UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA

REPRESENTAÇÕES EUCLIDIANAS DE DADOS

-

UMA ABORDAGEM PARA VARIÁVEIS HETEROGÉNEAS

Isabel Maria Tudela Reimão Pinto de França Doria

DOUTORAMENTO EM MEDICINA

BIOMATEMÁTICA

2008

UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA

REPRESENTAÇÕES EUCLIDIANAS DE DADOS
-
UMA ABORDAGEM PARA VARIÁVEIS HETEROGÉNEAS

Isabel Maria Tudela Reimão Pinto de França Doria

Sob a orientação de
Professora Doutora Helena Bacelar Nicolau
Professor Doutor Georges Le Calvé

DOUTORAMENTO EM MEDICINA
BIOMATEMÁTICA
2008

“A impressão desta dissertação foi aprovada pela Comissão Coordenadora do Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 17 de Julho de 2008.”

As opiniões expressas nesta publicação são da exclusiva responsabilidade do seu autor.

Aos meus filhos

À minha neta

Aos médicos da minha família

E em especial aos meus pais

RESUMO

Esta dissertação insere-se na área de Análise de Dados Multivariados, sob o tema de Representações de Dados. Os objectivos da tese são de ordem metodológica, incluindo desenvolvimento de *software* e aplicação a dados reais no âmbito da Biomatemática. O principal objectivo consiste na representação simultânea de variáveis de diferentes tipos ou seja de variáveis heterogéneas. Para este efeito foram usados coeficientes existentes na literatura, alguns dos quais ainda pouco estudados.

Em Biomatemática, como em outras disciplinas, as variáveis descritoras dos indivíduos são frequentemente de natureza diversa – esta situação é habitual, por exemplo, em investigação baseada em inquéritos e questionários. Contudo, a heterogeneidade das variáveis cria problemas matemáticos e estatísticos específicos, que são difíceis de resolver quando se pretendem obter representações euclidianas. A abordagem apresentada neste trabalho dá uma contribuição para a representação de variáveis heterogéneas.

Em geral, os dados em estudo são considerados sob a forma de uma matriz "numérica" que cruza um conjunto de indivíduos e um conjunto de variáveis. Esta matriz é transformada numa matriz de semelhanças ou de dissemelhanças. É esta matriz de proximidades que vai ser representada. Esta abordagem, mais geral do que a abordagem tradicional, permite tratar de uma maneira unificada os diversos métodos factoriais (Análise em Componentes Principais (ACP), a Análise Factorial das Correspondências e outros), por um lado, e por outro, os de Posicionamento Multidimensional (*Multidimensional Scaling*).

Tratamos também de casos em que não existe representação euclidiana exacta dos dados. Interessamo-nos em particular, pela categoria importante das transformações monótonas, que "perturbem ao mínimo" os dados originais. Mostramos que os métodos "da constante aditiva" não são fiáveis (a constante é, com frequência, extremamente grande em relação às proximidades iniciais). A transformação pela função potência, por outro lado, é uma proposta que parece ser muito promissora.

São abordados em pormenor os coeficientes s, s_{LC} e P_L (Le Calvé, 1977) e os coeficientes de afinidade – a, a_W, a_δ, VAL_{AW}, $VAL_{A\delta}$ e os generalizados (e.g., Bacelar-Nicolau, 1980, 1988, 2002; Nicolau *et al*., 2007) – que permitem tratar simultaneamente variáveis do mesmo ou de diferentes tipos. A abordagem probabilística destes coeficientes – VAL_{AW},

$VAL_{A\delta}$ e P_L – tem a mesma origem (ver Lerman, 1970). Quando se comparam variáveis do mesmo tipo, alguns daqueles coeficientes coincidem com coeficientes conhecidos, tal como o coeficiente de correlação de Pearson.

Generalizaram-se os coeficientes s, s_{LC} e P_L para serem utilizados com dados simbólicos (e.g., Doria *et al*., 2007) e variáveis de ordem sequencial. Verificou-se que, em certos casos, estes permitem comparar variáveis com dados omissos sem recorrer a métodos de imputação e são uma mais valia para o caso das variáveis com categorias parcialmente ordenadas (e.g., Doria *et al*., 2006b), assim como na comparação de ultramétricas (e.g., Doria *et al*., 2000). Paralelamente, desenvolveu-se *software* para o cálculo dos coeficientes s, s_{LC} e P_L e aplicaram-se estes coeficientes a problemas reais do domínio da Biomatemática (e.g., Doria *et al*., 2006a; Doria *et al*., 2007). As análises em componentes principais das matrizes de semelhanças S_{LC} e P_L conduziram, de forma geral, a boas visualizações das relações entre as variáveis. No caso particular de todas as variáveis serem métricas, a ACP da matriz de semelhanças S_{LC} coincide com a ACP clássica, a menos de uma translação. Também foi possível visualizar as relações entre as variáveis, através de dendrogramas que resultaram de análises classificatórias hierárquicas ascendentes aplicadas directamente sobre as matrizes de semelhanças S, S_{LC} e P_L.

Palavras-Chave: variáveis heterogéneas, variáveis simbólicas, coeficiente de semelhança, coeficiente probabilístico, distância, distância euclidiana, análise em componentes principais, análise classificatória hierárquica ascendente.

ABSTRACT

This thesis focuses on data representation within the scope of Multivariate Data Analysis. The main aim is to study the simultaneous representation of variables with different levels of measurement. Its objectives are largely methodological and include software development and testing with biomathematical data. Several coefficients presented in the literature, some of which understudied, were considered for this purpose.

In Biomathematics, as in other areas, it is often useful to simultaneously analyse variables with different levels of measurement – this situation is frequent, for example, in research based on surveys and questionnaires. However, the heterogeneity of the types of variables can lead to mathematical and statistical problems, which are difficult to solve when multivariate Euclidian representations are intended. This thesis presents a proposal for the representation of heterogeneous variables.

In general, the analysed data are considered as a 'numerical' matrix, which crosses a set of individuals with a set of variables. This matrix is transformed into the matrix of similarities or dissimilarities that will be represented. This approach, more comprehensive than commonly used procedures, allows a unified treatment of several factorial methods (Principal Component Analysis and Correspondence Analysis, among others) and of Multidimensional Scaling.

The cases where an exact Euclidian representation is not usually feasible are also considered. Particular attention was given to monotonous transformations, which have a minimal interference with the original data. It is known that "additive constant" methods are not particularly reliable (as the constant is frequently extremely large with regards to the initial proximities). However, power function transformations seem satisfactory.

The coefficients s, s_{LC} e P_L (Le Calvé, 1977) and the affinity coefficients (a, a_W, a_δ, VAL_{AW}, $VAL_{A\delta}$ and those generalised – e.g., Bacelar-Nicolau, 1980, 1988, 2002), which allow the treatment of variables of the same or of different types, were considered in detail. The probabilistic approach of coefficients VAL_{AW}, $VAL_{A\delta}$, e P_L has the same origin (see Lerman, 1970). When variables of the same type are compared, some of these coefficients take the form of well-known coefficients, such as the Pearson correlation coefficient.

The coefficients s, s_{LC} e P_L were generalised to be used with symbolic data (e.g., Doria *et al*., 2007) and with sequential order variables. It was found that, in some cases, these coefficients allow the comparison of variables with missing data without the need for imputation methods. It was also found that they are particularly useful for the analysis of variables of partially ordered categories (e.g., Doria *et al*., 2006b) and for the comparison of ultrametrics (Doria *et al*., 2000). In addition, software was developed for calculation of coefficients s, s_{LC} and P_L and these coefficients were applied to biomathematical data (e.g., Doria *et al*., 2006a; Doria *et al*., 2007). The principal component analyses (PCA) of the matrixes S_{LC} and P_L generally produced good visualisations of data relationships. In the case when all variables are metric, the PCA of the similarity matrix S_{LC} matches the traditional PCA. The visualisation of relationships between variables was also possible using dendrograms, which resulted from hierarchical cluster analyses applied directly to the similarity matrixes S, S_{LC} and P_L.

Keywords: heterogeneous variables, symbolic variables, similarity coefficient, probabilistic coefficient, euclidian distance, principal component analysis, hierarchical cluster analysis.

INDICES

Índice Geral

Índice de Definições

Índice de Exemplos

AGRADECIMENTOS

À Senhora Professora Doutora Helena Bacelar Nicolau, por quem tenho uma grande estima, apresento o meu profundo reconhecimento pelo incentivo e apoio que me tem dado ao longo de largos anos para exercer duas actividades de que gosto especialmente – o ensino e a investigação – nas áreas da Estatística e Análise de Dados Multivariados aplicadas às ciências humanas e à medicina. Agradeço-lhe a orientação desta dissertação, assim como a sua confiança e o apoio ao longo deste processo. Tenho tido o privilégio de acompanhar o papel pioneiro da Professora Doutora Helena Bacelar Nicolau, assim como o do Professor Doutor Fernando Nicolau, na investigação e ensino de Análise de Dados. Ao integrar-me nos vários projectos de investigação que tem coordenado, a Professora Doutora Helena Bacelar Nicolau deu-me assim a oportunidade de conhecer investigadores de renome, nacionais e estrangeiros, em especial nas áreas da Estatística e Análise de Dados. Agradeço-lhe ter-me apresentado o Professor Doutor Georges Le Calvé e as facilidades concedidas para as deslocações em trabalho a Rennes, no âmbito dos projectos luso-franceses de investigação que tem liderado. Agradeço-lhe também as críticas e sugestões oportunas que contribuíram para melhorar o meu trabalho, bem como o apoio que sempre dispensou aos meus estudos na área de Análise de Dados, que conduziram a esta tese.

Ao Senhor Professor Doutor Georges Le Calvé quero expressar a minha sincera estima e agradecimento por ter co-orientado esta dissertação com a mestria, a experiência e competência que lhe são próprias. Foi com gratidão que sempre acolhi as suas sugestões e ideias, as críticas construtivas e os conhecimentos que me transmitiu. Agradeço-lhe o acompanhamento empenhado e interessado, o estímulo, a sua inteira disponibilidade e presença, apesar da distância física. Agradeço-lhe ainda a forma como fui recebida no seu laboratório (Laboratoire d'Analyse et Traitements des Données, da Universidade de Rennes 2) durante os vários estágios que ali realizei, sob a sua supervisão, bem como na sua casa. As suas excepcionais qualidades humanas e o seu bom-humor em muito facilitaram os meus estágios em França e o desenvolvimento deste trabalho. À Madame Simone Le Calvé, que sempre me acolheu com muita amizade e carinho, estou sinceramente agradecida e recordarei sempre os bons momentos que passámos.

Este processo contribuiu em muito para aprofundar os meus conhecimentos científicos, por isso não poderia deixar de agradecer à Professora Doutora Helena Bacelar Nicolau e ao Professor Doutor Georges Le Calvé a sua ajuda.

Uma palavra particular de agradecimento para a Professora Doutora Ana Sousa Ferreira – que de perto e com grande estima sempre me acompanhou – para a Drª Otília Dias, para a Professora Doutora Margarida Mendes Leal e para a Drª Teresa Rodrigues, pela amizade, ajuda, conselhos e incentivos dados neste percurso. Agradeço também aos meus outros colegas do Laboratório de Biomatemática da FML e do LEAD (FPCEUL), sem esquecer os elementos dos respectivos secretariados, em particular a Srª D. Conceição Carita e a Srª D. Ana Marcos.

Quero agradecer ainda aos membros do Laboratoire d'Analyse et Traitements des Données da Universidade de Rennes 2 e aos elementos do secretariado – em especial à Madame Marinette Kerveno – pela simpatia com que me acolheram nas minhas passagens por Rennes. Em particular, ao Professor Doutor Serge Joly agradeço sinceramente a amizade, o acompanhamento, as trocas de impressões frutuosas, as discussões construtivas e as facilidades que me proporcionou na organização das minhas estadias em Rennes com o apoio de Madame Edith Joly. Também agradeço ao Professor Doutor Farid Beninel e ao Professor Doutor Mohammed Bennani Dosse, os conselhos, as longas conversas sobre os assuntos da investigação e a disponibilização de documentação.

Ao Professor Doutor Manuel Silvério Marques (IPOFG, FML) agradeço a cedência dos dados do Questionário SERVQUAL Modificado, que permitiram ilustrar algumas das aplicações a dados reais desta tese, assim como o interesse que manifestou sobre os resultados obtidos e o convite feito para a sua publicação. Agradeço também à Professora Doutora Eugénia Duarte Silva (FPCEUL), à Professora Doutora Maria Amália Botelho (FCM-UNL), à Professora Doutora Rosa Novo (FPCEUL), à Drª Margarida Albuquerque (HSM, FML), à Professora Doutora Áurea Sousa (Dpt. Mat.-UA), à Drª Ana Luísa Santos (HSM) e à Drª Filomena Sousa (HSM) a cedência dos dados reais que utilizei na tese e os esclarecimentos úteis que sempre me facultaram.

Agradeço reconhecidamente ao Professor Doutor Paulo Gonçalves (INRIA) pela ajuda imprescindível na programação em MATLAB e o interesse que manifestou pelo meu trabalho, assim como ao Professor Doutor Sergio Camiz (Università di Roma "La Sapienza") a cedência do programa EUCAPP.

Agradeço ao Dr. Jorge Revez, bibliotecário da FPCEUL, a preciosa ajuda que me deu na obtenção de grande número de artigos que fazem parte das referências bibliográficas desta dissertação.

Aos outros colegas da FPCEUL, da FML e do CEAUL e aos meus alunos agradeço as palavras de encorajamento.

Aos meus amigos agradeço o apoio incondicional e incentivo que me deram neste período, em especial à Drª Maria Luísa Bazenga, que me tem acompanhado mais de perto no meu trabalho, nas vicissitudes deste percurso e que me acolheu sempre com muita alegria em Rennes.

Expresso profunda gratidão à minha família, em particular aos meus pais, marido, filhos, nora, neta, irmãos e sobrinhos, pelo amor e apoio ao longo deste trabalhoso caminho. Merecedores de um destaque especial são os meus filhos - Miguel, Gonçalo e José Maria - a quem agradeço a compreensão, o estímulo e a ajuda preciosa que me deram, contribuindo para a realização desta tese.

Grande parte das minhas deslocações a França foram financiadas pelo Serviço Científico e de Cooperação da Embaixada de França e pelo Ministério da Ciência e do Ensino Superior do Estado Português, a quem agradeço. Agradeço à Fundação Calouste Gulbenkian e ao CEAUL pelo apoio dado para a participação em Congressos, nacionais e internacionais, onde tive a oportunidade de apresentar e discutir alguns dos resultados da minha investigação.

Por último, manifesto a minha gratidão a todas as outras pessoas, aqui não nomeadas, mas que de uma maneira ou de outra, contribuíram para que concluísse esta tese.

INTRODUÇÃO GERAL

No nosso dia a dia, são-nos frequentemente apresentadas bases de dados em que estão registadas as informações obtidas por indivíduos ou objectos, i.e., por unidades estatísticas, sobre variáveis do mesmo tipo ou de diferentes tipos, que são consideradas pelo investigador de interesse para o problema em estudo. Com o objectivo de extrair informação dos dados recorremos habitualmente a técnicas de Estatística, assim como à Análise de Dados. Segundo alguns autores, a Análise de Dados é uma Estatística Exploratória Multidimensional (e.g., Lebart, Morineau e Piron, 1995[1]) constituída por técnicas concebidas para explorar os dados, com o objectivo de reconhecer a sua estrutura, ou seja, de encontrar um modelo[2] não aleatório: "O objectivo de todas as técnicas é, lato senso, o de mostrar ou extrair o sinal nos dados na presença de ruído, e o de encontrar o que os dados nos podem mostrar no meio de um caos aparente" (Everitt e Dunn, 1997)[3].

O objectivo da Análise de Dados é a procura do modelo subjacente aos dados multivariados, no caso de ele existir, "o modelo deve seguir os dados, não o inverso"[4] (Benzécri, 1980[5]), ou seja a exploração dos dados, no sentido introduzido por John Tukey: deixar os dados multivariados falarem por si mesmos. Nesta perspectiva, a análise de dados permitirá encontrar perguntas que conduzirão à formulação de hipóteses, possivelmente interessantes para estudos futuros[6].

Diversos autores, incluindo Saporta (1990), Bouroche e Saporta (1998) consideram a Análise de Dados como um conjunto de técnicas de visualização de dados multidimensionais (n indivíduos descritos por p variáveis), que enriqueceram a estatística descritiva ou exploratória, permitindo um estudo global dos indivíduos e das variáveis.
Existe um largo consenso sobre a importância das representações gráficas em análise de dados (e.g., Greenacre, 1978; Everitt e Dunn, 1997). Daí que se possa considerar que "o

---

[1] Já existe uma 3ª edição publicada em 2000.
[2] *Pattern*, em inglês. Padrão, configuração, perfil, forma, estrutura, em português.
[3] Tradução livre do original em inglês "The aim of all techniques is, in a general sense, to display or extract the signal in the data in the presence of noise, and to find out what the data show us in the midst of their apparent chaos."
[4] Tradução livre do original em francês "le modéle doit suivre les données, non l'inverse."
[5] A 1ª edição data de 1973.
[6] Este também é o objectivo do *data mining*.

principal problema em Análise de Dados seja o da representação visual dos dados compreensível por todos" (Le Calvé, 1988) [7].

Gráficos bem escolhidos, de acordo com o tipo de dados a representar, permitem detectar *outliers*[8], identificar estruturas/formas, procurar um fenómeno novo e eventualmente inesperado. Em Everitt e Dunn (1997) são apresentadas diversas técnicas de visualização que permitem explorar os dados multivariados antes de serem utilizadas análises mais formais e complexas.

Apesar das inúmeras propostas, a visualização de estruturas multivariadas continua a ser um desafio.

Os principais métodos de análise de dados multivariados constituem dois grupos: os métodos de classificação[9] (análise classificatória hierárquica, não hierárquica e mista) e os métodos factoriais (análise em componentes principais, análise das correspondências simples e múltiplas, análise factorial[10]/análise em factores comuns e específicos de Spearman e Thurstone (e.g., Lebart, Morineau e Piron, 1995). Mais recentemente, Kettenring (2006), ao fazer o ponto da situação da Análise Classificatória com a sugestão de algumas prioridades de investigação, afirma que: "A Análise Classificatória (AC), a Análise em Componentes Principais (ACP), e a Análise Discriminante (AD) são três dos principais métodos da análise multivariada moderna"[11].

Diversos autores também referem a convergência e a complementaridade da Análise de Dados Exploratória (ADE) e da Estatística Matemática Confirmatória (EMC). Bacelar-Nicolau e Nicolau (1994) fazem uma breve resenha histórica sobre a evolução observada e a tendência de convergência que se verifica entre aquelas duas áreas; no âmbito da análise classificatória, exemplificam este movimento de convergência e complementaridade da ADE e a da EMC.

---

[7] Tradução livre do original em inglês "The main problem in data analysis is the problem of the representation of the data by a visuable display understandable by everybody".

[8] Observações aberrantes.

[9] Esta é a designação utilizada, de forma geral, pela escola francesa: *Classification* (e.g., Saporta, Lebart), *Classification Automatique* (e. g., Lerman). Não confundir com Análise Discriminante, tal como muitos autores, em particular os anglo-saxónicos, reconhecem aquela designação - *Classification*.

Classification, em francês; *cluster analysis*, em inglês. Análise de agrupamentos, análise de grupos, análise de aglomerados, análise de conglomerados, são outras designações em uso na língua portuguesa. No dicionário, classificação é "o acto ou efeito de classificar; classificar é distribuir em classes; arrumar; ordenar; qualificar; determinar as categorias (de um conjunto); atribuir valores a".

[10] Em inglês, *factorial analysis*. Utilizada, principalmente, pelos psicólogos e psicometristas.

[11] Kettenring (2006) afirma que, em 2003, existiam mais de 1100 artigos publicados incluindo análises classificatórias, mais de 1600 utilizando análises em componentes principais e cerca de 700 aplicavam análise discriminante.

Mas, como se sabe, é impensável falar em Análise de Dados e não falar em Informática – o desenvolvimento desta área só foi possível devido à enorme e rápida evolução dos meios informáticos, a que assistimos[12], e que permitem e facilitam extraordinariamente os cálculos e as representações gráficas, abrindo assim as portas a novas maneiras de pensar a estatística.

Aqui, a partir de um conjunto de dados descritos por uma matriz/tabela (unidades estatísticas x variáveis), por uma matriz de distâncias ou por uma matriz de semelhanças, procuramos obter uma representação euclidiana destes dados, que melhor traduza a informação neles contida, uma vez que a nossa visão limita-nos a observação à folha de papel ou ao ecrã do computador. "Embora tenham sido propostos outros espaços, a representação euclidiana garante que as configurações podem ser interpretadas de uma forma natural e permite a utilização de numerosos métodos numéricos com confiança" (Caillez e Kuntz, 1996).

Na Análise de Dados o que é importante não são os indivíduos ou as variáveis, mas as relações entre eles ou entre elas. É isso que a Análise de Dados faz: as suas técnicas permitem representar relações entre os dados, quer sejam unidades estatísticas, quer sejam variáveis! Esta ideia está bem patente, por exemplo, em Le Calvé (1988) ou em Lebart *et al*. (1995) quando introduzem a ACP ou em Caillez e Kuntz (1996) quando definem o *multidimensional scaling*[13] (MDS) (Subsecção 1.4.4).

Estamos pois interessados nas medidas de proximidade, semelhanças ou dissemelhanças, e, em particular, nos coeficientes que permitem comparar variáveis de tipos diferentes, isto é, variáveis que designamos como sendo heterogéneas[14], pois é uma situação muito comum e ainda pouco tratada na prática.

Por outro lado, para algumas das técnicas de visualização utilizadas, uma técnica frequente é a projecção num espaço euclidiano. Quando a matriz das distâncias não é semidefinida positiva há que recorrer a transformações que permitam obter distâncias euclidianas. É no desenvolvimento deste tema que se centrará a nossa tese, com aplicação a dados reais nas

---

[12] No início dos anos sessenta (do século XX) os computadores foram usados, pela primeira vez, na análise de proximidades (R. N. Shepard, L. Guttman, e outros) e a meio dos anos sessenta a equipa de J. P. Benzécri utiliza-os pela primeira vez, em França, na análise das correspondências. Há cerca de 35 anos, perfuravam-se cartões, em que cada cartão continha as instruções correspondentes a uma linha de um programa informático actual.

[13] *Analyse métrique* (Le Calvé, apresentação pessoal), *Positionnement multidimensionnel*, em francês. Escalonamento multidimensional ou posicionamento multidimensional, em português.

[14] *Variables d'un type quelconque* (Le Calvé, 1977; Lerman, 1987), *variables de types divers, variables de types différents* (Lerman, 1987; Lerman e Peter, 2003*), variables hétérogènes* (Le Calvé, 1977; Chah, 1984, 1985), *variables de types hétérogènes*, *heterogenous variables* (Lerman e Peter, 2003; Chah, 1985). *Variables mixtes* (Pagès, 2004). *Mixed variables* (Cox e Cox, 2000).

áreas de interesse da Biomatemática. Na nossa abordagem a este tema, temos sempre presente que não há soluções únicas, mas há diferentes caminhos que nos conduzem a essas soluções, em particular, quando abordamos coeficientes que permitem relacionar variáveis heterogéneas.

Objectivos

Esta dissertação insere-se na área de Análise de Dados Multivariados, sob o tema de Representações de Dados, no âmbito da Biomatemática[15]. No contexto dos aspectos apresentados anteriormente nesta introdução, os objectivos da tese são essencialmente de ordem metodológica.

O principal objectivo é o de estudar a representação (especialmente a euclidiana) de variáveis de diferentes tipos, em particular:

a) Variáveis atributo de descrição,
b) Variáveis nominais,
c) Variáveis com modalidades parcialmente ordenadas,
d) Variáveis com modalidades totalmente ordenadas,
e) Variáveis com modalidades estrita e totalmente ordenadas,
f) Variáveis número de ordem,
g) Variáveis de ordem sequencial,
h) Variáveis métricas, e
i) Variáveis simbólicas/complexas (variáveis categóricas com valores múltiplos, variáveis categóricas com todas as modalidades ordenadas pelas unidades estatísticas, variáveis intervalares e variáveis modais).

Para este efeito foram usados os coeficientes s, s_{LC} e P_L (Le Calvé, 1977).

Pretende-se também aplicar estes coeficientes a dados reais do domínio da Biomatemática, de forma a procurar solucionar problemas habitualmente encontrados na análise de dados deste domínio (i.e. de forma a comparar diferentes tipos de variáveis) e a ultrapassar algumas limitações frequentes (e.g. relacionadas com não-respostas e dados omissos).

Pretende-se ainda, sempre que possível, comparar os resultados obtidos com os coeficientes s, s_{LC} e P_L (Le Calvé, 1977), com os obtidos quando se utilizam os coeficientes de afinidade a, a_W, a_δ, VAL_{AW}, $VAL_{A\delta}$ e os generalizados (e.g., Bacelar-Nicolau, 1980, 1988, 2002; Doria, 1989; Nicolau *et al*., 2007).

---

[15] A Biomatemática, entendida como a aplicação de métodos e modelos matemáticos para a compreensão de fenómenos biológicos (Carlos Braumann, apresentação pessoal).

Esta tese tem ainda os seguintes objectivos:

a) sintetizar e rever o estado actual da arte;
b) generalizar os coeficientes s, s_{LC} e P_L para se comparar variáveis de ordem sequencial e variáveis simbólicas;
c) desenvolver *software* para o cálculo dos coeficientes s, s_{LC} e P_L;
d) representar matrizes de semelhanças recorrendo à análise em componentes principais e a técnicas de análise classificatória;
e) estudar a comparação de ultramétricas.

Plano de tese

O que acabamos de referir leva-nos a apresentar o seguinte plano de tese, organizado em seis capítulos:

Capítulo 1: Trata de uma apresentação unificada de diversos métodos de representação geométrica multidimensional, a saber: Análise em Componentes Principais, algumas Análises Ordinais e *Multidimensional Scaling*[16]. São dados um conjunto de elementos (unidades estatísticas ou variáveis) e uma relação de semelhança ou de dissemelhança entre os elementos. Discute-se o problema de encontrar uma imagem destas estruturas de dados, especialmente no caso euclidiano. Aproveita-se para relembrar as definições dos principais conceitos, resultados e teoremas que lhes dão suporte no âmbito das proximidades e das estruturas euclidianas.

Capítulo 2: Neste capítulo apresentam-se transformações monótonas que permitem passar de relações de dissemelhança não euclidianas a euclidianas, para resolver o problema da representação gráfica no espaço euclidiano. Estuda-se, em particular, o caso das transformações com constantes aditivas e o caso da transformação pela função potência.

Capítulo 3: Apresenta-se o coeficiente de afinidade introduzido por K. Matusita em Estatística, retomado e generalizado por Bacelar-Nicolau e F. Nicolau. Este coeficiente, que tem sido usado em Análise de Dados, foi generalizado recentemente para ser utilizado com dados simbólicos. Interessa-nos, em particular, a abordagem probabilística deste coeficiente.

---

[16] *Analyse métrique* (Le Calvé, apresentação pessoal), *Positionnement multidimensionnel*, em francês. Escalonamento multidimensional ou posicionamento multidimensional, em português.

Capítulo 4: Estudam-se os coeficientes s, s_{LC} e P_L, que permitem comparar variáveis do mesmo tipo, bem como variáveis de diferentes tipos. Estes índices partem de uma ideia antiga de Daniels, retomada por Lerman e generalizada por Le Calvé. Generalizar-se-ão estes coeficientes para serem utilizados com dados simbólicos. Paralelamente, desenvolve-se *software* de apoio.

Capítulo 5: As técnicas estudadas nos capítulos precedentes são aplicadas ao tratamento de dados reais no domínio da Biomatemática.

Capítulo 6: Terminaremos a dissertação com as conclusões finais e a referência às perspectivas futuras de desenvolvimento deste trabalho.

No final da tese encontram-se a bibliografia e anexos (em CD).

1 APRESENTAÇÃO UNIFICADA DE DIVERSOS MÉTODOS DE REPRESENTAÇÃO GEOMÉTRICA MULTIDIMENSIONAL

1.1 Introdução

Em Análise de Dados, os dados apresentam-se-nos, habitualmente, de três maneiras, como veremos na Secção 1.2, sendo muito frequente a tabela/matriz que cruza a informação de n indivíduos e p variáveis, heterogéneas ou não. Seguindo o exposto por Le Calvé[17], Le Calvé (1988) e Joly e Le Calvé (1994), perante uma matriz de dados, pretendemos extrair informação dela; mas, para descrever uma variável é preciso compará-la com alguma coisa. Numa das grandes vias de comparação, os elementos[18] são comparados entre eles (não há memória de referência) – é o que se faz em Análise de Dados Exploratória (ADE). Na outra via, a da Estatística dita clássica, a comparação faz-se com um modelo preestabelecido. Há duas maneiras de proceder a estas comparações: - através das parecenças/semelhanças entre os elementos, utilizando coeficientes de semelhança (comparação global); - ou através das diferenças, utilizando índices de dissemelhança (comparações locais).

Quando se representam os dados, o que se analisa são as relações entre eles, nomeadamente as semelhanças e dissemelhanças, daí a Secção 1.3. Convém relembrar as definições dos principais conceitos sobre as proximidades (Subsecções 1.3.1, 1.3.2, 1.3.3 e 1.3.4.1) assim como resultados e teoremas principais que lhe dão suporte (Subsecção 1.3.4.2), no âmbito das estruturas euclidianas (Subsecção 1.3.5). Também se faz referência à nomenclatura utilizada, por diferentes autores da especialidade, para os mesmos conceitos. A terminologia utilizada nesta área não é universal, nem consensual, por isso temos que ser cuidadosos. Devido à importância das aplicações nas mais diversas áreas, e em particular, na esfera de interesses da biomatemática, abordam-se os coeficientes de semelhança no caso dos dados binários (Subsecção 1.3.6).
Quando recorremos às semelhanças somos levados, por exemplo, no caso particular de dados quantitativos, a utilizar análises em componentes principais para representar os

---

[17] *Les Représentations Planaires des Données Revisitées* - Conferência realizada na FPCEUL, em Maio de 2006.
[18] Os elementos tanto podem ser unidades estatísticas, como variáveis.

dados. No caso de recorrermos a dissemelhanças, se pensarmos em distâncias e em ultramétricas, situando-nos em espaços métricos, somos levados a utilizar grafos com valores[19] ou não, como por exemplo, classificações hierárquicas e árvores aditivas. A Secção 1.4 trata de uma apresentação unificada de diversos métodos de representação geométrica multidimensional, a saber: a Análise em Componentes Principais, algumas análises ordinais e *Multidimensional Scaling*. O estudo incide especialmente no caso euclidiano. A ACP pode ser vista como um exemplo de procura de imagem do produto escalar.

1.2 Os dados

Grandes conjuntos de dados constituem, frequentemente, o ponto de partida de todos os métodos de Análise de Dados. Estes dados, quer sejam obtidos em estudos observacionais ou como resultado de experiências, apresentam-se-nos sob uma das três formas gerais, que passamos a descrever.

1. O conjunto finito (ou não) dos indivíduos ou sujeitos ou objectos, i.e., das unidades estatísticas, I, pode ser descrito por uma matriz de dados, X(nxp), em que se registam as observações de n unidades estatísticas em relação a p variáveis qualitativas e/ou quantitativas, X_i (i=1, ..., p): $X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \dots & x_{ij} & \dots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$. As variáveis, X_i (i = 1, ..., p), tomam os seus valores num conjunto F munido ou não de uma estrutura algébrica.

Ao investigador competirá escolher e definir as unidades estatísticas e as variáveis de forma a que lhe seja possível obter respostas sobre o assunto que motiva a sua investigação. Tarefa delicada e nem sempre fácil, pois as variáveis, segundo a sua natureza, podem ser codificadas de diversas maneiras e nem sempre é possível encontrar forma de as quantificar. Para assuntos específicos, existem escalas já estudadas e validadas em várias áreas, nomeadamente em medicina ou em psicologia (e.g., psicometria) e educação, que permitem avaliar os indivíduos. Na área da saúde, encontra-se em Streiner e Norman (2003) uma boa sensibilização para esta problemática. A influência da forma de apresentar as questões nas respostas dadas pelos sujeitos encontra-se bem patente, por exemplo, em Schwarz (1999).

---

[19] Grafos com pesos. *Graphes valués*, no original.

Ao estatístico competirá, entre outras coisas, ter em consideração a natureza das variáveis. Neste contexto, existem várias tipologias de variáveis, bem conhecidas, apresentadas por vários autores:

- Uma tipologia muito geral, baseada na possibilidade ou não de quantificar as varáveis, considera dois grandes grupos: variáveis quantitativas e variáveis qualitativas ou categóricas[20].
- Tipologia baseada no número de valores distintos que uma variável pode assumir (Anderberg, 1973): Variável contínua, variável discreta e variável binária ou dicotómica (é uma variável discreta que só pode tomar dois valores).
- Tipologia baseada no "nível de medição"[21] das variáveis: nominal, ordinal, de intervalo e de razão. Mais recentemente, também se utiliza a designação de variável métrica[22] que engloba as variáveis de intervalo e de razão.

Existem também outras tipologias de variáveis (e.g., Lerman, 1973, 1981, 1987; Chah, 1984, 1985; Térouanne, 1998), das quais destacamos uma tipologia mais detalhada, proposta por Le Calvé (1977) e a tipologia referente a variáveis simbólicas[23]/complexas (Bock, 2000a). Adoptamos a tipologia proposta por Le Calvé (1977) (muito semelhante à proposta por Lerman (1973)), à qual acrescentámos algumas variáveis simbólicas/complexas que tratámos:

- Variáveis nominais
    - variáveis atributos de descrição (Subsecção 4.3.2)
    - variáveis nominais/características descritivas (Subsecção 4.3.3)
- Variáveis ordinais
    - variáveis com modalidades parcialmente ordenadas (Subsecção 4.3.4)
    - variáveis com modalidades totalmente ordenadas (Subsecção 4.3.5)
    - variáveis com modalidades estrita e totalmente ordenadas (Subsecção 4.3.6)
    - variáveis número de ordem[24] (Subsecção 4.3.7)
    - variáveis de ordem sequencial (Subsecção 4.3.8)
- Variáveis métricas (Subsecção 4.3.9)
- Variáveis simbólicas/complexas
    - variáveis categóricas com valores múltiplos (Subsecções 4.3.10.1 e 4.3.10.4)

---

[20] Categorizada.
[21] Conceito introduzido por S. Smith Stevens (1946): "On the theory of scales of measurement", *Science 103, 677-680*.
[22] *Scale*, em inglês. Outros autores designam-na por variável numérica.
[23] *Symbolic variables*, em inglês.
[24] *Variables rang*, em francês.

- variáveis categóricas com todas as modalidades ordenadas pelas unidades estatísticas (Subsecção 4.3.10.4)
- variáveis intervalares (Subsecção 4.3.10.5)
- variáveis modais (Subsecção 4.3.10.8)

As definições destas variáveis são apresentadas nas subsecções indicadas. Tivemos dificuldade em traduzir as designações de algumas delas, em particular, a tradução da designação *variable rang*[25] suscitou-nos alguma reflexão. Somos confrontados com a existência de vários termos, quer em inglês, quer em francês, ligados ao aspecto ordinal das variáveis - *rank*/*rang*, *ranking*, *rate*/*ordre* – dos quais, segundo pensamos, não é habitual fazer grande distinção quando se escolhem os coeficientes de semelhança. O facto de Le Calvé considerar a classificação de variável ordinal, mais detalhada do que é habitual, põe em evidência esta distinção entre *rate* e *ranking*, i.e., entre atribuir pontuações de uma escala ordinal (*rate*) e, ordenar (*rank*), atribuindo o valor da posição ocupada pela unidade estatística na sequência, ou seja atribuindo números de ordem ou ordens ou postos segundo determinado critério (o definido pela variável). No Glossário Estatístico Inglês-Português[26] traduz-se, segundo o contexto, *rank* por posto[27] ou por ordinal, e *order* por ordem ou ordinal. No IATE (Inter Active Terminology for Europe), *rank* e *rang* são traduzidos por ordem ou por número de ordem (atribuído às unidades estatísticas numa lista ordenada ou sequência). A tradução destes termos para português é delicada, mas preferimos as últimas, que sempre utilizámos. Em todo o caso, trata-se apenas de uma convenção de nomenclatura, que sendo aqui clarificada fica explícita. Como se sabe, existem técnicas estatísticas, nomeadamente, algumas das não paramétricas, que se baseiam nos números de ordem atribuídos às unidades estatísticas, tendo em conta a natureza ordinal das variáveis, embora estas possam ser medidas em escalas métricas - daí a necessidade de utilizar a noção de ordem.

Ao investigador caberá a decisão de escolher a escala de medição da(s) variável(eis), tendo em atenção o objectivo da investigação. Por exemplo, o nível de colesterol sanguíneo total (mg/dL) é uma variável quantitativa, mas o investigador pode estar interessado apenas no aspecto ordinal desta variável, entrando em linha de conta com o risco de doença cardíaca (1. Nível desejável: menor risco de doença cardíaca, 2. Limiar de alto risco, 3. Nível não desejável: alto risco) ou considerá-la como binária (0. Nível desejável, 1. Nível indesejável).

---

[25] *Rang*, em francês; *rank*, em inglês.
[26] Encontra-se no *website* da Sociedade Portuguesa de Estatística (SPE). Neste dicionário, ainda em construção, encontra-se a tradução dos termos em inglês para português e brasileiro.
[27] Posto, em brasileiro.

2. O conjunto das unidades estatísticas, I, pode aparecer descrito por uma relação de semelhança S estabelecida sobre o conjunto das p variáveis X_i (i=1,...p) ou sobre os seus n elementos.

Na prática estes dados são apresentados sob a forma de uma matriz, S, geralmente simétrica. Os elementos desta matriz correspondem aos valores de uma relação de semelhança definida sobre o conjunto das variáveis ou sobre as unidades estatísticas, que foram calculados a partir da matriz de dados X ou obtidos directamente, de que se apresentam em seguida alguns exemplos.

Exemplo 1.2.1. Matrizes de semelhanças calculadas

- Matrizes de correlações e de covariâncias entre pares de variáveis.

Neste caso, os elementos da matriz são os valores da correlação ou da covariância entre pares de variáveis, calculados a partir da matriz de dados.

Exemplo 1.2.2. Matrizes de semelhanças obtidas directamente

- Em algumas situações a matriz de semelhanças pode ocorrer directamente, sem haver necessidade de realizar cálculos. É o caso das matrizes de proximidades que resultam de experiências em que se pede aos sujeitos para julgarem as semelhanças entre dois estímulos (matrizes sociométricas) ou, por exemplo, na percepção entre cores (Helm, 1964).
- Matriz de semelhanças entre nove sintomas de depressão (Streiner e Norman, 2003).
- Matrizes de contiguidade[28] (matrizes dos países da Europa, se têm fronteira comum (1) ou não (0), ou registar o comprimento da fronteira comum), matrizes de trocas económicas/comerciais entre dois países. Matrizes de co-ocorrência (n.º de ninhos infectados pelos cogumelos). Matrizes de ocorrência (n.º de clientes comuns a duas empresas, n.º de produtos idênticos que têm no catálogo).
- Matrizes de semelhanças genéticas (Perrier, 1998).
- Matrizes de semelhanças que se referem ao consenso[29]. Por exemplo, numa experiência de psico-linguística, a semelhança entre dois conceitos exprime-se pelo número de inquiridos que atribuem os conceitos ao mesmo grupo (Mirkin, 2008, descreve e analisa estes dados, com referência a Rosenberg, 1982).

3. O conjunto das unidades estatísticas, I, ou o das variáveis $\{X_i, i=1,...,p\}$, também pode aparecer descrito por uma relação de dissemelhança, D**,** entre os seus elementos:

---

[28] *Contiguité,* em francês.
[29] *Consensus*, no original.

$$D = \begin{bmatrix} d_{11} & \dots & d_{1n} \\ \dots & d_{ij} & \dots \\ d_{n1} & \dots & d_{nn} \end{bmatrix}$$ O elemento d_{ij} representa, geralmente, a dissemelhança ou a distância entre os indivíduos i e j, d_{ij}= d(i,j). Habitualmente, em linguagem comum, utiliza-se a designação de distância em vez de dissemelhança, sem a preocupação das suas propriedades matemáticas. É o que faremos nos exemplos que se apresentam em seguida, tal como outros autores.

Estas distâncias podem ser calculadas a partir da matriz de dados X ou podem ocorrer por observação directa, como se exemplifica de seguida.

Exemplo 1.2.3. Distâncias observadas ou avaliadas

- Cartas geográficas que representam os tempos de percurso (Love e Morris, 1972) ou a matriz de distância da duração do percurso aéreo entre 26 cidades francesas (Al Ayoubi, 1991); as clássicas matrizes de distâncias por estrada entre cidades (Everitt e Rabe-Hesketh, 1997) ou de distâncias aéreas entre cidades americanas (exemplo de Kruskal e Wish, 1978, apresentado por Everitt e Dunn, 1997).
- Dados de mutações. A matriz de dissemelhança apresentada em Hartigan (1975) que se refere às distâncias entre vinte espécies animais e vegetais, entre as quais se encontra o homem e o macaco. A distância entre duas espécies é dada pelo número de posições na molécula da proteína citocromo-c em que duas espécies têm diferentes aminoácidos (exemplo tratado por Beninel, 1987, e Al Ayoubi, 1991, entre outros).
- Distâncias sensoriais, matrizes sociométricas – distâncias afectivas. Matrizes de confusão (confusão entre as letras - número de vezes que se enganam entre duas letras; confusão entre formas, ou entre vozes).
- Distância das fronteiras entre os países (é uma dissemelhança).
- Distâncias morfológicas (entre duas caras, entre uma cara e uma fotografia de identidade).
- Distância em tempo de percurso (Rennes fica a 2h de Paris, em TGV. Rennes está mais próximo de Paris do que de Brest), o que permite construir mapas que se baseiam no tempo de percurso em vez da distância (anamorfose).

Exemplo 1.2.4. Distâncias calculadas

- Em biologia, o cálculo de distâncias entre os alinhamentos de sequências múltiplas constitui um objectivo importante de investigação (Rajasekaran *et al*., 2004).
- Matriz de distâncias calculadas entre os alinhamentos das sequências de proteínas, utilizando a distância de Felsenstein (Wiwanitkit, 2004).

- Distâncias genéticas (e.g., Lefort-Buson e Vienne, 1985).
- Distância euclidiana entre os alunos de um curso línguas, baseada na informação das suas classificações (Gower e Blasius, 2005).

Everitt e Rabe-Hesketh (1997) e Borg e Groenen (2005), entre outros autores, dão exemplos de proximidades (semelhanças/dissemelhanças) e apresentam maneiras de as obter.
Para os dados de proximidade, também existem várias tipologias, tais como, os esquemas de classificação propostos por Carroll e Arabie (1980) e Jacoby (1991), citados e exemplificados em Everitt e Rabe-Hesketh (1997).

1.3 Dissemelhanças e semelhanças

Em análise de dados multivariados recorre-se frequentemente ao cálculo de coeficientes de comparação[30] ou de proximidade[31] entre os pares de elementos do conjunto (unidades estatísticas ou variáveis) que se pretende descrever. Os coeficientes de comparação podem ser do tipo dissemelhança ou do tipo semelhança, conforme quantificam quão diferentes ou semelhantes são os elementos do conjunto a descrever (objectos, indivíduos, estímulos ou outras entidades que tenham interesse, assim como variáveis).

Nicolau (1980) utiliza o conceito geral de função de comparação, definindo-a da seguinte maneira:

Definição 1.3.1. Função de comparação

Uma função de comparação, $Y(i,j)$, entre elementos de I é uma aplicação Y, $Y: I \times I \rightarrow \mathbb{R}$, que pode ser:

- do tipo dissemelhança, em que pequenos valores de $Y(i,j)$ representam elevada "semelhança" entre os elementos. É, habitualmente, designada por d.

ou

- do tipo semelhança, em que grandes valores de $Y(i,j)$ representam elevada "semelhança" entre os elementos. É, habitualmente, designada por s.

---

[30] Medida de associação (Anderberg, 1973); índice de proximidade (Chandon e Pinson, 1981); medida, índice, coeficiente de semelhança, semelhança ou proximidade (Lerman, 1981); proximidade (Caillez e Kuntz, 1996)*; resemblance measure* (Esposito *et al*., 2000).

[31] O termo *proximidade* é utilizado, de maneira geral, para designar quer a dissemelhança, quer a semelhança entre pares de objectos, em MDS.

Daí a interpretação óbvia: s(i,j) > s(k,l) e d(i,j) < d(k,l) significam que i é mais parecido com j do que k com l. Quanto maior for o índice de dissemelhança entre os indivíduos, mais estes serão considerados diferentes.

Bacelar-Nicolau (1980) reúne as definições de índice de semelhança e de índice de dissemelhança e apresenta uma definição de coeficiente de comparação mais restrita do que a apresentada anteriormente, coincidente com a definição apresentada por Esposito, Malerba, Tamma e Bock (2000).

Definição 1.3.2. Coeficiente de comparação

Um coeficiente de comparação entre pares de elementos de I é uma aplicação C,

C: $I \times I \rightarrow \mathbb{R}$, tal que:

- $\forall i, j \in I$, $C(i,j) = C(j,i)$
- $\forall i, j \in I$, $C(i,j) \leq C(i,i)$, se C avalia uma semelhança

  $C(i,j) \geq C(i,i)$, se C avalia uma dissemelhança.

Para cada um dos tipos de função de comparação – dissemelhança, semelhança – são apresentadas algumas definições de interesse, nomeadamente a de vários coeficientes de dissemelhança e de semelhança ou, apenas, referências a eles. A escolha dos coeficientes de comparação constitui uma preocupação uma vez que esta escolha pode influenciar o resultado da análise. Não se pode dizer que um coeficiente é melhor do que todos os outros; no entanto, a escolha adequada depende certamente da natureza das variáveis e do contexto em que os dados estão inseridos. Nesta escolha também se deve ter em consideração o(s) método(s) de Análise de Dados que se pretende(m) usar. Por isso, há a preocupação de referir as propriedades métrica e euclidiana dos coeficientes de dissemelhança; só por abuso de linguagem se faz referência àquelas propriedades para os coeficientes de semelhança (as propriedades métrica e euclidiana são dos coeficientes de dissemelhança correspondentes).

1.3.1 Dissemelhanças particulares

Para as funções de comparação do tipo dissemelhança podem-se considerar as definições e propriedades que se seguem.

Definição 1.3.3. Dissemelhança

Seja I um conjunto finito, não vazio, com dimensão n ($n \in \mathbb{N}$).

Define-se, sobre $I \times I$, uma função d com valores reais, $d: I \times I \rightarrow \mathbb{R}$.

d é uma dissemelhança[32] sobre I, tal que:

$\forall i, j \in I,\ d(i,j) \geq 0$ (1.3.1)

$\forall i \in I,\ d(i,i) = 0$ (1.3.2)

$\forall i, j \in I,\ d(i,j) = d(j,i)$ (1.3.3)

Observações:

1. I é o conjunto dos elementos (unidades estatísticas ou variáveis) sobre os quais se pretende fazer a análise.
2. Certas medidas de dissemelhança podem conduzir a medidas não simétricas, como por exemplo, os tempos de percurso entre duas localidades (Perrier, 1998).

Dizemos que d é uma dissemelhança não simétrica quando verifica apenas as condições (1.3.1) e (1.3.2).

3. Também podem existir dissemelhanças negativas, como por exemplo, distâncias em astronomia.
4. Quando d(i,j)= 0, em geral, isto significa apenas a igualdade das variáveis observadas e não implica a identidade das duas unidades estatísticas.
5. A definição de dissemelhança permite ter ao mesmo tempo: d(i,j) = 0 e $d(i,k) \neq d(j,k)$. A definição de dissemelhança semi-própria, que se apresenta em seguida, permite ultrapassar esta incoerência. Interessa-nos, pois trabalhar com dissemelhanças semi-próprias.
6. Também pode acontecer que d(i, i)>0 (Esposito, Malerba, Tamma e Bock, 2000).

Definição 1.3.4. Dissemelhança semi-própria

Uma dissemelhança d, definida em I, é uma dissemelhança semi-própria[33], quando se verifica a condição: $\forall i, j \in I,\ d(i,j) = 0 \Rightarrow (\forall k \in I,\ d(i,k) = d(j,k))$ (1.3.4)

Definição 1.3.5. Dissemelhança própria

Uma dissemelhança d, definida em I, é uma dissemelhança própria[34], quando se verifica a condição: $\forall i, j \in I,\ d(i,j) = 0 \Leftrightarrow i = j$ (1.3.5)

---

[32] Índice de dissemelhança (Nicolau, 1980); indice d'écart (Barthélemy e Guénoche, 1988).

[33] *Semi-propre*, em francês (Fichet e Le Calvé, 1984); em inglês: *semi-proper* (Fichet), *even* (Critchley, 1986), semi-definite (Joly e Le Calvé, 1994).

Definição 1.3.6. Semi-distância

Uma dissemelhança d, definida em I, é uma semi-distância[35], quando respeita a desigualdade triangular[36]: $\forall i, j, k \in \mathrm{I}$, $d(i,j) \leq d(i,k) + d(k,j)$ (1.3.6)

Quando a igualdade é verificada, diz-se, por abuso de linguagem, que os três pontos estão alinhados, com k situado entre i e j.

Definição 1.3.7. Distância

Uma distância[37] d é uma semi-distância que é própria, i. e., é uma dissemelhança própria que respeita a desigualdade triangular:

$\forall i, j \in \mathrm{I}$, $d(i,j) \geq 0$

$\forall i \in \mathrm{I}$, $d(i,i) = 0$

$\forall i, j \in \mathrm{I}$, $d(i,j) = d(j,i)$

$\forall i, j \in \mathrm{I}$, $d(i,j) = 0 \Leftrightarrow i = j$

$\forall i, j, k \in \mathrm{I}$, $d(i,j) \leq d(i,k) + d(k,j)$

Também se pode dizer que a distância d é uma dissemelhança que verifica a desigualdade triangular e o axioma da separabilidade: $d(i,j) = 0 \Leftrightarrow i = j$ (Joly e Le Calvé, 1986).

Definição 1.3.8. Distância euclidiana

Uma distância d é euclidiana se e só se verifica a condição:

Para todo o $i \in \mathrm{I}$, existe um vector $x_i = (x_{i1}, ..., x_{ik}, ..., x_{ip}) \in \mathbb{R}^p$ tal que:

$$\forall i, j \in \mathrm{I}, \quad d(i,j) = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{1/2} \qquad (1.3.7)$$

- A distância euclidiana é um caso particular da distância de Minkowski de ordem 2; é pois uma distância de tipo M^2 (Tabela 1.3.1).

---

[34] *Propre*, em francês (Fichet e Le Calvé, 1984); *definite*, em inglês; índice de distância (Cailliez e Pagès, 1976; Nicolau, 1980); índice de distância ou dissemelhança (Barthélemy e Guénoche, 1988).

[35] *Écart* (Cailliez e Pagès, 1976; Barthélemy e Guénoche, 1988); métrica (Gower e Legendre, 1986); semi-métrica (Caillez e Kuntz, 1996); pseudo-métrica ou semi-distância (Esposito *et al*., 2000).

[36] Desigualdade métrica (Gower e Legendre, 1986; Everitt e Rabe-Hesketh, 1997).

[37] Métrica (Caillez e Kuntz, 1996); métrica ou distância (Esposito *et al*., 2000).

- A distância euclidiana é influenciada pelas diferentes unidades de medida presentes nos dados, daí a sugestão de se padronizar cada variável pelo seu desvio padrão ou mesmo pela amplitude total.
- Esta distância não deve ser usada para comparar locais com base em informação relativa à abundância das espécies (em ecologia), pois pode ocorrer o seguinte paradoxo descrito por Legendre e Legendre (2000), com referência ao exemplo de Orlóci (1978): dois locais que não tenham nenhuma espécie em comum podem estar a uma distância menor do que dois locais que partilhem espécies. Este problema também é referido como o problema dos "zeros duplos"[38]. Nesta situação é aconselhável recorrer à distância corda (Tabela 1.3.1).

A distância euclidiana é estudada com mais pormenor na Subsecção 1.3.5.

Definição 1.3.9. Ultramétrica

Uma dissemelhança d definida em I é uma ultramétrica, se verifica a condição:

$$\forall i, j, k \in I,\ d(i,j) \leq \max [d(i,k), d(j,k)] \qquad (1.3.8)$$

O que equivale a dizer, sob o ponto de vista geométrico, que num triângulo qualquer lado é menor ou igual do que o maior dos outros dois. Ou, ainda de forma equivalente, todo o triângulo ou é equilátero ou é isósceles; no caso de ser isósceles, a base é o menor lado.

- Desta definição resulta imediatamente que uma ultramétrica sobre I é uma distância sobre I.
- Entre outras propriedades, sabe-se que "Uma distância ultramétrica sobre I é uma distância euclidiana de dimensão (n-1), (sendo n=card (I))" (Teorema de Holman, 1972). Este teorema é recordado e demonstrado por Joly e Le Calvé (1986).

Sabe-se que (e.g., Perrier, 1998):

- Considerando as funções d que verificam as propriedades (1.3.2) e (1.3.3) apresentadas na Definição 1.3.3 de dissemelhança, o conjunto das funções d, $\mathfrak{D}$, munido da adição e da multiplicação por um escalar é um espaço vectorial de dimensão n(n-1)/2.
- As n(n-1)/2 funções $d^{\{i,j\}}$ formam uma base canónica de $\mathfrak{D}$, sendo $d^{\{i,j\}}$ definida para todo o par em que $i \neq j$, tal que:

[38] *Double-zeros*, no original.

$$\forall k,l \in \mathrm{I},\ \ d^{\{i,j\}}\left(k,l\right) = \begin{cases} 1 \text{ se } \{k,l\}=\{i,j\} \\ 0 \text{ senão} \end{cases} \quad (1.3.9)$$

- O conjunto $\mathfrak{D}_+$ de todas as dissemelhanças é o ortante[39] não negativo de $\mathfrak{D}$ na base canónica. Mostra-se que $\mathfrak{D}_+$ é um "cone pontiagudo"[40], convexo e fechado de $\mathfrak{D}$, ao qual pertence o elemento d_0 tal que: $\forall i, j \in \mathrm{I},\ d_0(i,j) = 0$ (1.3.10)
- O subconjunto $\mathfrak{D}_\infty$ das semi-distâncias também é um "cone pontiagudo", convexo e fechado de $\mathfrak{D}$.
- Um espaço de dissemelhança é um par ordenado (I, d), $d \in \mathfrak{D}_+$.
- Um espaço de dissemelhança (I, d) diz-se mergulhável num espaço métrico (X, ρ) se existe uma função ϕ de I em X tal que: $\forall i, j \in \mathrm{I}, \rho(\phi(i),\phi(j)) = d(i,j)$ (1.3.11)
- Se existir um espaço métrico (X, ρ) no qual o espaço (I, d) possa ser mergulhado, então d é uma semi-distância.
- A dissemelhança dir-se-á do tipo T quando o seu espaço de dissemelhanças tiver a propriedade de ser mergulhável num espaço métrico do tipo T. O espaço métrico poderá ser um espaço vectorial de dimensão finita munido de uma norma.

