser difícil só por si ajudar a estabelecer algum padrão nos dados. Deste modo, optou-se por agrupar os 11 elementos em 4 grupos, que correspondem ao tipo de elemento pela Tabela Periódica. Assim, os 4 grupos são designados de:

- elementos representativos (ER), grupo constituído por:
  - **Al** (Alumínio);
  - **Pb** (Chumbo);
- elementos de transição (ET), grupo constituído por:
  - **Fe** (Ferro);
  - **Co** (Cobalto);
  - **Cr** (Cromo);
  - **Ni** (Níquel);
  - **Cu** (Cobre);
  - **Zn** (Zinco);
  - **Hg** (Mercúrio);
- metais alcalinoterrosos (AT), grupo constituído por:
  - **Ca** (Cálcio);
- semimetais (SM), grupo constituído por:
  - **As** (Arsénio).

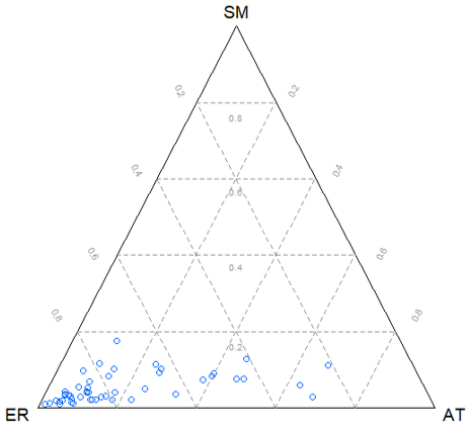
Deste modo, é possível representar os 11 elementos devido ao seu amalgamento em 4 grupos e, assim, há apenas a necessidade de realizar ${}^{4}C_3 = 4$ diagramas ternários.



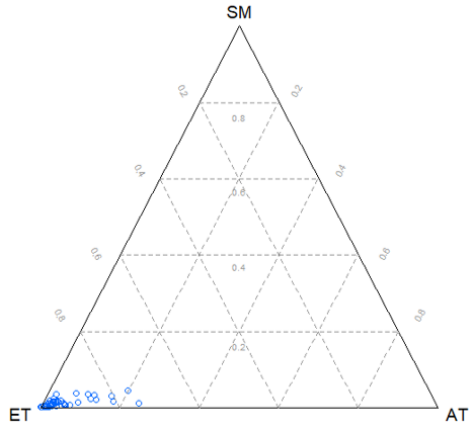
(a) Grupos ER, ET e AT



(b) Grupos ER, ET e SM



(c) Grupos ER, AT e SM



(d) Grupos ET, AT e SM

Figura 5.7: Diagramas ternários, por grupos, para os dados da qualidade de solos.

Pela observação dos diagramas ternários da figura 5.7, conclui-se que nos diagramas onde está presente o grupo ET (figuras 5.7a, 5.7b e 5.7d) tem este grupo de elementos como dominantes em relação aos restantes. Isso acontece não só pelo facto de este grupo ser constituído por mais elementos que os grupos AT, ER e SM, mas também pela quantidade individual de alguns deles (por exemplo, Cu) ser muito elevada. Na ausência do grupo ET (figura 5.7c), o grupo ER é o que se destaca como dominante. O grupo SM é aquele que apresenta menor concentração associada.

Para visualizar a distribuição dos dados numa perspetiva composicional comparando cada parte com o todo, tome-se os dados clr-transformados (pela sua maior facilidade na interpretação) e a representação destes em *box plots* (figura 5.8) e *bagplots* (figura 5.9).

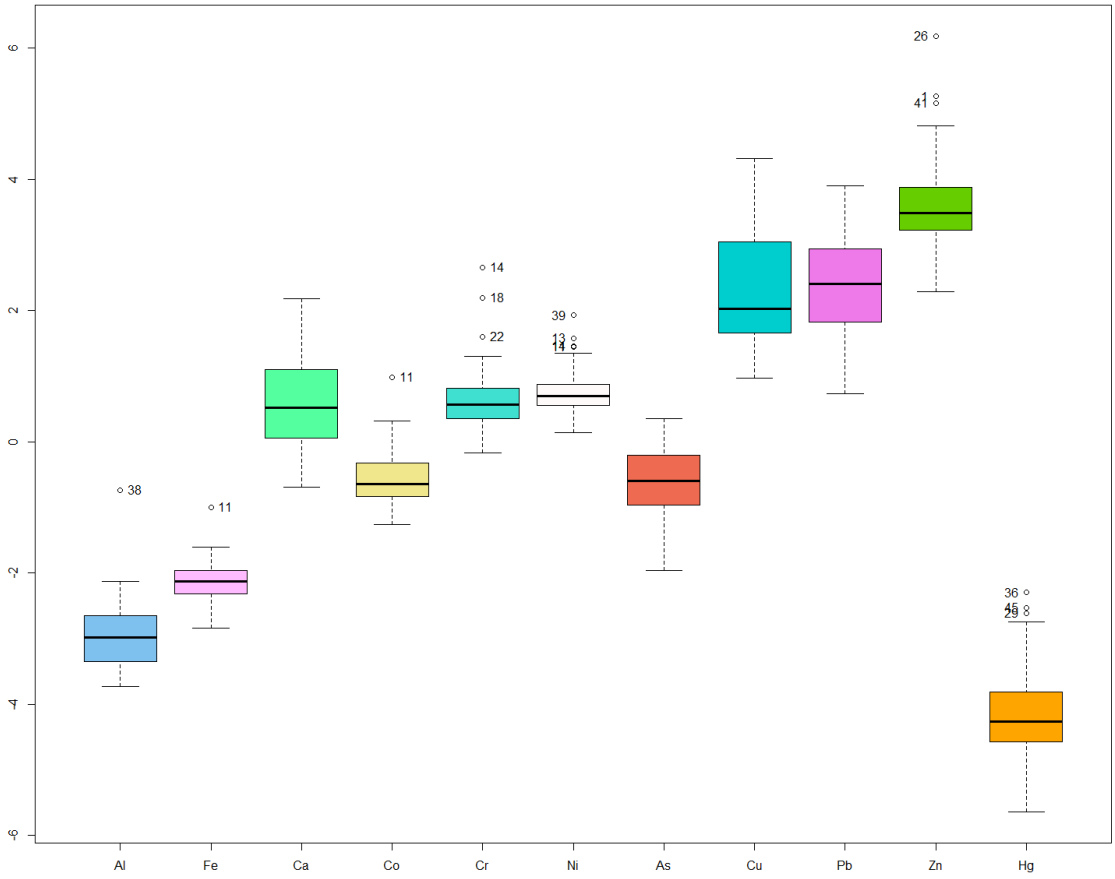


Figura 5.8: *Box plots* para as variáveis clr-transformadas do solo em Lisboa.

Na tabela 5.7 estão assinalados, de forma resumida, as observações clr-transformadas que foram declaradas como atípicas em relação a cada variável.