A escolha dos coeficientes de dissemelhança também é orientada pela natureza das variáveis presentes na base de dados. Tradicionalmente são propostas medidas de dissemelhança entre objectos descritos por variáveis do mesmo tipo, tais como:

- No caso das variáveis quantitativas: a distância de Minkowski e os seus casos particulares – euclidiana e *city block* -, as dissemelhanças de Camberra, de Bray-Curtis e corda (Tabela 1.3.1) (e.g., Gower e Legendre, 1986; Everitt e Rabe-Hesketh, 1997; Legendre e Legendre, 2000). As distâncias euclidiana, *city block* e de Minkowski são influenciadas pelas diferentes unidades de medida presentes nos dados, daí a sugestão de se padronizar cada variável pelo seu desvio padrão ou mesmo pela amplitude total. A distância corda foi proposta para ultrapassar o problema, surgido em ecologia, da utilização da distância euclidiana quando se comparam locais com base na abundância das espécies (ver Subsecção 1.3.1 e e.g., Legendre e Legendre, 2000).
- Adaptações daquelas distâncias para variáveis binárias são estudadas e comparadas por Gower e Legendre (1986) e Sarkar e Saiful Islam (1999), entre

[39] Tradução livre de *orthant*. Em geometria, um *orthant* fechado é um dos 2^n subconjuntos de um espaço euclidiano a n dimensões definido restringindo cada eixo de coordenadas cartesianas a ser não negativo ou não positivo. Isto é, um *orthant* fechado é análogo a um quadrante fechado no plano e a um octante fechado num espaço a três dimensões.

[40] Tradução livre de *cône pointu*.

outros autores. Os coeficientes de dissemelhança para dados binários são numerosos e são estudados na Subsecção 1.3.6.

- No caso das variáveis qualitativas é utilizada, entre outras, a distância qui-quadrado (distância χ^2), $d_{ii'} = \sqrt{\sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2}$ (1.3.12), em que f_{ij} designa a frequência relativa dos acontecimentos i e j (Lebart *et al*., 1995). Esta é a distância preservada na Análise das Correspondências. Mais geralmente, esta distância é usada para calcular a associação entre as linhas ou colunas de uma tabela de contingência.
- Quando as variáveis são simbólicas, são numerosas as propostas de coeficientes de dissemelhança (e.g., Esposito, Malerba e Tamma, 2000; Bock, 2000b), em particular, baseadas nas funções de distribuição das variáveis.

Tabela 1.3.1. Algumas dissemelhanças para dados quantitativos e suas propriedades métricas (semi-distância) (Gower e Legendre, 1986). A distância D_5 é euclidiana só para valores positivos

Dissemelhanças	Fórmula	Métrica/Semi-distância
Minkowski	$d_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^t \right)^{1/t}$, $t \geq 1$	*Sim*
City block (distância de tipo M^1)	$d_{ij} = \sum_{k=1}^{p} \left\| x_{ik} - x_{jk} \right\|$	*Sim*
Distância euclidiana (distância de tipo M^2)	$d_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{1/2}$	*Sim*
Distância euclidiana padronizada *Sendo: r_k=desvio padrão da variável x_k* *ou r_k=amplitute total da variável x_k*	$d_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \Big/ r_k^2 \right)^{1/2}$	*Sim*
Camberra	$d_{ij} = \sum_{k=1}^{p} \frac{\left\| x_{ik} - x_{jk} \right\|}{\left\| x_{ik} + x_{jk} \right\|}$	*Sim só para valores positivos*
Distância D_5 (Gower e Legendre, 1986) *(Distância associada à métrica de Camberra)*	$d_{ij} = \left(\sum_{k=1}^{p} \frac{\left(x_{ik} - x_{jk} \right)^2}{\left(x_{ik} + x_{jk} \right)^2} \right)^{1/2}$	*Sim só para valores positivos*
Bray-Curtis	$d_{ij} = \frac{\sum_{k=1}^{p} \left\| x_{ik} - x_{jk} \right\|}{\sum_{k=1}^{p} \left\| x_{ik} + x_{jk} \right\|}$	*Não*
Corda	$d_{ij} = \left(2 \left(1 - \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2}} \right) \right)^{1/2}$	*Sim*[1]

[1] Legendre e Legendre, 2000.

No caso das variáveis heterogéneas, dois coeficientes generalizam o coeficiente de Gower (1971) (Subsecção 1.3.2), outro surge no âmbito do reconhecimento de formas[41], assim como em análise simbólica:

- O coeficiente de dissemelhança geral definido no contexto da análise classificatória por Gordon (1990); este autor, baseando-se no coeficiente de semelhança geral de Gower (1971), propõe um coeficiente de dissemelhança geral entre objectos/unidades estatísticas descritos por variáveis de vários tipos (variáveis numérica, ordinal e categórica) dando especial atenção à problemática da ponderação das variáveis. Este coeficiente coincide, em casos particulares, com a métrica de *city block*, com o quadrado da distância euclidiana, ou com a métrica de Minkowski ponderada.
- O coeficiente de dissemelhança entre objectos descritos por variáveis misturadas[42] (quantitativas, binárias, nominais e ordinais) sugerido por Cox e Cox (2000) é uma extensão do coeficiente de Gower (1971) que também calcula simultaneamente as dissemelhanças entre as variáveis. A definição das dissemelhanças incorpora automaticamente a ponderação dos indivíduos e das variáveis. As dissemelhanças assim calculadas podem ser usadas em *multidimensional scaling* ou com qualquer outra técnica que utilize proximidades.
- No âmbito do reconhecimento de formas, Ichino (1988) e Ichino e Yaguchi (1994) propõem funções de distância gerais para comparar objectos descritos por variáveis mistas (quantitativas discretas e contínuas, qualitativas nominais e ordinais, e estruturais[43]) que se baseiam na métrica de Minkowski de ordem t generalizada. Nos exemplos que apresentam utilizam aquelas distâncias associadas a métodos de Análise de Dados (ACHA e ACP[44]).

De forma geral, Esposito, Malerba, Tamma e Bock (2000) referem dois métodos possíveis para tratar da comparação de objectos descritos por variáveis heterogéneas: combinação linear ponderada de medidas de dissemelhança (entre variáveis do mesmo tipo) e redução dos dados a variáveis do mesmo tipo.

Entre as várias dissemelhanças particulares que se podem definir[45] – distâncias de Minkowski (ou do tipo M_t), distâncias de cadeia, distâncias ao centro, distâncias

---

[41] *Pattern recognition*, no original.
[42] *Mixed variables*, no original.
[43] Pensamos que os autores se referem a variáveis simbólicas.
[44] Generalização da ACP que se baseia nas métricas de Minkowski generalizadas (H. Yaguchi e M. Ichino "A generalized principal component analysis for mixed measurement level data", *Trans. IEICE* Japan, vol. J75-A, no. 10, 1580-1589 (em japonês), 1992).
[45] Por exemplo, Beninel (1987), Al Ayoubi (1991), Perrier (1998), Le Calvé (1985).

ultramétricas, distâncias aditivas, dissemelhanças de Robinson, dissemelhanças sobre as partes de um conjunto (e.g., Capítulo 3, Subsecção 4.3.10), ... - vai-nos interessar estudar, em especial, a distância euclidiana na perspectiva da sua representação gráfica (Subsecção 1.3.5).

A escolha dos coeficientes de dissemelhança também é orientada pelas suas propriedades; o facto da dissemelhança ser euclidiana é um factor que pesa na sua escolha, atendendo aos métodos de Análise de Dados que serão usados (especialmente no caso do *Multidimensional Scaling*). Na Subsecção 1.3.4 sobre as funções de semelhança também se encontram sugestões para esta questão.

1.3.2 Semelhanças particulares

Das funções de comparação do tipo semelhança podemos considerar as definições e propriedades que se seguem.

Definição 1.3.10. Semelhança

Seja I um conjunto finito com dimensão n.

Define-se, sobre I x I, uma função *s* com valores reais, s: $I \times I \rightarrow \mathbb{R}$.

s é uma semelhança[46] sobre I, quando

$\forall i, j \in I,\ s(i,j) \geq 0$ (1.3.13)

$\forall i, j \in I,\ s(i,i) \geq s(i,j)$ (1.3.14)

$\forall i, j \in I,\ s(i,j) = s(j,i)$ (1.3.15)

Poder-se-ia também definir semelhança negativa. Bastaria pensar no coeficiente de correlação.

Definição 1.3.11. Semelhança própria

Uma semelhança s, definida em I, é uma semelhança própria[47], quando se verifica a condição: $\forall i, j \in I,\ s(i,i) > s(i,j)$ (1.3.16)

Ou, de forma equivalente, $s(i,j) = s_{max} \Rightarrow i = j$ (1.3.16')

---

[46] *Índice de semelhança* (Nicolau, 1980); semelhança normalizada a s_{max} (Chandon e Pinson, 1981), quando se substitui a condição (1.3.13) por $s(i,i)=s_{max}$.

[47] *Índice de proximidade* (Nicolau, 1980).

Definição 1.3.12. Semelhança normada

Uma semelhança s, definida em I, é uma semelhança normada, quando

$\forall i \in I,\ s(i,i) = 1$ (1.3.17)

Existem numerosos coeficientes de semelhança propostos, dependendo a sua escolha do tipo de variáveis descritoras, sendo estas geralmente do mesmo tipo (e.g., Lerman, 1981; Siegel e Castellan, 1989; Saporta, 1990; Legendre e Legendre, 2000) dos quais se referem alguns:

- Para relacionar variáveis quantitativas, o coeficiente de correlação linear de Pearson (expressão 4.3.29, Subsecção 4.3.9.2) sobressai, assim como, uma sua alternativa – o coeficiente de correlação ordinal de Spearman (expressão 4.3.24, Subsecção 4.3.7.2).
- O coeficiente de correlação ordinal de Spearman e o coeficiente tau de Kendall (expressão 4.3.17, Subsecção 4.3.6.2) são opções frequentes, quando se pretende relacionar variáveis ordinais.
- Das medidas de associação entre variáveis qualitativas, em tabelas de contingência, (e.g., Bishop *et al*., 1975) são várias as que se baseiam na estatística qui-quadrado (e.g., o coeficiente de Cramer, o coeficiente de contingência), outras generalizam o coeficiente de Jaccard (Legendre e Legendre, 2000) e outras são probabilísticas (e.g., coeficiente de Goodall). Recorrendo também a tabelas de contingência, Abdallah e Saporta (1998) propõem um novo critério de associação entre variáveis qualitativas que se baseia em medidas/critérios de associação entre variáveis qualitativas conhecidos: critério introduzido por Rand em 1971, critério do desvio à indeterminação com referência a Marcotorchino (1984), critério do desvio quadrado à independência, critério do qui-quadrado, critério de Jordan e critério de Belson.

  Também, com aplicações importantes em medicina, em psicologia,…não se pode deixar de referir, os coeficientes de concordância entre dois observadores que classificam, separadamente, um conjunto de objectos numa escala nominal: o coeficiente de concordância Kapa de Cohen (e.g., Cohen, 1960; Kraemer, 2006; Wood, 2007) e o coeficiente de concordância ponderado Kapa de Cohen (e.g., Cohen, 1968; Kraemer, 2006). Existem coeficientes de concordância entre mais do que dois observadores e para classificações em escalas ordinais (e.g., Soares, 1999).

  No caso das variáveis serem binárias os coeficientes de associação são numerosos – serão estudados com mais pormenor na Subsecção 1.3.6.

- No contexto da análise classificatória as propostas de coeficientes de semelhança são numerosas, principalmente quando as variáveis são do mesmo tipo (e.g., Lerman, 1973, 1981; Bacelar-Nicolau, 1980; Saporta, 1990). O coeficiente de afinidade (e.g., Bacelar-Nicolau, 1980) será alvo de estudo mais pormenorizado no Capítulo 3.

Outros coeficientes, tal como o coeficiente de correlação bisserial por pontos (Lev, 1949), permitem comparar uma variável binária com uma variável contínua.
Sobre alguns dos coeficientes acima referidos, além de outros, Kraemer (2006) dá-nos uma perspectiva interessante da escolha adequada dos coeficientes de semelhança em Medicina.

Os coeficientes de semelhança entre objectos descritos por variáveis heterogéneas, e entre variáveis heterogéneas, não são muito numerosos. Eis alguns deles:

- Dos primeiros, é bem conhecido o coeficiente de semelhança geral de Gower (1971). Este coeficiente permite relacionar objectos i e j descritos por variáveis dicotómicas, nominais, ordinais e quantitativas:

$$s_{ij} = \frac{\sum_{k=1}^{p} \alpha_{ijk} s_{ijk}}{\sum_{k=1}^{p} \alpha_{ijk}} \qquad (1.3.18)$$

  Sendo s_{ijk} um coeficiente de semelhança entre os objectos i e j descritos pela variável k e α_{ijk}= 1, se i e j podem ser comparados em relação à variável k, e α_{ijk}= 0 se não.
  - No caso da variável k ser binária (podendo tomar os códigos +/detecção ou -/não detecção), Gower sugere que ela seja tratada como atributo de descrição, i.e., só a codificação + é portadora de informação:

Objecto i	Objecto j	s_{ijk}	α_{ijk}
+	+	1	1
+	-	0	1
-	+	0	1
-	-	1	0

  - No caso da variável k ser nominal, Gower sugere que s_{ijk}= α_{ijk}= 1 se i e j tomam o mesmo valor e s_{ijk}= α_{ijk}= 0 se não.
  - Quando k é uma variável quantitativa, Gower propõe como coeficiente de semelhança

$$s_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{A_k}$$ sendo A_k a amplitude total da variável k.

No caso das variáveis dicotómicas, o coeficiente de Gower coincide com o coeficiente de semelhança de Jaccard (A_4, Tabela 1.3.4).

Gower (1971) prova que a matriz $S=(s_{ij})$ é semidefinida positiva (s.d.p.), quando não existem dados omissos; sabe-se também que as distâncias $d_{ij}=(1-s_{ij})^{½}$ correspondentes são provavelmente euclidianas (Legendre e Legendre, 2000).

O coeficiente de Gower, que é descrito e referenciado por diversos autores (entre eles, Legendre e Legendre, 2000), está disponível, por exemplo, no *software* Clustan. Além de ser usado em algoritmos de Análise Classificatória, também é associado à Análise em Coordenadas Principais (ACoP) introduzida por Gower (1966) para obter uma representação euclidiana.

- Em ecologia, o coeficiente de semelhança entre objectos proposto por Legendre e Chodorowski (1977) é uma generalização do coeficiente de Jaccard a variáveis heterogéneas.
- No âmbito da investigação em Classificação[48] do grupo de I.C. Lerman, destaca-se um índice de semelhança entre objectos descritos por variáveis de vários tipos (Lerman, 1987; Lerman e Peter, 2003) cuja programação (programa SIMOB) se deve a Peter (Lerman e Peter, 1985).
- No contexto da Análise Classificatória, Bacelar-Nicolau (2000) propõe um coeficiente de afinidade ponderado entre unidades de dados descritas por variáveis mistas que será apresentado no Capítulo 3.
- Dos segundos, os coeficientes de semelhança entre variáveis heterogéneas propostos por Le Calvé (1977) com aplicações em Análise de Dados (classificação hierárquica e ACP sobre as matrizes de semelhança), que serão estudados e generalizados no Capítulo 4.
- Também em Classificação, Chah (1984) compara variáveis heterogéneas (quantitativas, qualitativas nominais e qualitativas ordinais), utilizando a noção de preordenação no âmbito dos métodos algorítmicos introduzidos por Marcotorchino e Michaud (1979).
- Ainda no âmbito da investigação em Classificação do grupo de I.C. Lerman, destaca-se o coeficiente de associação entre variáveis qualitativas, variáveis relacionais, e entre variáveis simbólicas (Ouali-Allah, 1991a; Lerman, 1992a; Lerman, 1992b), cuja programação (programa AVARE) se deve a Ouali-Allah (1991b). O índice de

[48] *Classification Automatique* (Lerman).

semelhança bruto é calculado a partir de matrizes que se baseiam na noção de preordenação[49], entre os pares de indivíduos, induzida pela relação binária associada a cada uma das variáveis.

- Térouanne (1998) apresenta uma medida de covariação[50] comum a quatro tipos de estruturas (nominal, ordinal, numérica e métrica). Para este autor, a estrutura métrica engloba estruturas não triviais, tais como as ultramétricas associadas a uma árvore de classificação, ou o número de diferenças entre duas palavras (genoma) (exemplos de variáveis encontradas com frequência em biologia e em ciências humanas).
- Mais recentemente, com aplicação prática nas árvores de decisão e na escolha das variáveis independentes em regressão logística, Lee e Huh (2003) propõem uma medida de associação entre variáveis de diferentes tipos (dados de tipo complexo: contínuas e discretas) que se baseia numa medida de afastamento da independência entre as variáveis, usando o valor-*p* de testes de independência (teste sobre o coeficiente de correlação de Spearman, no caso de variáveis contínuas-contínuas; teste do qui-quadrado de Pearson, no caso das discretas-discretas; teste de Kruskal-Wallis, no caso das variáveis discretas-contínuas).

Certamente que não esgotámos este assunto pois tem surgido nas mais diversas áreas da Estatística (e. g., Cuadras e Arenas, 1990) e da Análise de Dados a preocupação de tratar dados heterogéneos (como se viu e se pode ver também na Subsecção 1.4.2.3).

1.3.3 Representação matricial das dissemelhanças e das semelhanças

Em Análise de Dados é habitual representar as dissemelhanças e as semelhanças sob a forma matricial (e.g., Gower e Legendre, 1986; Joly e Le Calvé, 1994).

Seja I um conjunto finito com n elementos, $I=\{1, 2, \ldots, n\}$, e as matrizes nxn, de dissemelhanças sobre I, $D = (d_{ij})$, e de semelhanças, $S = (s_{ij})$.

- A matriz de dissemelhanças D é uma matriz quadrada com termos reais não negativos, simétrica e com a diagonal principal nula, ($d_{ii}=0$ e $d_{ij} = d_{ji} \geq 0$).

  A dissemelhança D é pois, uma matriz de dissemelhanças sobre I, com elementos $d_{ij} = d(i,j)$.

---

[49] *Préordonnance*, em francês. Termo introduzido por Lerman. Preordenação sobre um conjunto E com n elementos é uma relação de preordem total definida sobre o conjunto dos pares de elementos de E.

[50] *Covariation*, no original.

No caso da dissemelhança D ser própria, D é uma matriz cujos termos fora da diagonal principal são sempre positivos.

A matriz de dissemelhanças D é semidefinida/semi-própria se e só se $d_{ij} = 0 \Rightarrow (\forall k \in I, d_{ik} = d_{jk})$.

- A matriz de semelhanças S é uma matriz com termos reais não negativos, simétrica, com todos os termos da diagonal principal maiores ou iguais do que os restantes termos.

  A matriz de semelhanças normada tem todos os termos da diagonal principal iguais a 1.

Todas as definições apresentadas, quer para as dissemelhanças quer para as semelhanças, são apresentadas de forma análoga para as respectivas matrizes.

Em Análise de Dados de proximidade os dois tipos de coeficientes (semelhanças e dissemelhanças) têm sido utilizados para representar as relações entre os objectos/unidades estatísticas ou entre variáveis. No entanto, todas as análises factoriais estão ligadas à análise das semelhanças e as análises classificatórias estão principalmente ligadas à análise das dissemelhanças (Joly e Le Calvé, 1994). Mas, as semelhanças e as dissemelhanças são noções que se complementam, relacionando-se através de funções – é disso que trata a subsecção seguinte.

1.3.4 Funções de semelhança

> *It is interesting to quote that the representations are mostly read in terms of dissimilarities, while they are constructed in terms of similarities. It is then very important to study the relations between these two kinds of association coefficients.*
> (Le Calvé, 1988)

1.3.4.1 Definições e exemplos

As dissemelhanças podem converter-se em semelhanças e vice-versa, recorrendo a vários tipos de transformações ou funções de transformação como, por exemplo (Chandon e Pinson, 1981; Caillez e Kuntz, 1996):

- $s(i,j) = \frac{s_{\max}}{s_{\max} + d(i,j)}$, $\forall (i,j) \in I^2$, sendo s_{max}, número positivo qualquer, o valor máximo de s.

- $d(i,j) = s_{max} - s(i,j)$ ou $d(i,j) = 1 - s(i,j)$, .

Beninel (1987), Le Calvé (1988) e Joly e Le Calvé (1994) apresentam um estudo sobre funções que ligam os índices de semelhança aos índices de dissemelhança – as funções de semelhança – na perspectiva das suas propriedades em relação aos métodos de representação da estrutura de dados (I, D). Sendo dada uma matriz de dissemelhança D sobre o conjunto I, determina-se a semelhança associada, S, por transformação decrescente de D: S = f(D).

As funções de semelhança – linear, homográfica, quadrática, exponencial e circular – são exemplos de aplicação do Teorema 1.3.1 (Joly e Le Calvé, 1994) que se apresenta em seguida.

Teorema 1.3.1 (Joly e Le Calvé, 1994)

Dada a função real decrescente, f, com f(0) = 1, e a dissemelhança $D=(d_{ij})$, seja $S=(s_{ij})$ tal que: $\forall(i,j) \in I^2$, $s_{ij} = f(d_{ij})$.
Então S é uma semelhança normada.
Se f é estritamente decrescente e D é definida, então S é própria.
Reciprocamente, seja g uma função real decrescente, com g(1) = 0, e S uma semelhança normada. Então D = g(S) é uma dissemelhança.
Se S é própria e g é estritamente decrescente, então D é definida.

Resultados obtidos sobre as funções de semelhança - linear, homográfica, quadrática, exponencial e circular - são apresentados na Tabela 1.3.2, considerando-se sempre que as semelhanças são normadas e as funções invertíveis ($g= f^{-1}$).

Esposito, Malerba, Tamma e Bock (2000), por sua vez, referem as transformações:
- $\forall(i,j) \in I^2$, $s_{ij} = d_{max} - d_{ij}$, em que d_{max} é o valor máximo de d.
- $\forall(i,j) \in I^2$, $s_{ij} = \sqrt{d_{\max} - d_{ij}}$.
- $\forall(i,j) \in I^2$, $s_{ij} = d_{\max}^2 - d_{ij}^2$.

Tabela 1.3.2. Funções de semelhança (FS) e representação da estrutura de dados (I,D)

Funções de semelhança (FS)	Expressão geral que liga S e D S=f(D), D=g(S)=f^{-1}(S)	Observações	Representação da estrutura de dados (I,D)
Função linear	*S + D = 1* *S = 1-D* *D = 1-S*	Esta é a FS mais usada para variáveis "presença-ausência".	Este tipo de FS ajusta-se bem a representações hierárquicas (classificações, pirâmides). Quando D é uma ultramétrica, a árvore hierárquica associada a (I,D) permite uma leitura gráfica simultânea de S e D. A dissemelhança entre dois indivíduos i e j é representada pelo índice do nó que os liga. Exemplo em Beninel (1987).
Função quadrática	$S = 1 - \frac{1}{2} D^2$ $D = \sqrt{2(1-S)}$	Ideia de base desta transformação - considerar S como uma matriz de produtos escalares.	Fórmula bem ajustada pela representação euclidiana. Quando S é s.d.p., a sua diagonalização conduz a uma representação esférica da estrutura (I,D), com centro na origem O. Neste caso, s_{ij} é igual ao produto escalar $\overrightarrow{Oi}.\overrightarrow{Oj}$.
Função homográfica	$S = \frac{1-D}{1+D}$ $D = \frac{1-S}{1+S}$	FS com interesse, principalmente, analítico. Vários índices para variáveis binárias são obtidos homograficamente a partir de outros (também referenciadas por Gower). Ex.º: A distância de Jaccard está ligada homograficamente à semelhança de Sokal-Sneath-Anderberg e a dist. de Czenakowski-Dice à semelhança de Jaccard.	
Função exponencial	$S = e^{-D}$ $D = -\ln S$ *ou, mais geralmente,* $S = e^{-D^p}$	Este tipo de FS é utilizado frequentemente pelos geneticistas, pois esta transformação é eficaz quando trata com distâncias bastante pequenas (NEI, 1978). Esta FS está bem adaptada a variáveis transformadas multiplicativamente, como por exemplo a razão de crescimento económico.	A exponencial e^{-D^p} é bem ajustada á representação em espaços L^p . Dado p=2 e a distância euclidiana D, $D^2 = \sum_i (x_i - y_i)^2$ pode-se observar a ligação forte que existe entre a geometria euclidiana e a distribuição normal.
Função circular	$S = \cos D$ $D = Arc\cos S$	Quando S é s.d.p. esta transformação permite visualizar a dissemelhança D por um ângulo.	As distâncias angulares estão bem adaptadas às noções de "distância aparente" (em astronomia) e às representações esféricas. Para um observador situado no centro da esfera sobre a qual estão representados os pontos representativos dos indivíduos, este ângulo é a distância aparente.

Como se viu as comparações entre os dados podem ser feitas através da análise das semelhanças ou das dissemelhanças. Embora aquelas análises não sejam contraditórias, as técnicas de Análise Exploratória de Dados estão mais relacionadas com umas ou com outras. Por exemplo, a ACP está relacionada com a análise das semelhanças (através das covariâncias). Contudo, podemos associar à covariância a distância euclidiana: - Esta associação é trabalhada basicamente por meio de uma função decrescente que, neste caso, é uma função quadrática (Tabela 1.3.2) (Joly e Le Calvé, 1994). Quando as variáveis estão padronizadas o coeficiente de correlação linear, r_{XY}, e a distância euclidiana, d_{XY}, estão pois relacionados: $d_{XY} = \left(2 - 2r_{XY}\right)^{1/2}$. Neste caso, quando se recorre ao MDS ordinal é irrelevante usar uma ou outra proximidade (Borg e Groenen, 2005).

Neste contexto relembra-se ainda o teorema demonstrado por Gower e Legendre (1986): Se a matriz de semelhanças S, em que $0 \leq s_{ij} \leq 1$ e $s_{ii}=1$, é semidefinida positiva então a matriz de dissemelhança com elementos $d_{ij} = \sqrt{1 - s_{ij}}$ é euclidiana (Teorema 1.3.3).

No caso de se utilizar uma função linear para a geometria euclidiana, os novos índices poderão não estar bem adaptados à representação correspondente (Joly e Le Calvé, 1994).

Foram apresentadas algumas das transformações utilizadas por diversos autores. Uma vez que existem várias transformações, põe-se a questão de saber como as escolher. Na Tabela 1.3.2 encontram-se algumas respostas, tendo em conta a representação da estrutura de dados (I, D). Pode-se acrescentar que, a escolha de uma transformação depende, em geral, da natureza dos dados e da óptica sob a qual se pretende ver os dados.

1.3.4.2 Formas-W associadas a uma dissemelhança

As formas-W associadas a dissemelhanças são uma família muito importante das funções de semelhança (Joly e Le Calvé, 1994). Pela importância que têm, no contexto das estruturas euclidianas, passamos a apresentá-las brevemente.

Definição 1.3.13. Forma-W de D relativa ao ponto M. Forma-W de D^p relativa ao ponto M

Considere-se uma estrutura de dados (I, D), sendo D a matriz das dissemelhanças d entre os n elementos de I. Seja M um ponto de I. A forma-W de D relativa ao ponto M, $W^M(D)$, é uma matriz nxn cujos elementos são:

$$\forall (i,j) \in I \times I, \quad w^M(D)_{ij} = \frac{1}{2}\left(d_{Mi} + d_{Mj} - d_{ij}\right)$$

Para $p \in \mathbb{R}^+$, define-se da mesma maneira, a forma-W de D^p relativa a um ponto M:

$$\forall (\mathrm{i,j}) \in \mathrm{I} \times \mathrm{I}, \quad \mathrm{w}^{\mathrm{M}}\left(D^{p}\right)_{ij} = \frac{1}{2}\left(d^{p}_{Mi} + d^{p}_{Mj} - d^{p}_{ij}\right)$$

A forma-W de D, $W^M(D)$, é uma função de semelhança linear generalizada.

Observações (Joly e Le Calvé, 1994):

- O ponto M desempenha o papel de origem da forma.
- $W^M(D)$ é uma matriz simétrica que apresenta uma linha e uma coluna nulas, pois $\forall i \in \mathrm{I}, \ \mathrm{w}^{M}_{Mi}(D) = 0$, e cujos elementos são positivos ou nulos se d é uma semi-distância.
- O valor $\mathrm{w}^{\mathrm{M}}(D)_{ij}$ pode ser visto como uma medida do desvio ao alinhamento dos três pontos M, i e j.
- D é uma distância se e só se $\mathrm{w}^{\mathrm{M}}(D)_{ij} \geq 0$ (Le Calvé, 1988).
- Habitualmente $W^M(D^p)$ não é uma semelhança.
- Sempre que D^p é uma distância, então, para qualquer M, $W^M(D^p)$ é um índice de semelhança.

Em particular, a forma-W, para p=2, desempenha um papel importante no domínio euclidiano. $W^M(D^2)$, é uma matriz nxn cujos elementos são:

$$\forall (\mathrm{i,j}) \in \mathrm{I} \times \mathrm{I}, \quad \mathrm{w}^{\mathrm{M}}\left(D^{2}\right)_{ij} = \frac{1}{2}\left(d^{2}_{Mi} + d^{2}_{Mj} - d^{2}_{ij}\right) \quad (1.3.19)$$

Torgerson (1958) propõe uma forma quadrática em que o "centro de gravidade", G, da estrutura de dados desempenha um papel preponderante. O que corresponde a completar a estrutura de dados (I, D) com um indivíduo fictício G definido por:

$$\forall i \in \mathrm{I}, \quad d^{2}_{Gi} = d^{2}_{i.} - \frac{1}{2} d^{2}_{..}$$

$$\text{sendo, } d^{2}_{i.} = \frac{1}{n}\sum_{j=1}^{n} d^{2}_{ij} \quad \text{e} \quad d^{2}_{..} = \frac{1}{n^{2}}\sum_{i,j=1}^{n} d^{2}_{ij}$$

A forma $W^G\left(D^2\right)$ associada à estrutura de dados ($\mathrm{I} \cup \{\mathrm{G}\}$, D) é pois definida por:

$$\forall (\mathrm{i,j}) \in \mathrm{I} \times \mathrm{I}, \quad \mathrm{w}^{G}\left(D^{2}\right)_{ij} = \frac{1}{2}\left(d^{2}_{i.} + d^{2}_{.j} - d^{2}_{..} - d^{2}_{ij}\right) \quad (1.3.20)$$

1.3.5 Estrutura de dados euclidiana

A distância euclidiana é muito utilizada em Análise de Dados, devido á facilidade e simplicidade da representação gráfica de um conjunto de indivíduos I munido de uma distância euclidiana D. É uma distância que está muito estudada (e.g., Gower, 1971; Fichet e Le Calvé, 1984; Joly e Le Calvé, 1986; Gower e Legendre, 1986; Beninel, 1987; Caillez e Kuntz, 1996). Aqui apresentam-se algumas das suas propriedades métricas.

Recorde-se que uma distância $D=(d_{ij})$ sobre um conjunto I de cardinal n é euclidiana, se e só se ela verifica a condição: $\exists p \in \mathbb{N}$ e X uma matriz real, nxp, tais que $\forall (i,j) \in I \times I,\ d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2}$.

Também se pode dizer que: "Uma dissemelhança $D=(d_{ij})$ sobre I é euclidiana se e só se existem n pontos, $M_1,\ldots, M_n$ de um espaço euclidiano que verificam: $\left\| \overrightarrow{M_i M_j} \right\| = d_{ij}$ (1.3.21)" (e.g., Joly e Le Calvé, 1986).

Daí a definição que se segue:

Definição 1.3.14. Imagem euclidiana

Seja E um espaço afim euclidiano de dimensão finita. Chama-se imagem euclidiana em E associada à estrutura de dados (I, D) um sistema de pontos $\{M_i,\ i \in I\}$ de E que verificam a igualdade (1.3.21): $\left\| \overrightarrow{M_i M_j} \right\| = d_{ij}$.

A representação euclidiana de uma estrutura de dados (I, D) consiste, pois, em representar todo o indivíduo $i \in I$ por um ponto M_i do espaço euclidiano de tal forma que a figura assim obtida dê conta da dissemelhança D, i.e., $\left\| \overrightarrow{M_i M_j} \right\| = d_{ij}$.

Certas estruturas de dados admitem uma imagem euclidiana, outras não (e.g., exemplificado em Gower e Legendre, 1986; Beninel, 1987):

- Sobre um conjunto I limitado a 3 elementos toda a semi-distância é euclidiana.

- Uma dissemelhança euclidiana é necessariamente uma semi-distância. A propriedade da desigualdade triangular é uma condição necessária para que (I, D) seja euclidiana, mas não é suficiente.

- A "dimensão" de D é a dimensão do espaço gerado por uma qualquer das suas imagens euclidianas, sendo a dimensão máxima (n-1). Se a dimensão é igual a 2, o espaço de representação é o plano com as suas regras habituais de leitura, em que a escolha da origem e dos eixos é arbitrária.
  Mostra-se que o conjunto das distâncias euclidianas (dissemelhanças mergulháveis num espaço vectorial de dimensão finita munido da norma de Minkowski de ordem 2), $\mathfrak{D}_e$, é um cone fechado, mas não é convexo (com um contra-exemplo, vê-se que a soma de duas distâncias euclidianas de dimensão 1 não pode ter representação euclidiana) (e.g., Perrier, 1998).
- Uma caracterização de distância euclidiana pode ser dada a partir da forma-W associada a D^2 (em que $d_{ij}^2 = d^2(i,j)$), sob a forma do seguinte teorema atribuído a Fréchet[51] (Joly e Le Calvé, 1986):

Teorema 1.3.2. (Fréchet, 1935)

A matriz de dissemelhanças D pode ser considerada como uma matriz de distâncias entre n pontos de um espaço euclidiano se e só se existe um ponto M tal que a matriz $W^M(D^2)$ é semidefinida positiva (s.d.p.).

Observações:

- Sendo D euclidiana, a forma $W^M(D^2)$ pode ser vista como a matriz dos produtos escalares $\overrightarrow{Mi} \cdot \overrightarrow{Mj}$ dos vectores $\overrightarrow{Mi}$ e $\overrightarrow{Mj}$ (e.g., Joly e Le Calvé, 1994).
- A matriz $W^M(D^2)$ sendo uma matriz de produtos escalares é por isso s.d.p. (e.g., Beninel, 1987).
- Demonstra-se que se existir um ponto $M \in I$ tal que $W^M(D^2)$ seja s.d.p. então $W^M(D^2)$ é s.d.p. para todo o $M \in I$ - O que motiva o uso frequente deste

[51] Joly e Le Calvé (1986) lembram que este teorema também pode ser atribuído a Gauss, Minkowski, Schoënberg, dependendo da forma como é enunciado por cada um deles. Por exemplo, Gauss enunciou-o em 1831 para três dimensões, e Fréchet para os espaços de Hilbert, em 1935.

teorema para mostrar que uma dissemelhança é euclidiana, pois basta mostrar que, para pelo menos um ponto M, a forma-W associada é s.d.p..

- Perrier (1998) refere o programa WinABCD que propõe uma verificação da natureza euclidiana de uma distância por procura dos valores próprios da matriz $W^M(D^2)$ e verificação da positividade do menor dos seus valores próprios. Refere também que a detecção de uma matriz francamente não euclidiana não é um problema. O problema maior é o de mostrar que uma distância, euclidiana por construção, é realmente euclidiana devido aos valores dos valores próprios poderem ser muito pequenos[52] neste caso.

- Quando se escolhe para o ponto M o centro de gravidade G, M=G, obtém-se a forma de Torgerson $W^G(D^2)$. Pode-se demonstrar que a estrutura de dados é euclidiana se e só se $W^G(D^2)$ é s.d.p. (e.g., Beninel, 1987), recorrendo ao Teorema de Fréchet.
    - "Se D não é euclidiana, vale a pena fazer notar que a escolha do centro de gravidade G seria completamente errada" (Joly e Le Calvé, 1994). Só no caso de D ser euclidiana é que a forma de Torgerson permite calcular a distância entre i e o centro de gravidade.
    - "Se D é euclidiana, quando G é a origem, o plano factorial corresponde ao máximo da inércia – nisto reside o principal interesse em se escolher G. Quando a origem é um ponto arbitrário M, o plano factorial corresponde ao máximo do momento centrado de ordem 2 em relação a qualquer M" (Joly e Le Calvé, 1994).

- Uma condição suficiente para que $D=(1-S)^{½}$ seja euclidiana é dada pelo seguinte teorema, muito usado nesta área.

Teorema 1.3.3. (Gower e Legendre, 1986)

Se S é uma matriz de semelhanças s.d.p. com elementos $0 \leq s_{ij} \leq 1$ e $s_{ii}=1$, então a matriz de dissemelhanças com elementos $d_{ij} = (1 - s_{ij})^{½}$ é euclidiana.

- É muito usada a seguinte condição necessária e suficiente de eucliniaridade, dada pela decomposição da dissemelhança d definida num conjunto finito I via uma forma

[52] Mesmo atingindo a precisão da máquina 10^{-15} com os reais em precisão dupla, não está assegurada a obtenção de uma estimação positiva do menor valor próprio.

bilinear simétrica q, relembrada por vários autores (e.g., Caillez e Kuntz, 1996) com referência a (e.g., Blumenthal, 1953):

"Sendo $d^2(i,j) = q_{ii} + q_{jj} - 2 \cdot q_{ij}$, uma condição necessária e suficiente para que d seja euclidiana, é a de q ser s.d.p."

Uma consequência deste resultado, que terá interesse ter em conta quando se escolhe o coeficiente de semelhança em Análise de Dados, refere-se à possibilidade de poder interpretar q_{ij} como um coeficiente de semelhança s_{ij} com o valor máximo igual a 1, $s_{max}=1$; esta situação reporta-nos ao teorema apresentado anteriormente (Teorema 1.3.3).

- Outro resultado a reter é apresentado por Beninel (1987):
  "No caso da matriz de semelhanças S ser s.d.p., a diagonalização de S fornece uma imagem euclidiana esférica representando $(I, \sqrt{2D})$". Verificando-se $\forall(i,j)\in IxI$, $d_{ij}=1-s_{ij}$ ou ainda $\forall(i,j)\in IxI$, $d_{ij} = \frac{1}{2}(s_{ii} + s_{jj} - 2s_{ij})$, S desempenha o papel da forma quadrática associada à estrutura $(I \cup \{\omega\}, \sqrt{2D})$ tomada no ponto ω.

Existem mais resultados importantes (por exemplo, Lema de Schur) sobre formas semidefinidas positivas e sobre a correspondência entre as semi-distâncias euclidianas e as formas semidefinidas positivas (e.g., Joly e Le Calvé, 1994; Al Ayoubi, 1991) – aqui limitámo-nos a apresentar alguns deles por nos parecerem úteis no prosseguimento deste trabalho.

Outro aspecto interessante a ter em conta, refere-se à ligação das distâncias euclidianas a outras dissemelhanças (já referenciada para o caso das ultramétricas na Subsecção 1.3.1). Este assunto é apresentado e bem exemplificado, por exemplo, em Beninel (1987) e Perrier (1998) que se basearam em trabalhos de Le Calvé.

Uma última observação leva-nos a relembrar que apenas algumas das dissemelhanças que são calculadas são euclidianas:

- Gower e Legendre (1986) apresentam um estudo sobre as propriedades métrica e euclidiana de dez coeficientes de dissemelhança entre variáveis quantitativas e verificam que apenas a distância D_5 (Tabela 1.3.1) é euclidiana para valores positivos da variável. Verificam também que algumas não são métricas. Relembra-se que estes autores consideram que uma dissemelhança é métrica quando verifica a desigualdade triangular, ou seja, métrica é uma designação equivalente a semi-distância (Definição 1.3.6). Legendre e Legendre (2000) actualizam aquele estudo.

- Sabe-se que a distância χ^2 (fórmula 1.3.12, Subsecção 1.3.1) é euclidiana (Legendre e Legendre, 2000).
- No caso de alguns dos coeficientes de dissemelhança entre objectos descritos por variáveis binárias, estudados por vários autores (e.g., Fichet e Le Calvé, 1984; Gower e Legendre, 1986), nenhum deles é euclidiano; mas, para todos eles, a sua raiz quadrada é euclidiana (Subsecção 1.3.6 e Tabela 2.2.1, Capítulo 2).

1.3.6 Coeficientes de semelhança (e de dissemelhança) para dados binários

Recorde-se que as variáveis binárias são variáveis qualitativas com duas modalidades ou classes que, habitualmente, são representadas por 0 e 1. Estas variáveis podem ser interpretadas em termos de 1-presença/0-ausência do carácter (atributo) ou 1-sim/0-não ou 1-detecção/0-não detecção. São numerosos os exemplos destas variáveis em ecologia, taxonomia, biologia, zoologia, medicina (e.g., psiquiatria), psicologia (e.g., etologia), etc. No caso de se considerar que só a presença do atributo é informativa, esta variável designa-se atributo de descrição[53].

Exemplo 1.3.1. Variáveis binárias

- Os descritores da diversidade genética são frequentemente variáveis binárias, representando-se por 1 a presença do descritor e por 0 a sua ausência.
- No caso dos descritores bioquímicos ou moleculares, os caracteres codificam geralmente a presença ou ausência de uma banda sobre um gel de electroforese (Perrier, 1998).
- O comportamento dos indivíduos é frequentemente medido por variáveis binárias.
- Os sintomas de certas doenças podem ser codificados como variáveis binárias: 1-o sintoma está presente, 0-o sintoma está ausente.
- Uma variável binária pode indicar o resultado positivo (1) ou negativo (0) de um teste de diagnóstico médico.
- Presença ou ausência de um factor de risco, presença ou ausência de doença,...
- Uma variável binária pode indicar se foi administrado tratamento ou controlo ou se é caso ou controlo.

[53] Outros autores, tal como Gower designam-na por variável dicotómica.

Sendo as variáveis X e Y duas variáveis binárias, podem-se dispor os dados de forma habitual na tabela de contingência 2x2 (Tabela 1.3.3). Nesta tabela designam-se por:

- ***a*** o número de sujeitos que apresentam simultaneamente os caracteres ou atributos de X e de Y, i.e., as concordâncias positivas ou co-presenças.
- ***b*** o número de sujeitos que apresentam o carácter ou atributo de X e não apresentam o carácter ou atributo de Y, i.e., as discordâncias.
- ***c*** o número de sujeitos que não apresentam o carácter ou atributo de X e apresentam o carácter ou atributo de Y, i.e., as discordâncias.
- ***d*** o número de sujeitos que não apresentam nenhum dos caracteres ou atributos de X e de Y, i.e., as concordâncias negativas ou co-ausências[54].
- ***n*** o número total de sujeitos.

Tabela 1.3.3. Tabela de contingência das variáveis binárias X e Y, em que: *a* representa o número de co-presenças, *b* representa o número de presenças-ausências, *c* o número de ausências-presenças e *d* o número de co-ausências

X \ Y	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	n

Convém notar que a interpretação das frequências ***a***, ***b***, ***c*** e ***d*** depende da disposição dos uns (presença/sim) e zeros (ausência/não) nas margens, linhas/colunas, da tabela de contingência (Tabela 1.3.3). Segundo vários autores (e.g., Hubálek, 1982; Everitt e Rabe-Hesketh, 1997; Legendre e Legendre, 2000), a tabela acima é apresentada cruzando a informação dos indivíduos i e i', em vez da informação sobre as variáveis X e Y.

São numerosos os coeficientes de associação[55] entre variáveis binárias propostos por vários autores, inicialmente em ecologia, taxonomia e noutras disciplinas da biologia. O estudo destes coeficientes tem atraído diversos autores (e.g., Lerman, 1970; Roux e Roux, 1980[56]; Blanc *et al*., 1976; Bacelar-Nicolau, 1980, 1987; Nicolau e Bacelar-Nicolau, 1981; Hubálek, 1982; Fichet e Le Calvé, 1984; Gower e Legendre, 1986; Joly e Le Calvé, 1994; Perrier, 1998; Legendre e Legendre, 2000), que sugerem como os escolher na perspectiva da sua utilização em Análise de Dados, embora concluam que não existe um que seja o melhor. Hubálek (1982) apresenta uma recensão – a mais numerosa – na qual constam 43

[54] *Double-zeros*, Legendre e Legendre (2000).
[55] No caso de se estudar a relação entre unidades estatísticas, aqueles coeficientes designam-se genericamente por coeficientes de semelhança.
[56] A 1ª edição reporta-se a 1973.

coeficientes para dados binários (Tabela 1.3.4) e faz um estudo comparativo. Le Calvé (1993) dá a conhecer este estudo e comenta-o. Legendre e Legendre (2000) também se referem àquele estudo e acrescentam àquela recensão o coeficiente de Faith (1983) (Tabela 1.3.4). Juntamos ainda àqueles coeficientes, os coeficientes A_{45} e A_{46}, generalizações do coeficiente de Ochiai (A_{11}) e do coeficiente de Kulczynski (A_7), respectivamente, embora não tenham suscitado até à época o interesse dos ecologistas (Blanc *et al*., 1976). Outros coeficientes omitem-se naquela tabela por corresponderem a coeficientes já apresentados, sendo frequente diferentes autores atribuírem designações diferentes a um mesmo coeficiente (exemplificado em, e.g. Blanc *et al*., 1976).

Àquela lista de coeficientes também se pode acrescentar alguns coeficientes propostos para dar resposta a problemas específicos da área tecnológica[57], em que os indivíduos são máquinas. Todos eles são generalizações do coeficiente de semelhança de Jaccard (A_4) apresentadas por Sarker e Saiful Islam (1999): as medidas de Seifoddini e Wolfe, o coeficiente de semelhança ponderado (Mosier e Taube; Wei e Kern[58]), a medida de Kusiak e Cho (1992) e outros três coeficientes de semelhança ponderados.

Estes coeficientes foram propostos para darem resposta a situações concretas, mas naturalmente que se põe a questão de como os escolher.

Um critério de escolha tem a ver com a importância que se pretende dar às co-ausências dos atributos e leva-nos a considerar dois grupos de coeficientes:

- Os que não incluem as co-ausências, i.e., aqueles em que ***d*** não faz parte da expressão do coeficiente. Estes são os coeficientes de associação entre atributos de descrição ou coeficientes assimétricos[59] em ***a*** e ***d***: coeficientes A_1 a A_{13} e A_{15} (Tabela 1.3.4).
- Os que incluem as co-ausências, i.e., os coeficientes que consideram ***d*** na sua expressão, também designados por coeficientes simétricos[60]: coeficientes A_{14} e A_{16} a A_{46} (Tabela 1.3.4). Os coeficientes A_{19}, A_{20}, A_{22} e A_{23} são generalizações dos coeficientes de Kulczynski (A_3), de Jaccard (A_4), de Dice (A_5) e de Sokal e Sneath (A_6), respectivamente (substitui-se nas suas expressões ***a*** por (***a***+***d***)).

---

[57] Mais especificamente, em *Group Technology* (GT), técnica usada para projectar sistemas de manufactura celular (*cellular manufacturing systems*).
[58] Os autores designam o coeficiente de Jaccard ponderado por *commonality score*.
[59] Designação utilizada por Legendre e Legendre (2000).
[60] Designação utilizada por Legendre e Legendre (2000).

Tabela 1.3.4. Coeficientes de semelhança para dados binários propostos por diversos autores, em que *a, b, c* e *d* referem-se às quatro células da tabela de contingência 2x2, designando respectivamente, o número de co-presenças, presenças/ausências, ausências/presenças e co-ausências dos atributos. São apresentados os valores mínimo e máximo de cada um dos coeficientes

Coeficiente de semelhança	Autor[61]
A_1<0, 1>= *a*/[max{(a+b), (a+c)}]	*Braun-Blanquet (1932)*
A_2<0, 1>= *a*/[min{(a+b), (a+c)}]	*Simpson (1943): "degree of faunal resemblance"*
A_3 <0, ∞>=*a*/(*b*+*c*). Não definido se b=c=0	*Kulczynski (1927)*
A_4 <0, 1>= *a*/(*a*+*b*+*c*)	*Jaccard (1901), Sneath (1957)*[62]
A_5 <0, 1>= *a*/[*a*+ ½(*b*+*c*)]	*Dice (1945), Sørensen (1948)*
A_6 <0, 1>= *a*/[*a*+2(*b*+*c*)]	*Sokal & Sneath (1963): "un_2"*
A_7 <0, 1>= ½[*a*/(*a*+*b*)+*a*/(*a*+*c*)]	*Kulczynski (1927)*
A_8<0, 1>= (a/2)[1/(a+b)+1/(a+c)]	*Driver & Kroeber (1932)*
A_9<0, 2>= a/(a+b)+a/(a+c)	*Johnson (1967)*
A_{10}<-1, 1>= (a^2-bc)/[(a+b).(a+c)]	*McConnaughey (1964)*
A_{11}<0, 1>= *a*/[(*a*+*b*).(*a*+*c*)]^{½}	*Driver & Kroeber (1932); Ochiai (1957)*
A_{12}<0, 1>= a^2/[(a+b).(a+c)]	*Sorgenfrei (1959): "correlation ratio"*
A_{13}<-∞, 0.5>= *a*/[(*a*+*b*).(*a*+*c*)]^{½}	*Fager & McGowan (1963)*
A_{14}<0, 1>= *a*/(*a*+*b*+*c*+*d*) = *a*/*n*	*Russell & Rao (1940)*
A_{15}<0, ∞>= *a*/[½(*ab*+*ac*)+*bc*]	*Mountford (1962)*
A_{16}<0, 1>= (ab+bc)/(ab+2bc+cd)	*Peirce (1884)*
A_{17}<-1, 0>= [a-(a+b).(a+c)]/[(a+b).(c+d).(a+c).(b+d)] Não definido se c=d=0 ou b=d=0	*Eyraud (1936)*
A_{18}<0, 1>= ¼[*a*/(*a*+*b*)+*a*/(*a*+*c*)+*d*/(*c*+*d*)+*d*/(*b*+*d*)]	*Sokal & Sneath (1963): "un_4"*
A_{19}<0, ∞>= (a+d)/(b+c). Não definido se b=c=0	*Sokal & Sneath (1963): "un_3"*
A_{20}<0, 1>= (a+d)/(a+b+c+d)=(a+d)/n	*Sokal & Michener (1958): "simple matching coef.."*[63]
A_{21}<0, 0.573>= $(50\pi)^{-1}$ arcsin [$(A_{20})^{½}$]	*Goodall (1967), Austin & Colwell (1977)*[64]
A_{22}<0, 1>= (a+d)/[a+½(b+c)+d]	*Sokal & Sneath (1963)*
A_{23}<0, 1>= (a+d)/[a+2(b+c)+d]	*Rogers & Tanimoto (1960): "un_1"*
A_{24}<-1, 1>= (a+d-b-c)/n	*Hamann (1961)*
A_{25}< 0, 1>= *ad*/[(*a*+*b*).(*c*+*d*).(*a*+*c*).(*b*+*d*)]^{½}	*Sokal & Sneath (1963): "un_5"*
A_{26}<-1, 1>= (ad-bc)/[(a+c).(b+d)]. Não definido se b=d=0	*Peirce (1884)*
A_{27}<0, ∞>= n.(ad-bc)^2/[(a+b).(c+d).(a+c).(b+d)]= χ^2	*Pearson (1905)*[65]
A_{28}<0, 1>= [χ^2/(n+χ^2)]^{½}. Não definido se a=d=0 ou b=c=0	*Pearson (1905): "coef. of mean square contingency"*
A_{29}<-1, 1>= √(2) (ad-bc)/[(ad-bc)^2-(a+b).(c+d).(a+c).(b+d)] ½ Não definido se a=d=0 ou b=c=0	*Cole (1949): "mean square contingency"*

[61] Os coeficientes A_1 a A_{43} estão referenciados em Hubálek (1982). O coeficiente A_{44} está referenciado em Legendre e Legendre (2000). Vejam-se estes autores para as referências apresentadas nesta tabela.

[62] Jaccard (1901): *coefficient de communauté*; Sneath (1957): *similarity index*, no original.

[63] *Simple matching coefficient*, no original. Coeficiente de emparelhamento simples ou coeficiente de concordância simples, tradução livre em português.

[64] *Angular transformation of simple matching coefficient*, no original.

[65] Qui-quadrado para tabelas de contingência 2x2.

Coeficiente de semelhança (cont.)	**Autor**[66] (cont.)
A_{30}<-1, 1>= $(ad-bc)/[(a+b).(c+d).(a+c).(b+d)]^{½}$. Não definido se c=d=0 ou b=d=0	*Yule (1912): "the product-sum correlation"; Pearson & Heron (1913): "fourfold point (tetrachoric) correl. coefficient"*
A_{31}<0, 1>= $(ad-bc)^2/[(a+b).(c+d).(a+c).(b+d)]= A_{30}^2$. Não definido se c=d=0 ou b=d=0	*Doolittle (1885), Pearson (1926)*
A_{32}<0, 1>= $(\sqrt{(ad)}+a)/(\sqrt{(ad)}+a+b+c)$	*Baroni-Urban & Buser (1976): "S**"*
A_{33}<-1, 1>= $(\sqrt{(ad)}+a-b-c)/(\sqrt{(ad)}+a+b+c)$	*Baroni-Urban & Buser (1976): "S*"*
A_{34}<-1, 1>= $\pm(\chi^2/\chi^2_{max})^{½}$ com o sinal de (ad-bc)	*Cole (1949): 'C_7'; Hurlbert (1969)*
A_{35}<-1, 1>= $\pm[(\chi^2-\chi^2_{min})/(\chi^2_{max}-\chi^2_{min})]^{½}$ com o sinal de (ad-bc)	*Hurlbert (1969)*
A_{36}<-1, 1>= $(ad-bc)/(ad+bc)$. Não definido se b=d=0 ou c=d=0	*Yule (1900): coeficiente de associação Q*
A_{37}<-1, 1>= $(\sqrt{(ad)}-\sqrt{(bc)})/(\sqrt{(ad)}+\sqrt{(bc)})$. Não definido se b=d=0 ou c=d=0	*Yule (1912): "ω, coefficient of colligation"*
A_{38}<-1, 1>= $\cos [180\sqrt{(bc)}/(\sqrt{(ad)}+\sqrt{(bc)})]$. Não definido se b=d=0 ou c=d=0	*Pearson & Heron (1913)*
A_{39}<-1, 1>= $4(ad-bc)/[(a+d)^2+(b+c)^2]$	*Michael (1920)*
A_{40}<0, ∞>= n.a/ $[(a+b).(a+c)]$	*Forbes (1907)*
A_{41}<-∞, ∞>= log a – log n - log[(a+b)/n] - log[(a+c)/n]. Não definido se a=0	*Gilbert & Wells (1966)*
A_{42}<-1,1>=(a-a')/(a_{max}+a')= $[n.a(a+b).(a+c)]/[n.\min\{(a+b),(a+c)\} -(a+b).(a+c)]$, sendo a_{max} o maior valor possível de ***a*** consistente com os totais marginais dados.	*Forbes (1925)*
A_{43}<-1, 1>= $(a-a')/(a+a')$ $=[n.a-(a+b).(a+c)]/[n.a+(a+b).(a+c)]$	*Tarwid (1960)*
A_{44}<0, 1>= $(a+d/2)/n$	*Faith (1983)*
A_{45}<0, 1>= $(a+d)/[(a+d+b).(a+d+c)]^{½}$	*Coeficiente de Ochiai (A_{11}) generalizado (e.g., Blanc et al., 1976)*
A_{46} <0, 1>= $½[(a+d)/(a+d+b)+(a+d)/(a+d+c)]$	*Coeficiente de Kulczynski (1927) (A_7) generalizado (e.g., Blanc et al., 1976)*

Nesta perspectiva, a escolha do coeficiente de associação dependerá de saber quão útil ou importante será a informação das co-ausências dos atributos para descrever a semelhança entre variáveis ou entre unidades estatísticas. Em ecologia esta questão é fundamental e são numerosos os exemplos (e.g., Legendre e Legendre, 2000). Em medicina, quando se realizam estudos sobre factores de risco terá interesse considerar as variáveis como atributos de descrição e ignorar as co-ausências; é o caso, por exemplo, de um estudo sobre factores de risco do enfarte do miocárdio em que dois sujeitos fumadores e diabéticos serão naturalmente mais parecidos do que dois não fumadores e não diabéticos. A escolha dos coeficientes incidirá, pois, sobre aqueles em que ***d*** não faz parte da sua expressão, tal como o coeficiente de Jaccard (A_4).

[66] Os coeficientes A_1 a A_{43} estão referenciados em Hubálek (1982). O coeficiente A_{44} está referenciado em Legendre e Legendre (2000). Vejam-se estes autores para as referências apresentadas nesta tabela.

Neste sentido, para Everitt e Rabe-Hesketh (1997) esta escolha resume-se, geralmente, à opção entre dois coeficientes, o coeficiente "simple matching" (A_{20}) e o coeficiente de Jaccard (A_4), conforme o investigador queira tratar as co-ausências.

Um olhar mais atento permite distinguir os coeficientes pela importância que se pretende dar aos acordos (co-presenças ou co-ausências) e aos desacordos entre as variáveis, como por exemplo, o coeficiente de Ochiai que representa uma média geométrica das razões entre ***a*** e o número de presenças de cada um dos atributos ($A_{11}=\{[\boldsymbol{a}/(\boldsymbol{a}+\boldsymbol{b})].[\boldsymbol{a}/(\boldsymbol{a}+\boldsymbol{c})]\}^{½}=\boldsymbol{a}/[(\boldsymbol{a}+\boldsymbol{b}).(\boldsymbol{a}+\boldsymbol{c})]^{½}$)[67] ou o coeficiente de Sokal e Sneath que atribui um peso duplo às discordâncias no denominador ($A_6= \boldsymbol{a}/[\boldsymbol{a}+2(\boldsymbol{b}+\boldsymbol{c})]$). Uma análise deste tipo encontra-se, por exemplo, em Legendre e Legendre (2000) e em Blanc *et al*. (1976) com referência ao trabalho de Roux e Roux (1973)[68].

Contudo, os estudos realizados sobre estes coeficientes chamam a atenção para diversos aspectos:

- No estudo comparativo dos primeiros 43 coeficientes (Tabela 1.3.4; Hubálek, 1982) pode-se observar que:
  - Existem ligações de vários tipos entre eles[69]:
    - Igualdades: $A_7=A_8$; $A_{34}=\pm A_{42}$, se $\boldsymbol{ad}>\boldsymbol{bc}$.
    - Linear: $A_9=2A_7$
    - Quadrados: $A_{12}=A_{11}^2$; $A_{31}=A_{30}^2$
    - Logarítmicas: $A_{41}=\log A_{40}$
    - Trigonométricas: $A_{21}=\text{arcsen } A_{20}$; $A_{38}=\text{sen } 90\ A_{37}$
    - Afins: $A_{10}=2A_7-1$; $A_{24}=2A_{20}-1$; $A_{33}=2A_{32}-1$ (corresponde a transformar o intervalo de definição de [0, 1] em [–1, +1]).

    A lista dos 43 coeficientes reduz-se assim a 30 coeficientes.
  - Há coeficientes que verificam condições de admissibilidade:
    - Não ser coeficiente de dissemelhança – Todos verificam esta condição.
    - Atingir o mínimo quando $\boldsymbol{a}=\boldsymbol{d}=0$ e o máximo quando $\boldsymbol{b}=\boldsymbol{c}=0$ – Não verificado por A_{13}, A_{16}, A_{27}, A_{28}, A_{29}, A_{31}.
    - Simétricos em i e j, $S_{ij}=S_{ji}$ – Não verificado por A_1, A_2, A_{13}, A_{16}, A_{26}.

[67] Também se pode considerar que se obtem dividindo ***a*** pela média geométrica de (***a***+***b***) e (***a***+***c***).
[68] A referência bibliográfica a que temos acesso é a da 3ª edição, i. e., de 1980.
[69] Na mesma ordem de ideias, verifica-se que $A_{29}=A_{30}/(A_{30}\ \text{max})$ (Blanc *et al*., 1976).

- Discriminação entre associação positiva e negativa, S(***a***>***a***') > S(***a***<***a***') sendo ***a***'= (***a***+***b***)(***a***+***c***)/***n*** – Não verificado por A_{27}, A_{28}, A_{31}.
- Ter limites não centrados ($S \in [0, 1]$, $S \in [0, +\infty[$ ou centrados, $S \in [-1, 1]$ – Não verificado por A_9, A_{13}, A_{17}, A_{21}, A_{41}.
- Nulidade, S(***a***=0)=0 – Não verificada por A_{17}, A_{18}, A_{19}, A_{20}, A_{21}, A_{22}, A_{23}, A_{24}, A_{26}, A_{30}, A_{39}.
- Linearidade com margens constantes, S(***a***+1)-S(***a***)= S(***a***+2)-S(***a***+1) – Não verificada por:

  - A_3, A_4, A_6, A_{12}, A_{15}, A_{19}, A_{23}, A_{25}, A_{36}, A_{38}, A_{39} (convexas, <).
  - A_{21}, A_{22}, A_{27}, A_{28}, A_{29}, A_{31}, A_{32}, A_{33}, A_{37}, A_{41}, A_{43} (côncavas, >).

▪ No final, Hubálek realiza um estudo empírico sobre 20 coeficientes de associação que considera serem "admissíveis" e obtém a partir de um algoritmo de Análise Classificatória Hierárquica Ascendente (Coeficiente de correlação de Pearson r + Ligação pela média[70]), no melhor nível segundo determinado critério, a partição: $\{A_{14}, A_{39}\}$, $\{A_3, A_4, A_5, A_6; A_7, A_{11}, A_{25}; A_{18}, A_{30}\}$, $\{A_{32}\}$, $\{A_{34}, A_{35}\}$, $\{A_{36}, A_{37}, A_{38}; A_{43}; A_{15}, A_{40}\}$. Realizou-se um estudo análogo (Secção 5.4 do Capítulo 5) com outros algoritmos, tendo-se obtido resultados em parte análogos.

- Sarker e Saiful Islam (1999) também propõem o coeficiente de *performance* relativa[71] para comparar coeficientes de semelhança e de dissemelhança. Os coeficientes que têm uma métrica estrutural semelhante são habitualmente considerados numa mesma classe. Este coeficiente foi aplicado a onze coeficientes para dados binários, permitindo encontrar dois grupos:
  - Um grupo inclui os coeficientes de semelhança de Jaccard (A_4), *simple matching* (A_{20}), Rogers e Tanimoto (A_{23}), Sokal e Sneath (A_{22}), Russel e Rao (A_{14}), Baroni-Urbani e Buser (A_{32}), Sorenson (A_5) e Ochiai (A_{11}).
  - Do outro grupo fazem parte os coeficientes de Hamann (A_{24}), Yule (A_{36}) e o coeficiente *phi* de Pearson (A_{30}).

- Diversos coeficientes para variáveis binárias obtêm-se a partir de outros homograficamente ($S = \frac{1-D}{1+D}$) (Tabela 1.3.2). É o caso da distância de Jaccard que está associada ao coeficiente de semelhança de Sokal-Sneath-Anderberg e da distância de Czenakowski-Dice que está associada ao coeficiente de semelhança de Jaccard (Joly e

[70] *Average linkage*, em inglês.
[71] *Relative performance coefficient* (RPC), no original.

Le Calvé, 1994). A transformação D=maxS-S é a função decrescente das semelhanças mais frequentemente usada com coeficientes para dados binários (Tabela 1.3.2).

- A natureza geométrica de alguns coeficientes de dissemelhança obtidos a partir da transformação D=1-S e da transformação pela função potência $D=(1-S)^{1/2}$, sob o ponto de vista das suas estruturas métrica e euclidiana, são apresentadas no Capítulo 2 (Tabela 2.2.1). Destacam-se os resultados (as demonstrações destes resultados encontram-se em Fichet e Le Calvé (1984) e Gower e Legendre (1986)):
    - Os índices de dissemelhança definidos para variáveis binárias: Jaccard ($1-A_4$), Czenakowski e Dice ($1-A_5$), Sokal e Sneath e Anderberg ($1-A_6$), Ochiai ($1-A_{11}$), Russel e Rao ($1-A_{14}$) e Rogers e Tanimoto ($1-A_{23}$) são distâncias *city block* (Joly e Le Calvé, 1992). Este resultado incita, vivamente, a utilizar a representação do tipo-M^1 para aquelas variáveis (Joly e Le Calvé, 1994).
    - Todos os índices de dissemelhança – $1-A_1$, $1-A_4$, $1-A_5$, $1-A_6$, $1-A_{11}$, $1-A_{14}$, $1-A_{20}$, $1-A_{23}$, $1-A_{24}$, $1-A_{25}$ e $1-A_{30}$ – são quase-hipermétricas, i.e., a sua raiz quadrada é uma distância euclidiana.

      Uma vez que as matrizes de semelhanças respectivas são s.d.p., os elementos poderão ser representados recorrendo à ACP das respectivas matrizes de semelhanças ou recorrendo à Análise em Coordenadas Principais, ACoP, (Gower, 1966) – situação que se encontra bem exemplificada em Blanc *et al*. (1976).
- O coeficiente de Kulczynski (A_3) não deverá ser usado em análises de inércia, pois $s_{ii}=\boldsymbol{a}/0$; assim como a sua generalização, o coeficiente de Sokal e Sneath (A_{19}), por motivo análogo (Blanc *et al*., 1976).
- Alguns autores propõem famílias de coeficientes de semelhança associados a uma mesma preordenação que englobam vários coeficientes para variáveis binárias (Tabela 1.3.5) (e.g., Fichet e Le Calvé 1984; Gower e Legendre, 1986; Beninel, 1987). Demonstra-se que coeficientes que pertencem a uma mesma família induzem uma mesma preordenação sobre I (e.g., Beninel, 1987). O facto de se saber que os coeficientes que pertencem à mesma família induzem uma mesma preordenação sobre I facilita a escolha do coeficiente, quando se recorre a análises ordinais dos dados, pois como se sabe neste caso os índices que pertencem a uma mesma classe de preordenação fornecem o mesmo resultado. Em particular, o que se acaba de dizer concretiza-se em Análise Classificatória Hierárquica Ascendente quando se utilizam aqueles coeficientes de semelhança juntamente com coeficientes de agregação entre classes monotonicamente invariantes, tais como, o *single linkage* e o *complete linkage*.

Tabela 1.3.5. Famílias de coeficientes de semelhança para dados binários associados a uma mesma preordenação e respectivos coeficientes

Família de semelhanças	Definição da família de semelhanças	Coeficientes de semelhança
$(S^{\theta})_{\theta\in R^{+}}$	$S_{ij}^{\theta}=\dfrac{a}{a+\theta(b+c)}$	Sokal e Sneath-Anderberg, A_6 (θ=1/2) Jaccard, A_4 (θ=1) Czekanowski e Dice, A_5 (θ=2)
$(S^{\alpha,\beta})_{(\alpha,\beta)\in R^{+}\,x\,R^{+}}$	$S_{ij}^{\alpha,\beta}=\dfrac{a-\alpha(b+c)+d}{a+\beta(b+c)+d}$	Rogers e Tanimoto, A_{23} (α=0, β=2) Sokal e Sneath, A_{22} (α=0, β=½) Hamman, A_{24} (α=1, β=1) *Simple matching*, A_{20} (α=0, β=1)
$(S^{\lambda,\gamma})_{(\lambda\in R^{+}\,e\,\gamma\in R^{+})}$	$S_{ij}^{\lambda,\gamma}=\dfrac{a+\frac{1}{2}d}{\lambda(a+\frac{1}{2}a)+\gamma(\frac{1}{2}d+b+c)}$	Vários coeficientes pertencem a esta classe, incluindo os índices de Marcotorchino e Michaud (1981).
(S^{K} e S^{L})	$S_{ij}^{L}=\dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{n}S_{ij}^{k}} e^{-\frac{t^2}{2}}\,dt$	S^{K} é o coeficiente de Yule (1912), Pearson e Heron (1913), A_{30}. S^{L} é o índice de Lerman.

- Fichet e Le Calvé (1984) definem a família de índices de dissemelhança, das quais o número de concordâncias negativas não faz parte, $\{d_{\theta}, \theta\in R^{+}\}$,

$$d_{\theta}(x,y)=\begin{cases}(b+c)/(\theta a+b+c) & \text{se } (x,y)\neq(0,0)\\ 0 & \text{senão}\end{cases}$$

ou, de forma equivalente, a família de semelhanças $(S^{\theta})_{\theta\in R^{+}}$. Demonstram que os coeficientes que pertencem a esta família induzem uma mesma preordenação sobre I.

Esta família $(S^{\theta})_{\theta\in R^{+}}$ também é proposta por Gower e Legendre (1986) e por Roux e Roux (1973).

A escolha entre os três coeficientes da família $(S^{\theta})_{\theta\in R^{+}}$ depende dos valores que (***b***+***c***)/***a*** toma. O valor de (***b***+***c***)/***a*** representa o grau de homogeneidade faunística da colheita[72] estudada – quanto menor for (***b***+***c***)/***a***, maior é a homogeneidade faunística. Para valores de (***b***+***c***)/***a*** >>1, o coeficiente de Czekanowski e Dice (A_5) é o mais discriminante, pois tende mais lentamente para zero. Para valores de (***b***+***c***)/***a***<<1, o coeficiente de Sokal e Sneath é o mais discriminante. Enquanto que, para valores (***b***+***c***)/***a*** intermédios será preferível utilizar o coeficiente de Jaccard (A_4) (Blanc *et al*., 1976).

[72] *Prélèvement*, no original.

- A família de semelhanças $(S^{\alpha,\beta})_{(\alpha,\beta)\in R^+ xR^+}$, das quais o número de concordâncias negativas faz parte, contém a família T_θ proposta por Gower e Legendre (1986), assim como a família $(S^\lambda)_{(\lambda\in R^+)}$, $S_{ij}^{\lambda} = \frac{1}{1+\lambda\left[(b+c)/(a+d)\right]}$, – quando λ=1/2, S^λ=A_{22}; para λ=1, S^λ=A_{20}; para λ=2, S^λ=A_{23} (Blanc *et al*., 1976) –. A escolha entre os coeficientes referidos depende da razão entre o número de discordâncias e o número de concordâncias ((***b***+***c***)/(***a***+***d***)) – esta escolha é analisada, por exemplo, em Blanc *et al*. (1976).

- Os coeficientes de Czekanowski e Dice (A_5), de Ochiai (A_{11}) e de Kulczynski (A_7) exprimem-se, respectivamente, como razões entre o número de concordâncias positivas, ***a***, e as médias aritmética, geométrica e harmónica dos números de caracteres presentes para cada um deles ((***a***+***b***) e (***a***+***c***)). As suas definições dependem, pois, da média considerada. Como a média harmónica é sempre inferior à média geométrica e esta é sempre inferior à média aritmética, os valores das semelhanças verificam sempre a desigualdade: $A_5 < A_{11} < A_7$ (Roux e Roux, 1980). Verifica-se a igualdade dos três coeficientes, $A_5 = A_{11} = A_7$, quando ***a***+***b*** = ***a***+***c***. O critério de escolha entre estes coeficientes depende da noção de dissimetria entre (***a***+***b***) e (***a***+***c***). Para um estudo mais pormenorizado destes coeficientes pode-se recorrer a Blanc *et al*. (1976). A passagem de uns coeficientes aos outros não se faz através de uma transformação monótona como vimos para outros coeficientes.

- Caillez e Kuntz (1996) consideram uma família de dissemelhanças $\{d_\alpha\}_{\alpha\in R}$ entre subconjuntos de um conjunto finito, associadas a médias de ordem-α, que generaliza a família de coeficientes de dissemelhança correspondentes aos apresentados no parágrafo anterior (Tabela 1.3.6). Estes autores demonstram que a raiz quadrada da dissemelhança d_α é euclidiana se $\alpha\geq 0$. Este resultado generaliza o demonstrado por Fichet e Le Calvé (1984) e Gower e Legendre (1986) para os coeficientes de Ochiai e Dice (caso de α= 0 e α= 1 com pesos iguais), que se apresenta na Subsecção 2.2.2 do Capítulo 2, (Tabela 2.2.1).

Tabela 1.3.6. Associação entre médias-α e dissemelhanças d_α, no caso de caracteres com igual peso e natureza euclidiana da raiz quadrada das dissemelhanças respectivas

α	Média-α, m_α	Coeficientes de dissemelhança d_α	Natureza euclidiana de $(d_\alpha)^{1/2}$
-1	harmónica	Kulczynski (1-A_7)	Não é euclidiana
0	geométrica	Ochiai (1-A_{11})	Euclidiana
1	aritmética	Dice (1-A_5)	Euclidiana
$-\infty$	mínimo	Simpson (1-A_2)	Não é euclidiana
$+\infty$	máximo	Braun-Blanquet (1-A_1)	Euclidiana

- Numa abordagem probabilística, Bacelar-Nicolau (1980, 1987, 1989) demonstra que os coeficientes de semelhança A_3, A_4, A_5, A_6, A_7, A_{11}, A_{14}, A_{20}, A_{23}, A_{24}, A_{30} e A_{36}, sob determinada hipótese de referência, têm a mesma distribuição exacta ou têm a mesma distribuição assintoticamente normal, i.e., são equivalentes a qualquer uma das estatísticas de base (S, U, V, T)[73] ou assintoticamente equivalentes do ponto de vista distribucional. Com base nestes resultados Bacelar-Nicolau (1980) propõe um coeficiente geral de semelhança do tipo VL para variáveis binárias (X,Y):

  $$p_{xy} = P(S^* \le s^*) \approx \Phi\left(s^*\left(X,Y\right)\right) \quad (1.3.22)$$

  p_{xy} é o valor da função de distribuição da normal reduzida, N(0,1), no ponto s*(X,Y) e s* é o coeficiente de associação de Pearson, S_P, a menos do factor $\sqrt{n-1}$, $s^*\left(X,Y\right) = \sqrt{n-1}\, S_P$ (S_p é o coeficiente A_{30} na Tabela 1.3.4).

  Este estudo foi realizado no contexto da análise classificatória, daí a importância daqueles resultados, uma vez que se sabe que coeficientes que pertençam à mesma classe de equivalência distribucional associados à noção de semelhança VL (coeficientes de semelhança probabilísticos) produzirão exactamente a mesma árvore de classificação, qualquer que seja o critério de agregação utilizado.

  No caso particular das tabelas de contingência 2x2 de margens fixas, os resultados obtidos sobre aqueles coeficientes de semelhança basearam-se na demonstração da convergência da distribuição hipergeométrica para a distribuição normal e nas consequências deste resultado para o estudo deste tipo de tabelas (Bacelar-Nicolau e Nicolau, 1978).

Como se viu não se pode dizer que um coeficiente é melhor do que todos os outros, mas pode-se dizer que a escolha do coeficiente de semelhança, quando as variáveis são

[73] S, U, V, T designam as variáveis aleatórias correspondentes às células da tabela de contingência 2x2 em que se encontram as frequências ***a***, ***b***, ***c*** e ***d***, respectivamente.

binárias, é determinada pelo grau de homogeneidade faunística do conjunto das colheitas analisadas (Blanc *et al*., 1976).
Finalmente, embora sem "esgotar" este assunto, do que se apresentou, pode-se ficar com uma ideia dos cuidados a ter quando se escolhe o coeficiente de associação (ou de dissemelhança), pois esta escolha influencia certamente os resultados obtidos quando se usam métodos de Análise de Dados.

No âmbito da Medicina, quer em clínica, quer em epidemiologia, utilizam-se, além de alguns dos coeficientes já apresentados (e.g., o coeficiente phi, os coeficientes kapa de Cohen), ainda outros coeficientes de associação, em tabelas de contingência 2x2, tais como: a sensibilidade, a espificidade, os valores preditivos, os riscos relativos, o *odds ratio* (e.g., Kraemer, 2006; Bruce *et al*., 2008),... . A escolha adequada daquelas medidas, sob o ponto de vista da significância clínica, é abordada, por exemplo, em Kraemer (2006).

1.4 Representações

A representação gráfica das estruturas de dados é um dos objectivos dos métodos de Análise de Dados. A sua importância deve-se ao facto de elas permitirem visualizar e transmitir as informações fornecidas pelos dados. Por isso o problema da representação gráfica é um assunto tão importante em Análise de Dados. Estas representações deverão satisfazer a duas propriedades (Le Calvé, comunicação pessoal):

- Fidelidade – representação tão exacta quanto possível. Esta propriedade está ligada à ideia de optimização de um critério de aproximação.
- Legibilidade – regras de leitura simples e fáceis de comunicar.

As representações gráficas tomam aspectos diferentes consoante resultam da análise dos dados com técnicas de análise factorial e de *multidimensional scaling* ou com técnicas de análise classificatória. No caso das análises factoriais e do *multidimensional scaling* as representações gráficas tomam o aspecto de nuvens de pontos num espaço $\mathbb{R}^p$ ($p \in \mathbb{N}$) ou $\mathbb{R}^n$ ($n \in \mathbb{N}$), enquanto que a análise classificatória permite obter, por exemplo, dendrogramas, árvores aditivas. Relembremos que o que se representa não são os dados mas as suas inter-relações; assim umas representações estão mais ligadas às semelhanças

(as que resultam dos métodos factoriais[74]), outras estão principalmente ligadas às dissemelhanças (as que resultam das análises classificatórias). Na Tabela 1.4.1 encontra-se uma síntese de alguns resultados sobre a representação das estruturas de dados (I, D) apresentados por Beninel (1987) e por Le Calvé (e.g., 1988 e em apresentações pessoais) que os acompanham com demonstrações e indicações bibliográficas valiosas sobre as metodologias utilizadas para obter essas representações. Como já foi referido aquelas dissemelhanças estão ligadas umas às outras e essas ligações podem ser apresentadas num esquema de inclusão feito por Critchley, Fichet e Le Calvé – este esquema encontra-se reproduzido, por exemplo, em Beninel (1987).

Tabela 1.4.1. Quadro resumo sobre a natureza de algumas dissemelhanças e a representação gráfica geralmente associada à estrutura de dados (I,D)

Natureza da dissemelhança D	Representação gráfica geralmente associada a (I,D)
Distância	Grafo com valores[75]. Representação em $\mathbb{R}^2$
Distância *city block* (distância de tipo M^1)	Árvore aditiva. Representação *city block* em $\mathbb{R}^2$
Distância euclidiana	Representação em $\mathbb{R}^2$
Distância ultramétrica	Dendrograma
Distância ao centro	Árvore ao centro ou "árvore estrelada"[76]
Dissemelhança quadrangular	Árvore aditiva
Distância de cadeia	Cadeia (árvore aditiva particular, i.e., árvore linear)
Dissemelhança piramidal	Representação piramidal (também usada para representar uma estrutura de dados qualquer)
Distância quase-hipermétrica	Não se conhece uma representação gráfica típica
Distância hipermétrica	Não se conhece uma representação gráfica típica

Um estudo mais aprofundado sobre a representação das distâncias de tipo M^1 é apresentado, por exemplo, por Al Ayoubi (1991) (análise factorial em métrica M^1 e algoritmo de florestas aditivas). Para a representação em métrica M^1 ou M^0 reportamo-nos a Cécil Favre (1999). Sobre a "estrutura euclidiana de um dendrograma" terá interesse consultar Critchley (1988).

---

[74] Convém esclarecer, pois o português pode-nos atraiçoar, que os métodos factoriais são usados com semelhanças. Como é sabido, as semelhanças são também usadas na primeira etapa da Análise Classificatória Hierárquica.

[75] *Graphe valué*, no original.

[76] *Arbre étoilé*, no original.

Como se sabe, estamos interessados nas representações no plano, tais como a folha de papel ou o ecrã do computador. As representações em $\mathbb{R}^n$ ou em $\mathbb{R}^p$ utilizam justaposições das representações em $\mathbb{R}^2$, mas elas só têm sentido se soubermos qual é a métrica do plano. Uma das possibilidades é a representação euclidiana[77], sendo esta a mais legível. Nas subsecções que se seguem preocupamo-nos com a representação euclidiana das estruturas de dados.

1.4.1 Apresentação geral do problema da representação euclidiana das estruturas de dados

A cada um dos três aspectos gerais que os dados (Secção 1.2) podem apresentar em análise de dados correspondem, respectivamente, as três estruturas[78] de dados seguintes:

(1) (I,X), i.e., o conjunto/espaço I munido da aplicação X, num conjunto/espaço F, X: $I \rightarrow F$.

(2) (H,S), sendo H= X_i ou I – conjunto de elementos (variáveis ou unidades estatísticas) sobre os quais foi calculada a semelhança S.

(3) (H,D), sendo H= X_i ou I – conjunto de elementos (variáveis ou unidades estatísticas) sobre os quais foi calculada a dissemelhança D.

Interessa-nos procurar uma representação euclidiana destas três estruturas de dados, "fazendo bem a distinção entre a estrutura de que os dados estão munidos e a estrutura do espaço de representação" (Le Calvé, 1976b).

Le Calvé (1976b) apresenta o problema da representação das estruturas de dados, de forma geral, em termos de conjuntos/espaços, estruturas, imagens exactas e imagens aproximadas, considerando duas situações gerais que se passam a descrever.

1.4.1.1 Caso A. Representação da estrutura de dados (I,X)

Para apresentar o problema da representação da estrutura de dados (I,X) tem que se saber o que se entende por imagens exacta e aproximada de uma estrutura de dados.

---

[77] A outra é a *city block*.
[78] Uma estrutura de dados é um conjunto munido de uma aplicação.

Definição 1.4.1. Imagens exacta e aproximada de uma estrutura de dados (I,X) (Le Calvé,1976b)

Seja a estrutura de dados (I,X), i.e., o conjunto/espaço I munido da aplicação X, num conjunto/espaço F, X: $\mathrm{I} \rightarrow$ F.

Seja (E,Y) outra estrutura de dados, tal que E é um conjunto finito e Y uma aplicação no conjunto/espaço F, Y: E $\rightarrow$ F.

Diz-se que Y é uma imagem exacta de X se e só se, no sentido de um critério $\mathcal{C}$, $\mathcal{C}$(Y,X) = 0.

No caso do critério $\mathcal{C}$ ser apenas minimizado, $\mathcal{C}(\mathrm{Y},\mathrm{X})_{\mathrm{Min}}$, diz-se que Y é uma imagem aproximada de X.

Habitualmente são utilizados três critérios:

- O critério do valor absoluto da diferença no espaço métrico L_1, $\mathrm{C}(\mathrm{Y},X) = \|X - \mathrm{Y}\|_{L_1}$.
- O critério dos mínimos quadrados no espaço métrico euclidiano, L_2, $\mathrm{C}(\mathrm{Y},X) = \|X - \mathrm{Y}\|_{L_2}$.
- O critério do máximo dos valores absolutos das diferenças no espaço métrico de Chebyshev, L_∞, $\mathrm{C}(\mathrm{Y},X) = \|X - \mathrm{Y}\|_{L_\infty}$.

A estrutura do espaço de representação dependerá pois, do critério $\mathcal{C}$ utilizado.

1.4.1.2 Caso B. Representação das estruturas de dados (H,S) e (H,D)

A definição de imagem, tal como no caso A, é importante.

Definição 1.4.2. Imagens das estruturas de dados (H,S) e (H,D)

Seja H o conjunto de unidades estatísticas ou de variáveis. Seja a estrutura de dados (H,M), em que o conjunto finito H está munido da aplicação M em $\mathbb{R}$, sendo uma dissemelhança ou uma semelhança, M=D ou M=S,

M: $H^2 \rightarrow \mathbb{R}$

Considere-se outra estrutura de dados, (E,P), em que o conjunto finito E, com estrutura algébrica frequentemente mais forte do que a de H, está munido da aplicação P em $\mathbb{R}$,

P: $E^2 \rightarrow \mathbb{R}$ (as aplicações M e P podem ser de natureza diferente).

Diz-se que os pontos e_i, i=1,...,h, de E formam uma imagem dos pontos i de H se e só se for respeitada em E a relação de semelhança ou dissemelhança entre os elementos de H,

$H \rightarrow E$

$i \rightarrow e_i$ tal que, M respeita P

i. e., $(i,j) \rightarrow (e_i,e_j)$ tal que, $P(e_i,e_j)=M(i,j)$ ou $P(f(i),f(j))=M(i,j)$

A imagem da estrutura (H,M) na estrutura (E,P) é tal que as aplicações são as mesmas, $(H,M) \rightarrow (E,P)$.

No caso geral, a estrutura (H,M) é dada e a estrutura de dados (E,P) é escolhida.
Mas, como proceder para a escolher? Uma vez que a imagem exacta da estrutura (H,M) não existe, procuram-se imagens aproximadas. Em vez de se ter $P(e_i,e_j)=M(i,j)$, tem-se $P(e_i,e_j) \approx M(i,j)$, o que corresponde a utilizar um critério $\mathcal{C}$ sobre P e M, tal que $\mathcal{C}(P(e_i,e_j),M(i,j))_{Min}$.

Os critérios mais utilizados são os três critérios já referidos no caso A.
Habitualmente, escolhe-se o critério dos mínimos quadrados. Esta escolha leva-nos a uma estrutura de dados euclidiana, o que facilita tudo. Sabemos trabalhar no espaço vectorial euclidiano, pois é um espaço em que o produto interno está definido, o que permite exprimir as normas, distâncias e ângulos entre vectores em termos de produto interno (espaços de Hilbert).

- Uma escolha coerente da estrutura (E,P) será:
  - $E=\mathbb{R}^n$
  - P, uma forma quadrática s.d.p.
  - $\mathcal{C}$, o critério dos mínimos quadrados em L_2, $\mathrm{C}(\mathrm{M},\mathrm{P}) = \|M-P\|_{L_2}$

Quando se fala em imagens euclidianas trabalhamos com formas quadráticas – caso A e caso B para a matriz de semelhanças, M=S. Nestes casos, o problema consiste em encontrar em $\mathbb{R}^n$ uma imagem que represente o melhor possível os produtos escalares de partida – a resposta é dada pela ACP (Subsecção 1.4.2).

No Caso B para a matriz de dissemelhanças, M=D, situamo-nos numa análise de dissemelhanças ou de distâncias, mais concretamente no *Multidimensional Scaling*[79] (MDS) ou Posicionamento Multidimensional.

[79] *Analysis of Proximity Data*, em inglês. *Análise de escalonamento multidimensional*, em brasileiro. *Analyse métrique* (Le Calvé) ou *positionemment multidimensionel*, em francês. *Análisis de proximidades* ou *representación multidimensional*, em espanhol.

Conclusão: Os dois primeiros problemas de representação das estruturas de dados, (I,X) e (H,S), serão resolvidos com a "ajuda" de formas quadráticas pois a representação dessas estruturas vai ser feita num espaço euclidiano. Enquanto que o problema da representação da estruturas de dados (H,D) corresponde à análise de uma matriz de distâncias.

1.4.2 Representação euclidiana da estrutura de dados (I,X)

1.4.2.1 Enunciado do problema da representação euclidiana da estrutura de dados (I,X)

O problema da representação euclidiana da estrutura de dados (I,X) pode-se enunciar da seguinte maneira:

- A estrutura de dados (I,X) é definida pelo conjunto finito I munido da aplicação X no conjunto finito F,

  $$X : I \rightarrow F$$

  Habitualmente, a estrutura (I,X) é definida da seguinte maneira:

  - Sejam p variáveis X_1, ..., X_p, quantitativas ou qualitativas que tomam valores num conjunto Y munido ou não de uma estrutura algébrica[80].

  - Seja I um conjunto (finito) de *n* indivíduos/unidades de dados descrito pelas p, p<n, variáveis numa matriz $X_{(n,p)}$.

  X é uma matriz de ordem r, $r \leq p$,

  $$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \dots & x_{ij} & \dots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

  O elemento genérico x_{ij} (i=1, ...,n; j=1, ...,p) de X representa o valor que a variável X_j toma para o indivíduo i.

- O problema consiste em encontrar uma matriz $\hat{X}_q$, q<r, de ordem inferior à de X, que seja o mais semelhante possível a X. Ou seja, pretende-se encontrar $\hat{X}_q$ tal que a distância entre X e $\hat{X}_q$ seja mínima: $\left\| X - \hat{X}_q \right\|_{Min}$.

[80] Esta ausência de estrutura impede-nos de nos enquadrarmos no esquema de dualidade.

Exemplo 1.4.1. Enunciado do problema da representação da estrutura de dados (I,X)

Temos a estrutura (I,X), sendo:

- $I=\mathbb{R}^p$, *card* I = n, $X= X_{(n,p)}$.

Pretendemos encontrar a estrutura (E,Y), tal que:

- $E=\mathbb{R}^p$
- $Y= Y_{(n,p)}$ de ordem r, $r \leq p$, imagem de (I,X) no sentido de um critério tal que: $\|X-Y\|_{Min}$. Se $X_{(n,p)}$ é de ordem r então $\|X-Y\| = 0$.

Na resposta a este problema põe-se, em primeiro lugar, a questão: - Que norma devemos utilizar?

A escolha da norma em L_2, $\|.\|_{L_2}$, resulta de forma natural, pois pretendemos obter uma representação euclidiana da estrutura (I,X).

A escolha do espaço vectorial euclidiano tem em conta o facto deste espaço ser rico em propriedades que nos facilitam quer os cálculos, quer a visualização das representações.

Contudo, ao escolhermos o espaço euclidiano, o critério[81] que utilizamos leva-nos a calcular quadrados de distâncias que correspondem a "grandes" erros que têm um peso muito grande. Este problema está ligado à utilização da norma L_p e agrava-se quando p aumenta . Sabe-se que a regressão mais robusta é obtida em L_1 .

- Uma vez escolhida a norma em L_2, põe-se o problema de minimizar $\|X-\hat{X}\|_{L_2}$, ou seja, de saber como obter a solução de $\|X-\hat{X}\|_{L_2 Min}$.

1.4.2.2 Solução do problema da representação da estrutura (I,X)

É possível encontrar uma solução directa utilizando o teorema de Eckart e Young e o de Kristoff. Le Calvé (1976b) apresenta-os com os enunciados que se seguem.

Teorema 1.4.1. Teorema de Eckart e Young

Para toda a matriz real A, existem duas matrizes ortogonais P e Q, tais que P^TAQ é uma matriz diagonal[82] real com elementos não negativos.

---

[81] Minimizar os quadrados das distâncias entre as duas estruturas de dados.

[82] Dizemos que uma matriz rectangular B é diagonal se $b_{ij} = 0$, $\forall$ $i \neq j$

Este teorema, enunciado por Eckart e Young (1963), foi demonstrado no caso geral por Johnson (1963). Le Calvé (1976b) retomou a demonstração de Johnson, rectificou-a e completou-a.

Convém referir que:

- A é uma matriz m x n, $m \geq n$, de ordem $r \leq n$.
- P resulta da diagonalização de AA^T.
- Q resulta da diagonalização de A^TA.
- Λ corresponde à matriz diagonal dos valores característicos/singulares, que é a mesma para P e para Q.

O teorema de Kristoff (Kristoff, 1970) apresenta o seguinte enunciado:

Teorema 1.4.2. Teorema de Kristoff

Seja $(X_i)_{i=1, \ldots, n}$, uma família de matrizes ortogonais (m,m) e $(\Gamma_i)_{i=1, \ldots, n}$, uma família de matrizes diagonais. Seja $\hat{\Gamma}_i$ a matriz que se obtém de Γ_i ordenando os valores absolutos dos elementos da diagonal de Γ_i por ordem decrescente. Então

$$\left| tr\left(X_1\Gamma_1 \ldots X_n\Gamma_n \right) \right| \leq tr\left(\hat{\Gamma}_1 \ldots \hat{\Gamma}_n \right)$$

Le Calvé (1976b) retomou a demonstração deste teorema e simplificou-a.

Dos teoremas de Eckart e Young (1936) e Kristoff (1970) resulta a seguinte proposição

Proposição 1.4.1.

Seja A uma matriz (n, p) de ordem r. A solução do problema de "encontrar uma matriz B(n,p) de ordem $q \leq r$, tal que $\|A-B\|_{L_2}$ seja mínima" é obtida quando B é a projecção de A sobre os seus primeiros *q* vectores próprios.

Uma vez que não podemos falar de valores próprios, nem de vectores próprios de matrizes rectangulares, estes termos[83] designarão os vectores e valores diagonais que intervêm na decomposição de Eckart e Young.

Façamos pois, $A=P\Lambda Q^T$

$B=R\,\Delta\,T^T$

Por definição de norma euclidiana de uma matriz pode-se escrever

$\|A-B\|_{L_2}^2$ = tr[(A-B)(A-B)^T]

= tr(AA^T+B B^T-2BA^T)

= tr($P\Lambda Q^TQ\Lambda^TP^T$ + $R\Delta T^T\Delta^TR^T$- 2 $R\Delta T^TQ\Lambda^TP^T$)

= tr($\Lambda\Lambda^T$ + $\Delta\Delta^T$ – 2$Q\Lambda^T$ $P^TR\Delta T^T$)

Do teorema de Kristoff resulta que: R=P , T=Q

$\|A-B\|_{L_2}^2 \geq$ tr[(Λ-Δ)(Λ-Δ)^T]

Este traço é minimizado, $\{tr[(\Lambda-\Delta)(\Lambda-\Delta)^T]\}_{Min}$, fazendo a restrição de Λ aos seus *q* primeiros valores próprios, $\Delta=[\Lambda]_q$, o que corresponde à projecção da matriz X sobre a matriz gerada pelos *q* primeiros vectores próprios, $\hat{X}_q = [\Lambda]_q X$.

Corolário 1.4.1. (Le Calvé, 1976b)

Se B_0 é a solução de $\|A-B\|_{Min}$, então $B_0^TB_0$ é a solução de $\left\|A^TA-B^TB\right\|_{Min}$.

Com efeito, tendo em conta as notações precedentes:

- $A^TA = Q\Lambda^T\Lambda Q^T$
- $B^TB = T\Delta^T\Delta T^T$

Pelo teorema de Kristoff o mínimo de σ = tr[(A^TA-B^TB) (A^TA-B^TB)^T] é atingido por $\Delta^T\Delta=[\Lambda^T\Lambda]_q$, T=Q, ou seja, por $B_0^TB_0$.

O Corolário 1.4.1 afirma que as abordagens são equivalentes porque dão a mesma solução.

Le Calvé (1976b) enunciou este resultado da seguinte maneira:

---

[83] Também designados por valores singulares e vectores singulares.

"A solução da melhor aproximação, no sentido dos mínimos quadrados, dos produtos escalares é a projecção."

Este resultado constitui o fundamento das análises factoriais - Análise em Componentes Principais, Análise Factorial das Correspondências Simples e Múltipla, Análise Canónica – e de certas análises ordinais (e.g., Marcotorchino e Michaud, 1979) em espaço euclidiano (Le Calvé, 1976b).

Uma abordagem análoga está patente, por exemplo, em Greenacre (1984) que afirma (e demonstra), referindo-se à Análise em Componentes Principais, aos *Biplot*, à Análise das Correspondências e às Análises Canónicas, que "Estas técnicas são todas variações de um mesmo tema, e esse tema é a álgebra e a geometria da decomposição em valores singulares".

1.4.2.3 Conclusão: A Análise em Componentes Principais, ACP, pode-se apresentar de três maneiras

O problema da ACP pode-se apresentar de três maneiras (Le Calvé, apresentação pessoal), que correspondem às três apresentações clássicas desta metodologia.

- **Problema 1**. Abordagem geométrica: Trata-se de encontrar uma aproximação, de dimensão dada, a uma configuração de pontos:
  - Sendo dada uma matriz de dados X, nxp, com n linhas e p colunas, representável num espaço euclidiano de dimensão q, pretende-se encontrar uma matriz $\hat{X}_r$, também nxp, representável num espaço euclidiano de dimensão r, r<q, o menos afastado de X no sentido dos mínimos quadrados.
  Este problema formaliza-se matematicamente da seguinte maneira:
  - *Sendo dada uma matriz X, nxp, de ordem q, pretende-se encontrar uma matriz F_r, nxp, de ordem r e tal que $\|X - F_r\|^2$ seja mínimo.*
    $\|.\|^2$ designa a norma euclidiana, i.e., $\|X - F_r\|^2 = \sum_{i,j} (X_{ij} - F_{r_{ij}})^2$.

    A solução, bem conhecida, é a projecção da matriz X sobre o espaço gerado pelos seus r primeiros vectores singulares (Teorema de Kristoff, Subsecção 1.4.2.2).

- **Problema 2**. Problema formalizado em termos de produtos escalares ou de matrizes de covariância:
  - Pretende-se encontrar *r* variáveis, G_j, independentes, que sejam combinações lineares das variáveis X_i que não são independentes, com variância máxima.
  - Ou de forma equivalente, pretende-se obter em $\mathbb{R}^q$ o conjunto dos pontos M_j que representem as variáveis e tal que $\langle OM_j, OM_{j'} \rangle = Cov(V_j, V_{j'})$

  Este problema formaliza-se matematicamente da seguinte maneira:
  - *Sendo dada uma matriz X de ordem q, pretende-se encontrar uma matriz G_r, nxp, de ordem r tal que $\left\| X^T X - G_r^T G_r \right\|^2$ seja mínimo*.

  A solução é a projecção sobre o espaço gerado pelos primeiros vectores próprios de X^TX ou pelos primeiros vectores singulares de X. Neste caso, $G_r = \hat{X}_r$.

- **Problema 3**. Procura dos eixos de maior inércia/variabilidade da nuvem de pontos.
  Este problema formaliza-se matematicamente da seguinte maneira:
  - *Sendo X uma nuvem de pontos num espaço de dimensão q pretende-se encontrar um sub-espaço P_r tal que a projecção H_r de X sobre P_r tenha uma inércia máxima.*

  A solução é a do Problema 1: $\mathrm{Pr}_{H_r}(X) = \hat{X}_r$.

É pois possível enunciar o seguinte resultado:

Teorema 1.4.3. (Le Calvé, 1976b)

Seja X uma matriz de dados com n linhas e p colunas representável num espaço euclidiano com dimensão q, a matriz $\hat{X}_r$, projecção de X sobre os seus r primeiros vectores singulares é a solução dos três problemas precedentes.

Vê-se bem que este resultado provém da conclusão apresentada: “a melhor aproximação, no sentido dos mínimos quadrados, dos produtos escalares é a projecção” (Le Calvé, 1976b). Mas esta propriedade deixa de ser válida se não se trabalhar com produtos escalares ou covariâncias ou correlações ou ainda se o espaço não for euclidiano. Para que não fiquem dúvidas sobre a possibilidade de se pensar em projecções em métrica não

euclidiana, Le Calvé mostra que "mínimos quadrados e projecção só coincidem em métrica euclidiana e que se deve emitir as maiores reservas quanto à análise factorial de uma matriz não definida positiva" (Le Calvé, 1976b).

Como consequência do que se viu, destaca-se:

- A ACP analisa a variância e as somas dos quadrados das distâncias (euclidianas) interpretam-se em termos de inércia. A decomposição da variância, σ^2, dá

$$\sigma^2(X) = \frac{1}{n(n-1)} \sum_{i,j} d_{ij}^2 = \frac{1}{n(n-1)} \sum_{k} \sum_{i,j} \left| x_{ik} - x_{jk} \right|^2$$

  e por conseguinte

$$\sigma^2(\mathrm{X}) = \sum_{k} \sigma_k^2(\mathrm{X})$$

- O cuidado a ter com a interpretação das figuras obtidas, pois interpretam-se de forma diferente:
  - Nuvem dos pontos que representam as unidades estatísticas (nuvem dos indivíduos): - A distância entre duas unidades estatísticas é a distância euclidiana habitual. A ACP preserva a distância euclidiana.
  - Nuvem dos pontos-variáveis (e.g., Lebart *et al.*, 1995) ou projecções dos eixos-descritores (e.g., Legendre e Legendre, 2000).
    - A projecção dos pontos-variáveis sobre as componentes principais mostra a covariância[84] dessas variáveis com as componentes principais e daí a sua contribuição (positiva ou negativa) à posição das unidades estatísticas ao longo dos eixos e à sua formação. As componentes principais são frequentemente designadas pelos nomes das variáveis que mais contribuem para a sua formação.
    - As proximidades entre os pontos-variáveis interpretam-se em termos de produtos escalares, ou seja covariâncias, tendo em conta a função de semelhança $d_{jj'}^2 = 2(1 - s_{jj'})$ (Joly e Le Calvé, 1994) apresentada na Tabela 1.3.2.

      No caso das variáveis terem sido padronizadas (i.e., centradas e reduzidas) viu-se que $d_{jj'}^2 = 2(1 - r_{jj'})$, sendo $r_{jj'}$ o coeficiente de correlação linear (Subsecção 1.3.4.1). Neste caso, no espaço $\mathbb{R}^n$, o coseno do ângulo entre os dois vectores-variáveis é o coeficiente de correlação entre as duas variáveis ($r_{jj'} = \cos(V_j, V_{j'})$). As proximidades entre pontos variáveis

[84] Correlação, no caso da ACP normada.

interpretam-se em termos de correlações (e.g., Lebart *et al*., 1995). Os estatísticos conhecem bem esta interpretação:- duas variáveis fortemente correlacionadas estão muito próximas uma da outra ou, pelo contrário, o mais afastadas possível conforme a relação linear que as liga é positiva ou negativa, respectivamente. Duas variáveis ortogonais (correlação nula) estão a uma distância média. A ACP normada facilita a interpretação.

Em face do que se acaba de mencionar, não podemos deixar de citar Le Calvé:

- *Enfin la figure obtenue s'interprète en faisant des groupes utilisant les distances. Or ce sont les produits scalaires qui ont été optimisés!!* (Le Calvé, apresentação pessoal).
- *It is interesting to quote that the representations are mostly read in terms of dissimilarities, while they are constructed in terms of similarities*. (Le Calvé, 1988).

- Como a ACP privilegia as grandes distâncias, o primeiro eixo é frequentemente um eixo de dimensão[85] (fazendo sobressair um ou alguns elementos extremos) o que leva, por exemplo, a usar a ACP normada (dados centrados e reduzidos) em vez da ACP.

  Como se sabe, os resultados obtidos por estas duas técnicas – ACP e ACP normada – sobre os mesmos dados quantitativos não são os mesmos. O interesse prático em utilizar a matriz de covariâncias ou a matriz de correlações é bem exemplificado, por exemplo, em Legendre e Legendre (2000).

Outro aspecto importante a abordar sobre a ACP refere-se à possibilidade de a adaptar e estender a sua aplicação a dados que não são quantitativos ou que são qualitativos e quantitativos, desde que pensemos nela apenas como técnica descritiva, assim como a situações particulares (e. g., Lebart *et al*., 1995; Jolliffe, 1998). Aqui mencionam-se algumas delas:

- Em Jolliffe (1986) encontra-se a referência com mais ou menos pormenor às adaptações da ACP a dados qualitativos, no caso de todos serem binários (Gower, 1966; Cox, 1972; Greenacre, 1984), no caso de todos serem números de ordem[86] (Gower, 1967) e à generalização "não métrica" da ACP para dados discretos (uma forma de ACP não linear) descrita por De Leeuw e Van Rijckevorsel (1980) que está

[85] *Taille*, em francês.
[86] *Ranks*, no original.

implementada no *software* estatístico SPSS (*Statistical Package for the Social Sciences*), no módulo PRINCALS (Gifi, 1983).

A referência à análise das correspondências como "adaptação" da ACP a dados nominais com duas variáveis (caso dos dados numa tabela de contingência) é reforçada com uma análise mais pormenorizada, assim como "a ideia da análise das correspondências como uma forma de ACP para dados nominais ser válida para qualquer número de variáveis binárias". As várias ligações entre a ACP e a análise das correspondências simples ou múltipla são analisadas por Jolliffe com referência a Greenacre (1984).

Neste sentido também podemos relembrar os resultados apresentados nesta secção com alusão a Le Calvé (1976b).

- A Análise em Componentes Principais usando o Simplex Regular para dados categóricos (Niitsuma e Okada, 2005)[87] é uma adaptação da ACP a dados categóricos que se baseia na diagonalização da matriz de covariâncias de variáveis categóricas que usa uma expressão simplex regular para categorias.
- A Análise das Ordens[88] (Lebart *et al*., 1995) é uma adaptação da ACP que se realiza sobre a matriz de dados transformada em números de ordem. Esta análise que se aplica quando os dados são ordinais, também é recomendada quando eles são heterogéneos.
- Quando as variáveis são de natureza diversa, quantitativas e qualitativas, um procedimento habitual consiste em transformar as variáveis quantitativas em qualitativas e aplicar uma análise de correspondências múltiplas (ACM) à matriz de dados homogénea assim obtida. Outra maneira consiste em obter uma matriz de dados "duplicada", apresentando as variáveis repetidas duas vezes, uma na sua forma original e outra na sua forma complementar, e aplicar uma análise das correspondências a esta matriz (e.g., Greenacre, 1984; Jolliffe, 1986).

  Baseando-se nas ideias de Escofier (1979) sobre a ACM para dados mistos e transpondo-as para o quadro da ACP, Pagès (2004) propõe a Análise Factorial de Dados Mistos[89]. Pagès mostra a equivalência desta análise com a análise factorial múltipla (Escofier e Pagès, 1998; Pagès, 2002) em que cada grupo se reduz a uma variável.

---

[87] *Principal Component Analysis using the Regular Simplex for categorical data* (RS-PCA), no original.
[88] *Analyse des rangs*, no original.
[89] *Analyse Factorielle de Données Mixtes* (AFDM), no original.

Com a designação de Análise em Componentes Principais Categórica[90] (e.g., Gifi, 1990; Heiser e Meulman, 1994; Meulman *et al*., 2004) ou análise em componentes principais não linear[91] (e.g., Gower e Blasius, 2005), esta ACP não linear do tipo "optimal scaling" (i.e., as categorias são substituídas por *scores* óptimos) aplica-se a dados qualitativos e quantitativos, e encontra-se implementada no módulo CATEGORIES (Meulman e Heiser, 1999) do SPSS.

- A Análise em Componentes Principais Simbólica em que os objectos são descritos por variáveis intervalares (Chouakria *et al.*, 2000) é uma das propostas para este tipo de dados simbólicos.
- Quando as observações das variáveis não são independentes é proposta a Análise em Componentes Principais para dados não independentes e em particular para séries temporais (e.g., Jolliffe, 1986; Rodrigues e Branco, 2007).
- Pelas suas aplicações (em particular, em Biomatemática) tem particular interesse a Análise em Componentes Principais modificada para Dados de Composição[92] por Aitchison (1983), segundo Jolliffe (1986). Estes dados consistem em vectores com soma igual a um, em que cada elemento é uma proporção; mais geralmente, aquela soma pode ser um valor fixo. Pensamos reconhecer nestas variáveis, as variáveis simbólicas modais (Subsecção 4.3.10.8 do Capítulo 4) que analisamos.
- Também são propostas análises em componentes principais robustas para ultrapassar, em particular, o problema dos *outliers* (e.g., Lebart *et al*., 1995; Jolliffe, 1986; Branden, 2005).
- Outros procedimentos são sugeridos para ultrapassar o problema dos dados omissos na ACP (e.g., Jolliffe, 1986).
- Vários autores (e.g., Lebart, 1995; Jolliffe, 1986) consideram incontornável a leitura de Rao (1964), em particular, pela sugestão de métodos derivados da ACP (a análise das correlações parciais ou análise com variáveis instrumentais).
- Com a designação de "ACP Generalizada"[93] encontram-se várias versões das quais se destacam, por exemplo, a apresentada por Vidal (2003) no âmbito da compressão de imagens, do reconhecimento de faces e da segmentação de imagens, e a apresentada por Ichino (1988), por Yaguchi e Ichino (1992) e por Ichino e Yaguchi

---

[90] *Categorical Principal Components Analysis* (CATPCA), no original.
[91] *Nonlinear Principal Components Analysis* (NLPCA), no original.
[92] *Principal component analysis for compositional data*, no original. Também poderíamos traduzir por dados compositivos. Composição, em Química, significa proporção em que os elementos entram num composto.
[93] *Generalized principal component analysis*, no original.

(1994) – generalização da ACP que se baseia nas métricas de Minkowski generalizadas – em reconhecimento de formas e análise de dados.

Embora a ACP seja uma metodologia muito antiga (Pearson, 1901; Hotelling, 1933) a sua aplicação, tal como a de outros métodos de análise de dados, só se tornou uma realidade ao alcance de todos com a possibilidade de usar os computadores. A ACP encontra-se disponível em diversos *softwares* estatísticos. Um deles é o SPSS, que se encontra disponível em muitas universidades e instituições portuguesas. No entanto, pensamos que a ACP é um ponto fraco deste *software*, pois é, frequentemente, confundida pelos utilizadores com a análise factorial (análise em factores comuns e específicos).

1.4.3 Representação euclidiana da estrutura de dados (H,S)

A representação euclidiana da estrutura de dados (H,S) coincide com o caso anterior.

No caso de S ser a matriz de covariâncias ou de correlações (de Pearson) situamo-nos na ACP e na Análise em Componentes Principais Normada (ACPN), respectivamente.
No caso de S ser a matriz de correlações de Spearman situamo-nos na "Análise das Ordens"[94] (Lebart *et al*., 1995), i.e., numa ACP das ordens ou ordinal.

Também se pode pensar, de forma geral, na matriz de semelhanças S simétrica, semidefinida positiva, que se escreve sob a forma de produto escalar, e utilizar directamente uma ACP sobre ela, isto é, diagonalizar a matriz S – é o que Le Calvé (1977) sugere, assim como outros autores (e.g., Chardy *et al*. 1976; Blanc *et al*., 1976; Greenacre, 1978; Benzécri *et al*., 1980; Greenacre, 1984; Jolliffe, 1986). No Capítulo 4 retoma-se esta ideia quando propomos análises em componentes principais baseadas nas matrizes de semelhanças S, S_{LC} e P_L e se visualizam as variáveis.

Contudo, para outros autores o procedimento habitual para visualizar semelhanças entre objectos é dado pelo *multidimensional scaling* (e.g., Borg e Groenen, 2005), em particular, por exemplo, pela Análise em Coordenadas Principais[95] (ACoP) proposta por Gower (1966). Aliás, se se pensar numa matriz de dados X_{nxp} e na matriz de distâncias euclidianas, D, calculadas entre os mesmos n objectos, os valores próprios obtidos pela ACoP da matriz D

[94] *Analyse des rangs*, no original.
[95] *Principal Coordinate Analysis* (PCoA), no original. Também é conhecida por *metric multidimensional scaling* ou *classical scaling* em inglês.

são iguais aos obtidos pela ACP da matriz $[x - \bar{x}]^T [x - \bar{x}]$ e são (n-1) maiores do que os valores próprios obtidos com a ACP da matriz de covariâncias de X, CovX = (1/(n-1)) $[x - \bar{x}]^T [x - \bar{x}]$ (Legendre e Legendre, 2000); a ACP é pois considerada como um caso especial da ACoP (e.g. Gower, 1966; Gower e Blasius, 2005). Em Blanc *et al*. (1976) encontra-se bem exemplificada a representação gráfica de dados binários recorrendo quer à ACoP das matrizes de semelhanças obtidas por diversos coeficientes para este tipo de dados, quer à "análise de inércia geral" dessas matrizes de semelhanças.