Elementos	**Al**	**Fe**	**Ca**	**Co**	**Cr**	**Ni**
Outliers	38	11	-	11	22, 18, 14	11, 14, 13, 19
Elementos	**As**	**Cu**	**Pb**	**Zn**	**Hg**	
Outliers	-	-	-	41, 1, 26	29, 44, 26	

Tabela 5.7: *Outliers* declarados para cada elemento, segundo os respetivos *box plots*.

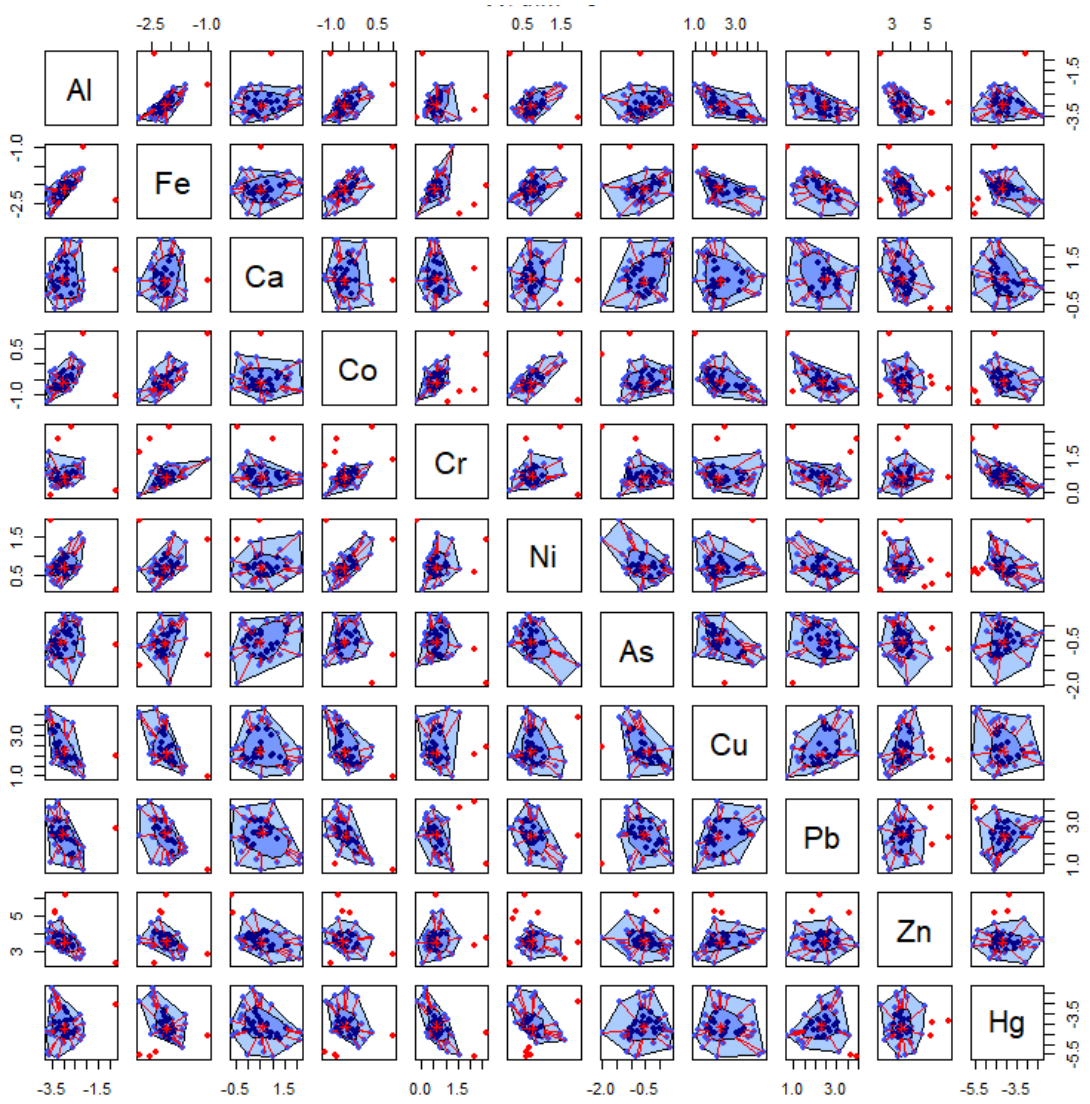


Figura 5.9: *Bagplot* bidimensional dos dados clr-transformados em relação às elementos das amostras do solo em Lisboa.

Pela análise conjunta dos *box plots* da figura 5.8 e dos *bagplots* da figura 5.9, é possível observar que os *bagplots* bivariados, por vezes, declaram um maior número de *outliers* que cada uma das variáveis por si só (como é o caso do par Alumínio e Cromo, em que, no caso univariado, cada variável apresenta um único *outliers* e, no caso bivariado, no respetivo *bagplot* são declaradas como atípicas 4 observações), mas, em outros casos, declaram um menor número de *outliers* que cada uma das variáveis no caso univariado (como, por exemplo, acontece no caso do par Ferro e Zinco que, no caso univariado têm 1 e 3 observações declaradas como *outliers*, respetivamente, mas no caso bivariado apenas apresentam uma observação como atípica). Com o auxílio dos *bagplots* é possível também estabelecer alguns padrões entre os diversos elementos, como acontece com o par Cromo e Mercúrio, que apresentam uma correlação negativa entre eles e as observações que não se adequam a este padrão são as declaradas como *outliers*.

Metodologias numéricas foram utilizados para identificar amostras de solo atípicas. Utilizou-se a Abordagem Comediana usando a transformação clr.

Neste caso, tomando o valor de $\alpha = 0.025$, que corresponde ao que se toma habitualmente na literatura para definir o quantil de corte (ver [56], [1]) tem-se:

$$\begin{aligned} cv &= 1.4826 \times \frac{\chi^2_{d;1-\alpha} \times \text{med}(rd_1, \ldots, rd_n)}{\chi^2_{d;0.5}} \\ &= 1.4826 \times \frac{21.92 \times 339.62}{10.34} \\ &= 1067.344 \end{aligned}$$

A distância de Mahalanobis robusta obtida para cada uma das amostras de solo recolhidas na área urbana de Lisboa está disposta na tabela 5.8.

Amostra	**Dist. Robusta**	**Amostra**	**Dist. Robusta**	**Amostra**	**Dist. Robusta**
1	**1175.25**	17	**1231.81**	33	321.15
2	354.26	18	**1813.96**	34	324.99
3	305.64	19	890.85	35	237.74
4	213.43	20	220.12	36	**1494.11**
5	121.14	21	288.89	37	147.91
6	261.27	22	**1101.92**	38	**1221.36**
7	320.73	23	563.06	39	**1886.11**
8	264.26	24	694.82	40	719.96
9	588.49	25	536.97	41	360.84
10	256.59	26	**1320.80**	42	291.98
11	**1747.70**	27	296.29	44	132.86
12	406.57	28	503.17	45	931.67
13	**1473.80**	29	1049.29	46	230.80
14	**2559.92**	30	256.08	47	183.13
15	323.74	31	84.40	48	156.20
16	488.85	32	203.67	49	117.21

Tabela 5.8: Distância robusta de Mahalanobis obtida para cada uma das amostras, pela Abordagem Comediana (a negrito destacam-se as distâncias superiores ao valor de corte).