Recentemente, Van Eck e Waltman (2006) propõem um método para visualizar semelhanças entre objectos – VOS (*Visualization Of Similarities*) – cujo objectivo é o de fornecer uma visualização num espaço euclidiano de baixa dimensão, em que os objectos estão situados de tal forma que a distância entre qualquer par de objectos reflecte, tanto quanto possível, a sua semelhança com exactidão, recorrendo à ideia de visualizar as semelhanças localizando os objectos próximo das suas coordenadas ideais. Os autores comparam este método a um outro análogo proposto por Davidson *et al*. (1998), que tem a vantagem de "permitir que o problema de optimização seja resolvido como um problema de valores próprios", mas com a desvantagem das visualizações não serem tão boas (como as obtidas com o VOS). O método VOS também foi comparado ao MDS sem ponderações, quer analiticamente, quer sobre dados conhecidos, tendo sido obtidos resultados diferentes – enquanto o MDS visualiza esses objectos com um aspecto de ferradura, o VOS não. Sob certas condições, Van Eck e Waltman mostram que o VOS é equivalente ao *mapping* de Sammon (1969).

1.4.4 Representação euclidiana da estrutura de dados (H,D)

O problema da representação euclidiana da estrutura de dados (H,D) pode-se enunciar da seguinte maneira:

- Sendo dada uma matriz de dissemelhanças D, pretende-se encontrar uma matriz euclidiana de distâncias $\hat{D}_r$ com dimensão r, e tal que $\left\| D - \hat{D}_r \right\|^2$ seja mínimo.

Não existe solução para este problema (Le Calvé, 1976b). Le Calvé mostra que não existe solução, apresentando também um exemplo simples, e afirma que:

- "Seja D uma matriz de distâncias e S a matriz de "produtos escalares" gerada por D. Então a melhor aproximação $\hat{S}$ de ordem r de S não gera a melhor aproximação $\hat{D}_r$ de ordem r de D, mesmo no caso em que S é definida positiva."

Esta afirmação é reforçada:

- " ... em vez de dizer que, "é falso que a análise factorial dê sempre uma solução ...", seria mais exacto dizer que "a análise factorial nunca dá uma solução óptima para o problema das distâncias", e isto quer os valores próprios sejam positivos ou negativos. .... Se a solução da melhor aproximação dos produtos escalares, no sentido dos mínimos quadrados, é a projecção, isto é falso para a distância"[96].

A representação euclidiana da estrutura de dados (H,D) é pois o caso da análise de uma matriz de distâncias ou de dissemelhanças - Análise Métrica[97] (AM) (Le Calvé, 1976b) -, habitualmente designada por *Multidimensional Scaling* (MDS) (e.g., Mardia *et al*., 1979; Saporta, 1990; Caillez e Kuntz, 1996; Borg e Groenen, 2005). Segundo Caillez e Kuntz (1996), "o *multidimensional scaling* representa as relações entre n dados (indivíduos, estímulos, conceitos, ...) em gráficos que se percebam facilmente, restringindo-se ao caso simples em que estas relações são quantificadas por n(n-1)/2 medidas de semelhança ou de dissemelhança entre os pares de dados. O processo de representação mais familiar consiste em mergulhar estas medidas (de semelhança ou de dissemelhança) num espaço euclidiano"[98]. O objectivo do *multidimensional scaling* é o mesmo da ACP mas a matriz de dados de partida é diferente – é uma matriz de dissemelhanças ou de distâncias entre os indivíduos (ou entre as variáveis).

Le Calvé (apresentação pessoal) apresenta de forma simples a procura da solução para o problema enunciado:

- Como não se sabe resolver aquele problema, enquanto que o da aproximação dos produtos escalares é simples, vários autores (e.g., Torgerson, 1958; Gower, 1966; Escoufier *et al*., 1978; Cailliez e Pagès, 1976) pensaram em substituir o problema enunciado pelo segundo problema, uma vez que, sendo dada uma matriz de dissemelhanças D, pode-se-lhe associar de diversas maneiras, uma matriz de semelhanças tal que:

$$d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$$

---

[96] Tradução livre de *«…*au lieu, «il est faux que l'analyse factorielle fournisse toujours une solution …», il serait plus exact de dire «l'analyse factorielle ne fournit jamais une solution optimal au problème des distances», et ceci que les valeurs propres soient positives ou négatives. … Si la solution de la meilleure approximation, au sens des moindres carrés, des produits scalaires, est la projection, ceci est faux pour la distance.» (Le Calvé,1976b).

[97] *Analyse métrique*, no original.

[98] Tradução livre de "Multidimensional Scaling represents relationships between n data (individual, stimuli, concepts, ...) on graphical displays wich are easily understandable. We restrict ourselves to the simple case where these relationships are quantified by n(n-1)/2 measures of pairwise similarities or dissimilarities among the data. The most familiar representation procedure consists of embedding these measures in a Euclidean space*"* (Caillez e Kuntz, 1996).

Recorrendo à fórmula de Torgerson (expressão 1.3.20, Subsecção 1.3.4.2), escolheu-se $S = \mathrm{W}^G(D^2)$, sendo $\mathrm{w}^G\left(D^2\right)_{ij} = \frac{1}{2}\left(d_{i.}^2 + d_{.j}^2 - d_{..}^2 - d_{ij}^2\right)$.

- O problema inicial é substituído pelo seguinte problema:
    - Sendo dada uma matriz de dissemelhanças D, pretende-se encontrar uma matriz de distâncias euclidianas $\hat{\mathrm{D}}_r$, de dimensão r, tal que $\left\|\mathrm{W}^G(D^2) - \mathrm{W}^G(\hat{D}_r^2)\right\|^2$ seja mínimo.

  É o que a Análise Factorial de uma Matriz de Distâncias[99] (AFMD) (e.g., Saporta, 1990) e a Análise em Coordenadas Principais[100] (ACoP) de Gower (1966) fazem – substituíram o problema inicial por este.
- A solução é dada pela diagonalização da matriz $\mathrm{W}^G(D^2)$. Sabe-se que os vectores próprios de $\frac{1}{\mathrm{n}}\mathrm{W}^G(D^2)$ são as componentes principais da nuvem dos n pontos. O melhor ajustamento a um espaço de representação de dimensão r (r<n-1) é pois dado pelos r vectores próprios de $\mathrm{W}^G(D^2)$ correspondentes aos r maiores valores próprios. Se D for euclidiana a matriz $\mathrm{W}^G(D^2)$ será s.d.p. de ordem (n-1).

  A dimensão do espaço de representação é igual à ordem de $\mathrm{W}^G(D^2)$, sendo a (ordem de $\mathrm{W}^G(D^2)$)< (n-1). Verifica-se que D é euclidiana se $\mathrm{W}^G(D^2)$ for s.d.p..

  No caso de $\mathrm{W}^G(D^2)$ ter valores próprios negativos, D não é euclidiana. Neste caso é habitual (nas duas análises, AFMD e ACoP) restringir-se à projecção sobre os vectores próprios correspondentes aos valores próprios positivos.

Mas, se a matriz $\mathrm{W}^G(D^2)$ tiver muitos valores próprios negativos – situação frequente na prática, pois é muito instável – pode-se recorrer a outros métodos de *multidimensional scaling* como, por exemplo, o *nonmetric multidimensional scaling*[101] (e.g., Legendre e Legendre, 2000; Everitt e Rabe-Hesketh, 1997). Outra das opções possível corresponde a deformar os dados, aplicando uma transformação monótona sobre as distâncias para as tornar euclidianas (e.g., Bénasséni *et al*., 2007; Beninel, 1999; Benayde e Beninel, 2002) – é o que se apresenta no Capítulo 2.

---

[99] *Analyse Factorielle sur Tableaux de Distances* (*AFTD*), no original.
[100] Também é designada por *classical multidimensional scaling* (e.g., Everitt e Rabe-Hesketh, 1997) ou por *metric multidimensional scaling*.
[101] *Non-metric scale* (e.g., Everitt e Rabe-Hesketh, 1997).

Na realidade, se a matriz $W^G(D^2)$ tiver muitos valores próprios negativos é porque a solução é má. E se a solução é má, todas as soluções euclidianas são más, o que faz pensar que não se deve escolher o espaço de representação euclidiano. Le Calvé (1976b) alerta precisamente para esta situação:

> "Todas as dificuldades resultam da obstinação em representar de forma euclidiana alguma coisa que o não é! Em vez de "violar" os dados para os fazer entrar no nosso modelo, seria muito mais simples representá-los no seu espaço natural, mesmo se não for euclidiano!"[102]

Algumas metodologias de MDS estão implementadas em diversos *softwares* estatísticos, entre eles o *SPSS*.

Mais recentemente, por exemplo, Greenacre (2005) propõe a "Weighted Metric Multidimensional Scaling"[103] (WMDS) que, a partir de qualquer medida de distância calculada entre os indivíduos de uma matriz de dados rectangular indivíduos-por-variáveis, permite visualizar os indivíduos e as variáveis utilizando *biplots*.

Em Gorban *et al*. (2008), de forma mais abrangente, apresentam-se vários métodos para visualizar os dados em espaços de dimensão reduzida dos quais fazem parte a ACP e o MDS, entre outros, com aplicações a várias áreas e particularmente à biologia.

Le Calvé e os seus colaboradores (e.g., Beninel, 1987, 1999; Bénasséni *et al*., 2007) têm dedicado uma parte importante da sua investigação a este problema da representação das distâncias (quer euclidianas, quer não euclidianas e em particular as *city block*), entre outros (e.g., Le Calvé, 2000). Os seus trabalhos são uma referência nesta área!

---

[102] Tradução livre de "Toutes les difficultés proviennent de ce que l'on s'obstine à représenter de façon euclidienne quelque chose qui ne l'est pas! Au lieu de "violer" les données pour les faire entrer dans notre modèle, il serait beaucoup plus simple de les représenter dans leur espace naturel, même s'il n'est pas euclidien!" (Le Calvé, 1976b).

[103] Conferência convidada da *International Conference of Classification and Data Analysis Group of Italian Statistical Society, University of Bologna, Italy*.

2 TRANSFORMAÇÕES EUCLIDIANAS DE DISSEMELHANÇAS

> *Lorsque le système des distances analysées n'est pas euclidien, […] d'importantes anomalies peuvent se présenter.*
> (Benzécri, 1967)

2.1 Introdução

O *Multidimensional Scaling* (MDS) propõe-se representar graficamente os dados de dissemelhança num espaço euclidiano com poucas dimensões, quer estes sejam euclidianos ou não. Este é o caso, por exemplo, de cartas geográficas que representam os tempos de percurso (Love e Morris, 1972), das clássicas matrizes de distâncias por estrada entre cidades (Everitt e Rabe-Hesketh, 1997) ou de distâncias genéticas já referidas na Secção 1.2 do Capítulo 1.

> *If d values are directly collected during an experimental process, the chance of d being Euclidean or even metric is very low.*
> (Caillez e Kuntz, 1996)

Se os dados de dissemelhança não são euclidianos, não existe imagem num espaço euclidiano. Relembremos que este tipo de problema de representação pode ser abordado sob dois pontos de vista, o do matemático e o do utilizador. O matemático procura geralmente a solução óptima tecnicamente, mesmo que seja uma imagem pouco legível por ter muitas dimensões. Regra geral, o utilizador prefere um conjunto de soluções claras e "aceitáveis", entre as quais ele poderá escolher, pois para ele vale mais uma "boa" imagem aproximada do que uma "má" imagem exacta. Quando se faz a abordagem a este tipo de problemas convém que se esclareça qual dos pontos de vista é utilizado.

No caso dos dados não serem euclidianos, existem vários modos de obter essa representação. Dois desses modos, geralmente mais utilizados são:

- Através da transformação dos dados para os tornar euclidianos – Consiste em substituir a dissemelhança de partida, d, por uma dissemelhança, δ=f(d), função de d que seja euclidiana e que não difira ‘‘muito’’ da dissemelhança inicial. Neste caso, é habitual utilizarem-se transformações monótonas, que têm a vantagem de conservar as preordenações, como por exemplo algumas das transformações simples que se pensa estarem mais adaptadas ao problema (função logarítmica, no caso dos fenómenos multiplicativos ou para separar os valores pequenos; função arcsen, no caso das distâncias angulares; função afim, no caso das mudanças de escala), assim como, as transformações pela constante aditiva, entre outras. No entanto, também se podem utilizar transformações não monótonas. Sobre este assunto encontramos, por exemplo, em Critchley (1986) coligidos resultados bem conhecidos nesta área. Naturalmente, surge-nos a pergunta “transformar para quê?”. Estas transformações surgem, pois, para responder à questão de encontrar uma representação euclidiana, independentemente da sua dimensão. Situamo-nos na perspectiva do matemático, i.e., procura-se uma solução óptima. Esta é a opção de vários autores de MDS.

- Procurar uma distância δ que minimize um certo critério de aproximação – A procura de δ, tal que o critério de aproximação C(δ,d) seja mínimo, constitui um problema difícil de resolver devido ao aspecto não linear das distâncias euclidianas. Esta procura constitui o objectivo, ou um dos objectivos, dependendo dos autores, do *Multidimensional Scaling*, tal como nos é apresentado, em geral, na literatura. Situamo-nos na perspectiva do utilizador.

Neste capítulo limitar-nos-emos ao primeiro modo, ou seja, a estudar algumas transformações monótonas que permitem obter uma distância euclidiana a partir de uma distância ou dissemelhança não euclidiana (Secção 2.2). Apresenta-se finalmente, na Secção 2.3, uma metodologia MDS, i.e. “Normed MultiDimensional Scaling” (e.g., Beninel, 1987, 1999), que utiliza naturalmente algumas destas transformações.

2.2 Transformações monótonas da distância

L'avantage des transformations monotones est de conserver l'ordre des distances.
(Le Calvé, apresentação pessoal)

O interesse das transformações monótonas de uma distância d reside no facto de elas respeitarem as preordens iniciais, modificando assim o menos possível as informações de d. As grandes dissemelhanças continuam a ser grandes e as pequenas dissemelhanças continuam a ser pequenas depois da transformação, sendo todas mais pequenas do que as iniciais. Este é o caso, por exemplo, das dissemelhanças sensoriais em que os indivíduos dão a conhecer os seus gostos; neste caso, o que interessa é a ordem das dissemelhanças e não o seu valor exacto.

Segundo diversos autores (por exemplo, Shepard, 1962a, 1962b; Benzécri, 1967), se uma distância não é euclidiana podemos construir a partir dela uma distância euclidiana que respeita as preordenações. Na literatura encontramos várias transformações monótonas para uma distância, como já referimos. Vai-nos interessar estudar mais em pormenor, por serem as mais usadas, as transformações com constantes aditivas, a transformação pela função potência e a transformação pela adição de distâncias[104].

2.2.1 Transformações com constantes aditivas

Como se sabe que,

> "Se d não é uma distância, existe uma constante c, tal que: $\delta_{ij} = d_{ij} + c$ e $\delta_{ii} = d_{ii} = 0$ é uma distância, sendo $c = \underset{i,j,k}{Max} \left| d_{ij} + d_{jk} - d_{ik} \right|$",

surgiu naturalmente, a pergunta: "Pode-se adicionar uma constante à dissemelhança d não euclidiana para a tornar euclidiana?" (Torgerson, 1952). Ou, por outras palavras, pretende-se encontrar a solução do "problema da constante aditiva" (e.g., Cailliez e Pagès, 1976):

> Sendo d uma dissemelhança num conjunto finito I com n elementos, pretende-se encontrar o menor número c* tal que a dissemelhança δ_c, definida por:

[104] *L'ajout de distances*, em francês.

$\delta_c(i,j) = \begin{cases} d_{ij} + c & \text{, se i} \neq \text{j} \\ 0 & \text{, se i = j} \end{cases}$ tenha uma representação euclidiana para todos os $c \geq c^*$.

A solução inicial proposta por Torgerson (1952), não resolve este problema.

No entanto, o problema da constante aditiva sobre d^2 é resolvido, em primeiro lugar. Lingoes (1971) resolve este problema, já posto inicialmente por Guttman, adicionando uma constante c ao quadrado da dissemelhança d, d^2, como podemos ver pelo teorema e corolário apresentados em seguida.

Teorema 2.2.1. (Lingoes, 1971)

Seja d uma dissemelhança não euclidiana e $W^G(D^2)$ a sua matriz de Torgerson[105]. Seja λ_n o seu menor valor próprio, $\inf_{i \in I} \lambda_i = \lambda_n$.

A solução do problema, encontrar a menor constante c ($c \in \mathbb{R}^+$) tal que a dissemelhança δ_c, definida por

$\delta_c^2(i,j) = \begin{cases} d_{ij}^2 + c & \text{, se i} \neq \text{j} \\ 0 & \text{, se i = j} \end{cases}$, seja euclidiana é dada por $c_0 = 2|\lambda_n|$.

Corolário 2.2.1. (Lingoes, 1971)

Sejam d uma dissemelhança e c_0 a constante precedente. Então, $\forall c > c_0$ a dissemelhança δ_C definida por $\delta_c^2(i,j) = d_{ij}^2 + c$, $\delta_c^2(i,i) = 0$ é euclidiana de ordem plena e por isso pode-se inscrever numa esfera de $\mathbb{R}^{n-1}$.

O sucesso desta transformação tem a ver com a simplicidade de cálculo e com o facto de ela não modificar os vectores próprios de $W^G(D^2)$, pois os vectores próprios das matrizes $W^G(D^2)$ e $W^G(\delta^2)$ são idênticos, sendo apenas modificado o peso dos eixos. Os valores próprios α_i de $W^G(D^2)$ são $\alpha_i = \sqrt{\lambda_i + |\lambda_n|}$, i= 1 ou 2, (Beninel, 1987). No entanto, não se percebe bem a que tipo de transformação geométrica ou a que deformação esta transformação faz referência.

Finalmente, Cailliez (1983) propõe os teoremas que apresentamos de seguida, para darem resposta ao problema da constante aditiva.

---

[105] A matriz de Torgerson está definida no Capítulo 1, Expressão (1.3.20).

Teorema 2.2.2. (Cailliez, 1983)

Seja δ_C a dissemelhança definida por

$$\delta_c(i,j) = \begin{cases} d_{ij} + c & \text{, se i} \neq \text{j} \\ 0 & \text{,se i = j} \end{cases}$$

Então existe uma constante c* tal que:

- para $\forall c \geq c^*$, a dissemelhança δ_C tem uma representação euclidiana.

- para $c = c^*$, a representação euclidiana de δ_{c^*} define um espaço com, no máximo, (n-2) dimensões.

Sendo o valor de c* obtido no teorema seguinte:

Teorema 2.2.3. (Cailliez, 1983)

A constante aditiva c* é o maior valor próprio da matriz quadrada 2nx 2n, B:

$$B = \begin{pmatrix} 0 & 2W_d \\ -I & -4W_{d^{1/2}} \end{pmatrix} \quad (2.2.1)$$

sendo: as matrizes $W_{d^{1/2}} = -\frac{1}{2}ADA$, $W_d = -\frac{1}{2}AD^2A$, $D^2 = \left(d_{ij}^2\right)$ e $A = I - \frac{1}{n}11'$ (I designa a matriz identidade e **1** o vector cujas coordenadas são todas iguais à unidade).

As matrizes $W_{d^{1/2}}$ e W_d, são as matrizes de Torgerson: $W_{d^{1/2}} = W^G(D)$ e $W_d = W^G\left(D^2\right)$ (Beninel, 1987).

"Analisando a sua demonstração, Cailliez também faz notar que a matriz $W^G(\delta_c^2)$ é igualmente a matriz de Torgerson associada à transformação $d_{ij} \rightarrow \left|d_{ij} - c\right|$" (Le Calvé, apresentação pessoal).

De forma análoga ao Teorema 2.2.3, Cailliez deduz o seguinte teorema:

Teorema 2.2.4. (Cailliez, 1983)

Se d não tem representação euclidiana existe uma constante negativa c**, tal que a distância δ_C definida por

$$\delta_c(i,j) = \begin{cases} \left|d_{ij} + c\right| & \text{, se i} \neq \text{j} \\ 0 & \text{, se i = j} \end{cases}$$

é euclidiana, de dimensão máxima (n-2), $\forall c \leq c^{**}$. A constante c^{**} é o menor valor próprio da matriz B (definida em 2.2.1).

Esta transformação, embora seja pouco utilizada na prática, tem uma interpretação geométrica interessante no caso em que $|c^{**}| \leq d_{ij}$, $\forall i \neq j$. Le Calvé mostra que, neste caso, a dissemelhança d_{ij} é a distância entre as superfícies das esferas de raio c^{**}, desenhadas à volta dos pontos M_i, representativos da distância $\delta_{C^{**}}$.

De forma equivalente pode-se pois concluir que (Cailliez, 1983):

- Para mergulhar uma dissemelhança d, definida num conjunto finito, num espaço euclidiano, uma das práticas mais comum, devido à simplicidade computacional, consiste em transformar d em ε definida por:

$$\varepsilon_{ij} = \begin{cases} \left(d_{ij}^2 + k\right)^{\frac{1}{2}} & \text{, se i} \neq \text{j} \\ 0 & \text{, se i = j} \end{cases}$$

  O menor n.º k* tal que ε tem uma representação euclidiana, para $\forall k \geq k^*$ é $(-2\lambda_n)$, em que λ_n é o menor valor próprio da matriz W_d.

"Quando se adiciona uma quantidade positiva e se faz uma deslocação paralela à diagonal, faz-se entrar d no cone das distâncias euclidianas." (e.g., Perrier, 1998).

Das duas transformações pela constante aditiva apresentadas, na prática utiliza-se preferencialmente a transformação pela constante aditiva associada a d^2 (Teorema 2.2.1, Lingoes, 1971) devido à dimensão da matriz B e por ela não ser simétrica (Beninel, 1987).

No caso particular dos índices de dissemelhança para dados dicotómicos, Fichet e Le Calvé (1984) estudam o impacto das transformações com as constantes aditivas sobre as famílias de índices d_θ (família que engloba, consoante os valores de θ, os índices de Jaccard, de Czekanowski e Dice e de Sokal e Sneath e Anderberg – Tabela 1.3.5) e índices $\sqrt{d_\theta}$.

Sobre as transformações com as constantes aditivas, encontram-se vantagens e desvantagens, que se passam a enunciar.

Vantagens:

- Tal como tem sido referido, estas técnicas preservam a preordenação induzida pela dissemelhança d.
- Estas transformações escrevem-se sob uma forma matemática simples e conduzem a uma solução analítica que se baseia num problema de valores próprios.

Desvantagens:

- As constantes aditivas introduzem, frequentemente, uma deformação importante nos valores de dissemelhança iniciais. Acontece com frequência que o valor da constante é muito grande em relação aos valores de d_{ij}, de tal maneira que a distorção é grande (Joly e Le Calvé, 1994). Benasséni (1994), utilizando outros métodos, exemplifica esta situação, com uma distância entre oito pontos, em que os valores destas variam entre 10 e 60, para a qual a constante de Lingoes é 507, enquanto que a matriz de Torgerson tem apenas dois valores próprios negativos.
- Mesmo que a constante seja menos importante do que a apresentada no exemplo, ela introduzirá uma deformação forte no caso das distâncias serem pequenas e uma deformação fraca no caso de elas serem grandes. Ou seja, adicionar a mesma constante a todas as dissemelhanças pode afectar de forma expressiva o padrão inicial dos valores, uma vez que as variações relativas das pequenas dissemelhanças são mais importantes do que as das maiores.
- Mardia (1978) afirma que a configuração obtida, quando se utiliza a técnica da constante aditiva sobre D, conduz geralmente a uma aproximação mais pobre do que a configuração que se obtém quando se utilizam apenas os valores próprios positivos (Everitt e Rabe-Hesketh, 1997).
- No caso dos índices de dissemelhança para variáveis dicotómicas, a utilização destas transformações torna-se delicada, pois põe problemas, devido à modificação da forma analítica do índice inicial (Fichet e Le Calvé, 1984).

Com o objectivo de ultrapassar as desvantagens referidas, Bénasséni, Bennani Dosse e Joly (2007) apresentam uma generalização da transformação pela constante aditiva, considerando uma classe de transformações introduzida por Critchley (1986), que se baseia em adicionar o quadrado de uma medida de dissemelhança, δ, que tem uma representação euclidiana com dimensão n-1, à dissemelhança inicial, d, de forma a obter $\tilde{d}$: $\tilde{d}_{ij} = \left(d_{ij}^2 + c\delta_{ij}^2\right)^{1/2}$, c >0. A existência de um valor $c \in \mathbb{R}^+$, tal que a distância $\tilde{d}$ seja euclidiana deve-se ao facto do conjunto dos quadrados de distâncias euclidianas ser um cone convexo cujo interior é formado por quadrados de distâncias euclidianas com ordem

maximal. Bénasséni *et al*. (2007) dão, para a distância euclidiana δ específica, o valor mínimo da constante c. São escolhidas para δ, entre outras, distâncias apresentadas nas subsecções seguintes (e.g., transformação pela função potência).

2.2.2 Transformação pela função potência

Sendo D uma matriz de dissemelhanças, considera-se que a dissemelhança D^{α} (geralmente, $0 \le \alpha \le 1$) é uma potência termo a termo da matriz D: $D^{\alpha} = \left(d_{ij}^{\alpha}\right)$. Naturalmente, surgem as perguntas: Sendo D uma distância, a potência da matriz de distâncias, D^{α}, também é uma distância? Como escolher α de forma a que D^{α} seja uma distância? Se D é uma distância euclidiana, D^{α} também será uma distância euclidiana? E, se D não é uma distância euclidiana, D^{α} será uma distância euclidiana?

Encontramos as respostas a estas perguntas e a muitas outras sobretudo em Joly e Le Calvé (1986), Fichet e Le Calvé (1984), Brossier e Le Calvé (1986) e Joly e Le Calvé (1994), e por exemplo, em Gower e Legendre (1986), Caillez e Kuntz (1996) e Gordon (1999).

Joly e Le Calvé (1986), estudam as propriedades das potências de uma distância, recordam resultados antigos e complementam-nos com demonstrações e proposições novas obtidas a partir de uma generalização do Lema de Schur, das quais se destaca a demonstração, mais simples, do resultado de Schoenberg sobre a potência de uma distância euclidiana. Daqui se destacam os seguintes resultados:

Teorema 2.2.5. (Schoenberg, 1937)

Se $D = \left(d_{ij}\right)$ é uma distância, então $\forall 0 \le \alpha \le 1,\ D^{\alpha} = \left(d_{ij}^{\alpha}\right)$ é uma distância.

Como corolário deste resultado, Joly e Le Calvé (1986) apresentam a noção de índice de distância:

Corolário 2.2.2. Índice de distância (Joly e Le Calvé, 1986)

Seja $D = \left(d_{ij}\right)$ uma dissemelhança sobre I.

Para toda a dissemelhança D existe um número real não negativo p, p ≥0, que depende de D, designado por "índice de distância", tal que $D^{\alpha} = \left(d_{ij}^{\alpha}\right)$ é uma distância para $\alpha \in$ [0, p], e não o é para α > p.

A partir deste resultado, fazem notar que:

- Uma dissemelhança D é uma distância se e só se o seu índice de distância é superior ou igual a 1.
- Uma dissemelhança D é uma distância ultramétrica se e só se o seu índice de distância é infinito (Brossier e Le Calvé, 1986).

A resposta à questão, "Como escolher α de forma a que D^{α} seja uma distância?", é-nos dada posteriormente, pelo seguinte teorema:

Teorema 2.2.6. (Joly e Le Calvé, 1994)

a) O conjunto de todos os α's tais que D^{α} é uma distância, é um conjunto fechado.

b) Se D é uma dissemelhança semidefinida, façamos

$$k = \sup_{i,j} d_{ij} \qquad q = \inf_{i,j} \left\{ d_{ij} : d_{ij} \neq 0 \right\} \qquad \alpha = \frac{\ln 2}{\ln k - \ln q}$$

então α pertence ao conjunto referido na alínea a).

c) Se D não é semidefinida, D^{α} é uma distância se e só se α=0.

Vemos assim que o índice de distância p admite um minorante, $p \geq \frac{\ln 2}{\ln k - \ln q}$.

Perrier (1998) desenvolveu, no programa WinABCD, uma função de cálculo numérico do índice de distância de uma matriz de dissemelhança, que permite obter um estimação numérica do índice de distância de uma matriz de dissemelhança dada.

O que se passa então se as distâncias forem euclidianas?

O resultado fundamental de Schoenberg (1937) é relembrado e apresentado com uma nova demonstração simplificada, baseada no Lema de Schur generalizado, por Joly e Le Calvé (1986):

Teorema 2.2.7. (Schoenberg, 1937)

Seja D uma distância euclidiana.

Então D^{α} é uma distância euclidiana, $0 \leq \forall \alpha \leq 1$, e de dimensão máxima[106], $\forall \alpha$<1.

[106] Se D é uma euclidiana a sua dimensão é igual à dimensão do espaço afim gerado pelos pontos de qualquer uma das suas imagens euclidianas. Para n pontos, a dimensão máxima é n-1 (Joly e Le Calvé, 1986).

Como caso particular deste teorema, resulta que:

- A raiz quadrada de uma distância euclidiana é sempre euclidiana (Schoenberg, 1937)[107].

Considerando ainda $\alpha = \frac{1}{2}$, Brossier e Le Calvé (1986) fazem notar que: "Se D admite uma imagem euclidiana de dimensão 1, então $\sqrt{D}$ admite uma imagem euclidiana de dimensão n-1."

Finalmente, vejamos qual é a resposta à questão: "Se D não é uma distância euclidiana, a potência D^{α} será uma distância euclidiana?"

Quando $\alpha = \frac{1}{2}$, i.e. para $\sqrt{D}$, surgem os resultados já conhecidos:

- Sobre quatro pontos a raiz quadrada de uma distância é sempre euclidiana (Propriedade dos quatro pontos – Blumenthal, 1936).
- Sobre mais de quatro pontos a raiz quadrada de uma distância nem sempre é uma distância euclidiana (Fichet e Le Calvé, não publicado).
- É falsa a conjectura "A raiz quadrada de uma distância é sempre euclidiana" (Joly e Le Calvé, 1986).
- A raiz quadrada de uma ultramétrica é euclidiana (Joly e Le Calvé, 1986).
- Seja D uma matriz nxn que verifica a desigualdade quadrangular. Então $\sqrt{D}$ é euclidiana e com dimensão exacta (n-1)" (Le Calvé, 1985).
- A raiz quadrada de uma distância ao centro é euclidiana (Le Calvé, 1985).
- Embora os índices de dissemelhança para dados dicotómicos não sejam distâncias euclidianas (muitos deles nem sequer são distâncias), alguns deles admitem potências de expoente ½ que são euclidianas (Tabela 2.2.1):
    - As raízes quadradas das dissemelhanças que estão associadas aos índices de semelhança, para variáveis dicotómicas, de Russel e Rao, de Kendall, Sokal e Michener, de Jaccard, de Czenakowski e Dice, de Sokal e Sneath e Anderberg, de Rogers e Tanimoto e de Hamman, através da transformação $d_{ij} = 1 - s_{ij}$ (D = 1 – S) são euclidianas e de dimensão máxima, (Fichet e Le Calvé, 1984).
    - O índice de Ochiai, que não é uma distância, é o quadrado de uma distância euclidiana (Fichet e Le Calvé, 1984; Gower e Legendre, 1986).

[107] Demonstração em Brossier e Le Calvé (1986) tirada de Joly e Le Calvé (1985).

Tabela 2.2.1. Coeficientes de semelhança para dados binários propostos por diversos autores, em que *a, b, c* e *d* referem-se às quatro células da tabela de contingência 2x2, designando respectivamente, o número de co-presenças, presenças/ausências, ausências/presenças e co-ausências dos atributos. Apresentam-se entre parêntesis as designações introduzidas por Hubálek (1982). Caracterização das respectivas matrizes de semelhanças S (d.p.-matriz definida positiva; s.d.p.-matriz semidefinida positiva) e natureza das dissemelhanças respectivas

Coeficiente	Semelhança s_{ij}	$\sqrt{S}$	Dissemelhança D $d_{ij}=1-s_{ij}$	Natureza de D	Natureza de $\sqrt{D}$
Russel e Rao (1940)	$\frac{a}{n}$ (A_{14}) **S é s.d.p.**[2]	d.p.[2]	$\frac{n-a}{n}$ ou (b+c+d)/n	Distância *city block* [1]	Distância euclidiana[2,3,4]
Sokal e Michener (1958) "simple matching coefficient" Kendall[3,4]	$\frac{a+d}{n}$ (A_{20})	d.p.[2]	$\frac{b+c}{n}$	*É* distância[3]	Distância euclidiana[2,3]
Jaccard (1901) Sneath (1957)	$\frac{a}{a+b+c}$ (A_4)	d.p.[2]	$\frac{b+c}{a+b+c}$	Distância *city block* [1]	Distância euclidiana[2,3,4]
Czenakowski e Dice, Dice (1945) Sørensen (1948)	$\frac{2a}{2a+b+c}$ (A_5) **S é s.d.p.**[6]	d.p.[2]	$\frac{b+c}{2a+b+c}$	Não é distância	Distância euclidiana[2,3,4,5,6]
Sokal e Sneath (1963) "un_2" Anderberg [3,4]	$\frac{a}{a+2(b+c)}$ (A_6) **S é s.d.p.**[6]	d.p.[2]	$\frac{2(b+c)}{a+2(b+c)}$ ou (b + c) / [$\frac{1}{2}$ a + (b + c)]	Distância *city block* [1]	Distância euclidiana[2,3,4,6]
Rogers e Tanimoto (1960)	$\frac{a+d}{a+2(b+c)+d}$ (A_{23}) **S é s.d.p.** [2, 6]	d.p.[2]	$\frac{2(b+c)}{a+2(b+c)+d}$ ou [2(b + c)] / [n + (b + c)]	Distância *city block* [1]	Distância euclidiana[2,3,4,6]
Hamann (1961)	$\frac{a+d-b-c}{n}$ (A_{24}) **S é s.d.p.** [6]	d.p.[2]	$\frac{2(b+c)}{n}$		Distância euclidiana[2,3,6]
Ochiai (1957) Driver e Kroeber (1932)	$\frac{a}{\sqrt{(a+b)(a+c)}}$ (A_{11}) **S é s.d.p.**[2]		$1-\frac{a}{\sqrt{(a+b)(a+c)}}$	Não é distância[2] Distância *city block* [4]	Distância euclidiana[2,4,5,6]
Sokal e Sneath (1963) 'un_5'	$\frac{ad}{\sqrt{(a+b).(c+d).(a+c).(b+d)}}$ (A_{25}) **S é s.d.p.**[6]		1-A_{25}	Não é distância[6]	Distância euclidiana[6]

Coeficiente (cont.)	**Semelhança s_{ij}** (cont.)	$\sqrt{S}$ (cont.)	**Dissemelhança D $d_{ij}=1-s_{ij}$** (cont.)	**Natureza de D** (cont.)	**Natureza de** $\sqrt{D}$ (cont.)
Yule (1912), Pearson e Heron (1913)	$\frac{ad-bc}{\sqrt{(a+b).(c+d).(a+c).(b+d)}}$ não é definido se c=d=0 ou b=d=0 (A_{30}) **S é s.d.p.**[6]		1- A_{30}	Não é distância[6]	Distância euclidiana[6]
Braun-Blanquet (1932)	$\frac{a}{\max\{(a+b),(a+c)\}}$ (A_1)		$1-\frac{a}{\max\{(a+b),(a+c)\}}$		Distância euclidiana[5]

[1]Joly e Le Calvé (1992), [2]Fichet e Le Calvé (1984), [3]Brossier e Le Calvé (1986), [4]Joly e Le Calvé (1994), [5]Caillez e Kuntz (1996), [6]Gower e Legendre (1986).

De forma mais geral, para alguns índices de dissemelhança sobre variáveis binárias sabemos encontrar uma potência euclidiana. Por exemplo, para o índice [108]D_θ = 1- S_θ , é fácil verificar que:

- Para $\alpha \to 0$, D_θ^α é euclidiana para todo o $\theta \in \mathbb{R}$.
- Para $\alpha = \frac{1}{2}$, D_θ^α é euclidiana para todo o $\theta \geq 1$.
- Para $\alpha = 1$, D_θ^α não é euclidiana.
- Para $\frac{1}{2}\alpha \leq \alpha \leq 1$, D_θ^α é uma distância.

As primeiras três transformações têm a vantagem de preservar a preordenação de partida, mas o inconveniente de repartirem uniformemente a deformação (Beninel, 1987). Com a preocupação de não repartir uniformemente a deformação, Beninel aconselha perturbar D com a ajuda de uma distância ao centro.

Partindo daqueles resultados, Gower e Legendre (1986) propõem a conjectura gráfica apresentada na Figura 2.2.1. Esta conjectura consiste em dizer que:

- A região limitada pelo traço ponteado e o eixo das abcissas contém os pares (α, θ), tal que D_θ^α é euclidiana.
- A região limitada pelo traço a cheio e o eixo das abcissas corresponde aos pares (α, θ), tais que D_θ^α é uma distância.

[108] $S_\theta = (a+d) / [a+d+\theta(b+c)]$

- A região limitada pelos traços a cheio e a ponteado corresponde aos pares (α, θ), tais que D_θ^α é uma distância não euclidiana.

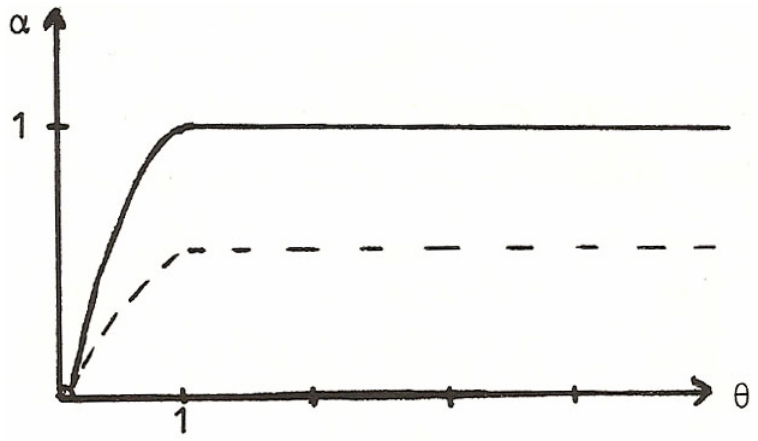

Figura 2.2.1. Conjectura gráfica sobre a natureza de D_θ^α, proposta por Gower e Legendre (1986).

Posteriormente, os resultados obtidos para os coeficientes de Ochiai e de Dice, por Fichet e Le Calvé (1984) e Gower e Legendre (1986), foram generalizados: "A raiz quadrada da dissemelhança d_α é euclidiana, se $\alpha \geq 0$" (Caillez e Kuntz, 1996). Esta proposição permite afirmar também que: A raiz quadrada do coeficiente de dissemelhança de Braun-Blanquet (1932) é euclidiana (Caillez e Kuntz, 1996).

Por outro lado, Joly e Le Calvé (1986) relembram que "toda a distância ultramétrica é euclidiana" (Holman, 1972), e afirmam que:

- "A distância ultramétrica pode-se caracterizar por ser a única distância para a qual todas as potências são euclidianas".

O Teorema 2.2.7 admite um corolário imediato, que permite dar uma resposta positiva à questão mais geral "Toda a distância admitirá uma potência que seja euclidiana?", introduzindo a definição de índice de Euclides:

Corolário 2.2.3. Índice de Euclides (Joly e Le Calvé, 1986)

Para toda a dissemelhança D = (d_{ij}) existe um número real positivo ***e***, $\boldsymbol{e} \in \mathbb{R}^+$, que se chama índice de Euclides, que depende de D, tal que:

- D^α é uma distância euclidiana, $\forall \alpha \leq \boldsymbol{e}$;
- D^α não é uma distância euclidiana, $\forall \alpha > \boldsymbol{e}$.

O espaço de representação é de ordem máxima (n-1) para $\forall \alpha < \boldsymbol{e}$.

Embora não se conheça uma solução analítica para o problema de como calcular o índice de Euclides, existem técnicas de aproximação. Perrier (1998) apresenta uma técnica, que está programada no programa WinABCD, que permite estimar o índice de Euclides através de bissecções sucessivas de um intervalo de procura até obter uma precisão de 10^{-3}. Sobre este assunto também podemos consultar, por exemplo, Camiz e Le Calvé (2001).
No entanto, pode-se dizer que o índice de Euclides é superior a 0.7/n, *e* >0.7/n, baseando-nos no seguinte teorema :

Teorema 2.2.8 (Deza e Maehara, 1990)

Se D é uma distância sobre n pontos, então D^{α} é uma distância euclidiana, $\forall 0 \le \alpha \le 0.72/n$.

A Tabela 2.2.2 permite-nos saber quais são os valores do índice de Euclides, ***e***, para algumas distâncias conhecidas (Joly e Le Calvé, 1986), de que destacamos:

- Para uma distância de árvore aditiva: ½ ≤ ***e*** <1.
- Para uma dissemelhança quase-hipermétrica: ½ ≤ ***e***.
- Para uma euclidiana de dimensão não máxima: ***e*** = 1.
- Para uma euclidiana de dimensão máxima: 1< ***e***.
- Para uma ultramétrica: ***e*** = ∞.

Encontramos as demonstrações em Joly e Le Calvé (1986) e em Beninel (1987).

Tabela 2.2.2. Valores dos índices de distância e de Euclides para algumas dissemelhanças conhecidas

Índice de distância P	Índice de Euclides e						
		$0 < e < ½$	$½ \leq e < 1$	$½ < e \leq 1$	$e = 1$	$e > 1$	$e = \infty$
	$0 < P < 1$	Dissemelhanças que não são distâncias					
	$0 < P \leq 1$			A distância gerada, D=1-S, não é euclidiana, mas a raiz quadrada é euclidiana (índices sobre variáveis dicotómicas)			
	$P = 1$						
	$P \geq 1$	Distâncias cujas raízes quadradas não são euclidianas	Distâncias que não são euclidianas, mas a raiz quadrada é euclidiana (dist. de árvores aditivas, dist. *city block*)		Distâncias euclidianas de dimensão não máxima		
	$P > 1$					Distância Euclidiana de dimensão máxima	
	$P = \infty$						Ponto das ultramétricas

2.2.3 Transformação pela adição de distâncias

O problema da transformação pela adição de distâncias pode-se formular da seguinte maneira: Seja D uma dissemelhança não euclidiana, pretendemos encontrar a função f, mínima, tal que a transformação aditiva $\delta_{ij} = d_{ij} + f(i, j)$ seja euclidiana.

Este problema não tem solução geral. No entanto, para alguns casos particulares encontramos soluções:

1 - Fazendo $f(i, j) = c$ (c é constante).

Sendo c constante, o problema corresponde ao da escolha da constante aditiva.

A adição de uma constante à dissemelhança D é pois, um caso particular daquela transformação aditiva.

2 - Uma outra escolha possível, corresponde a adicionar uma distância ao centro à dissemelhança D, de maneira a que ela se torne euclidiana.

Escolhe-se $f(i,j) = c_i + c_j$, $f(i,i) = 0$. A existência de tal distância ao centro é assegurada quando se faz $c_i = c^*$, sendo c^* a constante de Cailliez (Le Calvé, 1985).

Encontramos, entre outras, soluções dadas por Fichet (1983), por Beninel (1987) e por Beninel, Qannari e Qannari (1994):

- Para Fichet (1983), o vector $c_i + c_j = k(x_i + x_j)$, sendo o vector x conhecido, tem que se procurar o valor de k.
- Beninel (1987) conhece a solução deste problema, no caso de $W^n(D^2)$ admitir apenas um valor próprio negativo. Neste caso, o problema consiste em determinar $x \in \mathbb{R}^n_+$, com variância maximal, tal que a dissemelhança D definida por:

  $\forall (i,j) \in I \times I \quad \delta_C^2(i,j) = \begin{cases} d_{ij}^2 + x_i + x_j & \text{, se } i \neq j \\ 0 & \text{, se } i = j \end{cases}$ seja euclidiana.

  A restrição sobre a variância traduz a preocupação de perturbar sensivelmente apenas alguns elementos de I, não repartindo uniformemente a perturbação introduzida.
- Beninel *et al*. (1994) desenvolveram a matriz de Torgerson associada e dão condições que asseguram que esta matriz seja semidefinida positiva.

Sobre este assunto pode-se consultar também Camiz e Le Calvé (2001).

2.3 Aplicações

Estas transformações são utilizadas naturalmente no Posicionamento Multidimensional Normado[109], MDS Normada.

---

[109] *Positionemment Multidimensionnel Normé*, em francês. *Normed MultiDimensional Scaling*, em inglês.

Beninel (1987, 1999, 2002[110]) tem-se interessado na caracterização de medidas de dissemelhança sobre conjuntos finitos que admitam uma representação euclidiana esférica. Como consequência, propõe uma metodologia – *Normed MultiDimensional Scaling* – para determinar a representação euclidiana esférica de um conjunto de itens que restitua da melhor maneira as dissemelhanças entre os itens. Tendo em conta que "Sempre que D é euclidiana de ordem máxima (i.e., ordem de (D) = n-1) então (I, D) é do tipo esférico" (pois, toda a perturbação de uma dissemelhança d sobre I que conduza a uma dissemelhança euclidiana de ordem máxima é uma perturbação esférica) e os resultados, respectivamente, de Lingoes (1971, Teorema 2.1.1), de Cailliez (1983, Teorema 2.1.2) e de Joly e Le Calvé (1986,
Corolário 2.2.3), pode-se demonstrar que as famílias de transformações pelas constantes aditivas sobre D^2 e sobre D, assim como, as famílias de transformações pela função potência são esféricas (Beninel, 1999).
No caso de alguns índices de dissemelhança para variáveis binárias, Beninel (1987, 1999) aproveita o facto de $\sqrt{D}$ ser euclidiana (no caso de: D= 1-S e S s.d.p.), sendo por isso dissemelhanças esféricas, para os representar graficamente utilizando o Posicionamento Multidimensional Normado.

Neste âmbito, mais recentemente, encontramos aplicações das transformações, por exemplo, em Bénasséni *et al*. (2007) a matrizes de dissemelhanças estudadas por outros autores.

[110] Em co-autoria com M. Benayade.

3 O COEFICIENTE DE AFINIDADE E SUAS GENERALIZAÇÕES

3.1 Introdução

O conceito de afinidade entre duas distribuições de probabilidade discretas foi introduzido por Matusita (1951, 1955), no âmbito da estatística clássica, da seguinte forma:

- Se P=$\{p_i\}$ e Q=$\{q_i\}$, i=1,...k, são duas distribuições de probabilidade discretas sobre um mesmo suporte $\{1, 2, ..., k\}$, a afinidade entre P e Q é dada por:

$$a(P,Q) = \sum_{i=1}^{k} \sqrt{p_i q_i} \qquad (3.1.1)$$

Pode-se provar que a afinidade toma valores no intervalo [0,1].

O coeficiente de afinidade está associado à distância de Hellinger, $d(P,Q) = \left[\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2 \right]^{1/2}$, pela função quadrática apresentada na Tabela 1.3.2 (Subsecção 1.3.4): $d(P,Q) = \sqrt{2(1 - a(P,Q))}$ (e.g., Bacelar-Nicolau, 1980; Bacelar-Nicolau, 2000). A distância de Hellinger é a base da análise factorial esférica proposta por Domenges e Volle (1979) (e.g., Bacelar-Nicolau, 1980; Bacelar-Nicolau, 2000).

O coeficiente de afinidade é definido, de forma análoga, no caso das distribuições serem contínuas.

Matusita também utiliza o coeficiente de afinidade em Análise Classificatória (e.g., como critério de comparação entre classificações, em Matusita (1977)).

Constata-se que o coeficiente de afinidade (Expressão 3.1.1) coincide com o coseno da medida de divergência de Bhattacharyya (1943), Δ, uma vez que Bhattacharyya (1943) a definiu como sendo o ângulo entre as duas distribuições de probabilidade discretas P e Q, tal que: $\cos \Delta = \sum_{i=1}^{k} \sqrt{p_i q_i}$. Esta medida de divergência Δ também é definida, de forma análoga, no caso das distribuições de probabilidade serem contínuas.

Há autores que designam por coeficiente de Bhattacharyya, *Bh* , o coseno da medida de divergência de Bhattacharyya (1943) entre as duas distribuições de probabilidade, *Bh*=cosΔ, tal como, Krzanowski (1983) para o caso contínuo, e Sohail Khalid *et al*. (2006) no caso discreto. Este coeficiente de Battacharyya coincide, pois, com o coeficiente de afinidade. O

coeficiente de Bhattacharyya é, actualmente, muito usado em Reconhecimento de Formas (e.g., Sohail Khalid *et al*., 2006), assim como a distância de Bhattacharyya (e.g., Fukunaga, 1990) que é utilizada como uma medida de separabilidade de duas distribuições normais. Na literatura científica encontra-se o coeficiente de Bhattacharyya (*Bh*) associado a várias distâncias:

- através da função circular, $D = \cos^{-1} Bh$ (e.g., Krzanowski (1983) apresenta-a com referência a Bhattacharyya (1946)). Neste caso, naturalmente que D=Δ.
- $D^2 = \cos^{-1} Bh$ (e.g., Greenacre, 2005).
- através da função exponencial, $D = -\ln Bh$ (e.g., Ramakrishnan e Selvan, 2006).
- através da função quadrática, $D = \sqrt{2(1-Bh)}$ (e.g., Krzanowski (1983) apresenta-a com referência a Matusita (1956)).

No contexto da Análise Classificatória, Bacelar-Nicolau (1980, 1982, 1985, 1988):

- Introduziu o coeficiente de afinidade como um coeficiente de semelhança básico entre colunas ou linhas de uma matriz de dados M(DxV), sendo D o conjunto das unidades de dados e V o conjunto das variáveis (Subsecção 3.2).
- Estudou a distribuição assintótica do coeficiente de afinidade entre variáveis, sob diferentes hipóteses de referência (Subsecção 3.3), da qual resultou a definição de dois coeficientes de afinidade centrados e reduzidos (Subsecções 3.3.1 e 3.3.2).
- A partir deste coeficiente, definiu coeficientes de afinidade probabilísticos – validade da afinidade (Subsecção 3.3.3) –, o que lhe permitiu introduzir uma abordagem probabilística à Análise Classificatória. Os coeficientes de afinidade probabilísticos são do tipo de semelhança "validade da ligação" (VL)[111] (e.g., Lerman, 1970, 1981; Le Calvé, 1977; Bacelar-Nicolau, 1980, 1988; Nicolau, 1980).

Desde então, H. Bacelar-Nicolau generalizou o coeficiente de afinidade a dados inteiros (Secção 3.5), a dados simbólicos/complexos (Secção 3.6) e a dados mistos ou heterogéneos (Subsecção 3.6.3). O coeficiente de afinidade associado à distância *city block* é abordado na Secção 3.4.

Na Secção 3.7 far-se-á uma breve reflexão sobre a sua utilização em Análise de Dados.

---

[111] *Vraisemblance du lien*, em francês. A tradução portuguesa, introduzida por Bacelar-Nicolau, também se deve ao Prof. Tiago de Oliveira.

3.2 O coeficiente de afinidade básico

Situando-nos no enquadramento da Análise Classificatória, consideremos a matriz de dados M(D, V), nxp, em que:

- D é um conjunto de unidades de dados[112], correspondendo, no caso mais geral, a uma partição de um conjunto de indivíduos em n classes, U_1, ..., U_n ($\forall i,i' \in \{1, ..., n\}, U_i \cap U_{i'} = \varnothing$ e $\bigcup_{1 \le i \le n} U_i = D$). D também pode ser uma amostra ou um subconjunto com n indivíduos.
- V é um conjunto de p variáveis cujos valores são frequências (Caso A) ou variáveis métricas com valores positivos ou nominais ou binárias (Caso B).

Bacelar-Nicolau (1980) define a afinidade entre duas variáveis V_j e $V_{j'}$ (para simplificar a notação escrever-se-á frequentemente, j e j') da seguinte maneira:

Definição 3.2.1. Coeficiente de afinidade básico entre duas variáveis

O coeficiente de afinidade entre o par de variáveis V_j e $V_{j'}$ de M, com j≠j', é dado pelo valor

$$a(j, j') = \sum_{i=1}^{n} \sqrt{\frac{n_{ij}}{n_{.j}} . \frac{n_{ij'}}{n_{.j'}}} \qquad (3.2.1)$$

sendo:

- n_{ij} o número de vezes que a variável V_j foi observada na unidade de dados U_i e $n_{.j}$, $n_{.j} = \sum_{i=1}^{n} n_{ij}$, o número total de vezes que a variável V_j é observada no conjunto D (Caso A).

  Neste caso, a matriz de dados M(D,V) é uma tabela de frequências (nxp) e designa-se por perfil coluna j o correspondente vector de frequências relativas (ou condicionais) $n_{ij}/n_{.j}$. Assim sendo, o coeficiente de afinidade entre as variáveis V_j e $V_{j'}$ é o produto escalar das raízes quadradas dos respectivos perfis coluna, $\sqrt{n_{ij} / n_{.j}}$ e $\sqrt{n_{ij'} / n_{.j'}}$.
- n_{ij} = pontuação positiva ou código que o indivíduo i atribui à característica V_j ou presença-ausência; $n_{.j}$ representa a "pontuação" total que a variável V_j obteve no conjunto D (Caso B).

---

[112] Aqui adopta-se a designação habitualmente usada por Bacelar-Nicolau (e.g., 1980).

M(D,V) é, neste caso, uma matriz métrica com valores positivos ou nominal ou lógica[113], satisfazendo a determinadas condições (Subsecção 3.3.2). Neste caso, $n_{ij}/n_{.j}$ corresponde a uma recodificação dos dados.

Quando as unidades de dados têm pesos positivos diferentes w_i, $\sum_{i=1}^{n} w_i = 1$, a Expressão 3.2.1 adapta-se em conformidade (Nicolau e Bacelar-Nicolau, 2003).

No caso A, do ponto de vista da estatística "clássica", a sucessão de valores que constitui cada coluna da matriz M* (matriz das frequências relativas $n_{ij}/n_{.j}$ das colunas/variáveis j da matriz M) define uma lei de probabilidade discreta, em que o suporte D das p leis de probabilidade é o mesmo para todas elas. Na j'ésima coluna (j=1, ..., p) encontram-se os valores $p^*_{i/j}$ (i= 1, ..., n) da j'ésima lei de probabilidade: $\sum_{i} p^*_{i/j} = 1$, $0 \le p^*_{i/j} \le 1$. O coeficiente de afinidade pode então escrever-se sob a forma:

$$a(j, j') = \sum_{i=1}^{n} \sqrt{p^*_{i/j}.p^*_{i/j'}}$$

O coeficiente de afinidade é uma estimativa de máxima verosimilhança do parâmetro $\wp(j, j') = \sum_{i=1}^{n} \sqrt{p_{i/j}.p_{i/j'}}$, entre as distribuições de probabilidade condicionadas $\{p_{i/j}\}$ e $\{p_{i/j'}\}$, i= 1, ..., n, associadas às colunas j e j'.

No caso B, a sucessão de valores $n_{ij}/n_{.j}$ que constitui cada coluna da matriz M* é um "vector de probabilidade" cujas componentes representam a contribuição à "pontuação" total, que é igual a 1.

Como se referiu, o coeficiente de afinidade também se pode definir entre as unidades de dados, i.e., entre os "vectores perfis", e pode ser ponderado ou não. Apresenta-se de seguida as definições da afinidade entre os "vectores perfis" e da afinidade ponderada entre os "vectores perfis" (e.g., Bacelar-Nicolau, 1980; Sousa Ferreira *et al*., 2001), respectivamente.

[113] Designação atribuída por Chandon e Pinson (1981) às matrizes de dados binários.

Definição 3.2.2. Coeficiente de afinidade básico entre duas unidades de dados

O coeficiente de afinidade entre duas unidades de dados i e i' é dado pelo valor

$$a(i,i') = \sum_{j=1}^{p} \sqrt{\frac{n_{ij}}{n_{i.}} . \frac{n_{i'j}}{n_{i'.}}} \qquad (3.2.2)$$

sendo n_{ij} e $n_{i'j}$ o número de indivíduos que pertencem, respectivamente, às unidades de dados i e i' para as quais a variável j foi observada e $n_{i.} = \sum_{j=1}^{p} n_{ij}$.

Definição 3.2.3. Coeficiente de afinidade básico ponderado entre duas unidades de dados

O coeficiente de afinidade ponderado entre duas unidades de dados i e i' é dado pelo valor

$$a(i,i') = \sum_{j=1}^{p} \omega_j \sqrt{\frac{n_{ij}}{n_{i.}} . \frac{n_{i'j}}{n_{i'.}}} \qquad (3.2.3)$$

sendo n_{ij} e $n_{i'j}$ o número de indivíduos que pertencem, respectivamente, às unidades de dados i e i' para as quais a variável j foi observada e $n_{i.} = \sum_{j=1}^{p} n_{ij}$, $0 \leq \omega_j \leq 1$, $\sum_{j=1}^{p} \omega_j = 1$.

3.2.1 Algumas propriedades do coeficiente de afinidade básico entre variáveis

Prova-se que o coeficiente de afinidade tem as seguintes propriedades:

- É um coeficiente de semelhança (Definição 1.3.10) que toma valores no intervalo [0,1]. Quando os vectores de frequências relativas são iguais ou proporcionais, o coeficiente de afinidade toma o valor 1. Toma o valor 0, quando os vectores de frequências relativas são ortogonais (e.g., Bacelar-Nicolau, 1980; Doria, 1989). É uma semelhança normada, pois $\forall j \in V$, a(j,j) = 1. Não é uma semelhança própria sobre as variáveis iniciais, pois $a(j,j') = a_{max} \not\Rightarrow j = j'$; contudo, é uma semelhança própria sobre os perfis coluna (Bacelar-Nicolau, apresentação pessoal).
- Mede a tendência monótona entre os perfis coluna.
- Está relacionado com a distância de Hellinger, associada aos perfis coluna j e j',

$$d^2(j,j') = \sum_{i=1}^{n} \left(\sqrt{\frac{n_{ij}}{n_{.j}}} - \sqrt{\frac{n_{ij'}}{n_{.j'}}} \right)^2 ,$$

pela função quadrática $d^2(j,j') = 2\left(1 - a(j,j')\right)$ já apresentada anteriormente.

- Pode-se verificar que d(j,j') é uma distância euclidiana entre as raízes quadradas dos perfis representativos das colunas j e j', e que $0 \leq d^2(j,j') \leq 2$ (e.g., Bacelar-Nicolau, 1980; Doria, 1989).
- O coeficiente de afinidade (e a distância de Hellinger a ele associada) entre dois perfis coluna é uma semelhança (respectivamente, uma distância) na esfera-p com centro na origem e raio igual a 1; usando a função potência com expoente ½ permite-nos obter um vector dos perfis coluna com coordenadas esféricas positivas $\sqrt{\frac{n_{ij}}{n_{.j}}}$, i=1, ..., p, (e.g., Nicolau e Bacelar-Nicolau, 2003).

- Satisfaz ao "princípio de equivalência distribucional" (tal como foi definido por Benzécri (1980b) para a distância do qui-quadrado, isto é, podemos substituir duas ou mais linhas proporcionais por uma única linha, soma delas, pois a afinidade entre os pares de variáveis não se altera. Bacelar-Nicolau (1980, 1982) baseia a sua demonstração na noção de contribuição de cada elemento para a afinidade do par de variáveis. O mesmo se passa se juntarmos mais linhas (unidades de dados) proporcionais à matriz de dados.
- Os valores do coeficiente de afinidade entre os perfis coluna da matriz, que já existem, não se alteram quando se junta mais perfis coluna (e.g., Nicolau e Bacelar-Nicolau, 2003).
- É independente da dimensão do conjunto das unidades de dados D (e.g., Nicolau e Bacelar-Nicolau, 2003).
- Coincide com o coeficiente de Ochiai (A_{11}, Tabela 1.3.4), quando a matriz de dados é lógica, isto é, quando as variáveis V_j são binárias.

O coeficiente de afinidade relaciona-se, noutros casos particulares, com outros coeficientes conhecidos (Subsecções 3.3.2.1 e 3.3.2.2) e também pode ser generalizado, como se verá. Este coeficiente está programado, por exemplo, em Fortran, nos programas SIMIL2 (Bacelar-Nicolau, 1980) e AFIN (Doria, 1989).

3.3 Estudo distribucional do coeficiente de afinidade

Bacelar-Nicolau (1980, 1988) fez o estudo distribucional do coeficiente de afinidade entre variáveis, considerando duas condições experimentais distintas referidas na Definição 3.2.1

(Caso A, em que as unidades de dados são escolhidas aleatoriamente; Caso B, em que as unidades de dados podem ser ou não escolhidas aleatoriamente). A cada uma das condições experimentais vai corresponder um modelo probabilístico distinto (Modelo A e Modelo B) e respectivos coeficientes de semelhança que se baseiam no coeficiente de afinidade básico (Subsecção 3.3.1, Subsecção 3.3.2 e Subsecção 3.3.3).

Sousa (2005) e Sousa *et al*. (2005), na sequência de uma primeira abordagem desenvolvida por Nicolau (1994), realizaram o estudo empírico da distribuição assintótica do coeficiente de afinidade básico, com base em simulações pelo método de Monte Carlo. Neste estudo, concluíram que:

- A distribuição do coeficiente de afinidade básico depende do modelo de probabilidades básico (distribuições binomial, hipergeométrica, Poisson, uniforme (0,1), normal centrada e reduzida, exponencial (1), lognormal (0,1), logística (0,1) e de Pareto) e dos parâmetros desse modelo.
- O coeficiente de afinidade básico tende a seguir assintoticamente as distribuições normal e lognormal (considerando a aproximação boa quando as amostras têm dimensão superior ou igual a 30 e, em certos casos, 15), excepto no caso da distribuição de Pareto com os parâmetros (1, 1) e (1, 2), embora o nível de confiança seja, em geral, mais elevado no que diz respeito à distribuição normal.
- No caso particular dos dados binários, a média dos valores do coeficiente de afinidade entre pares de vectores independentes seguindo a distribuição de Bernoulli de parâmetro p tende assintoticamente para p.

Passemos agora em revista, sucintamente, cada um dos modelos.

3.3.1 Modelo A. O coeficiente de afinidade A_δ

No Modelo A:

- A matriz de dados M(D,V) é uma matriz de frequências. As unidades de dados são escolhidas aleatoriamente de uma população (supostamente) infinita.
- O conjunto das unidades de dados $D=\{U_1, ..., U_n\}$ está associado a um esquema de amostragem aleatória numa população (supostamente) infinita tal que, para um par de colunas/variáveis – $(n_{1j}, ..., n_{nj})$ e $(n_{1j'}, ..., n_{nj'})$ – da matriz M(D,V), aqueles dois conjuntos são amostras independentes associadas respectivamente às variáveis V_j e $V_{j'}$ e a cada uma daquelas amostras corresponde uma lei de probabilidade discreta:

$$\begin{pmatrix} U_1 & \dots U_n \\ p_{1/j} & \dots p_{n/j} \end{pmatrix} \quad \text{sendo } \sum_i p_{i/j} = 1,\ 0 < p_{i/j} < 1$$

$$\begin{pmatrix} U_1 & \dots U_n \\ p_{1/j'} & \dots p_{n/j'} \end{pmatrix} \quad \text{sendo } \sum_i p_{i/j'} = 1,\ 0 < p_{i/j'} < 1$$

Neste sentido,

- Os valores n_{ij} ($n_{ij'}$) da matriz M representam o número de observações da 1ª (2ª) amostra, que pertencem à unidade de dados U_i (i=1, ..., n).
- As frequências relativas/condicionadas $n_{i/j}$ ($n_{i/j'}$) são as estimativas de máxima verosimilhança $p^*_{i/j}$ ($p^*_{i/j'}$) dos parâmetros $p_{i/j}$ ($p_{i/j'}$).

Tendo em conta a Definição 3.2.1 e o que acabou de se apresentar, o coeficiente de afinidade a(j,j') é uma concretização da variável aleatória A(j,j') que é um estimador consistente do parâmetro afinidade, $\wp(j, j')$, entre as distribuições de probabilidade condicionadas $p_{i/j}$ e $p_{i/j'}$, i= 1, ..., n, associadas às colunas j e j' da população:

$$\wp(j, j') = \sum_{i=1}^{n} \sqrt{p_{i/j}.p_{i/j'}} \qquad (3.3.1).$$

Com base nas suposições acima indicadas, $(n_{1j},...,n_{nj})$ e $(n_{1j'},...,n_{nj'})$ são dois vectores aleatórios com distribuição multinomial. A variável aleatória A(j,j') converge em probabilidade para $\wp(j, j')$ (Bacelar-Nicolau, 1980).

Bacelar-Nicolau (1980) também enunciou e demonstrou o teorema sobre a distribuição assintótica de A(j,j'), com base na versão univariada do método δ (e.g., Tiago de Oliveira, 1982). Como consequência deste teorema resultou a definição do coeficiente de afinidade centrado e reduzido pelo método-δ.

Definição 3.3.1. Coeficiente de afinidade centrado e reduzido pelo método-δ, $A_\delta(j,j')$

O coeficiente de afinidade centrado e reduzido pelo método-δ, $A_\delta(j,j')$, resulta da centragem e redução do estimador A(j,j') utilizando o método δ no modelo A:

$$A_\delta(j, j') = \frac{A(j, j') - \wp(j, j')}{\sqrt{1 - \wp^2(j, j')}} . \frac{2\sqrt{\eta}}{\sqrt{a^2 + a'^2}} \qquad (3.3.2)$$

Sendo:

- $\wp(j, j')$ o parâmetro afinidade (Expressão 3.3.1),
- $\eta = \min(n_{.j}, n_{.j'})$

$$- a^2 = \lim_{n_{.j}\to\infty} \frac{\eta}{n_{.j}} \text{ e } a'^2 = \lim_{n_{.j'}\to\infty} \frac{\eta}{n_{.j'}}$$

A variável aleatória $A_\delta(j,j')$ é assintoticamente normal reduzida (resultado do teorema enunciado e demonstrado por Bacelar-Nicolau, 1980).

Contudo, sob o ponto de vista prático, a variável aleatória $A_\delta(j,j')$ não serve como coeficiente de comparação entre duas variáveis, pois na sua expressão intervém o parâmetro desconhecido $\wp(j,j')$. Para ultrapassar este problema, Bacelar-Nicolau aconselha, como hipótese de referência para o cálculo do coeficiente de afinidade centrado e reduzido pelo método-δ que se considere, entre outros: $\wp(j,j')$=1/2 ou $\wp(j,j')$=média de todas as realizações de A(j,j'). Resultam assim dois coeficientes de afinidade centrados e reduzidos pelo método-δ:

Definição 3.3.2. O coeficiente de afinidade $a_{\delta 1}(j,j')$

O coeficiente de afinidade centrado e reduzido pelo método-δ, $a_{\delta 1}(j,j')$, é uma realização da variável aleatória $A_\delta(j,j')$ considerando $\wp(j,j')$=1/2:

$$a_{\delta 1}(j,j') = \frac{2\sum_i \sqrt{n_{ij} n_{ij'}} - \sqrt{n_{.j} n_{.j'}}}{\sqrt{3/4\left(n_{.j} + n_{.j'}\right)}} \qquad (3.3.3)$$

Definição 3.3.3. O coeficiente de afinidade $a_{\delta 2}(j,j')$

O coeficiente de afinidade centrado e reduzido pelo método-δ, $a_{\delta 2}(j,j')$, é uma realização da variável aleatória $A_\delta(j,j')$ considerando $\wp(j,j')$=média de todas as realizações de A(j,j'):

$$a_{\delta 2}(j,j') = \frac{\sum_{i=1}^{n} \sqrt{\frac{n_{ij}.n_{ij'}}{n_{.j}.n_{.j'}}} - \overline{a(j,j')}}{\sqrt{1 - \overline{a(j,j')}^2}} . \frac{2\sqrt{n_{.j}.n_{.j'}}}{\sqrt{n_{.j} + n_{.j'}}} \qquad (3.3.4)$$

Os coeficientes $a_{\delta 1}(j,j')$ e $a_{\delta 2}(j,j')$ estão programados, por exemplo, em Fortran, nos programas SIMIL2 (Bacelar-Nicolau, 1980) e AFIN (Doria, 1989).

A partir da variável aleatória A_δ pode-se definir um coeficiente probabilístico como se verá na Subsecção 3.3.3.

É possível determinar intervalos de confiança assintóticos para o parâmetro afinidade $\wp(j, j')$ e efectuar testes de hipóteses sobre aquele parâmetro. O intervalo a (1-α)100% de confiança para a afinidade $\wp(j, j')$ é obtido a partir da expressão $P(|A_\delta| < z_{1-\alpha/2}) = 1-\alpha$, onde $z_{1-\alpha/2}$ designa o quantil de ordem 1- $\alpha/2$ da lei normal reduzida (Bacelar-Nicolau, 1980).

3.3.2 Modelo B. O coeficiente de afinidade A_W

Bacelar-Nicolau (e.g., 1980, 1988) definiu o Modelo B da seguinte maneira:

- O conjunto das unidades de dados $D=\{U_1, ..., U_n\}$ não está associado a um esquema de amostragem aleatória em população infinita.
  Pelo facto das unidades de dados poderem não ser escolhidas aleatoriamente, os dois subconjuntos – $(n_{1j}, ..., n_{nj})$ e $(n_{1j'}, ..., n_{nj'})$ – da matriz M(D,V), associados respectivamente às variáveis V_j e $V_{j'}$ não têm distribuições multinomiais como acontecia no Modelo A.
- A matriz de dados, M(D,V), é uma matriz de frequências, métrica com valores não negativos ou nominal, satisfazendo à seguinte condição:
  - Quaisquer dois vectores coluna da matriz M – $(n_{1j}, ..., n_{nj})$ e $(n_{1j'}, ..., n_{nj'})$ – associados respectivamente às variáveis V_j e $V_{j'}$, são sucessões de números reais não negativos, tais que as frequências condicionais/contribuições $n_{i/j}$ e $n_{i/j'}$, i=1, ..., n, são da mesma ordem de grandeza, n^{-1}, quando $n\to\infty$ ($n_{i/n} = O(n^{-1})$), isto é:

    $n_{i/j} = \varepsilon_{ij}n^{-1}$ e $n_{i/j'} = \varepsilon_{ij'}n^{-1}$, i=1, ..., n, em que $\{\varepsilon_{ij}\}$e $\{\varepsilon_{ij'}\}$ são conjuntos limitados de constantes positivas (próximas de zero), $0<\min\varepsilon_{ij}\leq\max\varepsilon_{ij'}<+\infty$.

    As sequências resultantes de números reais, $E_n = \sqrt{\frac{\varepsilon_{ij}}{n}}$ e $E'_n = \sqrt{\frac{\varepsilon_{ij'}}{n}}$, i=1, ..., n, são, respectivamente, os valores dos vectores aleatórios X_n e Y_n cujos espaços amostrais estão definidos como os conjuntos de todas as n! permutações de respectivamente E_n e E'_n com distribuição uniforme (isto é, em que cada permutação tem a mesma probabilidade 1/n!).

Naquelas condições, o coeficiente de afinidade é o produto interno daquelas sequências: $a(j,j') = \langle E_n, E'_n \rangle$.

Baseando-se neste modelo e recorrendo ao teorema limite de Wald e Wolfowitz (1944), Bacelar-Nicolau (1982, 1988) enunciou e demonstrou o seguinte teorema:

Teorema 3.3.1. Distribuição assintótica de A(j,j') sob as condições gerais do Modelo B

Sob as condições gerais do Modelo B, o coeficiente de afinidade A(j,j') segue assintoticamente uma distribuição normal com valor médio $\mu(j,j')$ e variância $\sigma^2(j,j')$:

$$\mu_{A(j,j')} = \frac{1}{n}\sum_{i,i'}\sqrt{n_{i/j}\ n_{i'/j'}} \qquad (3.3.5)$$

$$\sigma^2_{A(j,j')} = \frac{1}{n\text{-}1}\left[\sum_i\left(\sqrt{n_{i/j}} - \frac{1}{n}\sum_{i'}\sqrt{n_{i'/j}}\right)^2\right]\left[\sum_i\left(\sqrt{n_{i/j'}} - \frac{1}{n}\sum_{i'}\sqrt{n_{i'/j'}}\right)^2\right] \qquad (3.3.6)$$

Além disso, a variável aleatória A_W,

$$A_w(j,j') = \frac{A(j,j') - \mu_{A(j,j')}}{\sigma_{A(j,j')}} \qquad (3.3.7), \text{ tem distribuição limite N(0,1).}$$

Como consequência do Teorema 3.3.1 resulta que a variável aleatória A_W é uma estatística para o modelo B que pode ser usada como um coeficiente de semelhança, daí a definição que se segue.

Definição 3.3.4. Coeficiente de afinidade centrado e reduzido pelo método-WW , $a_W(j,j')$

O coeficiente de afinidade centrado e reduzido pelo método-WW, $a_W(j,j')$, é uma realização da variável aleatória $A_W(j,j')$:

$$a_W(j,j') = \frac{\dfrac{n\sum_i\sqrt{n_{ij}n_{ij'}} - \sum_i\sqrt{n_{ij}}\sum_i\sqrt{n_{ij'}}}{\sqrt{n\sum_i n_{ij} - \left(\sum_i n_{ij}\right)^2}}}{\sqrt{n\sum_i n_{ij'} - \left(\sum_i n_{ij'}\right)^2}}\sqrt{n-1} \qquad (3.3.8)$$

O coeficiente a_W está programado, por exemplo, em AFIN (Doria, 1989).

Sousa (2005) e Sousa *et al*. (2005), no estudo empírico da distribuição assintótica do coeficiente de afinidade A_W com base em simulações pelo *método de Monte Carlo*, concluíram que a convergência deste coeficiente para as distribuições normal centrada e reduzida e t-Student é rápida, excepto no caso do coeficiente A_W calculado para a distribuição de Pareto (1, 1); em particular, a aproximação à distribuição normal pode ser considerada boa para amostras com dimensão superior a 20 e em muitos casos (se a assimetria dos perfis coluna não for muito grande) para amostras com dimensão superior a 10. Verificaram também que o valor médio e o desvio padrão do coeficiente A_W não dependem, nem do tipo de distribuição dos vectores gerados, nem dos parâmetros da

distribuição. Foi realizado um estudo análogo com o coeficiente de afinidade centrado e reduzido utilizando a média e o desvio padrão empíricos (centragem e redução global). Os resultados foram análogos aos obtidos com o coeficiente A_W.

A partir da variável aleatória A_W pode-se definir um coeficiente probabilístico como se verá na Subsecção 3.3.3. Antes de se passar a essa definição vejamos mais algumas consequências do Teorema 3.3.1.

3.3.2.1 Relação entre o coeficiente de afinidade centrado e reduzido pelo método-WW e o coeficiente de correlação linear de Pearson

Verifica-se que o coeficiente de afinidade centrado e reduzido pelo método-WW, $a_W(j,j')$ (Expressão 3.3.8), coincide, a menos de um factor, com o coeficiente de correlação linear de Pearson, r(j,j'), entre as sucessões ($\sqrt{\varepsilon_{ij}}$, i=1, ...n) e ($\sqrt{\varepsilon_{ij'}}$, i=1, ...n): $a_W(j,j') = r(j,j')\sqrt{n-1}$.

3.3.2.2 O coeficiente de afinidade centrado e reduzido pelo método-WW no caso das variáveis serem binárias

Quando as variáveis são binárias, o coeficiente $a_W(j,j')$ é igual ao coeficiente de Ochiai centrado e reduzido:

$$O_{Cent\,Red} = \frac{na-(a+b)(a+c)}{\sqrt{(a+b)(a+c)(n-a-b)(n-a-c)}}\sqrt{n-1} \quad (3.3.9)$$

3.3.3 Os coeficientes de afinidade probabilísticos $VAL_{A\delta}$ e VAL_{Aw}

A partir das variáveis aleatórias A_δ e A_W podem-se definir coeficientes de afinidade probabilísticos que generalizam a semelhança VL.

Definição 3.3.5. O coeficiente $VAL_{A\delta}$

O coeficiente de validade da afinidade $VAL_{A\delta}$ é um coeficiente probabilístico,

$VAL_{A\delta}(j,j') = Prob(A_\delta(j,j') \leq a_\delta(j,j')) \approx \Phi(a_\delta(j,j'))$ (3.3.10)

sendo $a_\delta(j,j')$ uma realização da variável aleatória $A_\delta(j,j')$ e $\Phi(a_\delta(j,j'))$ o valor da função de distribuição da N(0,1) no ponto $a_\delta(j,j')$.

O coeficiente $VAL_{A\delta}$ (j,j') é, como se verifica, um coeficiente de semelhança probabilístico associado ao modelo A, generalizando assim a semelhança VL ao caso de variáveis numéricas que tomam valores no conjunto dos números naturais.

Analogamente se define o coeficiente de semelhança VAL_{Aw}.

Definição 3.3.6. O coeficiente VAL_{AW}

O coeficiente de validade da afinidade VAL_{Aw} é um coeficiente probabilístico,

$VAL_{Aw}(j,j') = Prob(A_W(j,j') \leq a_W(j,j')) \approx \Phi(a_W(j,j'))$ (3.3.11)

sendo $a_W(j,j')$ uma realização da variável aleatória $A_W(j,j')$ e $\Phi(a_W(j,j'))$ o valor da função de distribuição da N(0,1) no ponto $a_W(j,j')$.

A distribuição do coeficiente $VAL_{Aw}(j,j')$ pode ser aproximada pela distribuição uniforme no intervalo [0,1].

Os coeficientes $VAL_{A\delta}$ e VAL_{Aw} medem a validade do coeficiente de semelhança básico, no sentido probabilístico, i. e., quanto menor for a probabilidade de qualquer valor exceder o valor observado, respectivamente, $a_\delta(j,j')$, $a_W(j,j')$, tanto maior a sua validade ou a confiança dada à decisão de agrupar as colunas (j, j') ou as variáveis (V_j, $V_{j'}$) (Bacelar-Nicolau, 1980, 1988).

Os coeficientes $VAL_{A\delta}$ e VAL_{Aw} são coeficientes de semelhança que tomam valores no intervalo [0,1].

O coeficiente VAL_{Aw} também satisfaz ao "princípio de equivalência distribucional" e é independente da dimensão do conjunto de unidades de dados D, tal como o coeficiente de afinidade básico.

3.3.4 Conclusões sobre os modelos A e B

Os dois modelos probabilísticos podem ser usados em tabelas de contingência, quer para comparar variáveis, quer para comparar unidades de dados depois da manipulação adequada do conjunto das variáveis (Bacelar-Nicolau, 1988). Também podem ser generalizados a outros casos, tais como, a variáveis contínuas positivas (modelo A), a frequências negativas (modelo B) (Bacelar-Nicolau, 1988), a variáveis reais (e.g., Nicolau e Bacelar-Nicolau, 2003), a variáveis simbólicas e heterogéneas (e.g., Bacelar-Nicolau, 2000).

Entre as vantagens da passagem dos coeficiente de semelhança tradicionais à semelhança VL pode-se referir, por exemplo, a uniformização da escala de medida (Nicolau, 1981) e a

simplificação na escolha do coeficiente de semelhança (dissemelhança) entre elementos no caso de dados binários, como consequência da equivalência distribucional desses coeficientes (Bacelar-Nicolau, 1980; Subsecção 1.3.6).
O coeficiente probabilístico é, pois, uma medida que valida intrinsecamente o valor do próprio coeficiente básico.

Estes coeficientes têm sido usados em muitas aplicações práticas, em Análise Classificatória com bons resultados. Em particular, destaca-se aqui a utilização do coeficiente probabilístico VAL_{Aw} para comparar ultramétricas (classificações obtidas com vários coeficientes de semelhança e o mesmo critério de agregação entre classes) sobre o mesmo conjunto de dados, quer no contexto da Análise Classificatória (validação relativa), quer no contexto da ACP (Secção 5.4; Doria *et al*., 2000).

Note-se que a abordagem probabilística opera num contexto exploratório que usa o conhecimento prévio da estrutura dos dados como um instrumento para extrair informação acerca da sua estrutura de agrupamento hierárquico (e.g., Nicolau e Bacelar-Nicolau, 2003).

3.4 O coeficiente de afinidade associado à distância city block

No caso de M(D,V) = [n_{ij}: i=1, ..., n; j=1, ...p] ser uma matriz de frequências ou métrica ou nominal ou lógica, o coeficiente de afinidade associado à distância *city block*, a_1, foi inicialmente definido por Bacelar-Nicolau (e.g., Doria, 1989; Bacelar-Nicolau, 1991) da seguinte maneira:

Definição 3.4.1. O coeficiente de afinidade a_1 entre variáveis

O coeficiente de afinidade a_1 entre as variáveis $(V_j,V_{j'})$ é dado pela seguinte expressão

$$a_1(j,j') = 2 - \sum_{i=1}^{n} \max\left(\frac{n_{ij}}{n_{.j}}, \frac{n_{ij'}}{n_{.j'}}\right) \qquad (3.4.1)$$

O coeficiente a_1 está associado à distância *city block* entre os perfis representativos das colunas (j,j')/variáveis $(V_j,V_{j'})$, $d_1(j,j') = \sum_{i=1}^{n} \left|\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right|$ (3.4.2), através da função $d_1 = 2(1-a_1)$.

Verifica-se que o coeficiente de afinidade a_1 é um coeficiente de semelhança:

- $a_1(j,j')= a_1(j',j)$
- $a_{1max}(j,j)=1 > a_1(j,j')$
- $0 \leq a_1 \leq 1$

Para dados binários: $a_1 \leq a$, verificando-se a igualdade quando $n_{.j}=n_{.j'}$ (Bacelar-Nicolau, 1991).

O cálculo do coeficiente de afinidade a_1 encontra-se programado, por exemplo, em AFIN (Doria, 1989).
Foi usado, em ACHA, com dados reais positivos (e.g., Doria, 1989) com bons resultados.

3.5 Generalização do coeficiente de afinidade a dados inteiros

Considere-se a matriz de dados inteiros, M(D,V)= $[n_{ij}: i=1, ..., n; j=1, ...p]$, que pode ser uma tabela com valores que representam contagens positivas e/ou negativas; na prática, estes valores podem resultar de situações, tais como, saldos de empresas ou migrações entre regiões. Neste caso, não é possível utilizar o coeficiente de afinidade para comparar elementos, tal como foi definido anteriormente. Bacelar-Nicolau (1986) generalizou a noção de afinidade e o Modelo B a dados inteiros da forma que se apresenta de seguida.

Definição 3.5.1. O coeficiente de afinidade generalizado entre colunas/variáveis

O coeficiente de afinidade generalizado entre um par de colunas/variáveis $(V_j,V_{j'})$ da matriz de dados inteiros M(D,V) é dado por,

$$a_g(j,j') = \sum_{i=1}^{n} sign\left(\frac{n_{ij}}{n_{.j}}\right) sign\left(\frac{n_{ij'}}{n_{.j'}}\right) \sqrt{\left|\frac{n_{ij}}{n_{.j}} . \frac{n_{ij'}}{n_{.j'}}\right|} \text{, se } n_{ij} \leq 0 \text{ ou } n_{ij'} \leq 0 \qquad (3.5.1)$$

em que, *sign* designa "o sinal de", $n_{.j}=\sum_i |n_{ij}|$, $n_{.j'}=\sum_i |n_{ij'}|$, e se verifica $\Sigma_i\, n_{i/j}=1$, $\Sigma_i\, n_{i/j'}=1$.

Este coeficiente pode ser generalizado ao conjunto dos números reais, desde que a normalização utilizada faça sentido.

O coeficiente de afinidade generalizado é um coeficiente de comparação do tipo semelhança, que toma valores no intervalo [-1, +1]:

- $a_g(j,j) = 1 \geq a_g(j,j'), \forall (j,j), (j,j') \in V^2$.
- $a_g(j,j') = a_g(j',j), \forall (j,j') \in V^2$.

O coeficiente de afinidade generalizado a_g pode-se relacionar com a distância d_g na n-esfera através da função quadrática: $d_g(j,j') = \sqrt{2(1-a_g)}$ (e.g., Nicolau e Bacelar-Nicolau, 2003; Sousa, 2005).

Sousa (2005) e Sousa *et al*. (2005), no estudo empírico da distribuição assintótica do coeficiente de afinidade generalizado A_g com base em simulações pelo método de Monte Carlo, verificaram que, tal como o coeficiente de afinidade básico, a distribuição do coeficiente A_g depende do modelo de probabilidades básico e dos parâmetros desse modelo, e a sua variabilidade está naturalmente associada à dimensão da amostra, sendo maior quando a dimensão da amostra é reduzida. No entanto, o coeficiente de afinidade generalizado A_g tende assintoticamente para a distribuição normal; na prática, a aproximação é considerada boa para amostras com dimensão superior a 20, embora em alguns casos, seja verificada para amostras com dimensão superior a 10.

Sob as condições do Modelo B (Subsecção 3.3.2), atendendo à definição dos totais marginais das colunas $n_{.j}$, j=1, ..., p, prova-se, recorrendo ao teorema de Wald e Wolfowitz (1944), que o coeficiente de afinidade generalizado centrado e reduzido pelo método-WW, A_{gw},

$$A_{gw} = \frac{A_g - \mu_{A_g}}{\sigma_{A_g}} \qquad (3.5.2)$$ é assintoticamente N(0,1).

Sendo,

$$\mu_{A_{gw}}(j,j') = \frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} sign\left(\frac{n_{ij}}{n_{.j}}\right) sign\left(\frac{n_{i'j'}}{n_{.j'}}\right)\sqrt{\left|\frac{n_{ij}}{n_{.j}}\right|.\left|\frac{n_{i'j'}}{n_{.j'}}\right|}$$

$$\sigma^2_{A_{gw}}(j,j') = \frac{1}{n-1}\left[\sum_{i=1}^{n} sign\left(\frac{n_{ij}}{n_{.j}}\right)\sqrt{\left|\frac{n_{ij}}{n_{.j}}\right|} - \left(\frac{1}{n}\sum_{i'=1}^{n} sign\left(\frac{n_{i'j}}{n_{.j}}\right)\sqrt{\left|\frac{n_{i'j}}{n_{.j}}\right|}\right)^2\right]\left[\sum_{i=1}^{n} sign\left(\frac{n_{ij'}}{n_{.j'}}\right)\sqrt{\left|\frac{n_{ij'}}{n_{.j'}}\right|} - \left(\frac{1}{n}\sum_{i'=1}^{n} sign\left(\frac{n_{i'j'}}{n_{.j'}}\right)\sqrt{\left|\frac{n_{i'j'}}{n_{.j'}}\right|}\right)^2\right]$$

Definição 3.5.2. O coeficiente de afinidade generalizado centrado e reduzido pelo método-WW

O coeficiente de afinidade generalizado centrado e reduzido pelo método-WW, a_{gw} é realização da variável aleatória A_{gw} (Expressão 3.5.2):

$$a_{gw}(j,j') = \frac{A_g(j,j') - \mu_{A_g}}{\sigma_{A_g}} \qquad (3.5.3)$$

Tal como anteriormente, define-se um coeficiente de tipo probabilístico que é uma generalização da afinidade associada ao modelo B.

Definição 3.5.3. O coeficiente VAL_{Agw}

O coeficiente de validade da afinidade VAL_{Agw} é um coeficiente probabilístico,

$VAL_{Agw}(j,j') = Prob(A_{gW}(j,j') \leq a_{gW}(j,j')) \approx \Phi(a_{gW}(j,j'))$ (3.5.4)

sendo $a_{gW}(j,j')$ uma realização da variável aleatória $A_{gW}(j,j')$ e $\Phi(a_{gW}(j,j'))$ o valor da função de distribuição da N(0,1) no ponto $a_{gW}(j,j')$.

O cálculo dos coeficientes de afinidade generalizados a_g, a_{gw} e VAL_{Agw} encontra-se programado, por exemplo, em AFIN (Doria, 1989).

Estes coeficientes foram usados em ACHA, com os critérios de agregação entre classes tradicionais e com os critérios probabilísticos quando a semelhança VL foi utilizada, sobre dados verídicos, como por exemplo, os da área da economia que se referem à variação de *stocks* de diversos produtos (e.g., Doria, 1989; Bacelar-Nicolau e Doria, 1992; Bacelar-Nicolau e Nicolau, 1993).

3.6 Generalização do coeficiente de afinidade a dados complexos e heterogéneos

O coeficiente de afinidade foi generalizado a matrizes de dados complexos ou simbólicos (e.g., Bock e Diday, 2000; Subsecção 4.3.10) e mistos, no âmbito da Análise Classificatória (e.g., Bacelar-Nicolau, 2000; 2002) (Subsecção 3.6.1 e Subsecção 3.6.3). Neste contexto, o coeficiente de afinidade foi introduzido no SODAS Package que está disponível na Internet e que faz parte de um projecto – Projecto SODAS – que reúne os saberes de várias universidades e institutos de estatística (e.g., Bock e Diday, 2000). A generalização do coeficiente de afinidade entre unidades de dados descritas por variáveis intervalares foi concretizada posteriormente (e.g., Sousa, 2005; Sousa *et al*., 2007) e é apresentada na Subsecção 3.6.2; enquanto que a centragem e redução deste coeficiente usando o método-WW e a sua versão probabilística são estudadas na Subsecção 3.6.4.

3.6.1 Generalização do coeficiente de afinidade a dados simbólicos modais

Considere-se o conjunto de n unidades de dados, E={1, ..., n}, descritas por um conjunto de p variáveis Y={Y_1, ..., Y_p}. As unidades de dados podem ser sujeitos ou subconjuntos de sujeitos de alguma população/conjunto e as variáveis são simbólicas modais (e.g., Bock, 2000a; Subsecção 4.3.10.8), i.e., distribuições de probabilidade, histogramas (distribuições de frequências) ou frequências inteiras, que se podem representar numa matriz de dados tridimensional $(\xi_{kj})_{nxp}$ (Tabela 3.6.1); considera-se que a variável categórica modal Y_j com m_j modalidades, {1, ..., m_j}, toma o valor ξ_{kj}= (n_{kj1},…, n_{kjmj}) para a unidade de dados k, sendo n_{kjv} o número de indivíduos (na unidade k) que partilham a categoria v da variável Y_j. Esta matriz de dados também pode ser interpretada como uma matriz com n linhas e c colunas ($c=m_1+...+m_p$), sendo estas identificadas pelo índice duplo (j,v). Cada linha $x_k=(\xi_{k1}, ..., \xi_{kp})$ descreve um objecto simbólico modal (e.g., Bock, 2000a; Bacelar-Nicolau, 2002; Sousa, 2005) com p variáveis.

Tabela 3.6.1. Matriz de dados simbólica em que ξ_{kj} corresponde à distribuição de frequências que a variável Y_j toma para a unidade estatística k: $\xi_{kj}= (n_{kj1},\ldots, n_{kjmj})$

	Variáveis			
Unidades de dados	…	Y_j	…	Σ
…	…	…	…	…
k	…	$(n_{kj1},\ldots, n_{kjmj})$	…	$n_{k..}$
…	…	…	…	…
k'	…	$(n_{k'j1},\ldots, n_{k'jmj})$	…	$n_{k'..}$
…	…	…	…	…
Σ	…	$(n_{.j1},\ldots, n_{.jmj})$	…	$n_{...}$

Com a notação apresentada acima, Bacelar-Nicolau (2000, 2002) define o coeficiente de afinidade parcial que relaciona as distribuições de frequências de duas unidades de dados k, k'$\in$E .

<u>Definição 3.6.1. Coeficiente de afinidade parcial entre duas distribuições de frequências de unidades de dados descritas por uma variável simbólica modal Y_j</u>

O coeficiente de afinidade parcial ou específico[114] ou local da variável, que relaciona as distribuições de frequências ξ_{kj} e $\xi_{k'j}$ (correspondentes à variável Y_j) de duas unidades de dados k, k'$\in$E é dado por

$$aff(\xi_{k,j},\xi_{k',j}) = \sum_{\ell=1}^{m_j}\sqrt{\frac{n_{kj\ell}}{n_{kj.}}.\frac{n_{k'j\ell}}{n_{k'j.}}} \qquad (3.6.1)$$

[114] *Variable-specific*, no original.

sendo, $n_{kj.} = \sum_{\ell=1}^{m_j} n_{kj\ell}$ o número de indivíduos que pertencem (ou observados em) à unidade de dados (grupo) k para a qual a variável Y_j foi observada; m_j representa a dimensão do ℓ'ésimo "estrato".

Da definição resulta que a distribuição de frequências relativas $p_{kj}= n_{kj\ell}/n_{kj.}$, ℓ=1, ..., m_j, gera uma distribuição de probabilidade discreta, com um "perfil raiz quadrada"[115] $(\sqrt{p_{kj1}},...,\sqrt{p_{kjm_j}})$ (Bacelar-Nicolau, 2000).

Pode-se verificar que o coeficiente de afinidade *aff*(ξ_{kj},$\xi_{k'j}$) toma valores no intervalo [0,1], sendo igual a 1 se as distribuições de frequências ξ_{kj} e $\xi_{k'j}$ são idênticas ou proporcionais, e igual a zero se elas forem ortogonais.

Bacelar-Nicolau (2000, 2002) propõe ponderar a variável Y_j de que resulta, a partir da Definição 3.6.1, a definição de um coeficiente de semelhança afinidade ponderada entre unidades de dados que se passa a apresentar.

Definição 3.6.2. Coeficiente de afinidade ponderado entre unidades de dados descritas por uma variável simbólica modal

O coeficiente de afinidade ponderado, a(k,k'), entre as unidades de dados k, k'$\in$ E, resulta de ponderar a importância da variável modal Y_j (com m_j modalidades) com o peso w_j ($0\leq w_j \leq 1$ e $\sum_{j=1}^{p} w_j = 1$):

$$a(k,k') = \sum_{j=1}^{p} w_j . \, aff(\xi_{kj}, \xi_{k'j}) = \sum_{j=1}^{p} w_j . \sum_{\ell=1}^{m_j} \sqrt{\frac{n_{kj\ell}}{n_{kj.}} . \frac{n_{k'j\ell}}{n_{k'j.}}} \qquad (3.6.2)$$

sendo, $n_{kj.} = \sum_{\ell=1}^{m_j} n_{kj\ell}$ o número de indivíduos que pertencem (ou observados em) à unidade de dados (grupo) k para a qual a variável Y_j foi observada; m_j representa a dimensão do ℓ'ésimo "estrato".

[115] *Square root profile*, no original.

O coeficiente de semelhança a(k,k') mede a tendência monótona entre as raízes quadradas dos perfis das unidades de dados k e k' e verifica-se que toma sempre valores no intervalo [0,1] (Bacelar-Nicolau, 2000; Bacelar-Nicolau, 2002). Em Nicolau e Bacelar-Nicolau (1999) podem encontrar-se algumas escolhas possíveis para os pesos.

Bacelar-Nicolau (2000, 2002) chama a atenção para os seguintes casos particulares:

1. Se todas as variáveis modais tiverem o mesmo número m_j =m de categorias então c=pm.
2. Se todas as células da linha k tiverem o mesmo total de frequências, $n_{kj.}=n_k$, então $n_{k..}=pn_k$.
3. No caso de se verificarem 1 e 2 simultaneamente, então:

$$a(k,k') = \sum_{j=1}^{p} w_j . \sum_{\ell=1}^{m} \sqrt{\frac{n_{kj\ell}}{n_{k..}} . \frac{n_{k'j\ell}}{n_{k'..}}} \qquad (3.6.3)$$

4. Se todas as variáveis Y_j tiverem o mesmo peso, $w_j=w$, então $w_j=w=1/p$.
5. No caso de se verificarem 1, 2 e 4 simultaneamente, então:

$$a(k,k') = \sum_{j=1}^{p} \sum_{\ell=1}^{m} \sqrt{\frac{n_{kj\ell}}{n_{k..}} . \frac{n_{k'j\ell}}{n_{k'..}}} \qquad (3.6.4)$$

   Neste caso reencontra-se a definição do coeficiente de afinidade básico entre duas distribuições de frequências das unidades de dados k e k'.

O cálculo destes coeficientes de afinidade pode-se realizar recorrendo, por exemplo, ao programa em Fortran, *Afsimb* (Sousa, 2005).

A definição do coeficiente de afinidade entre variáveis simbólicas é análoga à apresentada quando se comparam objectos simbólicos (Nicolau *et al*., 2007):

Definição 3.6.3. Coeficiente de afinidade generalizado entre variáveis simbólicas ou complexas

O coeficiente de afinidade generalizado entre as variáveis k, k' do conjunto V é a média ponderada das afinidades locais entre k e k' sobre a j-ésima unidade de dados (j=1, ..., n), aff(k, k'; j):

$$a(k,k') = \sum_{j=1}^{p} \pi_j . \; aff(k, k'; j) = \sum_{j=1}^{n} \pi_j . \sum_{\ell=1}^{m_j} \sqrt{\frac{x_{jk\ell}}{x_{jk.}} . \frac{x_{jk'\ell}}{x_{jk'.}}} \qquad (3.6.5)$$

em que m_j representa o número de "modos" ou "situações" para a j-ésima unidade de dados, $x_{jk\ell}$ depende do tipo de variáveis e $0 \leq \pi_j \leq 1$, $\sum \pi_j = 1$.

Também se pode definir um coeficiente de afinidade global (e.g., Nicolau e Bacelar-Nicolau, 1999; Sousa *et al*., 2007), i. e., uma medida de afinidade p-multivariada entre os objectos simbólicos O_k e $O_{k'}$ (k, k'= 1, ...,n) descritos por p variáveis simbólicas, Y_j, j=1, ..., p, que é dada pela expressão:

$a(k,k') = \sum_{1\le j,j'\le p} \pi_{jj'} a_{jj'}^{kk'}$ (3.6.6), sendo $\pi_{jj'}$ os pesos ($\pi_{jj'} \ge 0$ e $\sum_{1\le j,j'\le p} \pi_{jj'} = 1$) e $a_{jj'}^{kk'}$ a afinidade local entre as distribuições de frequências ξ_{kj} e $\xi_{k'j'}$, correspondentes às variáveis j e j', extensão da Definição 3.6.1, $a_{jj'}^{kk'} = aff(\xi_{kj}, \xi_{k'j'}) = \sum_{\ell=1}^{m_j} \sqrt{\frac{n_{kj\ell}}{n_{kj.}} \frac{n_{k'j'\ell}}{n_{k'j'.}}}$, 1≤ j, j'≤ p, m_j=$m_{j'}$, e $n_{kj.} = \sum_{\ell=1}^{m_j} n_{kj\ell}$ ($n_{kj'.} = \sum_{\ell=1}^{m_j'} n_{kj'\ell}$) é o número de indivíduos pertencentes à unidade (grupo k) para a qual a variável Y_j ($Y_{j'}$) foi observada. Esta afinidade p-multivariada é um valor da tabela pxp associada ao par de objectos simbólicos, $A_{kk'} = (a_{jj'}^{kk'})$, 1≤ j, j'≤ p.

O coeficiente de afinidade global a(k,k') toma valores no intervalo [0, 1] (Sousa *et al*., 2007), podendo ser aplicado quando as variáveis são intervalares (ver subsecção seguinte) ou quando são de vários tipos (ver Subsecção 3.6.3).
De forma análoga se define o coeficiente de afinidade global entre duas variáveis de dimensão n, também designado por afinidade n-multivariada (Nicolau *et al*., 2007).

3.6.2 Generalização do coeficiente de afinidade a dados simbólicos intervalares

A definição do coeficiente de afinidade entre objectos simbólicos descritos por variáveis intervalares constitui um caso particular do coeficiente de afinidade generalizado ponderado entre unidades de dados descritas por variáveis modais (Definição 3.6.2 e Expressão 3.6.5), tal como é apresentada, por exemplo, em Sousa *et al*. (2007). Antes do cálculo do coeficiente de afinidade generalizado ponderado codifica-se a matriz de dados por discretização das variáveis intervalares, associando a cada um dos intervalos uma distribuição de frequências relativas apropriada; em Sousa *et al*. (2007) encontra-se descrito um dos processos possíveis (Sousa, 2005) para realizar essa codificação.

Estes cálculos também estão implementados no programa *Afsimb* (Sousa, 2005).

3.6.3 Generalização do coeficiente de afinidade a dados heterogéneos

Quando as unidades de dados são descritas por variáveis de vários tipos (frequências, binárias, reais, ...), Bacelar-Nicolau (2000, 2002) sugere que se disponham em grupos de variáveis homogéneas (o que corresponde a dissecar a matriz de dados total em submatrizes que correspondem a esses grupos). Então, pode-se definir um coeficiente de afinidade ponderado entre duas unidades de dados k e k' por uma combinação linear convexa das semelhanças que se referem às diferentes submatrizes, sendo todas elas casos particulares da mesma (i.e., a afinidade). Neste caso, a afinidade entre unidades de dados é uma média ponderada de afinidades parciais. Esta é precisamente a vantagem da utilização deste coeficiente – usa-se sempre a afinidade!

Esta generalização também se encontra, por exemplo, em Sousa (2005) e exemplos de aplicação, em Análise Classificatória, em Nicolau (2002) e Nicolau e Bacelar-Nicolau (2005).

3.6.4 Generalização do coeficiente de afinidade centrado e reduzido pelo método-WW ao caso de dados heterogéneos e de natureza complexa

No enquadramento da análise classificatória probabilística a três dimensões, é introduzida a generalização do coeficiente de afinidade centrado e reduzido pelo método-WW ao caso de dados heterogéneos e de natureza complexa (Nicolau *et al*., 2007):

- A variável aleatória aff(k,k';j) (Definição 3.6.3) tem distribuição assintoticamente normal. Sob a hipótese de referência permutacional R baseada no teorema limite de Wald e Wolfowitz, o valor médio e a variância assintóticos são, respectivamente:

$$\mu^{*}_{WW}(k,k') = \frac{1}{m_j}\sum_{\ell=1}^{m_j}\sqrt{\frac{x_{jk\ell}}{x_{jk.}}}\ \sum_{\ell=1}^{m_j}\sqrt{\frac{x_{jk'\ell'}}{x_{jk'.}}} \qquad (3.6.7)$$

$$\sigma^{*2}_{WW}(k,k';\ j)=\frac{1}{\mathrm{m_j}-1}\sum_{\ell=1}^{m_j}\left(\sqrt{\frac{x_{jk\ell}}{x_{jk.}}}-\frac{1}{m_j}\sum_{\ell'=1}^{m_j}\frac{x_{jk\ell'}}{x_{jk.}}\right)^2\times\sum_{\ell=1}^{m_j}\left(\sqrt{\frac{x_{jk'\ell}}{x_{jk'.}}}-\frac{1}{m_j}\sum_{\ell'=1}^{m_j}\frac{x_{jk'\ell'}}{x_{jk'.}}\right)^2 \qquad (3.6.8)$$

<u>Definição 3.6.4. O coeficiente de afinidade centrado e reduzido pelo método-WW generalizado entre variáveis simbólicas</u>

O coeficiente de afinidade centrado e reduzido pelo método-WW generalizado entre as variáveis k, k'$\in$V é definido pela expressão:

$$a_{WW}(k,k') = a^{*}(k,k') = \sum_{j=1}^{n}\pi_j \cdot aff^{*}_{WW}(k,k';\ j)$$

sendo, $aff^{*}_{ww}(k,k';\,j) = \dfrac{aff(k,k';\,j) - \mu^{*}_{ww}(k,k';\,j)}{\sigma^{*2}_{ww}(k,k';\,j)}$,

o valor médio $\mu^{*}_{ww}(k,k';\,j)$ e a variância $\sigma^{*2}_{ww}(k,k';\,j)$ definidos respectivamente pelas expressões 3.6.7 e 3.6.8.

3.6.5 O coeficiente de afinidade generalizado probabilístico

De forma análoga à que se apresentou anteriormente, também se define o coeficiente de afinidade generalizado probabilístico do tipo VL, i. e., o coeficiente de validade da afinidade (e.g., Bacelar-Nicolau, 2000, 2002; Sousa *et al*., 2007).

Definição 3.6.5. Coeficiente de afinidade generalizado probabilístico α_R entre unidades de dados

Considere-se que as unidades de dados são amostras aleatórias duma população e os valores da afinidade a(k,k') entre as unidades de dados descritas por variáveis simbólicas modais são realizações de uma variável aleatória A(k,k') com uma distribuição de probabilidade de referência *R*. O coeficiente de afinidade generalizado probabilístico entre as unidades de dados k, k' é definido por:

$$\alpha_R(k,k') = P_R\left(A(k,k') \le a(k,k')\right)$$

Recorde-se que este coeficiente também é designado por validade da ligação ou coeficiente de semelhança VL.

O coeficiente de semelhança probabilístico α_R mede a validade da afinidade entre as unidades de dados simbólicos k, k', numa escala probabilística. Quanto maior for o coeficiente α_R, maior é a probabilidade dos valores de semelhança A(k,k') serem inferiores ao valor observado, sob R.

No caso da hipótese de referência R coincidir com o conjunto de condições associadas ao teorema limite de Wald e Wolfowitz, α_R pode ser aproximado por:

$$\alpha_R(k,k') = P_R\left(A^{*}(k,k') \le a^{*}(k,k')\right) \cong \Phi\left(a^{*}(k,k')\right),$$

onde $\Phi(.)$ é a função de distribuição da distribuição normal reduzida.

De forma idêntica se define o coeficiente de afinidade probabilístico entre duas variáveis simbólicas (Nicolau *et al*., 2007). Um valor elevado do coeficiente probabilístico significa que

o valor “observado” da afinidade é “significativamente” maior do que o que se poderia esperar sob a hipótese R.

3.7 O coeficiente de afinidade em Análise de Dados

O coeficiente de afinidade tem sido utilizado nos diversos métodos de Análise de Dados.
Quer o coeficiente de afinidade básico, quer os coeficientes de afinidade generalizados e probabilísticos foram introduzidos no âmbito da Análise Classificatória com o objectivo de desenvolver modelos classificatórios, hierárquicos ou não hierárquicos, e medir a proximidade entre unidades de dados ou variáveis (e.g., Bacelar-Nicolau, 1980; Nicolau, 1980; Sousa, 2005; Nicolau *et al*., 2007).

H. Bacelar-Nicolau, F. Costa Nicolau e alguns dos seu alunos, de doutoramento ou de mestrado, têm complementado o estudo do coeficiente de afinidade com o objectivo de definir “métodos de classificação robustos” (e.g., Doria, 1989; Bacelar-Nicolau e Nicolau, 1993; Nicolau, 2002; Nicolau e Bacelar-Nicolau, 2003; Sousa, 2005), estando presente a preocupação da validação (e.g., Sousa *et al.*, 2007) e propostos modelos probabilísticos baseados na afinidade, em análise classificatória de dados complexos/simbólicos ou tridimensional (e.g., Nicolau *et al*., 2007). Como já se referiu, a abordagem probabilística opera num contexto exploratório que usa o conhecimento prévio da estrutura dos dados como um instrumento para extrair informação acerca da sua estrutura de agrupamento hierárquico (e.g., Nicolau *et al*., 2007).

O coeficiente de afinidade básico e o coeficiente de afinidade centrado e reduzido pelo método-WW incluídos em algoritmos de ACHA foram testados na presença de dados omissos e imputados revelando um comportamento melhor do que o coeficiente de correlação de Pearson (e.g., Silva, 2005).

Em análise classificatória, o coeficiente de afinidade também foi usado como medida de discrepância entre as distribuições que modelam cada classe da partição e usado como um critério para escolher o melhor número de classes num modelo de mistura de distribuições normais. Neste sentido, uma vez introduzida a noção de peso da classe, foi obtida uma nova expressão analítica do coeficiente de afinidade para misturas gaussianas, o coeficiente de afinidade ponderado (Soromenho e Bacelar-Nicolau, 1999). O desempenho do coeficiente de afinidade ponderado para misturas gaussianas revelou-se melhor do que o do coeficiente de afinidade de Matusita (1977); por outro lado, quanto maior for a diferença entre os pesos

das classes melhor é o desempenho daquele coeficiente, embora ambos os coeficientes tendam a sobrestimar o número de classes (Soromenho e Bacelar-Nicolau, 1999).

Alguns programas informáticos permitem calcular os coeficientes de afinidade. Além dos já referidos, mais recentemente, o programa *ProxMed* (Sousa, 2005) também permite calcular os coeficientes de afinidade "clássicos"; além disso, aqueles coeficientes estão programados em linguagem SAS (Nicolau, 2002), sendo a matriz de semelhanças afinidade transformada numa matriz de distância, através da função quadrática, de forma a que possa ser usado nos algoritmos de ACHA deste *software*.
Os coeficientes de afinidade foram também implementados no programa LEASP de análise estatística de dados, que está mais direccionado para o ensino e formação; este programa é a versão portuguesa completada do programa LEAS francês (desenvolvido em protocolo entre a Unité de Biométrie da Université de Montpellier, o Laboratório de Biometria do Departamento de Matemática da Universidade de Aveiro e o Laboratório de Estatística e Análise de Dados da Universidade de Lisboa) (e.g., Dias, 1994).

Contudo, a utilização do coeficiente de afinidade não se limita à análise classificatória. Como se verá, na Subsecção 4.3.10.9, permite definir a matriz *score* associada à variável modal, e na Secção 5.4 foi aplicada uma ACP directamente sobre a matriz de semelhanças VAL_{Aw}. Além disso, o coeficiente de afinidade básico, também tem sido usado em Análise Discriminante Discreta (e.g., Sousa Ferreira, 2000; Sousa Ferreira *et al*., 2001; Brito *et al*., 2006).

As medidas de distância associadas à afinidade entre populações, desenvolvidas por Matusita têm sido estudadas por diversos autores. Em particular, Krzanowski (1983) refere Matusita (1956, 1964, 1967, 1972)[116] e define uma distância entre populações utilizando variáveis mistas, a partir da afinidade ρ_{ij} entre populações π_i e π_j descritas por variáveis mistas – contínuas e categóricas – com distribuições normais multivariadas e multinomiais, respectivamente. Aquela distância resulta, como anteriormente, da transformação quadrática, $\Delta_M(\pi_i,\pi_j)=\sqrt{2(1-\rho)}$. No exemplo apresentado, Krzanowski (1983) aplica o *multidimensional scaling* métrico (com referência a Mardia *et al*. (1979, Capítulo 14)) às matrizes de distância que obteve.

[116] Estas referências encontram-se em Krzanowski (1983).

Em breve pretendemos obter representações euclidianas, quer dos indivíduos, quer das variáveis, recorrendo à ACP da matriz de afinidade, assim como a técnicas de *multidimensional scaling*. No caso dos dados serem compositivos, estas poderão ser alternativas à ACP proposta por Aitchison (1983) para este tipo de dados. Também é nosso objectivo futuro a visualização das relações entre indivíduos e entre variáveis através de *biplots*, utilizando a metodologia *multidimensional scaling* ponderada (Greenacre, 2005) sobre a matriz de distâncias obtidas a partir da afinidade através da função circular.

4 OS COEFICIENTES S, S_{LC} E P_L

Notre propos n'est pas de définir un indice de plus venant s'ajouter à la longue liste de ceux déjà connus, mais au contraire d'en trouver un pouvant recouvrir la plupart. Mais, surtout, il devra satisfaire à deux contraintes trop rarement prises en considération: la "structure" et l'hétérogeneité des caractères.
(Le Calvé, 1977)

4.1 Introdução

Quando pretendemos analisar informação registada em bases de dados, por exemplo de medicina, biologia e psicologia, recorremos frequentemente a métodos de análise de dados multivariados e somos, habitualmente, confrontados com o problema da heterogeneidade da natureza das variáveis. Isto porque a informação relevante para estas disciplinas engloba muitas vezes variáveis qualitativas (e.g., sexo, nível de educação), quantitativas (e.g., peso, idade) e/ou simbólicas/complexas (e.g., intervalos de temperaturas) o que torna uma análise global das suas relações particularmente complicada ou limitada[117]. O objectivo geral da abordagem que apresentamos neste capítulo é o de dar resposta a este problema. Neste contexto, são necessários coeficientes que permitam comparar variáveis heterogéneas[118].