Pela observação da tabela 5.8 e sabendo que o valor de corte é 1067.344, foram declaradas como atípicas 11 observações: 1, 11, 13, 14, 17, 18, 22, 26, 36, 38 e 39.

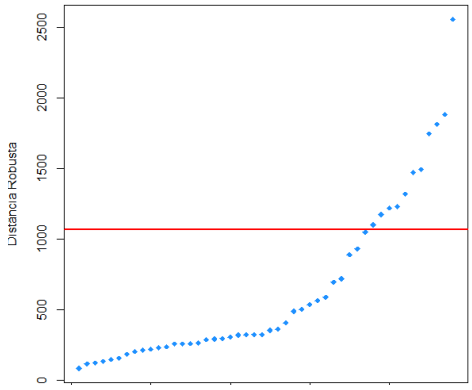


Figura 5.10: Representação gráfica das distâncias de Mahalanobis ordenadas das amostras de solo, obtidas pela Abordagem Comediana, com identificação no número das amostras.

Ao representar-se graficamente as distâncias obtidas e compiladas na tabela 5.8, é possível observar que a amostra 14 tem a maior distância, muito discrepante das demais.

A outra metodologia numérica explorada foi a da Atipicidade Ajustada. Esta foi aplicada a dados ilr-transformados, tomando as 11 pivotagens possíveis. De notar que o número sugerido de direções aleatórias a ser tomadas é $d = 250p$, onde p designa o número de variáveis, e, neste caso em particular, serão tomadas $d = 250 \times 10 = 2500$.

Amostra	Nº *outliers*	**Amostra**	Nº *outliers*	**Amostra**	Nº *outliers*
1	27	17	18	33	0
2	5	18	**1208**	34	1
3	1	19	18	35	0
4	0	20	2	36	9
5	11	21	1	37	0
6	70	22	545	38	**1975**
7	0	23	15	39	**1086**
8	3	24	99	40	18
9	12	25	6	41	245
10	0	26	393	42	7
11	**1182**	27	4	44	0
12	25	28	8	45	4
13	39	29	0	46	6
14	**1687**	30	1	47	3
15	2	31	0	48	38
16	0	32	6	49	0

Tabela 5.9: Contagem do número de vezes que cada observação foi declarada como *outlier* no método da Atipicidade Ajustada, ao aplicar 1000 vezes a cada pivotagem (a negrito estão assinalados as maiores frequências).

Pela observação da tabela 5.9, podem-se destacar cinco observações: 11, 14, 18, 38 e 39, com as que apresentam as mais elevadas frequências em que surgem como *outliers* no método da Atipicidade Ajustada. Uma vez que se realizaram as 11 pivotagens possíveis e, para cada uma delas, testou-se as 2500 direções com sementes (*seeds*) diferentes, de modo a aleatorizar as direções tomadas e, ao repetir o processo 1000, permitiu uma maior robustez de resultados. Cada observação foi avaliada 11000 se seria, ou não, *outlier*. Deste modo:

- a observação 11 foi declarada como atípica 10.7% das vezes;
- a observação 14 foi declarada como atípica 15.3% das vezes;
- a observação 18 foi declarada como atípica 11% das vezes;
- a observação 38 foi declarada como atípica 18% das vezes;
- a observação 39 foi declarada como atípica 9.9% das vezes;
- as restantes observações são declaradas como atípicas menos de 5% das vezes.

Com as metodologias numéricas utilizadas, pode-se compilar os resultados obtidos numa tabela.

Metodologia	*Outliers*
Abordagem Comediana	1, **11**, 13, **14**, 17, **18**, 22, 26, 36, **38**, **39**
Atipicidade Ajustada	**11**, **14**, **18**, **38**, **39**

Tabela 5.10: Observações declaradas como *outlier* em cada metodologia.

É possível utilizar representações gráficas com as coordenadas pivô, para, de modo análogo, ter uma percepção de como os dados se comportam, não levando em conta a sua distância, mas de forma univariada para cada possibilidade de pivô.