Como já referido, existem alguns coeficientes que dão resposta a este problema usando métodos baseados nas mesmas linhas de orientação (Daniels, 1944; e.g., Lerman 1973; e.g., Bacelar-Nicolau, 2000; Ouali, 1991a; Le Calvé, 1977). Neste capítulo apresentamos os coeficientes propostos por Le Calvé (1977) e por nós generalizados a outros tipos de variáveis.

---

[117] Um procedimento habitual consiste em transformar as variáveis quantitativas em qualitativas, i.e., em variáveis do mesmo tipo, como se viu no Capítulo 1.

[118] *Variables d'un type quelconque, variables de types divers, variables de types différents* (Lerman, 1987; Lerman e Peter, 2003), *variables hétérogènes* (Le Calvé, 1977; Chah, 1984, 1985), *variables de types hétérogènes, heterogenous variables* (Lerman e Peter, 2003; Chah, 1985), *variables mixtes* (Pagès, 2004).

Os coeficientes s, s_{LC} e P_L, propostos por Le Calvé, permitem relacionar variáveis, quer da mesma natureza, quer de natureza diferente (Le Calvé, 1977; Doria *et al*. 2006), entrando em consideração com a sua estrutura. Estes coeficientes também permitem relacionar dados sob a forma de matrizes, em particular, matrizes ultramétricas (Doria *et al*. 1999, 2000). Análises deste tipo são pouco contempladas nas abordagens tradicionais.

Nesta abordagem, a cada variável é associada uma matriz *score*, cuja definição depende da natureza da variável, assim como da natureza da variável com a qual a pretendemos comparar (ou seja, a definição depende da natureza das duas variáveis que se pretendem comparar). Surgem então as definições dos coeficientes s, s_{LC} e P_L que apresentamos na Secção 4.2. As definições das matrizes *score* associadas às variáveis a serem comparadas, quer sejam da mesma natureza (Secção 4.3) ou não (Secção 4.4), serão apresentadas com a interpretação que se poderá dar a estes coeficientes quando utilizados. Na Secção 4.5, a matriz *score* é definida no caso dos dados serem matrizes. A Secção 4.6 apresenta uma sugestão para o tratamento de dados omissos, em casos particulares, recorrendo àqueles coeficientes. Uma conclusão sobre a interpretação geral destes coeficientes é apresentada na Secção 4.7. Finalmente, na Secção 4.8 sugerimos a utilização dos coeficientes s_{LC} e P_L no contexto da inferência estatística. A definição destes coeficientes é acompanhada por exemplos didácticos e/ou exemplos com dados reais, sobretudo das áreas da Medicina, Psicologia e Sociologia.

Os coeficientes s, s_{LC} e P_L inspiraram-se numa ideia de Daniels (1944), retomada por Lerman (1973) e generalizada por Le Calvé (1977). Esta ideia enquadra-se na problemática da comparação de variáveis qualitativas, duas a duas, observadas sobre um conjunto de unidades estatísticas, E, e traduz-se em considerar as variáveis representadas por relações binárias sobre o conjunto E, sendo estas representadas pelas respectivas matrizes. A partir de um coeficiente de associação básico, s – o produto escalar daquelas matrizes –, a variável aleatória correspondente é centrada e reduzida. O coeficiente de Daniels (1944), tal como Le Calvé (1977) o apresentou, é a correlação entre duas variáveis, representadas pelas suas matrizes de zeros e uns vectorizadas e pode escrever-se sob a forma:

$$u_0 = \frac{\sqrt{n^2-1}\left(\sum_{ij} x_{ij}y_{ij} - \frac{\widetilde{X}\widetilde{Y}}{n^2}\right)}{\sqrt{\sum_{ij}\left(x_{ij} - \frac{\widetilde{X}}{n^2}\right)^2 \sum_{ij}\left(y_{ij} - \frac{\widetilde{Y}}{n^2}\right)^2}} \qquad (4.1.1) \quad \text{(Le Calvé, 1977)}$$

Sendo, “X e Y duas variáveis sobre o conjunto de unidades estatísticas E, isto é, duas relações binárias, que são representadas pelas suas matrizes $X = \left(x_{ij}\right)$ e $Y = \left(y_{ij}\right)$. Reproduzindo o método descrito acima, escolheremos *s* como sendo o cardinal da intersecção dos grafos de X e de Y em E^2, e consideremos a variável aleatória Card(A∩B), em que A e B são subconjuntos de E^2 que têm o mesmo número de elementos de X e de Y respectivamente, isto é, matrizes que têm o mesmo número de uns que X e Y. A variável aleatória S tem uma distribuição hipergeométrica com valor médio $\mu = \frac{\widetilde{X}\widetilde{Y}}{n^2}$ (4.1.2) e variância $\sigma^2 = \frac{\widetilde{X}\widetilde{Y}\left(n^2 - \widetilde{X}\right)\left(n^2 - \widetilde{Y}\right)}{\left(n^2 - 1\right)n^4}$ (4.1.3), designando por $\widetilde{X}$ (respectivamente $\widetilde{Y}$), o número de uns da matriz X (respectivamente Y) obtemos a expressão do índice utilizado e estudado por Daniels (1944)” (Le Calvé, 1977).

Da expressão do coeficiente de Daniels (1944), tal como ele o apresentou, surgem os coeficientes de correlação de Pearson, de Spearman e de Kendall, como casos particulares.

Le Calvé faz notar que, embora o coeficiente de Daniels (1944) entre em conta com o número de uns das matrizes X e Y, ele não entra em conta com a sua disposição, i.e., com a estrutura das duas variáveis: “Uma relação de ordem pode ter o mesmo número de uns que uma relação de equivalência e serem totalmente diferentes, uma das matrizes é simétrica e a outra não o é” (Le Calvé, 1977).

Lerman (1973) propõe um índice de proximidade geral entre variáveis do mesmo tipo (casos das variáveis “atributo de descrição”[119],”característica descritiva”[120], variável “característica com modalidades totalmente ordenadas”[121], variável ”número de ordem”[122] e a variável ”medida”[123]), utilizando a mesma ideia, a partir do coeficiente proposto por Kendall[124] (1970) para relacionar variáveis ordinais. Lerman (1973) demonstrou o teorema da dualidade para o caso das variáveis X e Y serem da mesma natureza e definiu os coeficientes, s, s_{LC} e P_L, entre as referidas variáveis, sob uma hipótese N de “ausência de ligação”[125], que corresponde a uma hipótese de independência.

Le Calvé (1977) demonstrou de uma forma simples o teorema da dualidade de Lerman (1973), generalizando-o para variáveis heterogéneas; calculou o valor médio e a variância

---

[119] *Attribut de description*, no original.
[120] *Caractère descriptif*, no original.
[121] *Caractère aux modalités totalement ordonnées*, no original.
[122] *Rang*, no original.
[123] *Mesure*, no original.
[124] *Coefficient de corrélation des rangs* τ *de Kendall*, no original.
[125] *Absence de liaison*, no original.

assintóticas da variável aleatória S; demonstrou que a distribuição de probabilidade do coeficiente S_{LC} (variável aleatória S padronizada) tende, sob condições bastante gerais, para a lei normal centrada e reduzida; deu-lhe uma forma explícita nos casos clássicos e obteve um algoritmo de cálculo nos outros casos; encontrou também, como casos particulares, diversos coeficientes conhecidos e mostrou como eles se podem aplicar a casos muito frequentes na prática, que não são contemplados pelos coeficientes tradicionalmente utilizados.

Enquanto que, a abordagem de Lerman (1973) é combinatória e trata da comparação de variáveis do mesmo tipo, a abordagem introduzida por Le Calvé (1977) é probabilística, o que lhe permitiu generalizar os coeficientes s, s_{LC} e P_L à comparação de variáveis heterogéneas. Na realidade o coeficiente s_{LC} definido por Le Calvé coincide com o definido por Lerman nos casos em que se comparam alguns tipos de variáveis da mesma natureza, como veremos na Secção 4.2.

Lerman (1977, 1981) retoma o estudo, de forma combinatória, do índice aleatório S_1 (variável aleatória S no caso de comparar um par de valores[126] e obtém uma expressão formal da distribuição de S_1. Em Lerman (1981), em Ouali (1991a) e Lerman (1992a) encontramos referência, respectivamente, a Le Calvé (1976) e a Le Calvé (1977)[127].

Como vimos na Subsecção 1.5.4, I.C. Lerman prosseguiu, e orientou outros colaboradores, na investigação em Classificação[128] e naturalmente foi conduzido a propor, quer um índice de semelhança entre objectos descritos por variáveis de vários tipos (Lerman, 1987; Lerman e Peter, 2003) cuja programação (programa SIMOB) se deve a Peter (Lerman e Peter, 1985), quer coeficientes de associação entre variáveis qualitativas, variáveis relacionais, e entre variáveis simbólicas (Ouali-Allah, 1991a; Lerman, 1992a), cuja programação (programa AVARE) se deve a Ouali-Allah (1991b). Embora a ideia geral destes coeficientes de associação entre variáveis seja a mesma que a apresentada por Le Calvé (1977), o índice de semelhança bruto é calculado a partir de matrizes que se baseiam na noção de preordenação[129], entre os pares de indivíduos, induzida pela relação binária associada a cada uma das variáveis.

---

[126] *Valuations*, no original.

[127] Este artigo de Le Calvé aparece referenciado em Lerman (1992) e em Ouali (1991a) com a data de 1976.

[128] Relembramos que a Classificação, segundo a escola francesa, também pode ser designada por Análise Classificatória (hierárquica ou não hierárquica). Não confundir com Análise Discriminante.

[129] *Préordonnance*, em francês. Termo introduzido por Lerman. Preordenação sobre um conjunto E com n elementos é uma relação de preordem total definida sobre o conjunto dos pares de elementos de E.

Neste capítulo, trata-se com maior detalhe do que Le Calvé, a comparação, quer de variáveis com modalidades parcialmente ordenadas (Subsecção 4.3.4), quer de variáveis heterogéneas (Secção 4.4) e a utilização destes coeficientes é por nós estendida e aplicada a alguns tipos de variáveis simbólicas (Subsecção 4.3.10). Generaliza-se assim a aplicação destes coeficientes a situações em que as unidades estatísticas são indivíduos/classes/grupos de indivíduos, descritas por variáveis intervalo, por variáveis de resposta múltipla e por variáveis modais (distribuições de probabilidades e de frequências relativas/barras e/ou histogramas).
Quanto maior for a dimensão da amostra ou da população a ser estudada, mais "pesados" se tornarão os cálculos que nos permitem obter os valores dos coeficientes s, s_{LC} e P_L . O facto de introduzirmos a possibilidade de analisar classes de indivíduos permite ultrapassar o problema da dimensão da matriz de dados, principalmente no que se refere ao número de indivíduos.

Estes coeficientes foram aplicados a dados reais com a utilização de *software* apropriado e programado para este efeito, inicialmente em Fortran7 e posteriormente em MATLAB 6.5 (*Toolbox Coeficientes s*, s_{LC} e P_L).

4.2 Os coeficientes s, s_{LC} e P_L

4.2.1 Introdução

A matriz ExV, (nxp), em que se descreve um conjunto de n unidades estatísticas (sujeitos ou grupos/amostras/populações de sujeitos), E, por um conjunto de p variáveis heterogéneas, V, constitui o ponto de partida de muitos dos estudos em análise multivariada (Tabela 4.2.1).

Tabela 4.2.1. Matriz de dados ExV, em que x_i é o valor que a variável X toma para a unidade estatística i. A unidade estatística i pode ser um indivíduo ou um conjunto de indivíduos

	Variáveis			
Unidades estatísticas	X	Y	…	W
1	x_1	y_1	…	w_1
2	x_2	y_2	…	w_2
...	...	...	...	...
i	x_i	y_i	...	w_i
...	...	...	...	...
n	x_n	y_n	…	w_n

Conforme o tipo de variável presente no estudo, a variável X pode tomar um único valor, x_i, para a unidade estatística i, ou vários valores, como por exemplo, $x_i = (x_{i1}, x_{i2}, ...x_{ik})$, no caso de X ser uma variável com k valores múltiplos, ou $x_i \in [a, b]$, no caso de X ser uma variável simbólica intervalar. Quando todas as variáveis são simbólicas/complexas (conjunto de valores) a matriz de dados ExV é uma matriz tridimensional[130] (Exemplo 4.3.28). Aqui optou-se por uma notação simplificada; contudo, quando as variáveis são simbólicas a notação torna-se um pouco mais "pesada" (e.g., Tabela 3.6.1, Capítulo 3).

A cada variável é associada uma matriz *score*, cuja definição depende da natureza da variável, assim como da natureza da variável com a qual a pretendemos comparar. O coeficiente básico, s, define-se como o produto escalar das matrizes *score*, o coeficiente s_{LC} é o coeficiente *s* padronizado, sob determinada hipótese de referência e o coeficiente P_L é o correspondente coeficiente probabilístico.

4.2.2 Definição de *score*

Seja a variável $X = (x_1, x_2, ...x_n)$, em que x_i é o valor que a variável X toma para a unidade estatística i. Como já foi referido acima, a variável X pode tomar um único valor, x_i, para a unidade estatística i, ou vários valores no caso da variável ser simbólica/complexa.
A representação da variável X depende da sua natureza. Se a variável for real/métrica pode ser representada por um ponto de $\mathbb{R}^n$; se a variável X for atributo de descrição será um valor de $\{0,1\}^n$.

Definição 4.2.1. Definição de *score*

Para cada uma das p variáveis define-se a função f de E x E em $\mathbb{R}$,

$$f: E \times E \to \mathbb{R}$$
$$(i, i') \to x_{ii'} \qquad (4.2.1)$$

tal que, $x_{ii} = 0$, $\forall i \in E$.

Esta função é designada por *score* e pode ser representada sob a forma de matriz (Tabela 4.2.2).

[130] *Three-way generalized data table*, em inglês (e.g., Nicolau *et al*., 2007).

Tabela 4.2.2. Matriz *score*, X (nxn), em que $x_{ii'}$ designa o *score* atribuído às unidades estatísticas i e i' pela variável X

	Unidades estatísticas				
Unidades estatísticas	1	...	i'	...	n
1	0	...	$x_{1i'}$	...	x_{1n}
...		...	…	...	…
i	x_{i1}		$x_{ii'}$		x_{in}
...	...		...		...
n	x_{n1}	...	$x_{ni'}$	...	0

O *score* da variável X é, pois, uma função sobre os pares de unidades estatísticas descritos pela variável X. Por exemplo, quando a variável X representa uma relação binária[131] sobre o conjunto das unidades estatísticas, E, tomaremos frequentemente, como *score* a matriz associada a esta relação, considerando sempre $x_{ii} = 0$. Neste caso, dizemos que duas unidades estatísticas i e i' estão em relação, para a matriz *score* X, se $x_{ii'} = 1$. Ou seja, dizer que duas unidades estatísticas, i e i', estão em relação significa que iRi' é verdadeiro.
A definição da matriz *score* da variável depende da sua natureza, assim como da natureza da outra variável com a qual a pretendemos comparar. A matriz de dados ExV (nxp) dá origem a, pelo menos, p matrizes *score*, (nxn), tantas ou em maior número do que as variáveis presentes no estudo, conforme a natureza dessas variáveis.

Como as relações finitas podem ser representadas por matrizes booleanas, isto é, por matrizes cujas entradas são constituídas por zeros e uns, algumas matrizes *score* sê-lo-ão, outras serão matrizes numéricas, de semelhança ou não. Mais à frente serão apresentadas as definições destas matrizes para cada uma das situações contempladas neste estudo.

4.2.3 Definição dos coeficientes s, s_{LC} e P_L

Uma vez definidas as matrizes *score*, podemos definir o coeficiente de semelhança básico s.

Definição 4.2.2. O coeficiente de semelhança *s*

O coeficiente de semelhança *s* entre as variáveis X e Y, $s_{X,Y}$, é definido pelo produto escalar das respectivas matrizes *score*:

$$s_{X,Y} = \sum_{i=1}^{n} \sum_{i'=1}^{n} x_{ii'} y_{ii'} \qquad (4.2.2)$$

sendo: $(i, i') \rightarrow x_{ii'}$, $(i, i') \rightarrow y_{ii'}$

[131] Uma relação binária definida sobre um conjunto E é uma parte R de ExE, $R \subseteq$ ExE. Se $(x,y) \in R$, dizemos que x está em relação com y e representa-se por xRy.

Ou, utilizando a notação matricial, para as funções $(i,i') \rightarrow x_{ii'}$ e $(i,i') \rightarrow y_{ii'}$, $s_{X,Y}$ é o traço da matriz produto XY^T (i.e., a soma dos elementos da sua diagonal principal):

$$s_{X,Y} = < X ,Y > = Tr\ (XY^T) \qquad (4.2.3)$$

em que Y^T designa a matriz transposta de Y e Tr o traço da matriz produto XY^T.

O coeficiente s toma valores num conjunto finito quando se relacionam, por exemplo, variáveis atributos de descrição ou variáveis nominais ou variáveis ordinais.

Para que se possa ter em conta a estrutura das duas variáveis, Le Calvé (1977) define matrizes do mesmo tipo, baseando-se em ideias de Lerman.

Definição 4.2.3. Matrizes do mesmo tipo

Seja $\Theta(E)$ o conjunto das permutações[132] definidas em E.

Seja θ, $\theta \in \Theta(E)$, uma matriz de permutação, nxn.

Sejam X e X' duas variáveis definidas sobre E, representadas pelas matrizes $X = (x_{ij})$ e $X' = (x'_{ij})$, respectivamente. Dizemos que elas são do mesmo tipo se se deduzirem uma da outra por uma permutação, i.e., se existe uma matriz de permutação θ sobre E tal que: $X' = \theta \times \theta'$.

A estrutura é a mesma quando se passa de X para X'. Assim, entramos em linha de conta com a estrutura de uns, além de termos em conta o número de uns. Como são permutações sobre as unidades estatísticas, as variáveis tomam os mesmos valores, têm as mesmas leis, mas não sobre as mesmas unidades estatísticas.

O índice de semelhança $s_{X,Y}$ é realização da variável aleatória $S_{X,Y}$. Para fazer o estudo desta variável é necessário ter presente a noção, já apresentada, de matrizes do mesmo tipo.

[132] Uma matriz de permutação só tem um 1 em cada linha e em cada coluna.

Definição 4.2.4. A variável aleatória $S_{x,y}$

Seja $\Theta(E)$ o conjunto das permutações definidas em E.

Seja θ, $\theta \in \Theta(E)$, uma matriz de permutação, nxn.

Seja $\Omega=\Theta(E) \times \Theta(E)$ o conjunto de todos os pares de permutações munido de uma medida de probabilidade uniformemente distribuída.

Para todo o par de variáveis, X e Y, associa-se a variável aleatória. $S_{X,Y}$:

$$S_{X,Y}(\omega) = S_{X,Y}(\theta,\theta') = \sum x_{\theta(i)\,\theta(i')} y_{\theta'(i)\,\theta'(i')} \quad , \quad \forall \omega = (\theta,\theta') \quad (4.2.4)$$

$S_{X,Y}(\theta,\theta')$ também se pode escrever sob a forma de produto interno:

$$S_{X,Y}(\theta,\theta') = \; < \theta X \theta^T,\ \theta' Y \theta'^T > \; = \; Tr\left(\theta X \theta^T . \theta' Y \theta'^T\right) , \forall \omega = (\theta,\theta') \qquad (4.2.5)$$

Le Calvé (1977) generaliza a todo o tipo de variáveis X e Y o teorema da dualidade de Lerman (1973), que o demonstrou sob uma outra forma, quando as duas variáveis são simultaneamente atributos ou equivalências ou preordens ou medidas[133]. Além disso, Le Calvé simplifica a sua demonstração por utilizar a forma matricial.

Teorema 4.2.1. Teorema da dualidade

$$S_{X,Y}(\theta,\theta') = S_{X,Y}\left(\theta'^T \circ \theta, I\right) \; = \; S_{X,Y}\left(I, \theta'^T \circ \theta\right) \forall X, \forall Y \quad (4.2.6)$$

Le Calvé demostra que acaba por ser a mesma coisa, escolher ao acaso as duas variáveis ou fixar qualquer uma das variáveis e escolher uma outra ao acaso entre as variáveis do mesmo tipo que a segunda.

Recorrendo ao teorema da dualidade, Le Calvé (1977) calculou o valor médio, μ_S, e a variância, σ_S^2, da variável aleatória $S_{X,Y}$, para variáveis de qualquer tipo:

$$\mu_S = \frac{\sum_{i,i'} x_{ii'} \sum_{i,i'} y_{ii'}}{n(n-1)} \qquad (4.2.7)$$

e

$$\sigma_S^2 = \frac{V_4 X . V_4 Y}{n(n-1)(n-2)(n-3)} + \frac{V_3' X . V_3' Y + V_3'' X . V_3'' Y + V_3''' X . V_3''' Y + V_3''' X . V_3''' Y}{n(n-1)n(n-2)} + \frac{V_2' X . V_2' Y + V_2'' X . V_2'' Y}{n(n-1)} - \mu_S^2$$

(4.2.8)

---

[133] *Mesures*, em francês.

Em que,

- $V_4 X = \sum_{i,j,k,l \in E^4} x_{ij} x_{kl}$ (4.2.9) designa o produto escalar das variáveis, em que os 4 índices são todos diferentes.

- $V_3 X$ representa o produto escalar, em que os 3 índices são todos diferentes:

$$V_3^{'} X = \sum_{i,j,l} x_{ij} x_{il} \quad , \quad V_3^{''} X = \sum_{i,j,k} x_{ij} x_{ki} \quad , \quad V_3^{'''} X = \sum_{i,j,l} x_{ij} x_{jl} \quad , \quad V_3^{''''} X = \sum_{i,j,k} x_{ij} x_{kj} \qquad (4.2.10)$$

- $V_2 X$ designa o produto escalar, em que os 2 índices são todos diferentes:

$$V_2^{'} X = \sum_{i,j} x_{ij} x_{ji} \quad , \quad V_2^{''} X = \sum_{i,j} x_{ij}^2 \qquad (4.2.11)$$

A fórmula da variância parece ser complicada mas, pelo facto de ser separável, permite calcular separadamente tudo o que se refere a X e a Y. O número de cálculos reduz-se de n^2 para n.

"A variável aleatória $S_{X,Y}$ = card(A∩B), em que A e B percorrem o conjunto das matrizes do mesmo tipo que X e Y, tem a mesma média que a variável de Daniels, mas a sua variância é diferente." (Le Calvé, 1977).

Os resultados são os mesmos que os encontrados por Lerman (1973) para variáveis do mesmo tipo.

Definição 4.2.5. A variável aleatória S_{LC}

Padronizando a variável aleatória $S_{X,Y}$ obtém-se a variável aleatória S_{LC}:

$$S_{LC}(\omega) = \frac{S_{X,Y}(\omega) - \mu_S}{\sigma_S} \qquad (4.2.12)$$

Quando n tende para o infinito, Le Calvé (1977) provou que, sob uma hipótese[134] de limites uniformes, a distribuição de probabilidade da variável aleatória S_{LC} tende para a lei normal centrada e reduzida, $S_{LC} \overset{a}{\cap} N(0,1)$.

[134] Hipótese H – Sejam A e B duas matrizes quadradas, nxn, tais que os elementos das respectivas diagonais principais sejam nulos, $a_{ii} = b_{ii} = 0$, $\forall i$. As quantidades a_{ij} e b_{ij} são uniformemente limitadas (isto é, quando n tende para o infinito, a distribuição dos valores pelas diversas

Teorema 4.2.2. Distribuição da variável aleatória S_{LC}

Designemos as modalidades das variáveis X e Y, respectivamente, por *a* e α. Se para todo o a (respectivamente, α), quando n tende para infinito, n_a/n (respectivamente, n_α/n) tendem para constantes k_a (respectivamente, k_α), então a distribuição de probabilidade da variável aleatória S_{LC}, $S_{LC}(\omega) = \frac{S_{X,Y}(\omega) - \mu_S}{\sigma_S}$, tende para a distribuição normal centrada e reduzida.

Definição 4.2.6. O coeficiente de semelhança s_{LC}

A partir da variável aleatória S_{LC}, considerando θ e θ‘ permutações identidade, $\theta=\theta$‘=I, obtém-se o coeficiente de semelhança s_{LC}:

$$s_{LC}(X,Y) = \frac{s_{X,Y} - \mu_S}{\sigma_S} \qquad (4.2.13)$$

O coeficiente s_{LC} é um coeficiente de semelhança entre variáveis que entra em consideração com a estrutura e a heterogeneidade das variáveis.

O coeficiente s_{LC} toma valores no conjunto dos números reais, $\mathbb{R}$, é pois um coeficiente de semelhança que toma valores positivos ou negativos, e verifica-se que:

- $s_{LC_{X,X}} \geq s_{LC_{X,Y}}$
- $s_{LC_{X,Y}} = s_{LC_{Y,X}}$

Quando a variância da variável aleatória $S_{X,Y}$ é nula, $\sigma^2_S=0$, o valor do coeficiente de semelhança s_{LC} é indeterminado. Esta situação ocorre, por exemplo, quando se pretende comparar variáveis que têm o mesmo valor para todas as unidades estatísticas. Assim sendo, estas variáveis deverão ser eliminadas do estudo. De qualquer forma, as variáveis que tomam o mesmo valor para todas as unidades estatísticas (o caso, por exemplo, de todas as unidades estatísticas darem a mesma resposta a essa variável) não têm interesse, sob o ponto de vista prático.

---

categorias é uniforme) e $\Sigma a_{ij}\, a_{il}$, $\Sigma b_{ij}\, b_{il}$, assim como as quantidades que delas se deduzem por permutação dos índices i, j, l, são de ordem n^3.

No entanto, podem existir outros casos em que a variância da variável aleatória S se anule, σ^2_S=0. Por exemplo, quando se relacionam duas variáveis heterogéneas, em que uma delas, X, é representada por uma matriz *score* simétrica[135] ($x_{ij}=1$ se iRj, $x_{ij}=0$ se não, $x_{ii}=0$) e a outra, Y, é representada por uma matriz *score* anti-simétrica[136] ($y_{ij}=1$ se iRj, $y_{ji}=-1$ se jRi, $y_{ij}=0$ se não, $y_{ii}=0$), a variância é nula, σ^2_S=0, e, neste caso, P_L=0.5.

Convém, pois, retirar do estudo, as variáveis que têm os valores todos iguais. Estes cálculos só deverão ser feitos entre variáveis que não apresentem variabilidade nula pois, além de trazerem problemas ao cálculo do coeficiente s_{LC}, também não têm interesse sob o ponto de vista prático.

A partir da variável aleatória S_{LC} pode definir-se um coeficiente de semelhança probabilístico.

Definição 4.2.7. O coeficiente P_L

O coeficiente de semelhança probabilístico, P_L, é o valor da função de distribuição de S_{LC} calculado em $s_{LC}(X,Y)$:

$$P_L(X,Y)=P(S_{LC}\le s_{LC}(X,Y))\cong\Phi(s_{LC}(X,Y)) \qquad (4.2.14)$$

em que, Φ designa a função de distribuição da lei normal centrada e reduzida.

O coeficiente probabilístico P_L é um coeficiente de semelhança que toma valores no intervalo [0, 1] e verifica-se que:

- $P_{L_{X,Y}}\ge 0$
- $P_{L_{X,X}}\ge P_{L_{X,Y}}$
- $P_{L_{X,Y}}=P_{L_{Y,X}}$

Também se verifica que é uma semelhança normada: $P_{L_{X,X}}=1$.

4.2.4 Interpretação geral dos coeficientes s, s_{LC} e P_L

Podemos então interpretar, de uma forma geral, o coeficiente *s* como a semelhança observada, sendo a variável aleatória S_{LC} a semelhança padronizada, que é análoga a uma covariância sobre um certo espaço de probabilidade, e P_L a função de distribuição dessa semelhança padronizada ser observada, isto é, a semelhança probabilística. O coeficiente

[135] $A=A^T$
[136] $A=-A^T$

P_L é, pois, um coeficiente de semelhança probabilístico do tipo "validade da ligação"[137], VL, (Lerman, 1970; Bacelar-Nicolau, 1980; Subsecção 3.3.3). Como tal, avalia a semelhança padronizada S_{LC} em termos probabilísticos: quanto maior é o valor daquela, tanto mais pequena é a probabilidade de observarmos valores de S_{LC} que o excedam. Tem sido demonstrado que coeficientes probabilísticos baseados na função de distribuição de um coeficiente de semelhança básico, têm propriedades importantes em análise classificatória de dados (e.g., Lerman, 1981; Bacelar-Nicolau, 1987; Nicolau e Bacelar-Nicolau, 1981; Sousa *et al*., 2005), como referido no Capítulo 3.

Como faz notar Lerman (1973), utilizando a nossa notação, "se s_{LC} é o valor da estatística de proximidade entre X e Y, P_L define uma medida da semelhança[138] entre as duas variáveis, em que a noção de semelhança é claramente substituída pela noção de verosimilhança[139], em relação à hipótese N[140]". Por outras palavras, para duas variáveis muito frequentes obtemos uma semelhança forte, o que é certo, pois são muito frequentes, mas a noção de verosimilhança permite questionar se esta semelhança é mais forte do que o acaso (P_L>0.5 ou P_L<0.5), ou não (P_L=0.5).

O coeficiente s_{LC} toma valores no conjunto dos números reais, $\mathbb{R}$. A sua interpretação é a habitual:

- Quanto maior for o seu valor absoluto, mais forte é a relação entre as variáveis.
- O sinal positivo (+) designa relação directa entre as variáveis e o sinal negativo (-) designa relação inversa entre as variáveis.

O coeficiente probabilístico P_L toma valores entre 0 e 1.

- Quando o coeficiente P_L toma valores muito próximos de 1, a semelhança s_{LC} é muito elevada positiva (s_{LC}>>0), i.e., a relação padronizada entre as variáveis é forte e directa. Neste caso, as unidades estatísticas que estão em relação numa das variáveis também o estão na outra variável.
- Quando o coeficiente P_L=0.5, a semelhança s_{LC} é nula, i.e., a relação padronizada entre as variáveis é nula. Neste caso, o valor do coeficiente *s* é igual ao valor médio da variável aleatória S.
- Quando o coeficiente P_L toma valores muito próximos de 0, a semelhança s_{LC} é muito elevada negativa, (s_{LC}<<0), i.e., a relação padronizada entre as variáveis é

[137] *Vraisemblance du lien*, em francês.
[138] *Ressemblance*, em francês.
[139] *Vraisemblance*, em francês.
[140] Hipótese de "ausência de ligação".

forte e inversa. Assim sendo, as unidades estatísticas que estão em relação numa das variáveis não o estão na outra variável.

A interpretação destes coeficientes depende, em particular, da escolha das matrizes *score*, que por sua vez depende, como já foi referido, da natureza das variáveis que descrevem as unidades estatísticas. Esta interpretação poderá ser mais clara se tivermos em conta a natureza das variáveis que estão a ser comparadas e as respectivas matrizes *score*, pois esclarece o que é "estar em relação". As matrizes score serão definidas na Secção 4.3, após algumas considerações sobre a representação gráfica das variáveis.

4.2.5 Representação gráfica das variáveis recorrendo aos coeficientes s, s_{LC} e P_L

O estudo comparativo das variáveis tem particular interesse, por permitir sintetizar informação sobre elas e obter perfis de resposta ou tipologias delas, que facilitam a caracterização das unidades estatísticas presentes no estudo.
Em Análise de Dados, a Análise em Componentes Principais (ACP) e a Análise Classificatória Hierárquica Ascendente (ACHA)[141] permitem obter representações gráficas simples dos dados. Estas duas técnicas de análise multivariada de dados complementam-se na informação que nos dão. Assim sendo, a ACP e a ACHA serão usadas sobre as matrizes de semelhanças S, S_{LC} e P_L (e.g., Le Calvé, 1976b), uma vez que aqueles coeficientes permitem relacionar quer variáveis homogéneas, quer variáveis heterogéneas. Quando as variáveis a relacionar são quantitativas com unidades de medida diferentes ou heterogéneas restringimos as análises aos coeficientes s_{LC} e P_L.

O objectivo da ACP clássica é o de reduzir um conjunto de variáveis quantitativas num menor conjunto de componentes não correlacionadas no espaço euclidiano (combinações lineares das variáveis originais), que representam a maior parte da informação encontrada nas variáveis originais. A ACP é a diagonalização do produto escalar. Ora, desde que haja um coeficiente de semelhança, cuja matriz que o representa seja semidefinida positiva, podemos considerá-lo como um produto escalar e assim diagonizá-lo, o que dará o mesmo resultado que uma ACP sobre os dados que ele representa (Le Calvé, 1976b; Capítulo 1, Subsecção 1.4.3).

---

[141] Este é um dos métodos da Análise Classificatória. Os métodos de Análise Classificatória podem ser divididos em quatro grupos principais: métodos de optimização-partição, métodos hierárquicos, métodos de densidade e outros métodos.

A designação de ACP aqui utilizada, embora não seja inteiramente correcta, é pois compreensível e utilizada por vários outros autores como se viu (Subsecção 1.4.3). Para a distinguirmos da ACP clássica, poderíamos designá-la por Análise em Componentes Principais de Matrizes de Semelhanças (ACPMS). A opção da designação ACP Generalizada (ACPG) foi excluída por já existir.

As ACP das matrizes S, S_{LC} e P_L permitem representar graficamente as relações entre as variáveis heterogéneas ou não, que serão apresentadas nas próximas secções. Em todos os casos estudados de dados verídicos, as matrizes S e S_{LC} são d.p. ou s.d.p., enquanto a matriz de semelhanças P_L não o é. Para casos particulares está demonstrado que a matriz S_{LC} é s.d.p., podendo-se escrever sob a forma de produto escalar de duas matrizes (Subsecções 4.3.2 e 4.3.9).
Quando se utilizam os coeficientes s_{LC} ou P_L, as representações gráficas respectivas são inevitavelmente diferentes. A ACP da matriz P_L permite obter uma representação gráfica aproximada. A transformação associada à função de distribuição da N(0,1) é irregular, existindo por isso uma tendência para aproximar os valores que estão próximos dos extremos e para afastar os valores próximos dos "valores centrais".
Na Subsecção 4.3.9.2, ver-se-á que se obtêm os mesmos resultados, a menos de uma translação, que a ACP Normada e a Análise das Ordens, sempre que se realiza a ACP das matrizes S_{LC} de variáveis quantitativas ou de variáveis número de ordem, respectivamente.

Para realizar as ACP sobre as matrizes S, S_{LC} e P_L (obtidas com a *Toolbox Coeficientes s*, s_{LC} e P_L) utiliza-se a *Toolbox ACPMS* programada em MATLAB 6.5. Também se utilizou o programa em Fortran EUCAPP[142]. Os dois programas calculam os vectores e os valores próprios de matrizes de semelhanças e as coordenadas principais. Nas representações gráficas das variáveis não utilizamos setas (convenção a que fizemos referência na Subsecção 1.4.2.3), por não acharmos necessário.

Como referido acima, para obter uma tipologia das variáveis recorreu-se à análise classificatória hierárquica ascendente (ACHA) utilizando na primeira etapa do seu algoritmo quer o coeficiente s, quer os coeficientes s_{LC} e P_L. Na segunda etapa do algoritmo, serão usados os critérios de agregação entre classes clássicos ("Ligação única"[143], "Ligação

[142] Agradecemos ao Professor Doutor Sergio Camiz o ter facilitado a utilização do programa da sua autoria, EUClidean APProximation.
[143] *Single linkage* (SL), *ultramétrique inférieure maximale* (uim), *critère du plus proche voisin*, ...

completa"[144], "Ligação média"[145]) e alguns dos probabilísticos da família VL: AVL (Algoritmo da validade da ligação), AVM (Critério da ligação pela média aritmética com transformação e com f.d.d.), AVB (avlm-bacelar, p0**sqrt(a*b)), AV1 (p0**((a+b)/2)) (e.g., Bacelar-Nicolau, 1972, 1980; Nicolau, 1980; Sousa, 2003; Sousa, 2005). As hierarquias de partições são bem visualizadas por dendrogramas. As hierarquias de partições, obtidas com algoritmos em que são usados os coeficientes de semelhança s_{LC} e P_L, com o mesmo critério de agregação entre classes, sobre os mesmos dados, serão análogas. No Capítulo 5, faremos o estudo comparativo das hierarquias obtidas com os coeficientes s, s_{LC} e P_L, a partir dos exemplos tratados.

Para realizar as ACHA sobre as matrizes S, S_{LC} e P_L utilizaram-se os programas, em Fortran, POLARB e HIERARQ inicialmente desenvolvidos por Fernando Nicolau e Helena Bacelar-Nicolau e, ao longo dos anos, pelos seus colaboradores (e.g., Nicolau e Soromenho, 1988) sobre aquelas matrizes. Também se usou o programa CLUSTER (Sousa, 2003), que é a última versão do programa CLUSTER (Sousa, 2000), sobre as matrizes S, S_{LC} e P_L previamente obtidas com a *Toolbox Coeficientes s, s_{LC} e P_L*.

Na Subsecção 4.3.9.2 veremos que se obtêm os mesmos resultados sempre que se recorrer a algoritmos de ACHA em que, na primeira etapa, se utilize sobre os mesmos dados, quer o coeficiente de correlação de Pearson, quer o coeficiente s_{LC}, no caso das variáveis serem métricas ou o coeficiente de correlação de Spearman e o coeficiente s_{LC}, no caso das variáveis número de ordem.

4.3 Definição das matrizes score quando se comparam variáveis do mesmo tipo - O que representam os coeficientes s, s_{LC} e P_L

4.3.1 Introdução

Quando pretendemos comparar/relacionar variáveis do mesmo tipo, a definição da matriz *score* depende apenas da natureza das variáveis em estudo, como já foi referido.

Vejamos como se definem as matrizes *score* associadas, respectivamente, às variáveis atributo de descrição (Subsecção 4.3.2), descritiva/nominal (Subsecção 4.3.3), ordinais – com modalidades parcialmente ordenadas (Subsecção 4.3.4), com modalidades totalmente ordenadas (≤) (Subsecção 4.3.5), com modalidades estrita e totalmente ordenadas (<)

---

[144] *Complete linkage* (CL), *ultramétrique supérieure maximale* (usm), *critère du voisin le plus éloigné*, ...

[145] *Average linkage* (AL), no original. Critério da ligação média entre duas classes (AM).

(Subsecção 4.3.6), números de ordem (Subsecção 4.3.7), ordem sequencial (Subsecção 4.3.8) -, real/métrica (Subsecção 4.3.9), assim como às variáveis simbólicas com valores múltiplos (Subsecções 4.3.10.1 e 4.3.10.4), variáveis simbólicas intervalares (Subsecção 4.3.10.5) e variáveis simbólicas modais (Subsecção 4.3.10.8), quando se comparam entre si estas variáveis do mesmo tipo. São aqui apresentadas, pela primeira vez, as definições das matrizes *score* das variáveis simbólicas/complexas referidas acima.

Veremos também o que representam o coeficiente bruto *s* e o coeficiente s_{LC}, quando se comparam estas variáveis do mesmo tipo, duas a duas. Nesta classificação, em que se distinguem vários tipos de variáveis ordinais, a interpretação dos coeficientes s, s_{LC} e P_L é análoga para todas elas.

Em alguns dos casos analisados encontraremos coeficientes conhecidos.

4.3.2 Variáveis atributo de descrição

A variável atributo de descrição, X, só toma os valores 1 e 0 (por convenção), interpretados respectivamente como presença ou ausência do atributo, em que só a presença do atributo dá informação. Trata-se de um caso particular da variável nominal.
Esta variável induz uma parte de E, que é constituída por todos os elementos de E que possuem o atributo X.

4.3.2.1 Matriz score, X (nxn), da variável atributo de descrição

Neste caso, como se considera que só a presença do atributo dá informação, a definição da matriz *score* é a seguinte:

Definição 4.3.1. *Score* da variável atributo de descrição

O *score* da variável atributo de descrição é a relação binária associada à variável, definida da seguinte maneira:

$$x_{ii'} = \begin{cases} 1 & \text{, se i possui o atributo X} \\ 0 & \text{, se não} \end{cases} \qquad (4.3.1)$$

$x_{ii} = 0$

Sob esta óptica, $x_{ii'} = x_i$ é uma constante em *i'*.

O *score* pode-se considerar como uma função indicatriz de um subconjunto A de E, $A \subseteq E$, que é constituído por todos os elementos de E que possuem o atributo X:

$$\begin{aligned} & iRi' \rightarrow x_i \\ & (i,i') \rightarrow x_{ii'} = x_i = 1 \quad se \quad i \in A \\ & \qquad x_{ii'} = x_i = 0 \quad se \quad i \notin A \\ & \qquad x_{ii} = 0 \end{aligned} \qquad (4.3.2)$$

Exemplo 4.3.1. *Score* da variável atributo de descrição

Seja a variável X= "Tem dores?" (1-sim, 0-não)

A variável X toma os valores $X = [1 \quad 0 \quad 1 \quad 1 \quad 0]^T$, para 5 doentes.

A matriz *score* da variável X é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

4.3.2.2 O que representam os coeficientes s, s_{LC} *e* P_L *quando se comparam variáveis atributo de descrição*

Sejam as variáveis X e Y dois atributos de descrição e a respectiva tabela de contingência 2x2, que se apresenta da forma habitual (Tabela 4.3.1) tal como se viu na Subsecção 1.3.6 do Capítulo 1.

Tabela 4.3.1. Tabela de contingência das variáveis atributo de descrição, X e Y. Em que: *a* representa o número de co-presenças, *b* representa o número de presenças/ausências, *c* o número de ausências/presenças e *d* o número de co-ausências

X \ Y	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	n

Convém notar que a interpretação das frequências, ***a***, ***b***, ***c*** e ***d***, depende da disposição dos uns (sim) e zeros (não) nas margens, linhas/colunas, da tabela de contingência (Tabela 4.3.1). Os resultados que apresentamos alteram-se, se alterarmos a disposição dos uns (sim) e zeros (não) das variáveis, na tabela de contingência.

Verifica-se que: s = (n-1)a (4.3.3).

O coeficiente s toma valores entre zero e n(n-1):

- O coeficiente s toma o valor zero quando a = 0. Por exemplo, quando nenhum indivíduo responde sim às duas questões.
- s toma o valor n(n-1) quando a=n. Por exemplo, quando todas as unidades estatísticas respondem sim às duas questões.

As variáveis que tomam o mesmo valor para todas as unidades estatísticas (casos em que, quer a=0 e d=n, quer a=n e d=0), devem ser retiradas do estudo pois, quando pretendemos relacioná-las, o valor da variância da variável aleatória S é nulo e a expressão do coeficiente s_{LC} é indeterminada, além de não ter interesse prático o seu estudo.

Quando as duas variáveis são atributos de descrição, a interpretação dos valores dos coeficientes P_{LC} e s_{LC} leva-nos a dizer que:

- No caso do coeficiente P_L tomar valores muito próximos de um, $P_L \approx 1$ ($s_{LC} >> 0$), os resultados obtidos pelas unidades estatísticas nas duas variáveis terão "padrões" de resposta idênticos nas categorias das duas variáveis, (0,0) ou (1,1), isto é, há uma tendência para emparelhar os zeros com os zeros e os uns com os uns das duas variáveis.
- No caso do coeficiente P_L tomar valores muito próximos de zero, $P_{LC} \approx 0$ ($s_{LC} << 0$), as unidades estatísticas às quais correspondem resultados que estão numa mesma categoria numa delas, também lhes correspondem resultados que não estão na mesma categoria na outra variável, (0,1) ou (1,0), isto é, há uma tendência para emparelhar os zeros com os uns, e os uns com os zeros das duas variáveis.

Pode-se verificar que o coeficiente s_{LC} é igual a $\sqrt{n-1}$ vezes o coeficiente de associação Φ[146] (Yule, 1912; Pearson e Heron, 1913; Coeficiente A_{30}, Tabela 1.3.4) e o seu quadrado relaciona-se, a menos de um coeficiente, com a estatística qui-quadrado, χ^2, para tabelas de contingência 2x2 (Pearson, 1905; Coeficiente A_{27}, Tabela 1.3.4):

$$s_{LC} = \sqrt{n-1}\ \Phi \qquad (4.3.4)$$

Sendo: $$\Phi = \frac{ad - bc}{\sqrt{(a+c)(a+b)(b+d)(c+d)}} \qquad (4.3.5)$$

[146] *Phi coefficient*, em inglês.

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(a+b)(b+d)(c+d)} \qquad (4.3.6)$$

Como $\Phi = \sqrt{\frac{\chi^2}{n}}$, escreve-se $s_{LC} = \sqrt{\frac{n-1}{n}}\sqrt{\chi^2}$. Elevando ao quadrado ambos os membros, obtém-se a relação:

$$s_{LC}^2 = \frac{n-1}{n}\chi^2 \qquad (4.3.7)$$

Quando n tende para o infinito, i.e., no caso de grandes amostras, na prática $\frac{n-1}{n} \approx 1$, $s_{LC} \approx \sqrt{n}\ \Phi$ e o quadrado do coeficiente s_{LC} "coincide" com o χ^2, $s_{LC}^2 \approx \chi^2$.

Neste caso, também se pode verificar que o coeficiente s_{LC} coincide com a estatística Q de Lerman (1973, 1981).

Sabe-se que a matriz Φ é s.d.p. (Tabela 2.2.1, Capítulo 2), podendo-se escrever sob a forma de produto escalar de duas matrizes (e.g., Blanc *et al*., 1976), daí a que a matriz S_{LC} também o possa ser e a ACP da matriz S_{LC}, neste caso, dê os mesmos resultados (a menos de uma translação) que a ACP da matriz Φ.

- Também se verifica que, neste caso, o coeficiente P_L coincide com o coeficiente geral de semelhança do tipo VL para variáveis binárias p_{xy} (Expressão 1.3.22, Subsecção 1.3.6) proposto por Bacelar-Nicolau (1980). Este resultado é importante em análise classificatória devido à equivalência distribucional (exacta ou assimptótica) dos coeficientes para dados binários, demonstrada por Bacelar-Nicolau (Subsecção 1.3.6).

4.3.3 Variáveis nominais

A variável nominal também designada por característica descritiva[147], X, apresenta um conjunto finito de k modalidades, $k \geq 2$, sobre o qual não existe nenhuma estrutura. Esta variável induz uma partição sobre o conjunto das unidades estatísticas, E: uma mesma classe da partição é constituída por subconjuntos de E que possuem uma certa modalidade da característica (Lerman, 1973). Ou seja, a variável nominal induz sobre o conjunto das unidades estatísticas uma relação de equivalência (excepto na diagonal principal).

---

[147] *Variable caractère descriptif* (Lerman, 1973; Le Calvé, 1977).

Em relação às unidades estatísticas, *a* e *b*, cujos valores/códigos em X são, respectivamente, x_a e x_b, só podemos dizer que: $x_a = x_b$ ou $x_a \neq x_b$.
A variável binária é um caso particular desta variável - é uma variável nominal com apenas duas categorias.

Exemplo 4.3.2. Variáveis nominais

- Origem da Patologia: a-Metabólica, b-Neoplásica, c-Infecciosa.
- Paridade: 0 - Primípara, 1- Multípara.

4.3.3.1 Matriz score, X(nxn), da variável nominal

A variável nominal será representada pela matriz, clássica, da relação de equivalência (excepto na diagonal principal).

Definição 4.3.2. *Score* da variável nominal

O *score* da variável nominal é a relação binária associada à variável, definida da seguinte maneira:

$$x_{ii'} = \begin{cases} 1 & \text{, se i e i' apresentam a mesma modalidade de X} \\ 0 & \text{, se não} \end{cases} \qquad (4.3.8)$$

$$x_{ii} = 0$$

Exemplo 4.3.3. *Score* da variável nominal

Seja a variável

X="Local de vigilância da gravidez", (CS-Centro Saúde, H-Hospital, P-Privado)

A variável X toma os valores X= [CS P H H]^T, para 4 mães.
A matriz *score* correspondente é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Por ter sido tomada a decisão, por convenção, de considerar sempre os elementos da diagonal principal nulos, $x_{ii}=0$, esta relação não é uma relação de equivalência clássica, pois

com esta convenção ela deixa de ser reflexiva. Coincide, no entanto, com a noção de equivalência de Bourbaki. A matriz *score* desta variável é simétrica.

4.3.3.2 O que representam os coeficientes s, s_{LC} e P_L *quando se comparam variáveis nominais*

Sendo as variáveis X e Y nominais, o coeficiente s representa o número de pares de indivíduos que estão em relação nas duas variáveis; ou seja, s representa, por exemplo, o número de pares de indivíduos que respondem da mesma maneira às questões X e Y.

Consideremos, para todo o par de variáveis X (com k≥2 modalidades) e Y (com k'≥2 modalidades), uma tabela de contingência k x k', representando por $n_{ii'}$ o número de unidades estatísticas que escolheram a modalidade *i* de X e a modalidade *i'* de Y. O coeficiente *s* é, neste caso, dado por:

$$s = \sum_{i,i'} n_{ii'} \left(n_{ii'} - 1 \right) \qquad (4.3.9)$$

Lerman (1973) divide por 2 a expressão (4.3.9) pois considera $i < i'$: sendo $n_{ii'}$ o cardinal da classe (i,i') da partição, cruzamento das duas partições π_0 e $\pi_0^{'}$, $s = \sum_{i<i'} n_{ii'} \left(n_{ii'} - 1 \right) / 2$, número de pares de objectos de E que são reunidos por cada uma das duas partições.

No caso particular de se relacionarem variáveis dicotómicas, a expressão do coeficiente bruto é dada por $s= a^2+ b^2+ c^2 + d^2 – n$ (4.3.10), utilizando aqui a notação já apresentada na Tabela 4.3.1.

O coeficiente s toma valores entre 0 e n(n-1):

- O coeficiente s toma o valor zero quando todos os pares de indivíduos respondem de uma maneira diferente a X ou de uma maneira diferente a Y ou de uma maneira diferente a X e de uma maneira diferente a Y. Isto é, quando um par de unidades estatísticas está em relação em X e não o está em Y (ou vice-versa), ou não está em relação nas duas variáveis.
- O coeficiente s toma o valor n(n-1) quando todas as unidades estatísticas dão a mesma resposta a cada uma das variáveis.

  Como referido anteriormente, nos casos em que a variância de S é nula e o valor do coeficiente s_{LC} é indeterminado, as variáveis são retiradas do estudo.

Quando as duas variáveis são nominais, a interpretação dos valores dos coeficientes P_L e s_{LC} leva-nos a dizer que:

- No caso do coeficiente P_L tomar valores muito próximos de 1, $P_L \approx 1$ ($s_{LC} >> 0$), os resultados obtidos pelas unidades estatísticas terão "padrões" de resposta idênticos nas categorias de cada uma das variáveis.
- No caso do coeficiente P_L tomar valores muito próximos de zero, $P_L \approx 0$ ($s_{LC} << 0$), as unidades estatísticas às quais correspondem resultados que estão numa mesma categoria numa delas, também lhes correspondem resultados que não estão na mesma categoria na outra variável. Significa que todos os pares de unidades estatísticas responderam de maneira diferente às duas variáveis.

Quando calculamos o coeficiente s_{LC}, reencontramos a estatística de Lerman para variáveis deste tipo (Le Calvé, 1977).
No caso em que as variáveis X e Y são binárias, s é um índice clássico já estudado (Lerman, 1973).

4.3.4 Variáveis com modalidades parcialmente ordenadas

A variável ordinal com modalidades parcialmente ordenadas, X, toma valores que estão associados às diferentes k, k>2, modalidades da variável, existindo entre algumas destas modalidades a possibilidade de as comparar com a relação de ordem. Isto é, em relação às unidades estatísticas *a* e *b*, cujos valores/códigos em X são, respectivamente, x_a e x_b, podemos dizer que: $x_a = x_b$ ou $x_a \neq x_b$, e, só para alguns deles, $x_a < x_b$ ou $x_a > x_b$.

Exemplo 4.3.4. Variáveis com modalidades parcialmente ordenadas

- Níveis de formação escolar e categorias socio-profissionais.
- Num questionário de "Autonomia Instrumental", a questão: Neste último mês, se tem telefone, como usou o telefone?: *0-Não tem telefone, 1- Não usou o telefone, 2-Atendeu e alguém lhe ligava os números, 3-Atendeu, não fazendo telefonemas, 4-Atendeu e ligou números, que conhecia, 5-Atendeu e ligou números, que soube procurar*. Nesta questão/variável a relação de ordem parcial é definida por: (0, 1<2<3<4<5) ou (0, 1<2<4<5, 1<3<4<5).
- Num questionário de "Autonomia Instrumental", a questão: Neste último mês, recorreu a algum apoio para usar transportes, deslocar-se a uma distância que não pode fazer a pé?: *1-Não usou transportes, 2-Sim: usou táxi e/ou carro acompanhado, 3-Sim: usou transportes acompanhado, 4-Sim: guiou carro acompanhado, 5-Não:*

usou táxi sozinho, 6-Não: usou transportes sozinho, 7- Não: guiou carro sozinho. Sendo a relação de ordem parcial definida, por exemplo, por: 1<2, 1<3, 1<4, 1< 5, 1< 6, 1<7, 2<5, 3<4, 3<6, 4<7.

- Num questionário de "Avaliação da Saúde", a pergunta: Neste último mês, como classifica a sua saúde, comparada com há 1 ano?: *1-Não sabe, 2-Pior, 3-Idêntica, 4-Melhor, 5-Muito melhor.* Sendo a relação de ordem parcial definida por: 1, 2<3<4<5.

É frequente encontrar as opções "Não sabe", "Não responde", "Não se aplica", e outras do mesmo género, entre as categorias, ordenadas ou não, de variáveis qualitativas de questionários. De resto é considerado boa prática a inclusão deste tipo de variáveis em questionários, excepto nos raros casos em que os objectivos do estudo levam a forçar uma resposta. No entanto, as técnicas de análise de dados tradicionais não consideram, habitualmente, o facto das modalidades serem parcialmente ordenadas. É o caso, por exemplo, da existência da categoria "Não sabe", entre as outras categorias da variável, para as quais se pode estabelecer uma relação de ordem e em que só se tem em conta o aspecto nominal da variável. Deste facto resulta, pois, um empobrecimento dos resultados da análise. Esta classificação, pelo contrário, permite manter essa informação. Noutras situações, as categorias "Não sabe"/"Não responde" são tratadas como dados omissos. Contudo, a utilização de métodos de imputação nestes casos não é ideal, pois o facto de não saber ou não responder é em si uma informação, que desta forma se perde. Também neste caso se pode dizer que a classificação aqui estudada permite manter essa informação.

4.3.4.1 Matriz score, X(nxn), da variável com modalidades parcialmente ordenadas

Neste caso, consideramos duas definições, sendo a segunda, a de ordem parcial estrita.

Definição 4.3.3. *Score* da variável com modalidades parcialmente ordenadas

O *score* da variável com modalidades parcialmente ordenadas é a relação binária associada à variável, definida da seguinte maneira:

$$\begin{aligned} x_{ii'} &= 1 \text{ , se } X(i) \leq X(i') \\ &\quad 0 \text{ , se não} \qquad (4.3.11) \\ x_{ii} &= 0 \end{aligned}$$

A relação binária associada a esta variável é a de uma preordem[148] não total[149] sobre o conjunto das unidades estatísticas (com excepção dos elementos da diagonal principal).

Definição 4.3.4. *Score* da variável com modalidades estrita e parcialmente ordenadas

O *score* da variável com modalidades estrita e parcialmente ordenadas é a relação binária associada à variável, definida da seguinte maneira:

$x_{ii'} = 1$, se $X(i) < X(i')$
$\quad\quad 0$, se não (4.3.12)
$x_{ii} = 0$

A escolha entre estas duas definições das matrizes *score* dependerá do peso (no sentido de importância) que se pretenda dar à relação "<", e à relação "=", entre as modalidades da variável. Para simplificar a nomenclatura, referiremos também estas variáveis como sendo variáveis de ordem parcial.

Exemplo 4.3.5. *Score* da variável com modalidades parcialmente ordenadas

Seja a variável,

X="Posição correcta de deitar o bebé" (1- Não sabe, 2- Barriga para baixo, 3- De lado, 4- Barriga para cima, 5- Outra posição).

Sendo a relação de ordem parcial definida por: 1, 2<4, 3<4, 5 ou por 1, 2<3<4, 5, conforme a opinião dos pediatras.

A variável X toma os valores: X= [2 1 3 4]^T, para 4 grávidas.

A matriz *score* correspondente à relação de ordem parcial, "1, 2<4, 3<4, 5, entre as modalidades da variável, considerando qualquer uma das definições, é dada neste caso por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Esta relação é irreflexiva, devido à convenção sobre a diagonal principal. Por ser anti-simétrica e transitiva para os elementos que estão em relação, é pois uma relação de ordem parcial.

[148] Uma preordem é uma relação binária reflexiva e transitiva.
[149] Uma relação binária em que quaisquer dois elementos são comparáveis diz-se total.

4.3.4.2 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis com modalidades parcialmente ordenadas

O coeficiente s representa o número de unidades estatísticas que estão em relação em cada uma das variáveis. Em particular, neste caso, o coeficiente s representa o número de pares de indivíduos tais que:

$x_a \leq x_b \ (x_a<x_b)$ e $y_a \leq y_b \ (y_a<y_b)$, ou seja, s representa o número de concordâncias.

O coeficiente P_L toma valores muito próximos de zero, $P_L \approx 0$ ($s_{LC}<<0$), quando a unidade estatística *a* está em relação com a unidade estatística *b* em X, $x_a \leq x_b \ (x_a<x_b)$, quaisquer que sejam *a* e *b*, e as unidades estatísticas *a* e *b* não estão em relação em Y, i.e., $y_a \leq y_b \ (y_a<y_b)$ é falso.

O coeficiente P_L toma valores muito próximos de um, $P_L \approx 1$ ($s_{LC}>>0$), quando as unidades estatísticas *a* e *b* que estão em relação em X, $x_a \leq x_b \ (x_a<x_b)$, também o estão em Y, $y_a \leq y_b \ (y_a<y_b)$.

Com excepção de Ouali (1991a), que aborda este assunto e propõe um coeficiente de semelhança, não temos conhecimento de outro(s) coeficiente(s) que permitam relacionar variáveis deste tipo.

Para percebermos melhor a interpretação dos resultados obtidos com estes coeficientes e a vantagem da sua utilização, em relação aos métodos de análise de dados tradicionais, apresentamos um exemplo com dados verídicos, que foi tema de uma comunicação oral (Doria *et al*., 2006b), e que tratamos detalhadamente no Capítulo 5 (Secção 5.2).

Exemplo 4.3.6. Semelhanças s_{LC} e P_L entre variáveis de ordem parcial

Um dos objectivos propostos na comunicação, "Os Coeficientes s, s_{Lc} e P_L na Análise de um Questionário de Qualidade e Satisfação na Saúde" (Doria *et al*., 2006b), foi o de comparar os sete itens que constituem a "Escala A. Elementos Tangíveis" e, separadamente, os nove itens da "Escala D. Interesse/Capacidade de Resposta" do Questionário SERVQUAL Modificado (Tabela 5.2.1, Secção 5.2, Capítulo 5).

As respostas, aos itens daquelas escalas, são dadas sob a forma: 1-Totalmente em desacordo, 2-Desacordo, 3-Des/Acordo, 4-Acordo, 5-Totalmente de acordo, 6-Não se aplica,

9-Não sabe/Não responde. Os itens destas escalas são variáveis de ordem parcial, uma vez que entre as suas modalidades se verifica: 1<2<3<4<5, 6, 9.

Na amostra estudada, todos os itens da "Escala A. Elementos Tangíveis", com excepção do "a2. Existe uma cópia do relatório clínico no seu domicílio..." e do "a7. A equipa de saúde facilita o acesso a outro equipamento de acesso", apresentam como resposta mais frequente a opção "5-Totalmente de acordo" (Anexo 1).

Estudámos as relações entre os itens utilizando os coeficientes *s*, s_{LC} e P_L, pois têm em conta a natureza de ordem parcial destes.

Das semelhanças s_{LC} (Tabela 5.2.3, Tabela 5.2.11) e P_L (Tabela 5.2.4, Tabela 5.2.12) entre as respostas dadas aos itens daquelas escalas, destacamos:

- A semelhança positiva entre as respostas dadas aos itens "a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento..." e "a5.A equipa de saúde fornece outra medicação...": $s_{LC}(a3, a5)= 3.1880$ e $P_L(a3, a5)= 0.9993$.

  Os itens "a3" e "a5" apresentam respostas em todas as categorias, mas revelam predominância nas categorias positivas (4-Acordo e 5-Totalmente de acordo) (Anexo 1). Os cuidadores que respondem "5-Totalmente de acordo" ao item "a3" (43 - 74.1%), são cuidadores que também respondem a todas as possibilidades de resposta a "a5", com predominância para as opções "4-Acordo" e "5-Totalmente de acordo", enquanto que as opções de resposta "6-Não se aplica" e "9-Não sabe/Não responde" do item "a3" são respondidas por cuidadores que estão "5-Totalmente de acordo" ou acham que "6-Não se aplica" ao item "a5" (Figura 4.3.1). Observa-se assim, a tendência para emparelhar "valores"/ordens baixas com "valores"/ordens baixas e "valores"/ordens elevadas com "valores"/ordens elevadas dos dois itens, distribuindo-se as opções "6-Não se aplica" e "9-Não sabe/Não responde" de um dos itens por algumas categorias do outro item.
- A relação entre os itens relativos à acessibilidade, "d4.Quando chamada a equipa de saúde dá resposta rapidamente " e "d6p.Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir... " da "Escala D. Interesse/Capacidade de Resposta", que é muito forte e positiva, $s_{LC}(d4, d6p)= 20.5080$ e $P_L(d4, d6p)\approx 1$, e se pode perceber facilmente no gráfico seguinte (Figura 4.3.2).

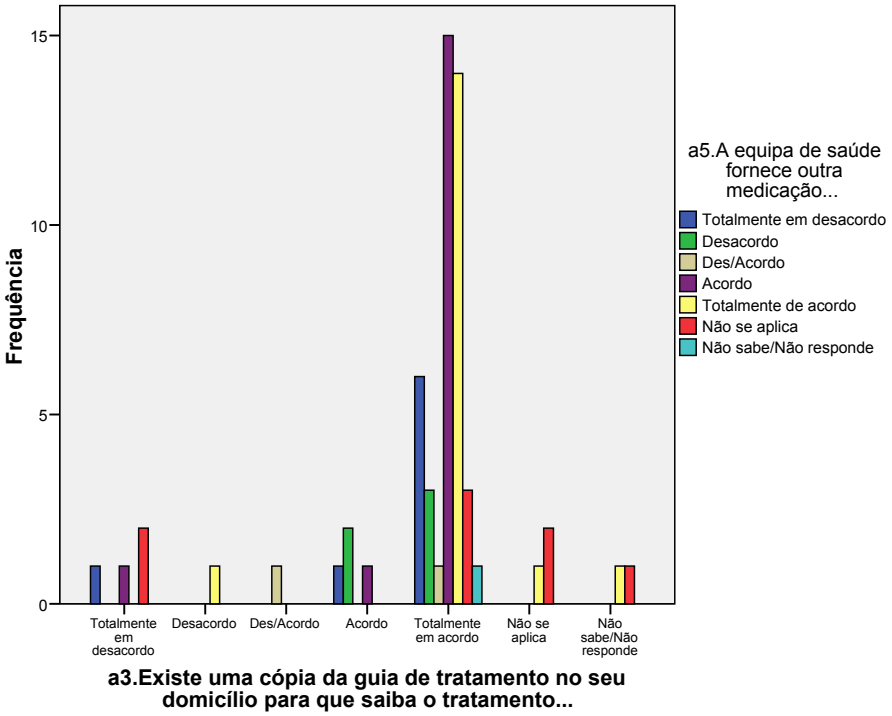

Figura 4.3.1. Gráfico de barras da distribuição de frequências das respostas ao item "a5.A equipa de saúde fornece outra medicação...", por cada categoria do item "a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...", (s_{LC}(a3, a5)= 3.19 e P_L(a3, a5)= 0.9993).

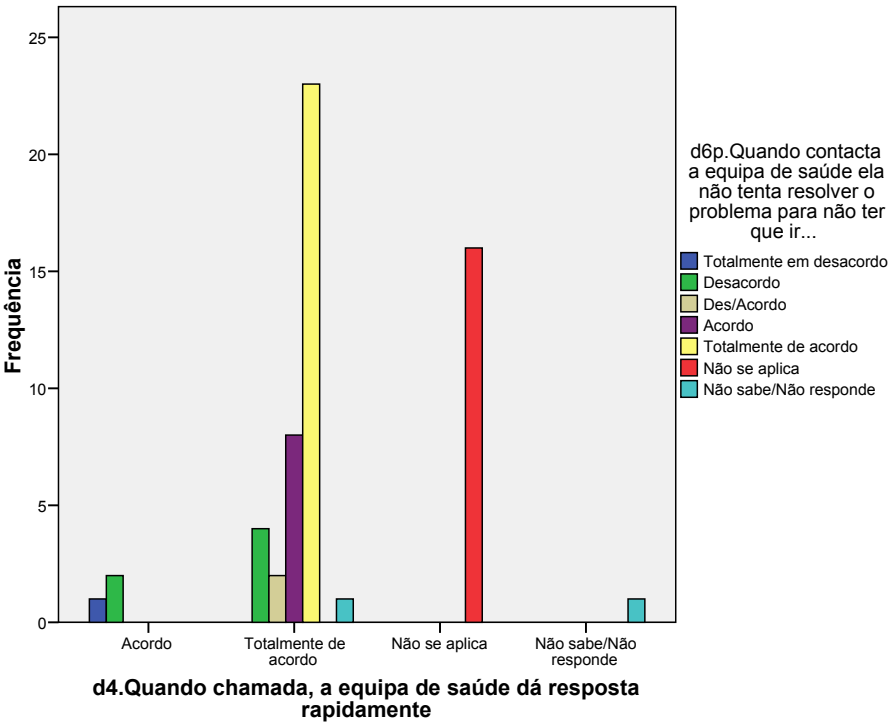

Figura 4.3.2. Gráfico de barras da distribuição de frequências das respostas ao item "d6p.Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir... " por cada categoria do item "d4. Quando chamada, a equipa de saúde dá resposta rapidamente ", (s_{LC}(d4, d6p)= 20.51 e P_L(d4, d6p)≈1).

- A semelhança negativa entre as respostas aos itens “a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento... “ e “a7. A equipa de saúde facilita o acesso a outro equipamento de acesso“: s_{LC}(a3, a7)= -1.6665 e P_L(a3, a7)= 0.0478.

  Os itens “a3” e ”a7”, apresentando respostas em todas as categorias, revelam predominância de respostas em categorias diferentes e “opostas”: “a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento... “ tem a moda nas categorias positivas (“4-Acordo” e “5-Totalmente de acordo”) (Anexo 1), enquanto que, “6-Não se aplica” é a resposta mais frequente (56.7%) a “a7. A equipa de saúde facilita o acesso a outro equipamento de acesso“ (Anexo 1). A categoria “6-Não se aplica”, moda do item “a7”, é maioritária na maior parte das possibilidades de resposta do item “a3” e todas as possibilidades de resposta do item “a7” são contempladas pelos cuidadores que respondem “5-Totalmente de acordo” ao item “a3” (Figura 4.3.3). Verificamos então, que o coeficiente P_L toma valores próximos de zero, $P_L \approx 0$ ($s_{LC} << 0$), quando a unidade estatística *a* está em relação com a unidade estatística *b* em X, $x_a \leq x_b$ $(x_a < x_b)$, as unidades estatísticas *a* e *b* não estão em relação em Y, i.e., $y_a \leq y_b (y_a < y_b)$ é falso, estando a(s) categoria(s) não “ordinais” de uma das variáveis distribuída(s) pelas categorias da outra variável. Verifica-se bem que todas as categorias que representam o desacordo encontram-se na categoria “Totalmente de acordo”, o que mostra que as ordens estão em sentido inverso; enquanto que a categoria “Não se aplica “ de “a7” não está igualmente distribuída pelas categorias de “a3”, mas está fortemente concentrada na categoria “Totalmente de acordo” de “a3”.
- A semelhança s_{LC} entre os itens “a1.Os equipamentos médicos são adequados” e “a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...“ é quase nula: s_{LC}(a1, a3)= - 0.0068 e P_L(a1, a3)=0.4973. Enquanto o item “a1” não apresenta respostas negativas e nele predomina a modalidade “5-Totalmente de acordo”, além da “6-Não se aplica”, o item “a3”, apresentando respostas em todas as categorias, revela predominância nas categorias positivas (“4-Acordo” e “5-Totalmente de acordo”), que se distribuem por todas as categorias do item “a1” (Figura 4.3.4).

  Não se observa, pois, um “padrão” de resposta: – os cuidadores deram respostas que permitem emparelhar o valor o elevado “5-Totalmente de acordo” do item “a1” com a maioria dos valores do item “a3”, assim como a categoria “6-Não se aplica” do item “a1” com várias categorias do item “a3”.

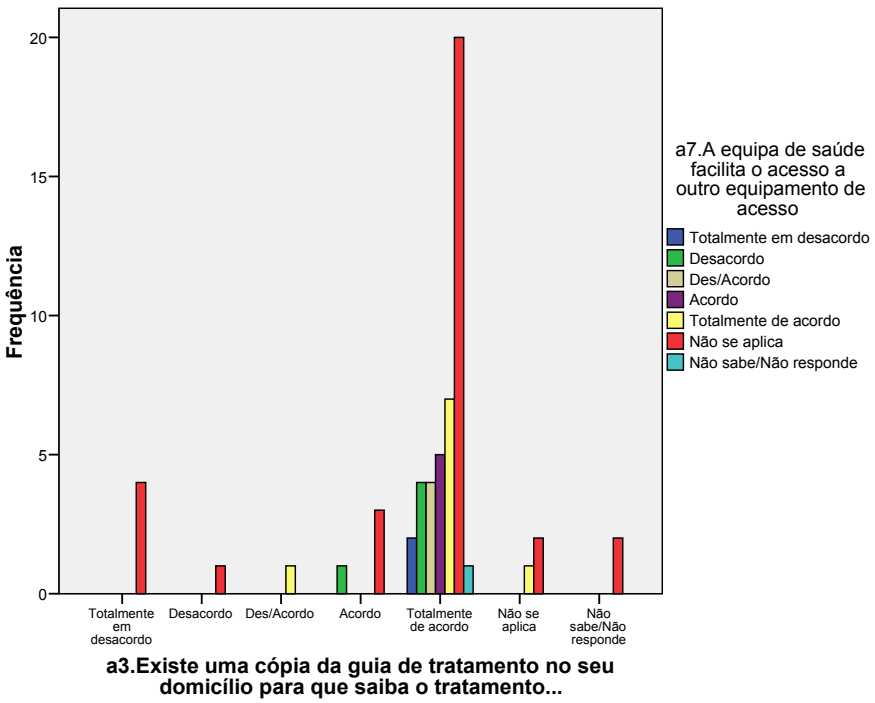

Figura 4.3.3. Gráfico de barras da distribuição de frequências das respostas ao item "a7. A equipa de saúde facilita o acesso a outro equipamento de acesso" por cada categoria do item "a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento... ", (s_{LC}(a3, a7)= -1.67 e P_L(a3, a7)= 0.0478).

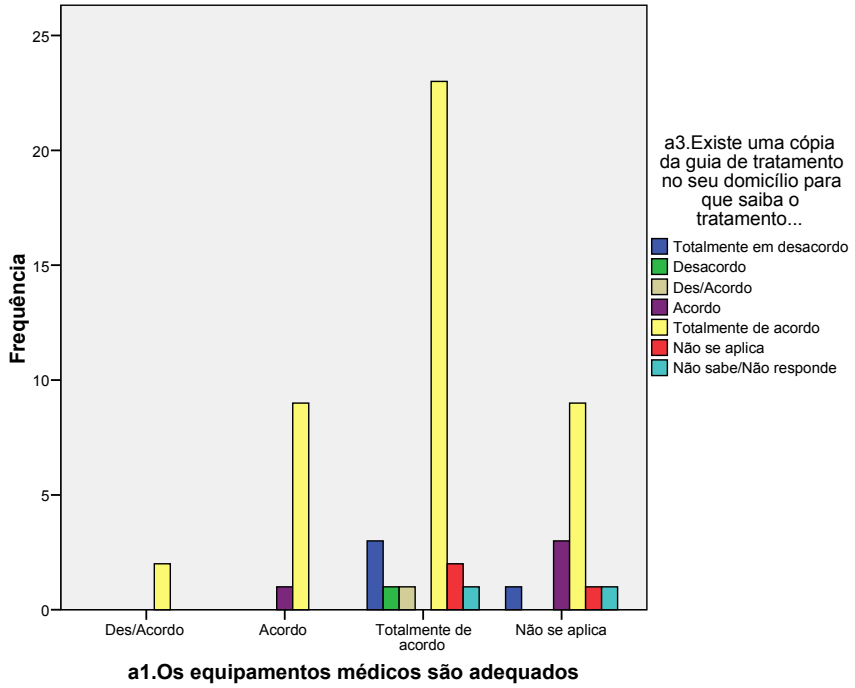

Figura 4.3.4. Gráfico de barras da distribuição de frequências das respostas ao item "a1.Os equipamentos médicos são adequados" por cada categoria do item "a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...", (s_{LC}(a1, a3)= -0.0068 e P_L(a1, a3)= 0.4973).

4.3.5 Variáveis com modalidades totalmente ordenadas

A variável ordinal com modalidades totalmente ordenadas ($\leq$), X, toma valores que estão associados a um número finito k, k>2, de categorias/modalidades da variável, existindo entre todas elas a possibilidade de as comparar com a relação de ordem total. Isto é, em relação às unidades estatísticas *a* e *b*, cujos valores/códigos em X são, respectivamente, x_a e x_b, podemos dizer que: $x_a \leq x_b$ ou $x_a \geq x_b$, e não só $x_a = x_b$ ou $x_a \neq x_b$, como na variável nominal.

Exemplo 4.3.7. Variáveis ordinais com modalidades totalmente ordenadas

- Num questionário sobre nutricionismo, a pergunta/variável: "Gosta de chocolate?" (1-Detesto, 2-Não gosto, 3-Indiferente, 4-Gosto, 5-Gosto muito).
- Num Inventário de Saúde Mental, a pergunta: "Neste último mês, durante quanto tempo se sentiu só?" (1-Sempre, 2-Quase sempre, 3-A maior parte do tempo, 4-Durante algum tempo, 5-Quase nunca, 6-Nunca).
- Num Questionário de Caracterização Sociográfica, a pergunta: "Mantém relações de convívio com amigos ou vizinhos?" (1-Não, 2-Raramente, 3-Sim,regularmente).
- A variável "Nível socioeconómico da família (Índice de Graffar: 1-I, 2-II, 3-III, 4-IV, 5-V)".

4.3.5.1 Matriz score, X (nxn), da variável com modalidades totalmente ordenadas

O *score* da variável com modalidades totalmente ordenadas define-se da seguinte maneira:

Definição 4.3.5. *Score* da variável com modalidades totalmente ordenadas

O *score* da variável com modalidades totalmente ordenadas é a relação binária associada à variável, definida da seguinte maneira:

$$x_{ii'} = \begin{cases} 1 & \text{, se } X(i) \leq X(i') \\ 0 & \text{, se não} \end{cases} \qquad (4.3.13)$$

$$x_{ii} = 0$$

A relação binária associada a esta variável é a de uma preordem[150], no sentido lato, sobre o conjunto das unidades estatísticas, E.

"A variável, "característica com modalidades totalmente ordenadas", induz uma preordem total sobre o conjunto dos indivíduos; a característica apresenta aqui um conjunto finito de modalidades sobre o qual é dada uma estrutura de ordem total. Uma mesma classe da pre-ordem é constituída pelo conjunto dos objectos/indivíduos de E que possuem uma certa modalidade da característica. Relembremos que, o dar uma ordem total é equivalente a dar uma partição e uma ordem total sobre o conjunto das classes desta última." (Lerman, 1972).

Exemplo 4.3.8. *Score* da variável com modalidades totalmente ordenadas

Seja a variável, X=" Nível Socioeconómico da família" (Índice de Graffar:1-I, 2-II, 3-III, 4-IV, 5-V)[151]]

Esta variável toma os valores X= $[1 \quad 2 \quad 2 \quad 4]^T$, para 4 mães.

A matriz *score* correspondente é dada por: $X = (x_{ii'}) = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

Esta variável ultrapassa o problema dos empates.

4.3.5.2 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis com modalidades totalmente ordenadas

O coeficiente s representa o número de pares que estão dispostos na mesma ordem em X e em Y (com os empates), isto é, s representa o número de concordâncias[152], n_c: $s = n_c$, incluindo os empates.

O coeficiente s toma valores entre 0 e n(n-1):

- O coeficiente s toma o valor 0, quando não há nenhuma concordância, ou seja, quando a todas as unidades estatísticas correspondem resultados/modalidades diferentes e as

[150] Com a ressalva da irreflexividade (devido à convenção da diagonal principal ser nula), esta relação é uma relação binária de preordem total sobre o conjunto das unidades estatísticas (todos os valores estão em relação e é transitiva).

[151] A escala foi recodificada de tal forma que:1.V<2.IV<3.III<4.II<5.I.

[152] Neste caso, diz-se que existe concordância entre as duas classes/categorias das duas variáveis, quando dois indivíduos i e i' estão na mesma ordem, para as duas variáveis: $x_i \leq x_{i'}$ e $y_i \leq y_{i'}$.

categorias/modalidades das duas variáveis estão dispostas por ordem inversa. Por exemplo, para 4 indivíduos: X=[2 3 4 5]^T, Y=[5 4 3 2]^T, s_{XY}=0.

- O coeficiente s só toma o valor n(n-1), quando a todas as unidades estatísticas corresponde o/a mesmo/a resultado/modalidade numa e noutra variável, não sendo necessário que o resultado seja o mesmo nas duas variáveis. Por exemplo, para cinco indivíduos, X=[1 1 1 1 1]^T, Y=[2 2 2 2 2]^T, $S = \begin{pmatrix} 20 & 20 \\ 20 & 20 \end{pmatrix}$. Este caso, como já foi referido, não será contemplado, pois as variáveis deverão ser retiradas do estudo.

Quando as variáveis são de tipo ordinal, podemos dizer que:

- Para valores do coeficiente s_{LC} fortemente negativos, $s_{LC}<<0$ e $P_L\approx 0$, as categorias/modalidades das duas variáveis estão dispostas por ordem inversa, ou seja, os resultados obtidos pelas unidades estatísticas correspondem a categorias opostas em cada uma das variáveis.
- Para valores do coeficiente s_{LC} fortemente positivos, $s_{LC}>>0$ e $P_L\approx 1$, as categorias/modalidades das duas variáveis ocupam posições que são idênticas na escala ordinal, ou seja, os resultados obtidos pelas unidades estatísticas correspondem a categorias que ocupam posições semelhantes na escala ordinal das duas variáveis. Por exemplo: X=[1 1 1 1 2]^T e Y=[2 2 2 2 3]^T, ou X=[1 2 4 1]^T e Y=[3 7 11 3]^T.

Neste caso, o coeficiente s_{LC} coincide com o coeficiente de Lerman (1973) para preordens (Le Calvé, 1977).

4.3.6 Variáveis com modalidades estrita e totalmente ordenadas

A variável ordinal com modalidades estrita e totalmente ordenadas (<), X, toma valores que estão associados às k, k>2, diferentes modalidades da variável, existindo entre todas elas a possibilidade de as comparar com a relação de ordem estrita. Em relação às unidades estatísticas *a* e *b*, cujos valores em X são, respectivamente, x_a e x_b, podemos dizer que: $x_a<x_b$ ou $x_a>x_b$.

As unidades estatísticas são ordenadas segundo determinado critério e não existem empates. Há pois, tantas unidades estatísticas quantas as modalidades da variável (k= número de unidades estatísticas = número de categorias da variável).

Esta variável é um caso particular da preordem, em que cada classe da preordem contém exactamente um elemento. As questões podem ser as mesmas que no caso anterior, mas a maneira de olhar para os resultados não é a mesma, pois não são permitidos empates. Esta situação é mais rara de encontrar na prática.

Exemplo 4.3.9. Variável com modalidades estrita e totalmente ordenadas

Num concurso, os "Resultados do concurso (sem empates)", correspondem aos resultados de uma variável com modalidades estrita e totalmente ordenadas.

4.3.6.1 Matriz score, X (nxn), da variável com modalidades estrita e totalmente ordenadas

Definição 4.3.6. *Score* da variável com modalidades estrita e totalmente ordenadas

O *score* da variável com modalidades estrita e totalmente ordenadas é a relação binária associada à variável, definida da seguinte maneira:

$x_{ii'}$= 1 , se $X(i) < X(i')$

0 , se não (4.3.14)

$x_{ii} = 0$

Não existem empates. A cada indivíduo corresponde um número e todos eles são diferentes.

Esta variável induz uma relação irreflexiva e transitiva, isto é, uma relação de ordem estrita[153] total sobre o conjunto das unidades estatísticas, E.

Exemplo 4.3.10. *Score* da variável com modalidades estrita e totalmente ordenadas

Seja a variável X= "Resultados de um concurso, sem empates (escalões D<C<B<A)".

A variável X toma os valores: $X=[B\ \ C\ \ A\ \ D]^T$, para 4 indivíduos.

A matriz *score*, associada à variável X, é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

[153] Uma relação é uma ordem estrita (quase-ordem) se ela é irreflexiva e transitiva (por isso, anti-simétrica).

4.3.6.2 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis ordinais com modalidades estrita e totalmente ordenadas

O coeficiente s representa o número de pares que estão dispostos na mesma ordem em X e em Y, isto é, s representa o número de concordâncias, n_c: $s = n_c$ (4.3.15).

Neste caso, o coeficiente *s* obtém-se do coeficiente tau de Kendall, τ, a partir de uma transformação afim: $s = \frac{N(\tau + 1)}{2}$ (4.3.16), sendo $N = n(n-1)/2$ e o coeficiente tau de Kendall (Siegel e Castellan, 1989) dado por $\tau = \frac{n_c - n_d}{n(n-1)/2}$ (4.3.17),

em que, n_c representa o número de concordâncias e n_d o número de discordâncias[154].

Sabe-se ainda que: $(n_c - n_d)_{max} = -(n_c - n_d)_{min} = n(n-1)/2$ (4.3.18) (Saporta, 1990) e $N = n_c + n_d$.

O coeficiente s toma valores entre 0 (as classificações estão invertidas nas duas variáveis) e $N = n(n-1)/2$ (as classificações são idênticas nas duas variáveis), enquanto o coeficiente τ toma valores entre -1 (classificações invertidas) e +1 (classificações idênticas).

A teoria assintótica permite afirmar que a distribuição do coeficiente T pode ser aproximada por uma distribuição normal com valor médio nulo, $\mu_T = 0$, e desvio-padrão

$\sigma_T = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$ (4.3.19). Isto é, $T \overset{ap}{\cap} N\left(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$ (4.3.20).

Sendo esta aproximação muito boa para $n \geq 8$ (Saporta, 1990).

Neste caso, o coeficiente tau de Kendall centrado e reduzido coincide com o coeficiente s_{LC}:

$s_{LC} = \frac{\tau}{\sigma_\tau}$ (4.3.21).

O coeficiente τ de Kendall deve ser utilizado com variáveis que induzem uma ordem total estrita sobre o conjunto das unidades estatísticas (não existem empates). Lerman (1973) provou que o coeficiente τ de Kendall não deve ser utilizado com variáveis ordinais com modalidades totalmente ordenadas, pois "o coeficiente τ com a correcção para empates (Kendall, 1970) é uma estatística enviesada".

---

[154] Neste caso, diz-se que existe concordância entre as duas classes/categorias das duas variáveis, quando dois indivíduos i e i' estão na mesma ordem para as duas variáveis: $x_i < x_{i'}$ e $y_i < y_{i'}$. Diz-se que duas classificações discordam, quando $x_i < x_{i'}$ e $y_i > y_{i'}$.

4.3.7 Variáveis número de ordem

A variável número de ordem[155], tal como a variável com modalidades estrita e totalmente ordenadas, induz uma relação de ordem total sobre o conjunto das unidades estatísticas, E. Pode-se dizer que esta variável é, algebricamente, um caso particular da variável com modalidades totalmente ordenadas, em que cada classe da preordem contém exactamente um elemento.

A variável número de ordem, X, faz corresponder a cada unidade estatística o seu número de ordem, ou seja, a sua posição na lista ordenada dos valores da variável.

Exemplo 4.3.11. Variáveis número de ordem

- Ordenação/posicionamento de indivíduos, segundo determinado critério, com a possibilidade de existirem empates.
- Sempre que "juízes" ordenam as unidades estatísticas segundo determinado critério, a ordenação obtida é uma variável número de ordem.
- Número de ordem de entrada num concurso a uma escola.
- Número de chegada numa prova de ciclismo.

Nesta variável a ordem é definida sobre as unidades estatísticas, enquanto que nas variáveis ordinais anteriores, a ordem é definida sobre as suas modalidades.

4.3.7.1 Matriz score, X (nxn), da variável número de ordem

Definição 4.3.7. *Score* da variável número de ordem

O *score* da variável número de ordem é definido da seguinte maneira:

$x_{ii'}$= número de ordem da unidade estatística i (4.3.22)

x_{ii}= 0

Passaremos a representar o número de ordem da unidade estatística i por R_i. No caso de existirem empates, consideramos que a ordem dos elementos empatados é igual à média das ordens que lhes seriam atribuídas no caso destes não se considerarem como

[155] *Variable rang* (Le Calvé, 1977), no original. *Rank*, em inglês. Ordem, número de ordem, posto, em português. Veja-se a nota introdutória no Capítulo 1 (Secção 1.2).

empatados. Como se sabe, esta é uma possibilidade, entre outras conhecidas, para tratar os empates.

Exemplo 4.3.12. *Score* de variáveis número de ordem

- A matriz *score* associada à variável X, X= [11 20 20 45]^T para 4 indivíduos, é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 2.5 & 0 & 2.5 & 2.5 \\ 2.5 & 2.5 & 0 & 2.5 \\ 4 & 4 & 4 & 0 \end{bmatrix}$$

- Considere-se a variável "Opinião do cirurgião X" sobre quatro técnicas cirúrgicas (a, b, c, d). Solicita-se que as disponha por ordem de preferência (1<2<3<4), tendo-se obtido:

Técnicas cirúrgicas	Opinião do Cirurgião X (R_i)
a	3
b	1
c	4
d	2

A matriz *score* de X é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 3 & 3 & 3 \\ 1 & 0 & 1 & 1 \\ 4 & 4 & 0 & 4 \\ 2 & 2 & 2 & 0 \end{bmatrix}$$

4.3.7.2 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis número de ordem

O coeficiente *s* representa, a menos de um factor que depende da dimensão da amostra/população, um produto escalar entre as ordens das unidades estatísticas nas duas variáveis j e j': s = (n-1) $\sum_i R_{ij} R_{ij'}$ (4.3.23).

Neste caso, o coeficiente s_{LC} coincide com o coeficiente de correlação de Spearman, r_S, a menos de $\sqrt{n-1}$: $s_{LC} = \sqrt{n-1}\, r_S$ (Le Calvé, 1977), sendo

$$r_S = 1 - \frac{6\sum_{i=1}^{n}\left(R_{ij} - R_{ij'}\right)^2}{n\left(n^2-1\right)} \qquad (4.3.24)$$

R_{ij} designa o número de ordem da observação i da variável j.

A interpretação dos valores dos coeficientes, s_{LC} e P_L, coincide com a que foi feita anteriormente para as outras variáveis de tipo ordinal.

No caso em que as variáveis são todas número de ordem, os resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre aquelas variáveis são os mesmos que os obtidos com a ACP normada da matriz das ordens dos dados da matriz original, a menos de uma translação, que depende da dimensão n da amostra/população em estudo. Neste caso, os valores próprios são $\sqrt{n-1}$ vezes maiores e as coordenadas são $\sqrt{\sqrt{n-1}}$ vezes maiores. As representações gráficas das variáveis obtidas com as duas ACP são pois as mesmas, a menos de uma translação, que depende da dimensão da amostra/população em estudo.

4.3.8 Variáveis de ordem sequencial

A variável de ordem sequencial, X, é uma variável em que os indivíduos têm sempre um indivíduo antes (predecessor) ou/e depois (sucessor), isto é, em que se consideram apenas os pares sucessivos.

Exemplo 4.3.13. Variáveis de ordem sequencial

- Exemplos de casos clássicos:
  - A disposição dos indivíduos numa fila de espera. Em geral, há um primeiro e um último na fila.
  - A escolha do curso, por ordem de preferência.
  - Uma escolha dos hospitais, segundo a escolha preferencial de determinada especialidade, tal como, medicina geral ou cirurgia.
- Exemplos de casos que não são clássicos:
  - A sucessão dos meses do ano, considerada num plano sazonal. Não há um primeiro, nem último – todos são sucessores ou predecessores. É uma relação binária "circular". Na ordem sequencial, Dezembro e Janeiro "tocam-se", enquanto que, para a variável ordem sequencial clássica, eles estão afastados 12 meses.
  - A disposição dos nós na árvore filogenética.

4.3.8.1 Matriz score, X (nxn), da variável de ordem sequencial

Propomos duas definições de matriz *score* da variável de ordem sequencial, de acordo com a definição clássica e a não clássica.

Definição 4.3.8. *Score* da variável de ordem sequencial (definição clássica)

O *score* da variável ordem sequencial é a relação binária associada à variável, definida da seguinte maneira:

$x_{ii'}$= 1 , se n.º de ordem de i' = n.º de ordem de i +1

0 , se não (4.3.25)

$x_{ii} = 0$

Neste caso, quando existem valores empatados, a ordem atribuída às unidades estatísticas é a mesma para todas elas, sendo esta igual à menor ordem que lhes corresponderia se os seus valores não estivessem empatados.

Definição 4.3.9. *Score* da variável de ordem sequencial (definição não clássica)

O *score* da variável de ordem sequencial é a relação binária associada à variável, definida da seguinte maneira:

$x_{ii'}$= 1 , se i e i' são "vizinhos"

0 , se não (4.3.26)

$x_{ii} = 0$

Exemplo 4.3.14. *Score* da variável de ordem sequencial (caso clássico)

A matriz *score* associada à variável X, X= [A B B D]^T para 4 indivíduos, é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Exemplo 4.3.15. *Score* da variável de ordem sequencial (caso não clássico)

A matriz *score* associada à variável X="mês da cirurgia do indivíduo i",

X= [Jan, Fev, Fev, Dez]^T para 4 indivíduos, é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

4.3.8.2 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis de ordem sequencial

Quando as variáveis que se pretendem comparar são variáveis de ordem sequencial (definição clássica), o coeficiente s representa o número de pares de indivíduos cujas ordens se seguem/são consecutivas em X e em Y, isto é, duas unidades estatísticas são consecutivas para X e para Y também.

O coeficiente *s* toma valores entre 0 e n-1:

- O coeficiente *s* toma o valor 0, quando os valores das duas variáveis não estão dispostos por ordem sequencial.
- O coeficiente *s* toma o valor n-1, no caso da definição clássica, quando todos os valores das duas variáveis ocupam posições sequenciais/consecutivas e diferentes na escala ordinal; no caso da definição não clássica, o coeficiente *s* toma o valor 2n-1.

A interpretação dos valores dos coeficientes, *s*, s_{LC} e P_L, é análoga à que já foi feita para as variáveis de tipo ordinal. Quando as variáveis são de ordem sequencial, podemos dizer que:

- Para valores do coeficiente s_{LC} fortemente negativos, $s_{LC}<<0$ ($P_L\approx0$), as categorias/modalidades das duas variáveis estão dispostas por ordem inversa, ou seja, os resultados obtidos pelas unidades estatísticas correspondem a categorias sequenciais/consecutivas opostas em cada uma das variáveis.
- Para valores do coeficiente s_{LC} fortemente positivos, $s_{LC}>>0$ ($P_L\approx1$), as categorias/modalidades das duas variáveis ocupam posições que são idênticas e sequenciais/consecutivas na escala ordinal, ou seja, os resultados obtidos pelas unidades estatísticas correspondem a categorias que ocupam posições consecutivas na escala ordinal das duas variáveis.

Como veremos, na Secção 4.4, estas variáveis (definição clássica) ser-nos-ão muito úteis, quando relacionarmos variáveis do tipo ordinal com outros tipos de variáveis.

4.3.9 Variáveis métricas

A variável métrica[156], X, é uma variável que toma valores reais e que, relativamente às unidades estatísticas *a* e *b*, cujas pontuações em X são, respectivamente, x_a e x_b, podemos

[156] *Variable réelle* (Le Calvé, 1977).

dizer que: além de $x_a = x_b$ ou $x_a \neq x_b$, $x_a > x_b$ ou $x_a < x_b$, também se pode dizer que *a* é $x_a - x_b$ unidades diferente de *b*, podendo existir ou não o zero absoluto.

A variável métrica afecta a cada unidade estatística de E, um número real, sendo por isso também designada de variável real.

Exemplo 4.3.16. Variáveis métricas

- Idade gestacional (semanas).
- Peso (gramas) da placenta.
- Comprimento (cm) do bebé.

4.3.9.1 Matriz score, X (nxn), da variável métrica

> *Lui associer une matrice sur E^2 est évidemment quelque peu artificiel. Cependant le résultat justifie la méthode puisque l'on retrouve, à un coefficient près le coefficient de corrélation*. (Le Calvé, 1977)

Definição 4.3.10. *Score* da variável métrica

O *score* da variável métrica é definido da seguinte maneira:

$$\begin{aligned} x_{ii'} &= X(i) \quad \text{, se } i \neq i' \\ x_{ii} &= 0 \end{aligned} \qquad (4.3.27)$$

Sendo X(i) o valor que a variável X toma para a unidade estatística i.

Obtemos matrizes *score* com o aspecto peculiar que se pode observar no exemplo que se segue.

Exemplo 4.3.17. *Score* da variável métrica

Seja a variável X=" Idade (anos)".

A variável X toma os valores X= [21 22 24 34]^T, para quatro indivíduos.

A matriz *score* correspondente é dada por:

$$X = (x_{ii'}) = \begin{bmatrix} 0 & 21 & 21 & 21 \\ 22 & 0 & 22 & 22 \\ 24 & 24 & 0 & 24 \\ 34 & 34 & 34 & 0 \end{bmatrix}$$

4.3.9.2 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis reais

Quando se comparam duas variáveis métricas, o coeficiente s toma valores no conjunto dos números reais, sendo: $s = (n-1)\sum_i x_i y_i$ (4.3.28), em que x_i e y_i designam, respectivamente, os valores de X(i) e Y(i).

Neste caso, o coeficiente *s* representa, a menos de um factor que depende da dimensão da amostra/população, o produto escalar entre as duas variáveis.

No caso de duas variáveis métricas, o coeficiente s_{LC} coincide, a menos de um factor que depende da dimensão da amostra/população, com o coeficiente de correlação linear de Pearson, r : $s_{LC} = \sqrt{n-1}\, r$, sendo

$$r = \frac{\sum_i x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\left(\sum_i x_i^2 - n\bar{x}^2\right)\left(\sum_i y_i^2 - n\bar{y}^2\right)}} \qquad (4.3.29).$$

O coeficiente s_{LC} toma valores entre $\pm\sqrt{n-1}$.

A interpretação dos valores obtidos pelos coeficientes s_{LC} e P_L é a habitual para este tipo de variáveis:

- Quando a semelhança s_{LC} é fortemente negativa ($s_{LC}<<0$), i.e., a relação padronizada entre as variáveis é forte e inversa, o coeficiente P_L toma valores muito próximos de 0 ($P_L \approx 0$). Assim sendo, as unidades estatísticas que estão em relação numa das variáveis não o estão na outra variável. Há pois, uma tendência para emparelhar valores elevados de uma das variáveis com valores baixos da outra e vice-versa.
- Quando a semelhança s_{LC} é nula, i.e., a relação padronizada entre as variáveis é nula, $P_L=0.5$. Neste caso, o valor do coeficiente *s* é igual ao valor médio da variável aleatória S.
- Quando a semelhança s_{LC} é fortemente positiva ($s_{LC}>>0$), isto é, a relação padronizada entre as variáveis é forte e directa, o coeficiente P_L toma valores muito próximos de 1 ($P_L \approx 1$), Neste caso, as unidades estatísticas que estão em relação numa das variáveis também o estão na outra variável. Há uma tendência para emparelhar valores elevados (baixos) de uma das variáveis com valores elevados (baixos) da outra variável.

Neste caso pode-se dizer que a matriz S_{LC} é s.d.p., pois como se sabe a matriz R (matriz das correlações) é s.d.p..

Os resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre variáveis métricas são os mesmos que os obtidos com a ACP normada da matriz de dados original, a menos de uma translação que depende da dimensão n da amostra/população em estudo. Os valores próprios são $\sqrt{n-1}$ vezes maiores e as coordenadas são $\sqrt{\sqrt{n-1}}$ vezes maiores (Anexo 2). As representações gráficas das variáveis obtidas com as duas ACP são pois as mesmas, a menos de uma translação que depende da dimensão da amostra ou população em estudo. Verificamos que os resultados que obtivemos são os mesmos que os obtidos com a ACP clássica sobre as escalas de percepções (em percentagem) do Questionário SERVQUAL Modificado (Bacelar Nicolau *et al.,* 2005).
Quando utilizamos o coeficiente s_{LC} não precisamos pois, de nos preocupar com a centragem e redução das variáveis que são medidas em escalas com unidades de medida diferentes.

Neste caso, obtêm-se os mesmos resultados com os algoritmos de ACHA que utilizem na primeira etapa, quer o coeficiente de correlação de Pearson, quer o coeficiente de semelhança s_{LC} (Anexo 2).

As escalas de percepções do Questionário SERVQUAL Modificado perdem informação ao serem descritas como variáveis métricas. Sugerimos, com vantagem, a sua descrição de forma simbólica modal, como veremos mais à frente (Subsecção 4.3.10.10).

4.3.10 Variáveis simbólicas/complexas

Até agora considerámos que as unidades estatísticas se referem a indivíduos/objectos descritos por variáveis que tomam um único valor; Bock e Diday (2000), entre outros, designam estes objectos por *objectos de 1ª ordem* e estas variáveis por *variáveis clássicas*. Nesta secção, consideramos que:

- as unidades[157] estatísticas também podem ser classes ou grupos/amostras/populações de indivíduos, i.e., *objectos de 2ª ordem*.
- o conjunto das unidades estatísticas (quer sejam objectos de 1ª ou de 2ª ordem), E, pode ser descrito por variáveis simbólicas[158],Y_1, ..., Y_p, categóricas com valores

[157] Entidades (Bock e Diday, 2000).
[158] *Symbolic variable* (Bock e Diday, 2000); variável complexa (Bacelar-Nicolau, 2000; Sousa, 2005).

múltiplos[159], categóricas com todas as modalidades estrita e totalmente ordenadas pelas unidades estatísticas, intervalares[160] ou modais[161].

A matriz de dados, que cruza a informação das unidades estatísticas e das variáveis que as descrevem, é, neste caso, mais complexa do que as matrizes de dados "tradicionais" que temos vindo a apresentar. Na intercepção de cada linha com cada coluna da matriz de dados podemos encontrar um conjunto finito de valores ou um intervalo ou uma distribuição de frequências/frequências relativas/probabilidades, i.e., barras ou histogramas.

Definição 4.3.11. Variável simbólica

Uma variável simbólica Y com domínio $\mathcal{Y}$ é uma aplicação, $E \rightarrow B$, definida sobre um conjunto E de entidades/unidades estatísticas (indivíduos, classes, objectos,...), que toma os seus valores no conjunto imagem B.

Bock (2000a) dá-nos uma síntese dos tipos de variáveis simbólicas, que depende da especificação do conjunto imagem B em termos do domínio $\mathcal{Y}$: Variável clássica com um único valor, variável *conjunto de valores*[162] – quer seja variável intervalar, quer seja variável com valores múltiplos (categórica ou quantitativa) – e variável modal. Também podem ser considerados tipos de dados mais gerais em que *B* pode ser um conjunto de taxonomias ou hierarquias em $\mathcal{Y}$.

Se $B=\mathcal{Y}$, temos a variável "clássica" com um único valor[163], que é um caso particular da variável *conjunto de valores* quando o cardinal de Y(k) é igual a 1, $|Y(k)|=1$, para todo o k de E.

Definição 4.3.12. Variável *conjunto de valores*

Uma variável Y, com domínio $\mathcal{Y}$, definida para todos os elementos k do conjunto E de entidades estatísticas (objectos de 1ª e de 2ª ordem), designa-se variável *conjunto de*

---

[159] *Multi-valued variable* (Bock, 2000a), variável categórica multi-estado (Sousa, 2005), variável categórica multi-valor (Brito, 2008). Designação proposta, por nós, por analogia com "perguntas de respostas múltiplas".
[160] *Interval variable* (Bock, 2000a).
[161] *Modal variable* (Bock, 2000a).
[162] *Set-valued variable* (Bock, 2000a).
[163] *Single-valued variable* (Bock, 2000a).

valores se toma os seus valores, Y(k), no conjunto potência[164] de $\mathcal{Y}$, $P(\mathcal{Y})$={U\U $\subseteq\mathcal{Y}$ }: Y(k) $\subseteq\mathcal{Y}$, $\forall k\in E$, possivelmente com algumas restrições na dimensão ou na estrutura de U. É de notar que o conjunto vazio U, U= { }, dever ser excluído na maioria das aplicações.
Neste caso, $B= P(\mathcal{Y})$.

Quatro tipos de variáveis conjunto de valores são apresentados nas subsecções que se seguem: variáveis categóricas com valores múltiplos (Subsecção 4.3.10.1), variáveis categóricas com todas as modalidades ordenadas pelas unidades estatísticas (Subsecção 4.3.10.4), variáveis intervalares (Subsecção 4.3.10.5) e variáveis modais (Subsecção 4.3.10.8).

4.3.10.1 Variáveis categóricas com valores múltiplos

A variável categórica com valores múltiplos é um dos tipos de variável *conjunto de valores* que consideramos.

Definição 4.3.13. Variável com valores múltiplos. Variável categórica com valores múltiplos

Uma variável (categórica ou quantitativa) com valores múltiplos, Y, é uma variável *conjunto de valores*, em que os seus 'valores' Y(k) são todos subconjuntos finitos do domínio subjacente $\mathcal{Y}$: Y(k) $\subseteq\mathcal{Y}$ e $|Y(k)|< \infty$, $\forall$ k$\in$E.

No caso da variável *conjunto de valores* possuir uma gama finita $\mathcal{Y}$ de categorias, tais que *a fortiori*[165], toma valores finitos, Y(k), para todos os elementos k de E, dizemos que é uma variável categórica com valores múltiplos.

O exemplo seguinte permitirá esclarecer a definição apresentada.

Exemplo 4.3.18. Variável categórica com valores múltiplos

A questão B.47 é uma das questões de resposta múltipla, sobre hábitos alimentares de um Questionário de Caracterização Sociográfica:

> B.47**.** No último mês... Quais as refeições que fez habitualmente por dia?
> 5.Pequeno almoço, 4.Almoço, 3.Lanche/merenda, 2.Jantar, 1.Outras.

[164] Conjunto de todos os subconjuntos não vazios de $\mathcal{Y}$.
[165] "com maior razão".

No contexto da análise simbólica, esta questão é uma variável categórica com valores múltiplos, que podemos enquadrar da seguinte maneira: - Consideremos as respostas dadas pelos dez idosos(as) de uma amostra E. Sendo $\mathcal{Y}$ = {5.Pequeno almoço, 4.Almoço, 3.Lanche/merenda, 2.Jantar, 1.Outras} o conjunto de todas as refeições possíveis, então a variável Y(k)='As refeições que o indivíduo k fez, habitualmente, por dia, no último mês' pode dar origem a uma tabela de dados do tipo:

Tabela 4.3.2. Matriz de dados simbólicos de uma variável categórica com valores múltiplos, Y

k	Y(k)='As refeições que o indivíduo k fez, habitualmente, por dia, no último mês'
Maria	{5.Pequeno almoço, 4.Almoço, 2.Jantar}
António	$\mathcal{Y}$
João	{5.Pequeno almoço, 4.Almoço, 2.Jantar}
Rita	{5.Pequeno almoço, 4.Almoço, 3.Lanche/merenda, 2.Jantar}

Com o objectivo de analisar variáveis deste tipo, esta matriz de dados (Tabela 4.3.2) pode apresentar uma forma equivalente, em que a presença ou ausência do "valor"/categoria é registada, respectivamente, com uns e zeros numa matriz de dados (Tabela 4.3.3).

Tabela 4.3.3. Matriz de dados de uma variável categórica com valores múltiplos

k	Y(k)=' As refeições que o indivíduo k fez habitualmente por dia, no último mês'				
	5.Pequeno almoço	4.Almoço	3.Lanche/merenda	2.Jantar	1.Outras
Maria	1	1	0	1	1
António	1	1	1	1	1
João	1	1	1	1	0
Rita	1	1	0	1	0

4.3.10.2 Matriz score, X (nxn), da variável categórica com valores múltiplos

Seja X uma variável categórica com q valores múltiplos. Para cada unidade estatística i a variável X toma q valores que são uns ou zeros, conforme o "valor"/categoria da variável esteja presente ou não nessa unidade estatística:

$$X(i)= (x_{i1}, ..., x_{iq}) \text{ , sendo: } \begin{cases} x_{i\ell}=1 & \text{se o "valor" } \ell \text{ está presente} \\ x_{i\ell}=0 & \text{se não} \end{cases} , \ell =1, ...,q .$$

De forma equivalente e mais geral, podemos representar, para cada unidade estatística i, uma variável categórica, X_j, com q valores múltiplos, da seguinte maneira:

$$X_j(i)= (x_{ij1}, ..., x_{ijq}) \text{ , sendo: } \begin{cases} x_{ij\ell}=1 & \text{se o "valor" } \ell \text{ está presente} \\ x_{ij\ell}=0 & \text{se não} \end{cases} , \ell =1, ...,q .$$

Uma vez que, para cada variável deste tipo, cada unidade estatística é representada por um vector de zeros e/ou uns, pensamos que faz todo o sentido que o *score* da variável seja

definido pelas semelhanças entre os n vectores de dados "atributos de descrição". Essas semelhanças serão então calculadas a partir dos coeficientes de semelhança para dados binários (Capítulo 1, Subsecção 1.3.6). Aqui sugerimos a utilização, quer do coeficiente de Pearson, Φ (Definição 4.3.14), quer do coeficiente de Ochiai/afinidade, s_{ochiai}, (Definição 4.3.15).

Definição 4.3.14. *Score* da variável categórica com valores múltiplos

O *score* da variável categórica com valores múltiplos é definido da seguinte maneira:

$$x_{ii'} = \Phi(X(i),X(i')), \text{ se } i \neq i'$$
$$x_{ii} = 0 \qquad (4.3.30)$$

Sendo $\Phi = \frac{\text{ad-bc}}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$, tal como foi definido em (4.3.5). Em que, para estar de acordo com a terminologia utilizada por Hubálek (1982) e já apresentada na **Tabela 4.3.1**, ***a*** representa o número de co-presenças, ***b*** representa o número de presenças-ausências, ***c*** o número de ausências-presenças e ***d*** o número de co-ausências.

O coeficiente de Pearson, Φ, toma valores entre -1 e 1. A fórmula deste coeficiente torna-se indefinida quando, ou um ou ambos vectores de dados são todos zeros ou todos uns. Neste caso:

- O valor do coeficiente Φ é igual a 1 quando b+c=0, o que significa que as unidades estatísticas estão totalmente de acordo.
- O coeficiente Φ toma o valor -1 quando a+d=0, o que significa completo desacordo entre os vectores que estão a ser comparados.
- Quando o numerador se anula, ad-bc=0, o coeficiente Φ toma o valor 0.

Definição 4.3.15. *Score* da variável categórica com valores múltiplos

O *score* da variável categórica com valores múltiplos é definido da seguinte maneira:

$$x_{ii'} = s_{Ochiai}\,(X(i),X(i')) \text{ , se } i \neq i'$$
$$x_{ii} = 0 \qquad (4.3.31)$$

Relembremos que o coeficiente de Ochiai, s_{Ochiai}, não entra em conta com as co-ausências (Tabela 4.3.1):

$$s_{Ochiai} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

O coeficiente de semelhança de Ochiai toma valores entre 0 e 1. A fórmula do coeficiente é indefinida quando um dos vectores que estão a ser comparados, ou os dois, são constituídos só por zeros. Quando os dois vectores são nulos, o valor do coeficiente de Ochiai é 1. Quando apenas um dos dois vectores é nulo, o valor do coeficiente de Ochiai é 0.

As matrizes *score* assim obtidas são matrizes de semelhanças constituídas pelos valores do coeficiente de Pearson Φ, ou pelos valores do coeficiente Ochiai (como vimos coincide com o coeficiente de afinidade para variáveis binárias).

Exemplo 4.3.19. *Score* da variável categórica com valores múltiplos

Seja a variável X definida no Exemplo 4.3.18 (Tabela 4.3.3)

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

As matrizes *score* da variável X são, respectivamente, de acordo com as duas definições apresentadas:

$$X_{\Phi} = \begin{bmatrix} 0 & 0 & -0.2500 & 0.6124 \\ 0 & 0 & 0 & 0 \\ -0.2500 & 0 & 0 & 0.6124 \\ 0.6124 & 0 & 0.6124 & 0 \end{bmatrix}$$

$$X_{Ochiai} = \begin{bmatrix} 0 & 0.8944 & 0.7500 & 0.8660 \\ 0.8944 & 0 & 0.8944 & 0.7746 \\ 0.7500 & 0.8944 & 0 & 0.8660 \\ 0.8660 & 0.7746 & 0.8660 & 0 \end{bmatrix}$$

4.3.10.3 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis categóricas com valores múltiplos

Quando se comparam duas variáveis categóricas com valores múltiplos utilizando o coeficiente bruto *s*, ele representa o produto escalar de duas matrizes de semelhanças simétricas, quer tenham valores entre -1 e 1, quer tenham valores entre 0 e 1.

Pensamos que nesta interpretação temos de ter em conta o coeficiente de semelhança utilizado para definir a matriz *score*, pois nele se vai reflectir o peso/importância que se dá às co-presenças e co-ausências. A restante interpretação é análoga à que foi feita para as outras variáveis clássicas:

- Valores positivos e muito elevados do coeficiente centrado e reduzido, $s_{LC}>>0$, (o coeficiente probabilístico P_L toma valores muito próximos de 1, $P_L \approx 1$) significam que as unidades estatísticas tendem a obter padrões de resultados idênticos nas categorias de cada uma das variáveis.
- Valores negativos e muito baixos do coeficiente centrado e reduzido, $s_{LC}<<0$, (o coeficiente probabilístico P_L toma valores muito próximos de 0, $P_L \approx 0$) significam que as unidades estatísticas tendem a obter padrões diferentes de resultados nas duas variáveis.