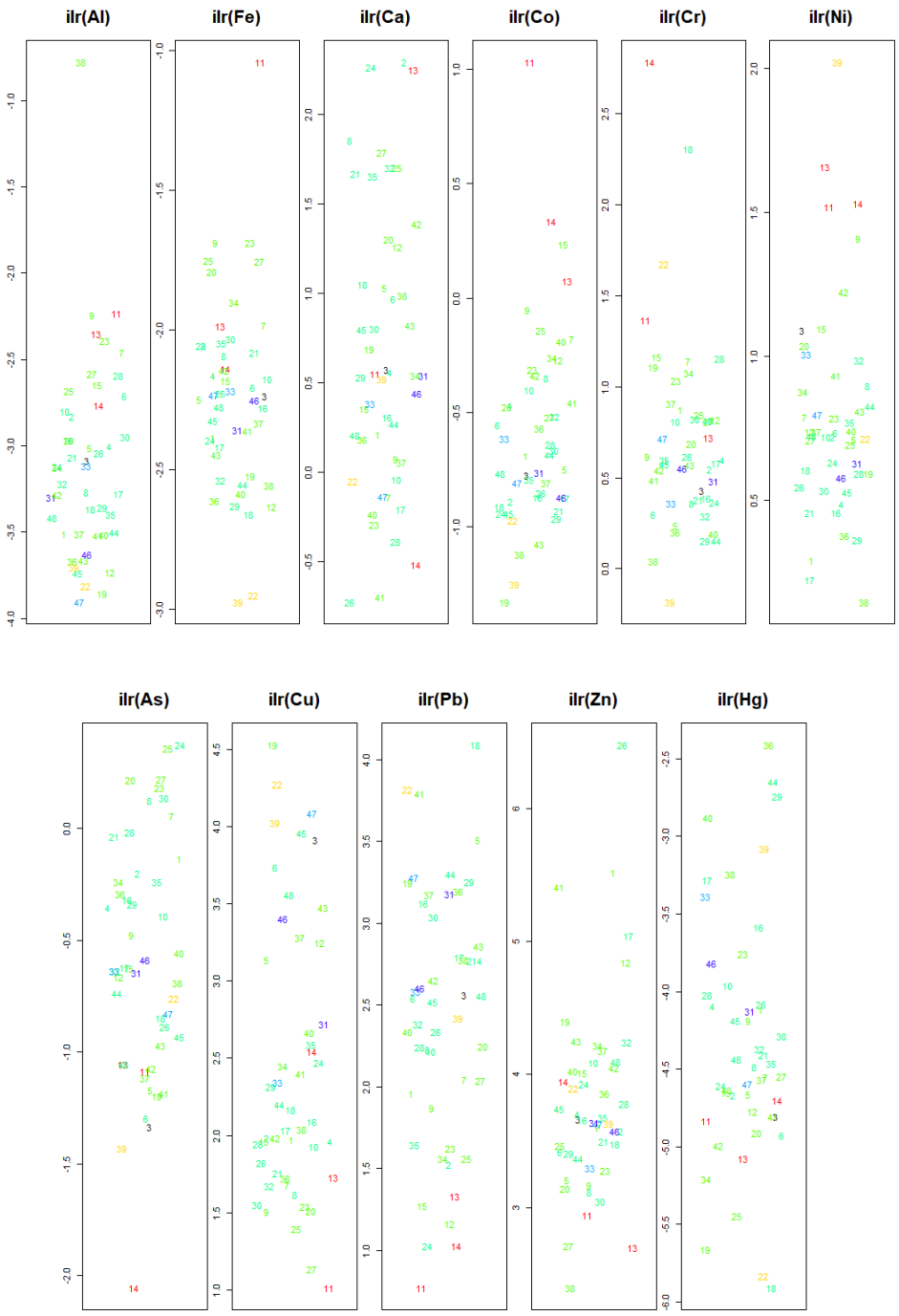


Figura 5.11: Gráficos de dispersão univariados para os elementos das amostras de solo.

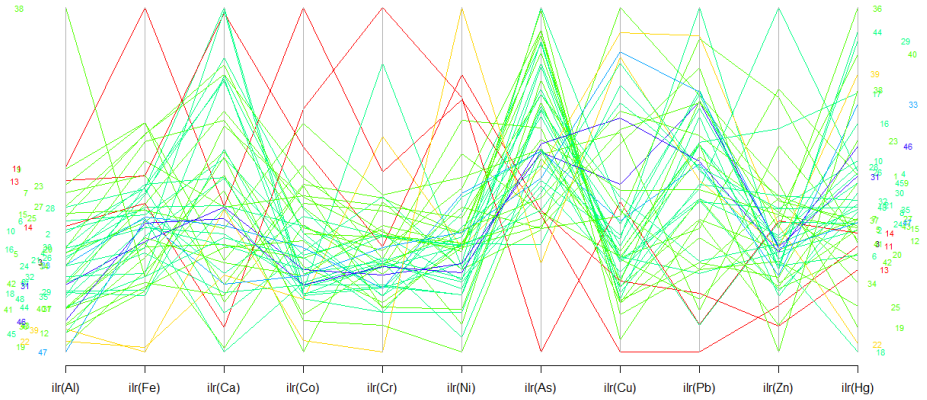


Figura 5.12: Gráfico de dispersão das coordenadas paralelas para os elementos das amostras de solo.

Pela análise da tabela 5.10 pode-se concluir que observações 11, 14, 18, 38 e 39 são, então, os *outliers* deste conjunto de dados e pelas figuras 5.11 e 5.12, compreender o motivo pela qual são declaradas como atípicas.

- As observações 11 e 14 correspondem a amostras extraídas numa mesma área verde com amplo espaço, mas ambas perto de estradas. Pelos gráficos de dispersão univariados, para a observação 11 há, em termos relativos, elevada predominância dos elementos Ferro e Cobalto e baixa predominância dos elementos Cobre e Chumbo. A observação 14 apresenta elevada predominância do elemento Cromo e baixa para Arsénio.

- A observação 39 corresponde igualmente a uma amostra extraída em área verde de reduzida dimensão e perto de estradas e apresenta a log-razão de Ferro em relação às restantes variáveis e de Cromo em relação às demais variáveis muito inferior às demais, isto é, presença destes elementos nessa amostra de solo em razão é inferior às demais amostras de solo e, em contrapartida, a log-razão de Níquel em relação às restantes variáveis muito superior às demais, ou seja, esta amostra está localizada numa zona com prevalência baixa de Cromo e Ferro e elevada de Níquel na composição do seu solo.

- As observações 18 e 38 correspondem a amostras extraídas em áreas urbanas. A observação 18 apresenta uma log-razão de Mércurio em relação às restantes variáveis inferior às demais observações e, por outro lado, a log-razão de Chumbo em relação às restantes variáveis muito alta. A observação 38 apresenta a log-razão de Alumínio em relação às restantes variáveis muito superior às demais e, em contrapartida, a log-razão de Zinco em relação às restantes variáveis e de Níquel em relação às restantes variáveis inferior às demais observações. Embora estas duas amostras correspondam ao mesmo tipo de localização, a composição dos seus solos é bastante distinta: enquanto que numa se destaca a presença de Chumbo, na outra destaca-se a presença de Alumínio.

Na tabela 5.11 apresentam-se, de forma concisa, os elementos que contribuem, de forma dominante, para os *outliers* detetados pelos Método da Atipicidade Ajustada.

Amostra	11	14	18	38	39
+	Co, Fe	Cr	Pb	Al	Ni
-	Cu, Pb	As	Hg	Ni, Zn	Cr, Fe

Tabela 5.11: Composição dominante (pela positiva e pela negativa) das amostras declaradas como *outliers*

.

Para um estudo mais aprofundado dos dados seria conveniente ter mais informações sobre a localização da amostra que possam influenciar a qualidade dos solos, bem como subdividir o tipo de local de onde foram colhidas as amostras, o que, em alguns tipos de localização, implicaria uma maior quantidade de observações para que, de facto, se pudessem aplicar as metodologias descritas e que os seus resultados permitissem conclusões mais consistentes entre elas.

Capítulo 6

Conclusões e Trabalho Futuro

O propósito desta dissertação centrou-se na aplicabilidade de métodos de deteção de *outliers* baseados em distância robusta a dados composicionais log-transformados. Até ao momento, a deteção de obervações atípicas em dados composicionais tem sido baseada na distância de Mahalanobis robusta, usando a estimativa MCD para a matriz de covariâncias e usadas representações gráficas, como dos diagramas ternários ou gráficos de dispersão univariados, que usam o MCD para tentar explicar o porquê dos possíveis *outliers*.

O método MCD é um estimador altamente robusto para a localização e dispersão, mas que tem como inconveniente a elevada complexidade computacional que envolve a procura da matriz de covariância de determinante mínimo. No que aos dados composicionais diz respeito, a menos que se façam testes estatísticos ou análises gráficas prévias, como *box plots* ou histogramas, não se pode assumir que eles sejam simétricos. Assim, se se optar por utilizar métodos que assumam a normalidade nos dados analisados, as conclusões obtidas podem ser imprecisas uma vez que os seus pressupostos não são cumpridos.

De modo a ultrapassar esse inconveniente, utilizou-se duas outras abordagens, igualmente baseadas em distância robusta, mas que usam medidas de localização (mediana) e dispersão (comediana, MAD) que são menos influenciadas pela presença de observações atípicas na Abordagem Comediana, ou o uso da *medcouple*, medida de enviesamento, no caso na Atipicidade Ajustada.

Todavia, a Atipicidade Ajustada utilizada com um elevado número de repetições e direções, como foi executado nesta dissertação, é um processo moroso e, por isso, seria importante realizar um estudo comparativo que permitisse descobrir o número de direções e iterações que obtenham os melhores resultados possíveis no menor tempo (equilíbrio entre eficácia e custo computacional), para um elevado conjunto de bases de dados na perspetiva composicional.

Além de um estudo aprofundado sobre o desempenho dos métodos utilizados neste trabalho, baseados em distância, como objetivo futuro pretende-se estudar métodos não tradicionais de deteção de *outliers* numa perspetiva composicional, como por exemplo métodos de *clustering* e que estes possam, eventualmente, ser usados em articulação

com outros não tradicionais ou com métodos baseados em distância robusta.

Bibliografia

[1] Filzmoser, P., Hron, K., Templ, M., *Applied Compositional Data Analysis*, Springer, Switzerland, 2018

[2] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., *Lecture Notes on Compositional Data Analysis*, May 28, 2007

[3] Aitchison, J., *The Statistical Analysis of Compositional Data*, Chapman and Hall, New York - USA, 1986

[4] Sousa, R., *Análise Estatística de Dados Composicionais*, Dissertação de Mestrado, Universidade de Aveiro, 2016.

[5] Leite, L., *Técnicas Exploratórias na Deteção de Outliers em Dados Composicionais*, Dissertação de Mestrado, Universidade de Aveiro, 2019.

[6] Sajesh, T. A., Srinivasan, M.R., *An Overview of Multiple Outliers in Multidimensional Data*, Sri Lankan Journal of Applied Statistics, Vol (14-2), 2013

[7] Barnett, V., Lewis, T., *Outliers in Statistical Data*, John Wiley and Sons, Chichester, England, 1994. DOI:10.1016/0169-2070(95)00625-7

[8] Grubbs, F.E., *Procedures for Detecting Outlying Observations in Samples*, Technometrics, 11: pp. 1-21, 1969. DOI: 10.1080/00401706.1969.10490657

[9] Hawkins, D.M., *Identification of Outliers*, Chapman and Hall, London, 1980. DOI: 10.1002/bimj.4710290215

[10] Figueira, M .M. C., *Identificação de Outliers*, MILLENIUM n°12, Outubro de 1998.

[11] Cantin, Guillaume; Silva, Cristiana J., *Influence of the topology on the dynamics of a complex network of HIV/AIDS epidemic models*, AIMS Mathematics, 4(4): pp. 1145-1169, 2019

[12] Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, New York, USA, 1977.

[13] Dietz, Tt, Kalof, L., *Introduction to Social Statistics: The Logic of Statistical Reasoning*, Wile-Blackwell, pp. 133-568, 2009.

[14] Gaio, A. R., Apontamentos das Aulas Teóricas, Unidade Curricular de Estatística Aplicada, Licenciatura em Matemática, Faculdade de Ciências da Universidade do Porto, 2017

[15] Hoaglin, David C., Iglewicz, Boris, Tukey, John W., *Performance of Some Resistant Rules for Outlier Labeling*, Journal of the American Statistical Association, Vol. 81, No. 396, pp. 991-999, dezembro de 1986

[16] Hoaglin, D.C., Iglewicz, B., *Fine Tuning Some Resistant Rules for Outlier Labeling*, Journal of the American Statistical Association, 82: pp. 1147-1149, 1987. DOI: 10.1080/01621459.1987.10478551

[17] Kimber, A.C., *Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions*, Applied Statistics, 39: pp. 21-30, 1990.

[18] Van der Loo, M.P.J., *Distribution Based Outlier Detection in Univariate Data*, Statistics Netherlands, 2010.

[19] Gao, S., Li, G., Wang, D., *A New Approach for Detecting Multivariate Outliers*, Communications in Statistics - Theory and Methods, 34:8, pp. 1857-1865, 2005. DOI: 10.1081/STA-200066315

[20] `https://pt.wikipedia.org/wiki/Dist%C3%A2ncia_de_Mahalanobis`, consultado pela última vez em: 23/04/2020

[21] Farcomeni, Alessio, Ventura, Laura, *An overview of robust methods in medical research*, Statistical Methods in Medical Research 21(2), pp. 111–133, 2010. DOI: 10.1177/0962280210385865

[22] Lopuhaä, H.P., Rousseeuw, P.J., *Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices*, The Annals of Statistics, 19: pp. 229-248, 1991. DOI:10.1214/aos/1176347978

[23] Aitchison, J., *A Concise Guide to Compositional Data Analysis*, United Kingdom, 2003

[24] `https://brainly.com.br/tarefa/11236697`, consultado pela útlima vez em 06/05/2020.

[25] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., *Modeling and Analysis of Compositional Data*, John Wiley & Sons, Lta, United Kingdom, 2015

[26] Templ, M., Hron, K., Filzmoser, P., *Exploratory tools for outlier detection in compositional data with structural zeros*, Journal of Applied Statistics 44.4, pp. 734-752, 2017

[27] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., *Isometric Logratio Transformations for Compositional Data Analysis*, Mathematical Geology, Vol. 35, No. 3, Abril 2003

[28] Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., *Compositional Data Analysis in the Geosciences: From Theory to Pratice*, Geological Society, London, Special Publications n° 264, 2006

[29] Falk, M., *On MAD and Comedians*, Annals of Institute of Statistical Mathematics, Vol.49, No. 4, 615-644, 1997

[30] Hall, P., Welsh, A.H., *Limit Theorems for the Median Deviation*, Annals of Institute of Statistical Mathematics, 37: 27-36, 1985. DOI:10.1007/BF02481078

[31] Sajesh, T.A., Srinivasan, M.R., *Outlier Detection for High Dimensional Data using Comedian Approach*, Journal of Statistical Computation and Simulation, 82(5): pp. 745-757, 2012. DOI:10.1080/00949655.2011.552504

[32] Hubert, M., Van der Veeken, S., *Outlier detection for skewed data*, Journal of Chemometrics, 22: pp. 235-246, 2008

[33] Stahel, W.A., *Robuste Schatzungen: Infinitesimale Optimalitat und Schatzungen von Kovarianzmatrizen*, Tese de Doutoramento, ETH Zurich, Zurique, Suíça, 1981

[34] Donoho, D.L., *Breakdown Properties of Multivariate Location Estimators*, Ph.D Qualifying Paper, Department of Statistics, Harvard University, Cambridge, MA, 1982.

[35] Verardi, V., Vermandele, C., *Univariate and multivariate outlier identification for skewed or heavy-tailed distributions*, The Stata Journal 18, Number 3, pp. 517-532. 2018

[36] Maronna, R. A., Yohau, V. J., *Robust and efficient estimation of multivariate scatter and location*, arXiv:1504.03389 [math.ST], Cornell University, 2015.

[37] `https://medium.com/@phuctrt/loss-functions-why-what-where-\or-when-189815343d3f`, consultado pela última vez em: 29/04/2020

[38] `https://en.wikipedia.org/wiki/Order_statistic`, consultado pela última vez em: 04/04/2020

[39] Hubert, M., Vandervieren, E., *An adjusted boxplot for skewed distributions*, Computational Statistics and Data Analysis 52, pp. 5186-5201, 2008

[40] Chaves, M. G., Apontamentos das Aulas Teóricas, Unidade Curricular de Álgebra Linear e Geometria Analítica I, Licenciatura em Matemática, Faculdade de Ciências da Universidade do Porto, Ano Letivo 2013/2014

[41] von Eynatten, Hilmar, Pawlowsky-Glahn, Vera, Egozcue, d Juan José, *Understanding Perturbation on the Simplex: A Simple Method to Better Visualize and Interpret Compositional Data in Ternary Diagrams*, Mathematical Geology, Vol. 34, No. 3, April 2002

[42] `https://cran.r-project.org/web/packages/mvoutlier/mvoutlier.pdf`, consultado pela última vez em: 04/06/2020

[43] Rousseeuw, Peter J., Ruts, Ida, Tukey, John W., *The Bagplot: A Bivariate Boxplot*, The American Statistician, vol. 53, no. 4, pp. 382-387, 1999

[44] Freitas, A., Apontamentos das Aulas Teóricas, Unidade Curricular de Estatística Multivariada, Mestrado em Matemática e Aplicações, Universidade de Aveiro, Ano Letivo 2018/2019

[45] Aitchison, J., Greenacre, M., *Biplots of compositional data*, Journal of the Royal Statistical Society. Series C: Applied Statistics 51.4, pp. 375-392, 2002

[46] Eckart, C., Young, G., *The approximation of one matrix by another of lower rank*, Psychometrika 1.3, pp. 211–218, 1936

[47] https://www.aquare.la/o-que-sao-outliers-e-como-trata-los-em\
-uma-analise-de-dados/, consultado pela última vez em: 21/05/2020

[48] Cachada, Anabela F. O., *Contaminantes orgânicos em solos urbanos: fontes e potenciais riscos*, Tese de Doutoramento, Universidade de Aveiro, 2014

[49] Hahn, Beatrice H., Sharp, Paul M., *Origins of HIV and the AIDS Pandemic*, Cold Spring Harbor perspectives in medicine. 1, Cold Spring Harbor Laboratoty Press, setembro de 2011, doi: 10.1101/cshperspect.a00684

[50] https://pensapositivo.pt/o-vih/conhecer/o-vih/?gclid=
Cj0KCQjwzZj2BRDVARIsABs3l9KJSyKZuilmjkFx6HW4qSpHfwPJEP8zi-pvEI_
v1gBsMK9eNCNM6IUaAnAVEALw_wcB, consultado pela última vez em: 21/05/2020

[51] https://www.atlasdasaude.pt/publico/content/
hiv-e-sida-tudo-sobre-suas-diferencas, consultado pela última vez em: 21/05/2020

[52] Silva, Cristiana J., Torres, Delfim F. M., *A SICA compartmental model in epidemiology with application to HIV/AIDS in Cape Verde*, Ecological Complexity, vol. 30, pp.70-75, 2017

[53] Epidemia de VIH nos países de língua oficial portuguesa, 4a edição, 2018

[54] Cantin, Guillaume; Silva, Cristiana J., *Influence of the topology on the dynamics of a complex network of HIV/AIDS epidemic models*, AIMS Mathematics, 4(4): 1145?1169, 2019

[55] Comité de Coordenação do Combate a Sida, *Rapport de Progrès de la riposte VIH/-SIDA Cabo Verde*, 2015.

[56] Di Palma, M. A., Gallo M., *A co-median approach to detect compositional outliers*, Journal of Applied Statistics, 2016. DOI: 10.1080/02664763.2016.1163525

[57] http://www.spcs.pt/, consultado pela última vez em 26/05/2020

[58] Bünemann, Else K. *et al.*, *Soil quality - A critical review*, Soil Biology and Biochemistry, 120, pp 105-125, 2018

[59] Solos Contaminados - Guia Técnico, Valores de Referência para o Solo, Agência Portuguesa do Ambiente, Amadora, janeiro de 2019

Apêndice A

Matriz dos Dados da Qualidade dos Solos

	Al	Fe	Ca	Co	Cr	Ni	As	Cu	Pb	Zn	Hg
1	0.27	0.79	9.43	3.99	17.67	10.16	6.75	50.47	49.81	1480.01	**0.1514778**
2	0.11	0.23	14.51	0.70	2.75	3.25	1.35	10.83	6.99	49.25	0.019
3	0.33	0.74	10.83	2.99	9.41	17.70	1.75	261.35	71.74	206.16	0.064
4	0.13	0.29	3.88	1.46	4.03	3.63	1.63	14.76	31.97	77.88	0.046
5	0.29	0.60	13.68	2.51	6.43	10.09	1.68	101.66	145.08	109.06	0.06
6	0.29	0.47	9.71	2.27	5.13	7.77	1.12	135.98	43.37	100.03	0.035
7	0.14	0.22	1.28	1.23	4.34	3.10	1.54	7.23	10.25	45.19	0.019
8	0.15	0.46	19.83	2.43	4.77	7.99	3.82	15.89	28.36	66.01	0.047
9	0.30	0.51	2.74	2.43	4.59	9.78	1.62	10.73	15.16	52.19	0.047
10	0.27	0.49	3.76	2.66	8.44	7.75	2.68	24.57	32.22	192.20	0.089
11	0.55	1.71	7.80	12.36	16.97	19.69	1.64	12.15	9.65	76.46	0.046
12	0.20	0.57	23.34	5.43	15.31	14.18	3.73	155.60	21.33	709.00	0.074
13	0.39	0.55	31.25	3.94	7.27	17.78	1.34	19.07	13.04	48.06	0.029
14	1.10	2.00	9.42	21.22	218.96	66.25	2.17	173.81	41.08	663.66	**0.1739698**
15	0.25	0.39	4.38	3.91	9.48	8.87	1.72	20.28	10.50	142.68	0.037
16	0.27	0.52	6.12	1.99	6.63	7.06	3.37	33.49	89.46	149.44	0.15
17	0.26	0.59	4.87	2.58	10.33	7.35	3.28	41.31	84.73	728.19	0.26
18	0.26	0.51	17.56	2.71	58.12	11.50	2.88	50.83	318.34	177.65	0.023
19	0.23	0.82	17.53	2.56	26.15	16.01	2.90	681.06	200.67	602.01	0.041
20	0.16	0.49	9.38	1.72	5.23	7.28	3.33	11.49	23.09	54.16	0.025
21	0.09	0.23	8.23	0.69	2.40	2.59	1.62	8.94	23.51	46.92	0.025
22	0.44	1.00	15.87	6.59	82.45	32.92	8.07	977.95	633.86	682.86	0.064
23	0.22	0.43	1.63	1.60	5.77	4.56	2.56	9.34	10.14	48.92	0.0598
24	0.14	0.28	23.81	1.12	3.89	5.02	3.93	29.01	7.32	116.46	0.034
25	0.14	0.34	9.15	1.58	4.04	3.50	2.54	6.83	8.00	49.06	0.01
26	0.17	0.37	1.55	1.37	5.58	5.21	1.33	17.60	28.75	1491.33	0.063
27	0.30	0.66	19.38	2.14	7.61	6.92	4.35	10.40	24.53	46.54	0.046
28	0.09	0.15	0.74	0.58	3.22	1.88	1.05	6.81	9.06	39.12	0.023
29	0.29	0.58	11.80	2.84	8.23	10.05	5.15	64.89	158.01	183.29	0.52

	Al	Fe	Ca	Co	Cr	Ni	As	Cu	Pb	Zn	Hg
30	0.23	0.55	8.23	2.03	8.39	6.37	4.36	16.75	69.35	69.85	0.064
31	0.33	0.81	12.83	3.70	12.11	13.95	4.15	102.81	158.79	246.34	0.15
32	0.12	0.23	13.06	1.58	3.41	6.62	1.41	12.74	25.05	147.90	0.04
33	0.31	0.73	8.73	3.37	8.52	15.83	3.30	56.73	70.97	140.23	0.24
34	0.19	0.61	6.26	2.92	10.44	8.63	2.98	38.68	16.60	208.79	0.026
35	0.08	0.29	9.89	0.96	3.61	4.26	1.63	24.13	9.80	68.10	0.029
36	0.19	0.52	7.48	3.64	7.58	8.99	4.75	32.48	131.89	247.44	0.63
37	0.33	1.01	9.86	4.33	22.23	18.90	3.23	213.96	193.19	503.74	0.12
38	3.88	0.71	20.83	2.80	8.46	9.36	4.22	56.75	114.94	79.74	0.37
39	0.21	0.42	11.75	2.17	6.01	49.20	1.83	331.81	72.08	227.79	0.38
40	0.24	0.58	5.47	5.70	8.19	13.86	4.03	86.59	63.48	316.87	0.44
41	0.58	1.75	8.57	10.75	26.44	40.52	5.37	163.39	615.93	2891.54	0.17
42	0.22	0.65	18.85	3.64	8.44	16.17	1.81	33.27	62.83	238.95	0.043
44	0.22	0.70	15.81	2.59	12.45	15.61	2.86	197.90	110.17	414.16	0.087
45	0.28	0.69	10.20	4.10	9.12	17.30	3.90	64.09	182.37	194.87	0.63
46	0.20	0.77	15.13	2.88	12.23	11.70	2.91	307.94	78.00	250.00	0.13
47	0.24	0.89	11.59	3.32	12.90	13.23	4.36	195.65	91.16	229.53	0.2
48	0.27	1.33	9.84	5.18	22.09	23.93	5.08	547.75	254.53	354.31	0.14
49	0.37	1.10	11.74	4.63	20.88	19.22	3.52	285.68	110.12	477.87	0.14

Tabela A.1: Matriz de dados da qualidade dos solos (a negrito estão assinalados os valores que foram imputados pelo algoritmo $k-$NN).

Apêndice B

Poster

HIV/SIDA em Cabo Verde: uma abordagem por dados composicionais

Marta Maltez[1], Letícia Leite[1], Adelaide Freitas[1,2]

[1]Departamento de Matemática, Universidade de Aveiro

[2]Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA)

Resumo

Usando a distribuição de ocorrências de infeção pelo HIV, segundo os parâmetros do modelo epidemiológico SICA [2], nas 9 ilhas habitáveis do arquipélago de Cabo Verde, no ano de 2015, explora-se a abordagem composicional na análise exploratória deste tipo de dados, usuais em Epidemiologia.
Aplicam-se métodos estatísticos multivariados, clássicos vs robusto, em contexto de dados composicionais, com vista a encontrar observações que se desviem marcadamente das demais e tentar compreender a razão pela qual tal acontecer.

Introdução

O Síndrome da Imunodeficiência Adquirida (SIDA) é causado pela infeção com o Vírus da Imunodeficiência Humana (HIV). O HIV continua a ser um dos maiores flagelos de saúde pública a nível mundial, tendo causado a morte a mais de 35 milhões de pessoas. África é o continente que registra o maior número anual (2017) de novos casos a nível global. Em relação à SIDA, e segundo o modelo epidemiológico SICA [2], a população é dividida em quatro grupos de indivíduos mutuamente exclusivos:

- indivíduos suscetíveis (S);
- indivíduos infectados pelo HIV sem sintomas clínicos de SIDA, mas capazes de transmitir o HIV a outros (I);
- indivíduos infetados por HIV em tratamento com carga viral remanescente baixa (estado crónico) (C);
- indivíduos infetados com HIV com sintomas clínicos de SIDA (A).

O objetivo é, vendo a população das 9 ilhas habitáveis do arquipélago de Cabo Verde, dividida neste quatro grupos, perceber se alguma das ilhas tem um comportamento atípico numa perspetiva composicional.

Ilhas	S	I	C	A
Santo Antão	40388	10	93	9
São Vicente	80763	32	186	19
São Nicolau	12381	7	29	3
Sal	33642	22	78	8
Boa Vista	14404	10	33	3
Maio	9657	5	16	2
Fogo	35735	15	82	8
Brava	5681	5	13	1
Santiago	293084	303	676	87

Figura 1: Distribuição da população de Cabo Verde por ilhas e grupos em relação ao HIV.

Métodos

Tranformações

Dados composicionais são definidos por vetores D-dimensionais (D>1), onde as D componentes (variáveis) traduzem informação relativa (e.g., percentagens ou proporções), isto é, onde cada variável representa a contribuição para um todo (sendo a informação sobre o total irrelevante).
Seja $\mathbf{X} = [x_{ij}]_{n\times D}$ a matriz de dados composicionais . Para a i-ésima linha, $x_{i+} = (x_{i1}, \ldots, x_{iD})$, tem-se:

$$x_{ij} \in \mathbb{R}^+ \text{ e } x_{i1} + \cdots x_{iD} = k.$$

para algum $k \in \mathbb{R}^+$. Os dados composicionais definem uma geometria própria, não coerente com a geometria Euclidiana. Contudo, existem transformações das coordenadas composicionais para a geometria Euclidiana, nomeadamente:

Transformação - clr:
A transformação clr do vetor x_{i+} é dada pelo vetor $\mathrm{clr}(x_{i+}) = (y_{i1}, \ldots, y_{iD})'$, com $y_j = \ln \frac{x_{ij}}{\sqrt[D]{\prod_{k=1}^{D} x_{ik}}}$, $j = 1, \ldots, D$.

$$[x_{ij}] \mapsto [y_{ij}]$$

Transformação - ilr:
A transformação ilr do vetor x_{i+} é dada pelo vetor $\mathrm{ilr}(x_{i+}) = (z_{i1}, \ldots, z_{i,D-1})'$, com $z_j = \frac{\sqrt{D-j}}{\sqrt{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}$, $j = 1, \ldots, D-1$.

A transformação ilr do vetor x_{i+} com coordenadas pivot é definida por um vetor cujas j-ésima coordenada corresponde à primeira coordenada ilr - o pivot - do vetor permutação $x_{i+}^{(j)} = (x_{ij}, x_{i1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{iD})$, com $j = 1, \ldots, D$.

$$[x_{ij}] \mapsto \left[z_1^{(j)}\right]$$

Para evitar a singularidade observada com os dados clr-transformados, na deteção de *outliers* em dados composicionais, com recurso à inversa da matriz de covariâncias, é utilizada a transformação ilr com coordenadas pivot.

Gráficos

Gráficos de dispersão univariados: Para cada coordenada pivot j, representam-se os n valores da primeira coordenada ilr-transformada para cada $j = 1, \ldots, D$ versus um índice (os números de 1 a 9 correspondem às ilhas pela ordem apresentada na Figura 1.) das observações.
Para os dados da Tabela 1 resulta a Figura 2.

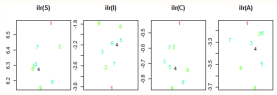


Figura 2: Gráficos de dispersão univariados (cores de acordo com o seu afastamento; cf. [3]).

Biplots: Representação simultâneas de linhas e colunas de uma matriz com característica 2.
A interpretação dos biplots depende sobre qual matriz é aplicada a decomposição em valores singulares para a representação dos pontos e setas no gráfico. Para avaliar o comportamento das 9 ilhas, importa interpretar os pontos.
Biplot clássico [4] matriz dos dados originais da Tabela 1, (Fig. 3). Interpretar as projeções dos pontos sobre as setas (variáveis originais).
Biplot de variação relativa [1]: matriz dos dados clr-transformados (Fig. 4, esq.). Interpretar as projeções dos pontos sobre os *links* (rácios entre duas variáveis).
Biplot composicional robusto [3]: matriz dos dados clr-transformados (Fig. 4, dir.). Interpretar as projeções dos pontos sobre as setas (componente dominante).

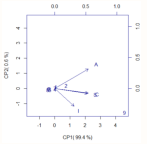


Figura 3: Biplot clássico para os dados originais normalizados.

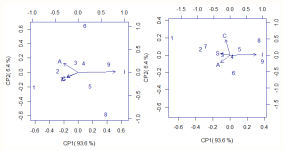


Figura 4: Biplot de variação relativa (à esquerda) e composicional robusto (à direita).

Método Comedian

Utiliza a *comedian*, uma medida alternativa de dependência entre pares de variáveis, substituindo, numa versão robusta, a matriz de correlação usual [5]. É calculada a estimativa robusta para a localização e para a dispersão e, usando essas medidas para o cálculo da distância de Mahalanobis robusta, é considerado *outlier* se:

$$\mathrm{cv} > 1.4826 \frac{\chi^2_D(0.95)\ \mathrm{mediana}(\mathrm{rd}_1, \ldots, \mathrm{rd}_n)}{\chi^2_D(0.5)}$$

onde cv é o valor de corte, D é o número de variáveis e rd_i corresponde à distância de Mahalanobis robusta da observação x_i.

Resultados

Gráficos de dispersão univariados. A ilha 1 é um *outlier*, uma vez que o peso das componentes S, C e A em relação às demais é muito superior, comparativamente às demais e, inversamente, quando se compara I com as restantes variáveis, o seu peso é muito menor.
Biplot clássico. A ilha 9 destaca-se das restantes, apresentando números de S, I, C e A muito acima da média das 9 ilhas. Em termos absolutos a ilha 2 é a segunda ilha com maiores números de S, I, C, A.
Biplot de variação relativa. As ilhas 8 e 9 com menor rácio A/I (i.e, menor proporção de doentes com sintomas relativamente ao número de infetados) enquanto que a ilha 1 é a que apresenta maior rácio A/I. O mesmo sucede para o rácio C/I.
Biplot composicional robusto. A ilha 1 destaca-se com maiores valores para os rácios em que no numerador se tem o número de S, C e A. As ilhas 9 e 8 apresentam os valores de I com maior componente dominante (i.e., rácios em que o numerador é o número de I). Estes resultados estão de acordo com os gráficos de dispersão univariados.
Todos os biplots construídos sobre as duas primeiras componentes principais apresentam elevada percentagem de variabilidade explicada.
Usando a **abordagem *Comedian***, apenas a ilha de Santo Antão foi classificada como atípica, o que vai de encontro às metodologias gráficas utilizadas.

Conclusão

As ilhas com maior densidade populacional (Santiago e São Vicente) apresentam (naturalmente) o maior número de indivíduos S, I, C e A.
Contudo, na perspetiva composicional, a ilha de Santo Antão é, de facto, aquela que tem um comportamento desviante das demais: o peso relativo do número de indivíduos S, C e A é mais acentuado e, de forma inversa, sendo que o número de indivíduos I tem um menor peso relativo.

Agradecimentos

Este trabalho foi parcialmente suportado pelo Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA) através da FCT - Fundação para a Ciência e a Tecnologia - referências UIDB/04106/2020 e UIDP/04106/2020.

Referências

[1] Aitchison, J., Greenacre, M., Biplots of compositional data. *Appl. Statist.*, 51(4): 375–392, 2002

[2] Cantin, Guillaume; Silva, Cristiana J., Influence of the topology on the dynamics of a complex network of HIV/AIDS epidemic models, *AIMS Mathematics*, 4(4): 1145–1169, 2019

[3] Filzmoser, P., Hron, K., Templ, M., *Applied Compositional Data Analysis*. Springer, Switzerland, 2018

[4] Gabriel, K. The Biplot-Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, 58: 453–467, 1971

[5] Sajesh, T. A., Srinivasan, M.R., *An Overview of Multiple Outliers in Multidimensional Data*, Sri Lankan Journal of Applied Statistics, Vol (14-2), 2013

Figura B.1: Poster apresentado nas 1as Jornadas de Estatística Médica, na Faculdade de Ciências da Universidade de Lisboa, realizadas a 12 e 13 de fevereiro de 2020.