Exemplo 4.3.20. Semelhanças s, s_{LC} e P_L entre variáveis categóricas com valores múltiplos

Pretende-se relacionar os resultados obtidos pela variável B.47 (Exemplo 4.3.18), na altura em que foi aplicado o Questionário de Caracterização Sociográfica pela primeira vez, com os resultados obtidos pela mesma variável uns anos depois, B.47_2, sobre os mesmos 67 indivíduos.

As matrizes *score* foram obtidas com o coeficiente Φ e os resultados foram os seguintes:

$s(B47, B47\_2) = 2226$, $s_{LC}(B47, B47\_2)= 1.2962$, $P_L(B47, B47\_2)= 0.9025$.

Verifica-se que os indivíduos tendem a dar os mesmos padrões de resposta às duas questões, não entrando em conta com as co-ausências. Neste caso, as opções *5.Pequeno almoço*, *4.Almoço* e *2.Jantar* foram referidas por todos (67-100%), ou com excepção de um (66-98.5%), antes e uns anos depois. As opções *3.Lanche /merenda* (respostas sim na fase 1: 49-73,1%) e *1.Outras* (respostas sim na fase 1: 25-37.3%) também não sofreram alterações importantes. ”Todos” tomam as refeições principais, a maioria lancha (respostas “Sim” na fase 1: 49-73,1%) e não opta por outras refeições (respostas “Não” na fase 1: 42-62,7%), embora se verifique nesta opção, de *outras* refeições, um ligeiro aumento (2 indivíduos deixaram de lanchar e passaram a optar por outras refeições, enquanto que 3 dos que lancham passaram também a optar por outras refeições).

4.3.10.4 Variáveis categóricas com todas as modalidades ordenadas pelas unidades estatísticas (escala ipsativa)

Um caso particular da variável com valores múltiplos é-nos dado pela variável categórica com todas as modalidades ordenadas pelas unidades estatísticas[166]. Esta variável é um outro tipo de variável conjunto de valores (Definição 4.3.13), em que o domínio $\mathcal{Y}$ é estrita e totalmente ordenado munido da relação de ordem estrita, <, (Definição 4.3.6).

[166] Pensamos que este caso não está contemplado em Bock (2000a).

Na prática, esta variável surge-nos frequentemente quando nos pedem para dispor por ordem[167] de preferência, ou seguindo um critério, todas as modalidades/opções de uma variável, sem possibilidade de empates[168]. Streiner e Norman (2003) referem este procedimento como *ranking/*ordenação duma série de alternativas e distinguem-no bem dos casos apresentados anteriormente, em que os sujeitos/unidades estatísticas atribuem pontuações de uma escala ordinal a itens/características/variáveis em estudo, designando este procedimento por *rate*.

Exemplo 4.3.21. Variáveis categóricas com todas as modalidades ordenadas pelas unidades estatísticas

- Pretende-se saber como os sujeitos valorizam dez qualidades, tais como, Saúde, Felicidade e Prosperidade, pedindo-lhes para as ordenarem por ordem de preferência.
- Considere-se a variável "Técnica cirúrgica" com quatro modalidades (a, b, c, d). Solicita-se a dois cirurgiões, i e i', que as disponham por ordem de preferência. Os *rankings*/ordenações obtidas foram:

	Técnica cirúrgica			
Cirurgiões	a	b	c	d
Cirurgião i	3	1	4	2
Cirurgião i'	3	2	4	1

A vantagem de usar este tipo de variáveis, ou seja, de ordenar[169] as categorias da variável segundo um critério, contrariamente a atribuir valores de uma escala ordinal[170], é a de obrigar os sujeitos, que respondem, a diferenciar melhor as possibilidades de resposta. Contudo, podem existir "problemas" na análise multivariada destes dados, com a utilização deste tipo de variáveis, que nos são apresentados, nomeadamente, por aqueles autores.
Quando se utilizam todos os valores da ordenação, a escala torna-se *ipsativa*[171]. Ou seja, a soma das ordens atribuídas aos itens/categorias é sempre a mesma e a mesma para todas as unidades estatísticas. Se os sujeitos tiverem que ordenar k itens/categorias a soma das ordens será k(k+1)/2 e a média (k+1)/2.

[167] *Rank,* em inglês.
[168] Também podemos pensar na possibilidade de existirem empates …
[169] *Ranking,* em inglês.
[170] *To rate,* em inglês.
[171] *Ipsative*, em inglês.

Streiner e Norman (2003) apresentam a situação de se pretender saber como os sujeitos valorizam a Saúde em relação a outras nove qualidades, tais como, Felicidade e Prosperidade, pedindo-lhes para as ordenarem por ordem de preferência e registando apenas as pontuações/ordens obtidas pela Saúde. É uma forma de usar a ordenação e de ultrapassar o problema estatístico referido por estes autores.

Neste caso, sugerimos que o *score* desta variável seja definido pelo coeficiente de correlação de Spearman, r_S, ou, eventualmente, pelo coeficiente tau de Kendall, τ.
Este assunto merece um estudo mais aprofundado, tendo em atenção a literatura que se debruça sobre ele. Pensamos resolver o problema da representação gráfica multivariada deste tipo de variáveis, utilizando os coeficientes s_{LC} e P_L, num futuro próximo.

4.3.10.5 Variáveis intervalares

A variável intervalar é uma variável conjunto de valores, que se pode exprimir sob a forma de intervalo $[\alpha,\beta]$: $\forall k \in E, Y(k) = [\alpha,\beta]$, quando uma relação de ordem $\preceq$ está definida no domínio $\mathcal{Y}$.

Definição 4.3.16. Variável intervalar

Uma variável conjunto de valores Y é uma variável intervalar se, para qualquer $k \in E$, o subconjunto $U=Y(k) =[\alpha,\beta]$ é um intervalo de $\mathbb{R}$ ou um intervalo de $\mathcal{Y}$, segundo a ordem $\prec$ estabelecida em $\mathcal{Y}$ (i.e., a ordem natural para $\mathcal{Y}=\mathbb{R}$): $Y(k)=[\alpha,\beta]$ com $\alpha,\beta \in \mathcal{Y}$, com $\alpha \leq \beta$ e $\alpha \preceq \beta$, respectivamente (possivelmente com exclusão de um ou ambos os limites do intervalo).
Assim $B = P(\mathcal{Y})$ é o conjunto $\mathfrak{I}$ de todos os intervalos em $\mathcal{Y}$.

Em Medicina, encontramos um exemplo desta variável quando se registam diariamente as temperaturas mínima e máxima dos doentes. Noutras situações, os intervalos podem ser definidos antecipadamente, tal como se apresenta no exemplo seguinte.

Exemplo 4.3.22. Variável intervalar

Num Questionário de Caracterização Sociográfica:

A48.Com que dinheiro conta mensalmente?:

1.]0, 300 euros], **2.**]300, 600 euros], **3.**]600, 1200 euros], **4.**]1200 euros, ...]

4.3.10.6 Matriz score da variável intervalar

Quando a variável simbólica é uma variável intervalar, propomos que a matriz *score* associada a esta variável seja a matriz, não simétrica, associada à distância não simétrica de Hausdorff. A distância de Hausdorff entre dois conjuntos é a distância máxima entre um dos conjuntos e o ponto mais próximo do outro conjunto.

Definição 4.3.17. Distância não simétrica de Hausdorff

A distância não simétrica de Hausdorff entre um conjunto A e um conjunto B, H(A,B), é uma função max-min definida por:

$$H(A,B) = \max_{a \in A} \left\{ \min_{b \in B} \{ d(a,b) \} \right\} \qquad (4.3.32)$$

Sendo *a* e *b*, respectivamente, os pontos dos conjuntos A e B, e d(A, B) uma métrica entre esses pontos. Consideramos que d é a distância euclidiana.

Informalmente, pode-se calcular a distância do conjunto A para o conjunto B, encontrando um ponto *a* de A, que seja o mais afastado de B, e então calcula-se a distância de *a* ao ponto mais próximo de B.

Se a distância de Hausdorff de A a B é igual a um valor *e*, $H(A,B)=e$, então, qualquer que seja o ponto de A, existe um ponto de B com distância inferior a *e*. Se estamos num ponto de A temos a certeza que com um "salto", *e*, chegamos a B.
Habitualmente, é uma dissemelhança não simétrica, $H(A,B) \neq H(B,A)$.

Exemplo 4.3.23. Matriz *score* da variável intervalar

A matriz *score* correspondente ao vector X= [[37, 38.5], [39, 40], [36, 40], [37, 39], [38, 40]]^T das temperaturas [min, max] dos 5 indivíduos, registadas num dia, é dada pela matriz de distância não simétrica:

$$X = \begin{bmatrix} 0 & 2.0 & 0 & 0 & 1.0 \\ 1.5 & 0 & 0 & 1.0 & 0 \\ 1.5 & 3.0 & 0 & 1.0 & 2.0 \\ 0.5 & 2.0 & 0 & 0 & 1.0 \\ 1.5 & 1.0 & 0 & 1.0 & 0 \end{bmatrix}$$

Pois:

H([37, 38.5], [39, 40])= 39-37= 2, H([39, 40], [37, 38.5])= 40-38.5=1.5
H ([37, 38.5], [36, 40])= 0, H([36, 40], [37, 38.5])= 40-38.5= 1.5
H ([37, 38.5], [37, 39])= 0, H([37, 39], [37, 38.5])= 39-38.5= 0.5
H ([37, 38.5], [38, 40])= 38-37= 1, H([38, 40], [37, 38.5])= 40-38.5= 1.5, ...

A distância simétrica de Hausdorff, $d_H(A,B)$, é, habitualmente, definida por $d_H(A,B)=\max(H(A,B), H(B,A))$ (4.3.33). Contudo, utilizaremos a distância não simétrica pois os resultados obtidos nas aplicações mostraram-se muito satisfatórios.

Convém chamar a atenção para o facto das matrizes *score,* que representam estas variáveis, serem matrizes de distâncias. Como vimos, habitualmente, as matrizes *score* são matrizes de semelhança. Por isso, quando compararmos variáveis intervalares com variáveis cujas matrizes *score* são matrizes de semelhanças, utilizaremos a seguinte transformação afim: $S_H(A,B)= \max H(A,B) – H(A,B)$ (4.3.34).
Se se compararem apenas variáveis intervalares, não é necessário utilizar aquela transformação. Os resultados serão os mesmos, mas teremos que ter isso em atenção quando fizermos a sua interpretação.

4.3.10.7 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis intervalares

O coeficiente bruto s é o produto escalar de duas matrizes de distâncias não simétricas ou das correspondentes matrizes de semelhanças, caso se tenha usado a transformação referida acima (Expressão 4.3.44). O coeficiente s_{LC} é a semelhança padronizada e o coeficiente P_L o coeficiente probabilístico.

- Quando a semelhança s_{LC} é fortemente positiva ($s_{LC}>>0$), i.e., a relação padronizada entre as variáveis é forte e directa, o coeficiente P_L toma valores muito próximos de 1 ($P_L \approx 1$). Neste caso, as unidades estatísticas que estão em relação numa das variáveis também o estão na outra variável. Há uma tendência para emparelhar valores elevados (baixos) de uma das variáveis com valores elevados (baixos) da outra variável, entrando também em consideração com a amplitude dos intervalos.
- Quando $P_L=0.5$, a semelhança s_{LC} é nula, i.e., a relação padronizada entre as variáveis é nula.
- Quando a semelhança s_{LC} é fortemente negativa ($s_{LC}<<0$), i.e., a relação padronizada entre as variáveis é forte e inversa, o coeficiente P_L toma valores muito próximos de

0 ($P_L \approx 0$). Assim sendo, as unidades estatísticas que estão em relação numa das variáveis não o estão na outra variável. Há pois, uma tendência para emparelhar valores elevados (baixos) de uma das variáveis com valores baixos (elevados) da outra e vice-versa, entrando também em consideração com a amplitude dos intervalos.

É importante ver como é que os coeficientes s, s_{LC} e P_L se "comportam" quando os aplicamos a dados reais. Optámos por recorrer a dois exemplos. Um exemplo com uma interpretação óbvia dos resultados – os nossos resultados deverão estar de acordo com os que seriam de esperar, baseando-nos no conhecimento que temos sobre o assunto. O outro exemplo apresentado foi escolhido por haver a possibilidade de comparar os resultados que obtivemos com os obtidos por Chouakria *et al.* (2000), utilizando uma análise em componentes principais simbólica.

Exemplo 4.3.24. Comparação de variáveis intervalares - Temperaturas (Guru *et al.*, 2004; Sousa, 2005)

Os dados consistem nas temperaturas mínimas e máximas, em graus centígrados, registadas durante um determinado ano, em 20 cidades consideradas pelos observadores muito semelhantes e por isso fazendo parte de uma mesma classe, C: C= {0.Amesterdão, 1.Atenas*, 7.Copenhaga, 9.Francoforte[172], 10.Genebra, 13.Lisboa*, 14.Londres, 16.Madrid, 20.Moscovo, 21.Munique, 24.NY, 25.Paris, 26.Roma, 27.S. Francisco*, 28.Seúl*, 30.Estocolmo, 33.Tóquio*, 34.Toronto, 35.Viena, 36.Zurique} (Anexo 3). Estas cidades estão situadas a uma latitude entre 40º e 60º, com excepção das assinaladas com um asterisco, *. As cidades assinaladas foram incluídas nesta classe porque, embora se situem a uma latitude entre 0º e 40º, por estarem próximo da costa marítima, têm temperaturas baixas que se assemelham às das cidades que estão situadas a uma latitude entre 40º e 60º.

Interessa-nos comparar os meses do ano, tendo em conta a informação das suas temperaturas mínimas e máximas registadas nestas cidades, e representá-los graficamente. As matrizes de semelhanças S, S_{LC} e P_L entre as temperaturas foram obtidas e estão apresentadas no Anexo 3. Observando as semelhanças s_{LC} verificamos que os valores são todos positivos, o que indica a tendência para que, nas cidades consideradas, se emparelhem temperaturas baixas com temperaturas baixas e temperaturas elevadas com temperaturas elevadas, aumentando ou decrescendo os valores dos coeficientes de acordo com a estação do ano a que se refere o mês em que foram registadas as temperaturas. O que acabamos de referir encontra-se bem exemplificado na tabela seguinte, no que se refere, por exemplo, ao mês de Dezembro:

[172] *Frankfurt*.

Tabela 4.3.4. Valores do coeficiente de semelhança s_{LC} entre as temperaturas [min,max] de Dezembro e dos restantes meses do ano retirados da matriz S_{LC} (Anexo 3)

	Jan	Fev	Mar	Abr	Maio	Jun	Jul	Ago	Set	Out	Nov	Dez
Dez	7.9768	7.2235	7.0873	5.9717	2.4038	3.4828	3.2104	3.8774	5.4443	6.8342	7.1005	8.3912

Tal como seria previsto, a semelhança das temperaturas [min,max] registadas em Dezembro é máxima quando se relacionam estas com as do próprio mês (s=19321, s_{LC}=8.39, $P_L \approx 1$), é forte quando se relacionam com as dos meses de Inverno, decrescendo progressivamente nos meses correspondentes à Primavera e atingindo os valores mais baixos nos meses de Verão; a semelhança volta a aumentar, progressivamente, quando se relacionam as temperaturas de Dezembro com as obtidas nos meses de Outono, e o aumento mais elevado observa-se, de novo, quando se voltam a relacionar com temperaturas obtidas em Novembro. Este ciclo de semelhanças, que cresce e decresce, progressivamente, com as estações do ano, só é "alterado" pelo mês de Maio. As temperaturas [min,max] registadas em Maio apresentam, de forma geral, uma semelhança mais fraca com as registadas nos outros meses. Pensamos que resulta do facto de Maio ser o que apresenta: - a menor variabilidade inter-quartis de amplitude térmica (tal como o mês de Junho, com o qual apresenta uma semelhança mais elevada, $s_{LCJun,Mai}$=4.44), – a amplitude térmica média (12.0 ºC) mais elevada, – e maior variabilidade de amplitude térmica (d.p.=5.84 ºC), devido à existência de *outliers* introduzidos pelas amplitudes térmicas das cidades 4.Francoforte, 8.Madrid, 20.Zurique e, especialmente, 18.Toronto, neste mês (Figura 4.3.5 e matriz S_{LC} em Anexo 3).

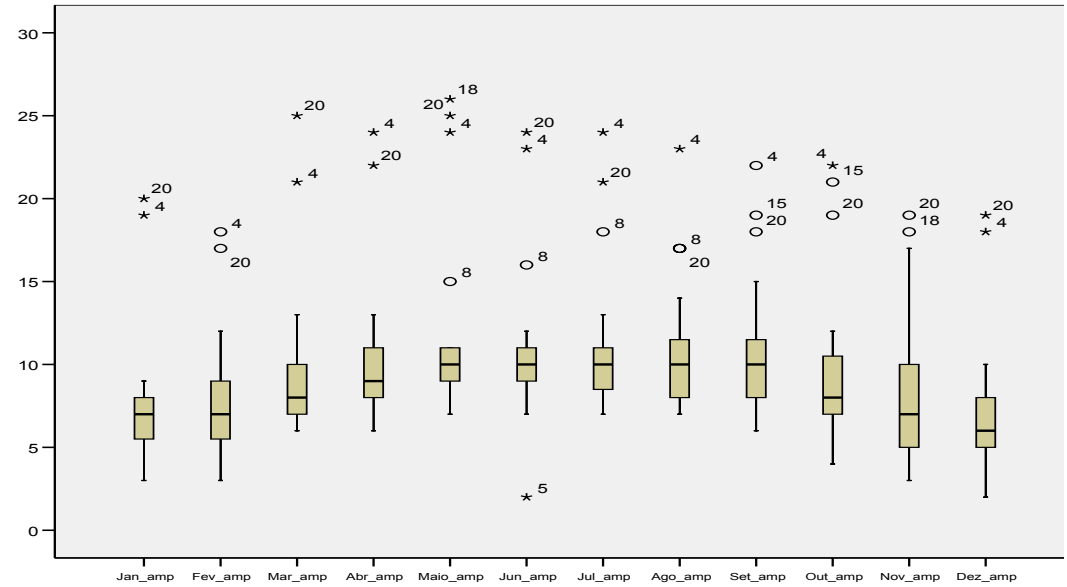

Figura 4.3.5. Diagramas em caixas de bigodes da distribuição das amplitudes térmicas registadas ao longo dos meses do ano, nas cidades indicadas na classe C. Os *outliers* são assinalados com os códigos atribuídos pelo SPSS às cidades: 4.Francoforte, 5.Genebra, 8.Madrid, 15.Seúl, 18.Toronto e 20.Zurique.

A interpretação comparativa dos valores do coeficiente s é mais difícil. Basta observar os valores apresentados na tabela que se segue para nos apercebermos disso.

Tabela 4.3.5. Valores do coeficiente de semelhança *s* entre as temperaturas [min,max] de Dezembro e dos restantes meses do ano retirados da matriz S (Anexo 3)

	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Dez	20738	20191	17424	12990	11237	10113	11282	11984	13218	15608	17109	19321

Para completar o estudo do comportamento destes coeficientes, sob o ponto de vista multivariado, realizámos análises em componentes principais e análises classificatórias hierárquicas ascendentes sobre as matrizes de semelhanças S, S_{LC} e P_L (Anexo 3) cujos resultados apresentamos, em parte, na Tabela 4.3.6, na Figura 4.3.6, na Figura 4.3.7 e na Figura 4.3.8. Os gráficos obtidos permitem-nos visualizar as relações entre as variáveis, assim como a existência eventual de grupos de variáveis.

As matrizes S e S_{LC} têm todos os valores próprios positivos, são, pois, matrizes definidas positivas.

Tabela 4.3.6. Resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre as temperaturas [min,max] dos meses do ano nas cidades da classe C

Eixos	Valores próprios	% de inércia	% de inércia acumulada
1	69.62	67.6	67.6
2	18.01	17.5	85.1
3	5.28	5.1	90.2

O 1º factor é um factor geral, que explica 67.6% da variabilidade total dos dados. O 2º factor, que explica 17.5% da variabilidade total, é o factor que opõe o Verão (meses com temperaturas mais elevadas) ao Inverno (meses com temperaturas mais baixas). No 1º plano factorial (85.1% de variabilidade total explicada) observamos várias associações: a dos meses mais quentes (Junho, Julho, Agosto), a dos meses mais frios (Dezembro, Janeiro, Fevereiro, Março, Novembro, Abril), a dos meses com temperaturas mais moderadas que os anteriores (Outubro e Setembro), enquanto Maio se afasta destes meses devido às características peculiares das suas amplitudes térmicas apresentadas acima e no Anexo 3.

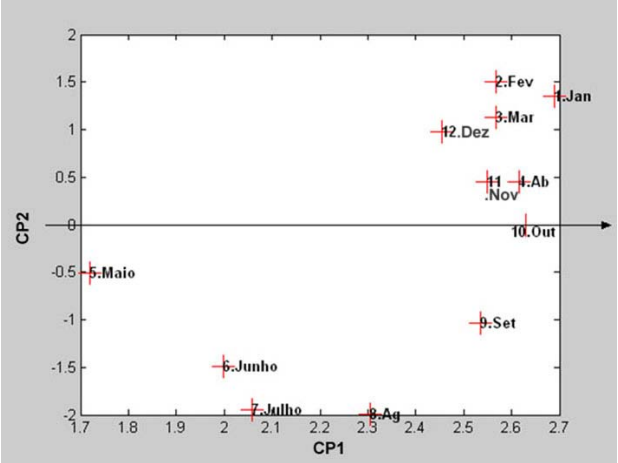

Figura 4.3.6. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças S_{LC} entre as temperaturas [min,max] dos meses do ano nas cidades da classe C.

Consideramos que o 3º factor, embora explique apenas 5.1% da variabilidade total, é importante. É o factor que representa as amplitudes térmicas observadas nos 12 meses daquele ano: opõe Maio (mês que apresenta, de forma geral, as maiores amplitudes térmicas) a Dezembro (mês que apresenta, de forma geral, as menores amplitudes térmicas). A conjugação deste factor com o 1º (Anexo 3) e com o 2º factor (Figura 4.3.7) faz sobressair a disposição "circular" das temperaturas dos meses do ano, nos respectivos planos factoriais.

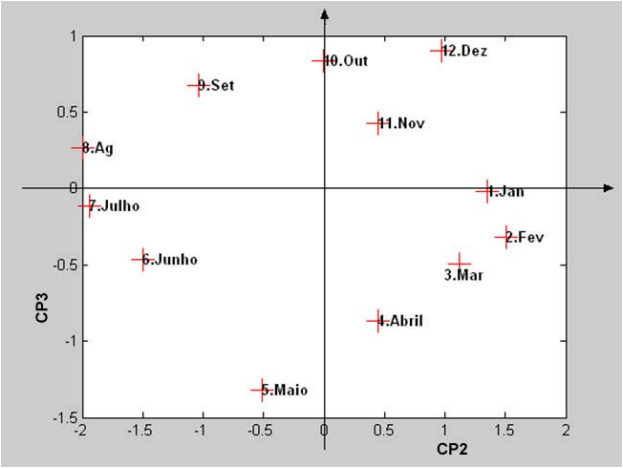

Figura 4.3.7. Representação gráfica do plano factorial (2,3) obtido com a ACP da matriz de semelhanças S_{LC} entre as temperaturas [min,max] dos meses do ano nas cidades da classe C.

O algoritmo de ACHA (s_{LC}+Ligação completa) permitiu obter a hierarquia e, em particular, no nível 9, a partição que traduz bem o que se passa no espaço factorial constituído pelos três primeiros factores: {Jan, Fev, Mar, Abr, Nov, Dez, Set, Out}, {Mai}, {Jun, Jul, Ag} (Figura 4.3.8).

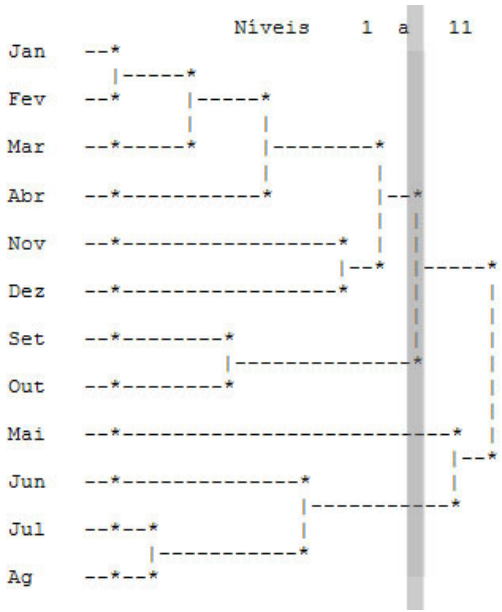

Figura 4.3.8. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+*Ligação completa*). O nível 9 é o mais importante, de acordo com o critério "estatística de níveis" (Lerman (1970); Bacelar-Nicolau (1972, 1980)), STAT(9)=6.2377.

Constatamos, assim, que o coeficiente s_{LC} entrou em consideração, não só com os valores mínimo e máximo dos intervalos das temperaturas, mas também com a sua amplitude, permitindo representar bem a continuidade "circular" das temperaturas dos meses, ao longo das estações do ano, nestas cidades.

O coeficiente P_L, como habitualmente, conduziu a semelhanças mais elevadas e como que "esmagou" as distâncias entre as temperaturas registadas ao longo dos meses do ano (Anexo 3). O coeficiente básico s permitiu comparar as temperaturas de forma interpretável. No entanto, neste caso, é preferível usar o coeficiente s_{LC} para comparar estas variáveis intervalares.

Não nos é possível comparar os resultados que obtivemos com os obtidos por Guru *et al*. (2004) e Sousa (2005), pois estes autores interessaram-se pela classificação das cidades.

Exemplo 4.3.25. Comparação de variáveis intervalares - Óleos e gorduras (e.g., Chouakria *et al.*, 2000; Ichino, 1988)

Os dados referem-se a oito óleos e gorduras descritos por quatro características quantitativas do tipo intervalar: "Gravidade específica", "Ponto de congelação", "Valor de iodo" e "Saponificação" (Tabela 4.3.7). Os valores, $[x_{ij}, x^{ij}]$ que se encontram na célula (i,j) desta tabela indicam que a j'ésima característica de um óleo que pertença à classe i toma valores entre x_{ij} e x^{ij}.

Tabela 4.3.7. Descrição de oito classes de óleos e gorduras por quatro características intervalares: gravidade específica (*gra*), ponto de congelação (*fre*), valor de iodo (*iod*) e saponificação (*sap*) (Chouakria *et al.*, 2000)

Óleos e gorduras	Gravidade específica (*gra*)	Ponto de Congelação (*fre*)	Valor de Iodo (*iod*)	Saponificação (*sap*)
Linhaça	[0.93 , 0.94]	[-27.00 , -18.00]	[170.00 , 204.00]	[118.00 ,196.00]
Perilla	[0.93 , 0.94]	[-5.00 , -4.00]	[192.00 , 208.00]	[188.00 ,97.00]
Semente de Algodão	[0.92 , 0.92]	[-6.00 , -1.00]	[99.00 , 113.00]	[189.00, 98.00]
Sésamo	[0.92 , 0.93]	[-6.00 , -4.00]	[104.00 , 116.00]	[187.00 ,193.00]
Camélia	[0.92 , 0.92]	[-25.00 , -15.00]	[80.00 , 82.00]	[189.00 , 193.00]
Azeite	[0.91 , 0.92]	[0.00 , 6.00]	[79.00 , 90.00]	[187.00 , 196.00]
Sebo de bovinos [173]	[0.86 , 0.87]	[30.00 , 38.00]	[40.00 , 48.00]	[190.00 , 199.00]
Sebo de suínos	[0.86 , 0.86]	[22.00 , 32.00]	[53.00 , 77.00]	[190.00 , 202.00]

Com o objectivo de reduzir o número de características descritivas, Chouakria *et al.* (2000) usaram dois métodos de análise em componentes principais simbólica. Apresentamos a representação gráfica das correlações entre as características iniciais e as componentes principais (Figura 4.3.9), obtidas com o método dos vértices[174] por Chouakria *et al*. (2000).

[173] Gordura de vaca. *Beef tallow*, em inglês.
[174] *Vertices method*, em inglês.

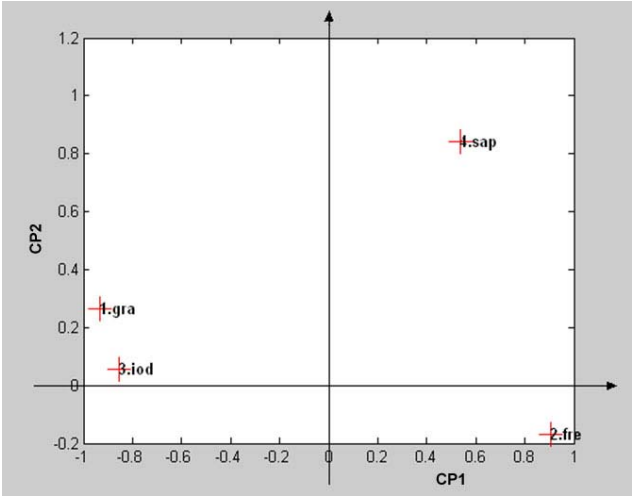

Figura 4.3.9. Método dos vértices – Diagrama de dispersão da correlação entre as componentes principais, cp1, cp2, e as características originais. (Valores próprios: 2.732 e 0.809, respectivamente; % de variabilidade: 68.29 e 20.23, respectivamente).

Em contrapartida, para atingir aquele objectivo, utilizámos análises em componentes principais e análises classificatórias hierárquicas ascendentes sobre as matrizes de semelhanças S_{LC} e P_L (Tabela 4.3.8 e Anexo 4, respectivamente) entre as quatro características referidas, uma vez que as unidades de medida das quatro características descritivas são diferentes. Utilizando estes coeficientes não é necessário centrar e reduzir as variáveis.

Tabela 4.3.8. Matriz de semelhanças S_{LC} entre quatro características intervalares dos óleos e das gorduras

	Gra	*Fre*	*Iod*	*Sap*
Gra	4.4605			
Fre	3.7010	4.3997		
Iod	1.7202	1.0573	4.6064	
Sap	0.1651	0.8432	1.5975	2.6567

A matriz S_{LC} é definida positiva, enquanto a matriz P_L não o é (Tabela 4.3.9 e Anexo 4, respectivamente). O 1º plano factorial explica 84.7% da variabilidade total (Tabela 4.3.9). A 1ª componente principal caracteriza-se essencialmente pela "Gravidade específica", pelo "Ponto de congelação", e pela oposição destas características a "Saponificação" e a "Valor de Iodo" (Figura 4.3.10).

Tabela 4.3.9. Resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre quatro características intervalares dos óleos e das gorduras

	Componente			
	1	*2*	*3*	*4*
Unidades de inércia (valores próprios)	9.28	4.38	2.06	0.41
Variabilidade Explicada (%)	57.53	27.20	12.75	2.52
Var. Exp. Acumulada (%)	57.53	84.73	97.48	100.00

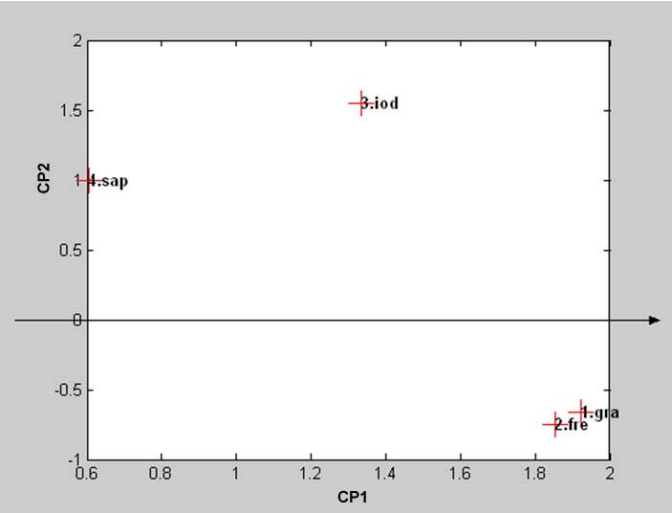

Figura 4.3.10. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças S_{LC} entre as quatro características intervalares dos óleos e gorduras.

Por sua vez os resultados obtidos com o algoritmo de ACHA (s_{LC}+Ligação única) complementam bem o que vimos no primeiro plano factorial. A partição obtida no nível 2, a melhor segundo o critério da estatística de níveis, também foi obtida por todos os outros algoritmos que utilizámos: {Gra, Fre}, {Iod}, {Sap} (Figura 4.3.11).

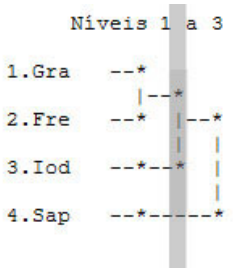

Figura 4.3.11. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+*Ligação única*). O nível 2 é o mais importante, de acordo com o critério "estatística de níveis" (Lerman (1970); Bacelar-Nicolau (1972, 1980)), STAT(2)=1.5275.

Os resultados obtidos com as duas análises – método dos vértices e ACP da matriz S_{LC} – apresentam em comum a "oposição" das variáveis "Gravidade específica" e "Saponificação".

A diferença reside na proximidade de “Gravidade específica” a “Valor de iodo”, opondo-as a “Ponto de congelação”, no método dos vértices (Figura 4.3.9).

Estes dados foram utilizados por diversos autores (e.g., Ichino, 1988; Ichino e Yaguchi, 1994; Chouakria *et al.*, 2000; Sato-Ilic e Oshima, 2006) para exemplificarem novos tipos de análises em componentes principais e de análises classificatórias sobre dados simbólicos. Sousa (2005) também os utiliza para obter partições dos óleos e das gorduras, utilizando a ACHA baseada nos coeficientes de afinidade generalizados para dados intervalares.

4.3.10.8 Variáveis modais

No caso mais simples, uma variável modal, Y, com domínio $\mathcal{Y}$ definido num conjunto de objectos E={a, b,...} é uma variável *conjunto de valores* com uma medida ou uma distribuição (de frequências, de probabilidade ou de pesos) associada ao conjunto das categorias Y(k) de domínio $\mathcal{Y}$, Y(k)$\subseteq \mathcal{Y}$, definidas para todo o objecto k do conjunto E.

Definição 4.3.18. Variável modal (Bock, 2000a)

Uma variável modal Y, com domínio $\mathcal{Y}$, definida sobre um conjunto de objectos E, E={a,b, ...}, é uma aplicação $Y(k)=\left(U(k),\ \pi_k\right)$, para $k \in E$ (4.3.35)

em que, π_k é uma medida não negativa ou uma distribuição de frequências ou de probabilidade ou uma ponderação sobre o domínio $\mathcal{Y}$ de possíveis valores observados e $U(k) \subseteq \mathcal{Y}$ é o suporte de π_k no domínio $\mathcal{Y}$.

De uma forma mais simples, podemos dizer que, uma variável modal Y com domínio $\mathcal{Y}$ é uma aplicação do conjunto das unidades estatísticas, E, na família $M(\mathcal{Y})$ de todas as medidas não negativas π em $\mathcal{Y}$, $Y: E \rightarrow B= M(\mathcal{Y})$, com valores $Y(k) = \pi_k$.

Se a medida π_k é especificada por um histograma, Y é uma variável histograma. Y é uma variável diagrama ou barra, se o domínio $\mathcal{Y}$ é finito e a medida π_k é descrita por um diagrama de barras.

Exemplo 4.3.26. Variável modal barras

Consideremos o conjunto E={a, b, c, d, e, f} constituído por 6 serviços de pediatria. A variável modal Y* descreve a frequência relativa de doenças do domínio $\mathcal{Y}$={C, D, S, O= outras} observadas nas crianças que acorreram à consulta desses serviços, em 2003:

$Y^* = \left((0.7, 0.2, 0.1, 0.0), (0.4, 0.4, 0.1, 0.1), (0.3, 0.4, 0.2, 0.1), (0.3, 0.4, 0.1, 0.2), (0.5, 0.2, 0.2, 0.1), (0.4, 0.3, 0.2, 0.1)\right)^T$

4.3.10.9 Matriz score da variável modal

A matriz *score* de uma variável modal é a matriz de proximidades entre as distribuições de frequências absolutas ou relativas, de percentagens ou de probabilidades, $\pi_i = (\pi_{i1}, ..., \pi_{ic})$ e $\pi_{i'} = (\pi_{i'1}, ..., \pi_{i'c})$, das unidades estatísticas i e i', associadas à variável com c categorias.

Entre as diversas medidas de proximidade propostas na literatura para distribuições – tais como o coeficiente de divergência-ϕ ou divergência-*I*, a distância de Hellinger, o coeficiente de divergência Kullback-Leibler, o coeficiente de divergência-χ^2 (Bock, 2000b) – optámos pelo coeficiente de afinidade (Bacelar-Nicolau, 2000, 2002; Capítulo 3) para definir o *score* da variável modal. A escolha do coeficiente de afinidade deve-se a ele ser, essencialmente, um coeficiente de semelhança que mede a extensão da relação monótona entre as raízes quadradas dos perfis de probabilidade correspondentes à variável.

Definição 4.3.19. *Score* da variável modal

O *score* da variável modal é definido da seguinte maneira:

$$\begin{aligned} x_{ii'} &= aff(\pi_i, \pi_{i'}) \text{ , se } i \neq i' \\ x_{ii} &= 0 \end{aligned} \qquad (4.3.36)$$

Sendo aff(π_i, $\pi_{i'}$), o coeficiente afinidade básico, definido por:

- $aff(\pi_i, \pi_{i'}) = \sum_{j=1}^{c} \sqrt{\pi_{ij}\pi_{i'j}}$ (4.3.37), no caso de $\pi_i = (\pi_{i1}, ..., \pi_{ic})$ e $\pi_{i'} = (\pi_{i'1}, ..., \pi_{i'c})$

serem distribuições de probabilidade ou de frequências relativas das unidades estatísticas, i e i', associadas à variável com c categorias.

- $aff(\pi_i, \pi_{i'}) = \sum_{j=1}^{c} \sqrt{\frac{n_{ij}}{n_{i.}} . \frac{n_{i'j}}{n_{i'.}}}$ (4.3.38), no caso de $\pi_i = (\pi_{i1}, ..., \pi_{ic})$ e $\pi_{i'} = (\pi_{i'1}, ..., \pi_{i'c})$

serem distribuições de frequências absolutas das unidades estatísticas, i e i', associadas à

variável com c categorias, e $n_{i.} = \sum_{j=1}^{c} n_{ij}$ é o número de unidades estatísticas que pertencem a, ou foram observadas na unidade estatística/classe i.

O coeficiente de afinidade básico toma valores entre 0 e 1, $0 \leq aff\left(\pi_i, \pi_{i'}\right) \leq 1$. A afinidade entre as distribuições π_i e $\pi_{i'}$ é igual a 1, $aff\left(\pi_i, \pi_{i'}\right) = 1$, se elas forem idênticas ou proporcionais, e é nula, $aff\left(\pi_i, \pi_{i'}\right) = 0$, se elas forem ortogonais.

4.3.10.10 O que representam os coeficientes s, s_{LC} e P_L quando se comparam variáveis modais

O coeficiente bruto s representa o produto escalar de duas matrizes de semelhanças cujas entradas são valores que variam entre 0 e 1 (inclusive). O coeficiente s_{LC} é a semelhança padronizada. P_L é o coeficiente probabilístico.
A interpretação é idêntica à dada para variáveis categóricas:

- Valores positivos e muito elevados do coeficiente centrado e reduzido, $s_{LC}>>0$ (o coeficiente probabilístico P_L toma valores muito próximos de 1, $P_L \approx 1$), significam que as unidades estatísticas tendem a obter "perfis"/distribuições idênticas ou proporcionais em cada uma das variáveis.
- Valores negativos e muito baixos do coeficiente centrado e reduzido, $s_{LC}<<0$ (o coeficiente probabilístico P_L toma valores muito próximos de 0, $P_L \approx 0$), significam que as unidades estatísticas tendem a obter padrões diferentes de perfis/distribuições nas duas variáveis. Significa que nenhum par de unidades estatísticas deu o mesmo perfil de respostas a uma e a outra variável, i.e., todos os pares de unidades estatísticas responderam de uma maneira diferente às duas variáveis. As unidades estatísticas, às quais correspondem perfis/distribuições idênticas ou proporcionais de uma das variáveis, também lhes correspondem perfis/distribuições diferentes na outra variável.

A utilização de um exemplo vai-nos permitir compreender melhor a interpretação dos coeficientes s, s_{LC} e P_L, quando aplicados a variáveis modais. Os dados[175], que escolhemos para exemplificar a aplicação dos coeficientes a este tipo de variáveis, têm a vantagem de terem sido analisados com outra técnica estatística (Sousa, 2005), o que permite comparar

[175] Estes dados são os únicos utilizados que não estão no âmbito da biomatemática.

os resultados obtidos. Esta vantagem é reforçada pelo facto da técnica estatística utilizada ser uma extensão da metodologia VL à Análise Classificatória Hierárquica Ascendente de variáveis simbólicas, com base no coeficiente de afinidade generalizado (Sousa, 2005).

Exemplo 4.3.27. Comparação de variáveis modais – Crenças religiosas (Sousa, 2005)

Os 517 inquiridos, residentes na Ilha de S. Miguel (Açores), foram agrupados de acordo com a frequência com que vão à Missa, em 8 grupos/classes/unidades estatísticas: "Nunca" (Nunca), "Raramente" (Rar), "Festas principais" (FesP), "Casamentos, baptizados e funerais" (CBF), "Uma ou duas vezes por mês" (UD), "Todos os domingos" (D), "Todos os domingos e até mesmo durante a semana" (DS) e "Quando sentem necessidade" (QSN). Estas unidades estatísticas têm, respectivamente, 14, 60, 29, 62, 41, 228, 14 e 69 indivíduos. As respostas dadas pelos inquiridos dizem respeito à opinião (C-Crê , D-Duvida, NC-Não crê, NS/NR-Não sabe ou não responde) acerca de dez afirmações, que fazem parte do discurso de doutrinação da Igreja católica: *V1- Deus é uno, mas em três pessoas*, *V2- Cristo é Deus*, *V3- Cristo realizou autênticos milagres*, *V4- O Papa não se engana quando fala verdades de fé*, *V5- Existe alguma coisa depois da morte, V6- Cristo salvou-nos morrendo pelos nossos pecados*, *V7- O diabo existe*, *V8- Existe um prémio para os bons e um castigo para os maus na outra vida*, *V9- O sacramento da confissão perdoa os pecados*, *V10- Todos nascem com o pecado original*. No processo de generalização inicialmente descrito, isto é, da constituição dos 8 grupos/unidades estatísticas, as variáveis qualitativas passaram a ser descritas por distribuições de frequências, tendo sido obtida uma matriz de dados simbólicos com 8 objectos simbólicos e 10 variáveis simbólicas modais (Anexo 5).

As afirmações V1, V2, V3, V6, V7, V9 e V10 são verdades da fé católica e as restantes são conteúdos da fé na Igreja. Interessa-nos, em primeiro lugar, comparar as opiniões sobre as afirmações, com base nas distribuições de frequências obtidas para cada grupo/unidades estatísticas e visualizá-las. Nesse sentido, foram calculadas as semelhanças s, s_{LC} e P_L entre as afirmações, e foram realizadas análises em componentes principais e análises classificatórias hierárquicas ascendentes baseadas naqueles coeficientes (Anexo 5).

A observação das semelhanças entre a afirmação *V2- Cristo é Deus* e as restantes variáveis (Tabela 4.3.10), complementada pelas tabelas de dados, e diagramas em caixas de bigodes (Figura 4.3.15, Anexo 5), vai permitir perceber melhor que a interpretação a dar a estas semelhanças é a que foi apresentada acima.

Tabela 4.3.10. Valores dos coeficientes de semelhança *s,* s_{LC} e P_L, respectivamente, entre as opiniões sobre *V2- Cristo é Deus* e as opiniões sobre as restantes afirmações, retirados do Anexo 5

Coef. Sem.		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
s	V2	49.16	51.02	49.34	48.84	50.06	49.47	50.54	50.10	48.97	48.76
s_{LC}	V2	2.413	3.594	2.118	0.087	0.555	1.842	2.542	0.272	2.544	- 0.227
P_L	V2	0.9920	0.9999	0.9829	0.5347	0.7106	0.9673	0.9945	0.6070	0.9945	0.4101

Estas relações encontram-se bem visualizadas no 1º plano factorial da ACP da matriz de semelhanças S_{LC}, cujos resultados apresentamos na Tabela 4.3.11 e na Figura 4.3.12 seguintes.

Tabela 4.3.11. Resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre as afirmações sobre a religião

	Componente		
	1	*2*	*3*
Unidades de inércia (valores próprios)	24.07	7.16	1.18
Variabilidade Explicada (%)	69.97	20.82	3.42
Var. Exp. Acumulada (%)	69.97	90.79	94.21

A matriz S_{LC} tem todos os valores próprios positivos, é, pois, uma matriz definida positiva.

O 1º factor explica 69.97% da variabilidade total dos dados. Com excepção da opinião sobre *V2- Cristo é Deus*, nele reconhecemos as classes da partição considerada como mais significativa por Sousa (2005): "Foi realizada a ACHA das 10 variáveis simbólicas com base no coeficiente de afinidade generalizado ponderado ... Como partição mais significativa, optámos pela partição: {V7, V9, V4, V5, V8, V10}, {V1, V2, V3, V6}, ...".

O facto da afirmação *V2- Cristo é Deus*, inesperadamente, à primeira vista, encontrar-se "afastada" da afirmação *V1- Deus é uno, mas em três pessoas*, traduz a existência de alguma diferença entre os perfis de respostas dadas pelas unidades estatísticas a estas questões. Neste caso, destaca-se a unidade estatística 3-DS (Todos os domingos e até mesmo durante a semana) que, pelas suas respostas, nos levam a questionar se os indivíduos desta classe sabem quem são "as três pessoas", pois uma grande percentagem, para este grupo, não reconhece Cristo como uma das "três pessoas" (Crê - 71%, Duvida - 7%, Não crê - 7%, NS/NR - 14%), crendo contudo, na sua maioria (93%), que *V1- Deus é uno, mas em três pessoas*.

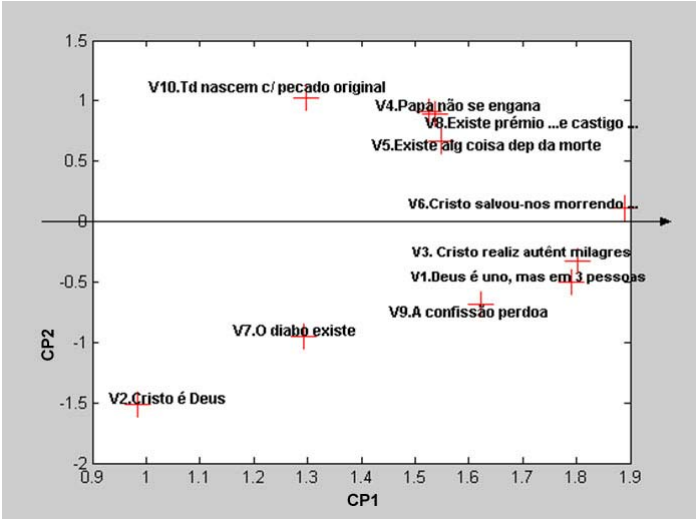

Figura 4.3.12. Representação das variáveis/afirmações no plano factorial (1,2) obtido com a ACP da matriz S_{LC} entre as afirmações religiosas.

O 2º factor, que explica 20.82% da variabilidade total, é o factor que opõe a afirmação *V2- Cristo é Deus* à afirmação *V10-Todos nascem com o pecado original* e no qual encontramos bem visualizadas as semelhanças descritas, anteriormente, entre a afirmação *V2- Cristo é Deus* e as restantes variáveis.

No 1º plano factorial, que explica 90.79% da variabilidade total dos dados, observamos a proximidade das respostas dadas às afirmações (V1, V3, V9, V7, V2) e (V6, V4, V5, V8, V10). Esta proximidade é encontrada também na melhor partição obtida, no nível 8, com o algoritmo de ACHA (s_{LC}+Ligação completa): {V1, V3, V9, V7, V2} e {V6, V4, V5, V8, V10} (Figura 4.3.13).

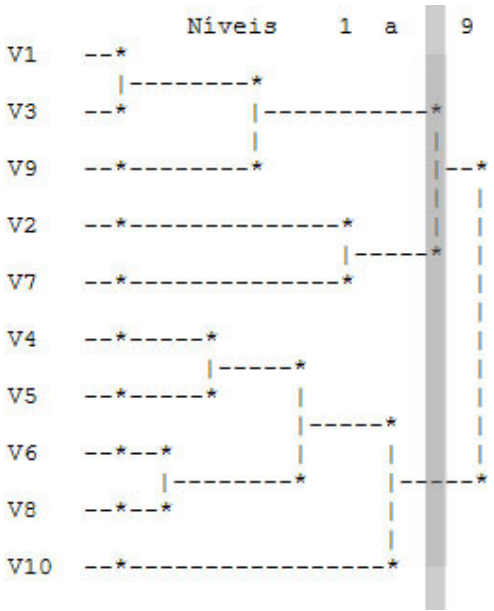

Figura 4.3.13. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+*Ligação completa*). O nível 8 é o mais importante, de acordo com o critério "estatística de níveis" (Lerman, 1970; Bacelar-Nicolau, 1972, 1980), STAT(8)= 4.2485.

As associações obtidas são coerentes, sob o ponto de vista da religião:

- Classe 1={V1, V3, V9, V7, V2}. Afirmações que fazem parte do discurso de doutrinação da Igreja, com incidência cultural forte:

V1- Deus é uno, mas em três pessoas

V3- Cristo realizou autênticos milagres

V9- O sacramento da confissão perdoa os pecados

V7- O diabo existe

V2- Cristo é Deus

Estas afirmações caracterizam-se por obter, de forma geral, maior percentagem de respostas "Crê" (P_{75}=77.77%).

- Classe 2={V6, V5, V8, V4, V10}. Esta classe é coerente e nela se concentram as afirmações que se referem aos conteúdos da fé na Igreja. Embora *V4- O Papa não se engana quando fala verdades de fé* seja uma afirmação mais "desgarrada" do contexto, a inclusão nesta classe mostra que os grupos reconhecem o Papa como referência da autoridade do ponto de vista doutrinal e moral:

V6- Cristo salvou-nos morrendo pelos nossos pecados

V5- Existe alguma coisa depois da morte

V8- Existe um prémio para os bons e um castigo para os maus na outra vida

V4- O Papa não se engana quando fala verdades de fé

V10- Todos nascem com o pecado original

Esta classe caracteriza-se por afirmações com percentagem de respostas "Crê" (P_{75}=63.3%) inferior à da Classe 1, e maior percentagem de respostas "Duvida", "Não crê", "Não sabe ou Não responde".

Tem também interesse observar a partição obtida no nível 6 do algoritmo de ACHA (s_{LC}+*AM*) (Figura 4.3.16): {V1, V3, V6, V9}, {V7}, {V2} e {V4, V5, V8, V10}. Esta partição complementa bem o que se passa na 1ª componente principal:

- Classe 1={ *V1-Deus é uno, mas em três pessoas*, *V3-Cristo realizou autênticos milagres*, *V6-Cristo salvou-nos morrendo pelos nossos pecados*, *V9-O sacramento da confissão perdoa os pecados*}. É constituída por verdades da fé católica que apresentam, de forma geral, maior consenso de respostas "Crê".

- Classe 2={*V7-O diabo existe*}, com distribuição de respostas "Crê" aproximada pela distribuição de respostas "Não crê" (Figura 4.3.14).

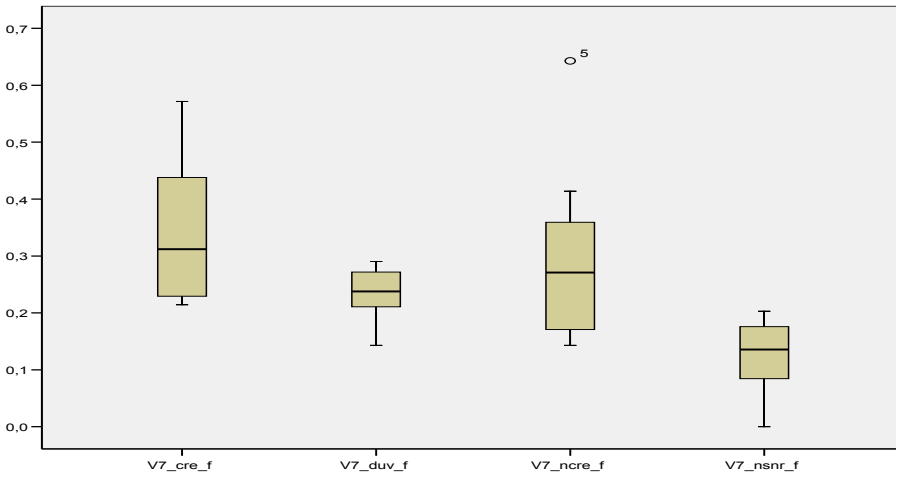

Figura 4.3.14. Diagramas em caixas de bigodes da distribuição de respostas "Crê", "Duvida", "Não crê", "Não sabe/Não responde", em proporção, à afirmação *V7-O diabo existe*. Está assinalada a resposta *outlier* dos que Nunca vão à Missa (5).

- Classe 3={*V2-Cristo é Deus*} que se caracteriza pela distribuição elevada da percentagem de respostas "Crê", com excepção das dadas pelos que "Nunca vão à Missa" (Figura 4.3.15), e mais baixa do que seria de esperar, as respostas dadas pelos que "Vão todos os domingos e até mesmo durante a semana à Missa" (71%).

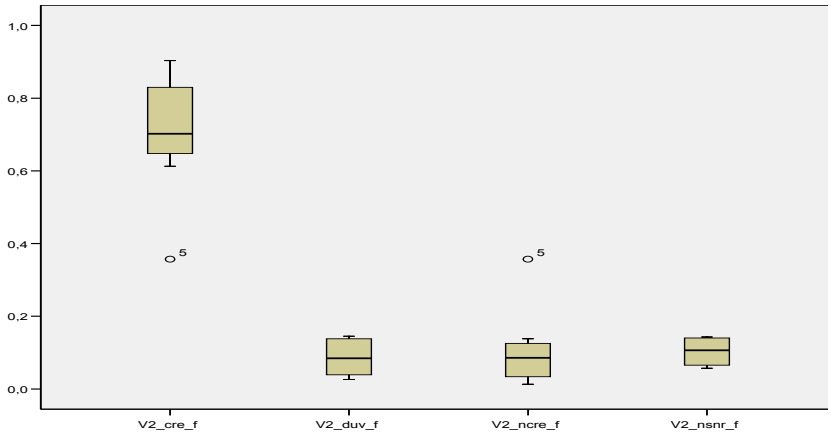

Figura 4.3.15. Diagramas em caixas de bigodes da distribuição de respostas "Crê", "Duvida", "Não crê", "Não sabe/Não responde", em proporção, à afirmação *V2- Cristo é Deus*. Está assinalada a resposta *outlier* dos que Nunca vão à Missa (5).

- Classe 4={*V4-O Papa não se engana quando fala verdades de fé*, *V5-Existe alguma coisa depois da morte*, *V8-Existe um prémio para os bons e um castigo para os maus na outra vida*, *V10-Todos nascem com o pecado original*}. É constituída, essencialmente, por conteúdos da fé na Igreja com percentagem de respostas "Crê" inferior à da Classe 1, e maior percentagem de respostas "Duvida", "Não crê", "Não sabe ou Não responde".

As associações obtidas são coerentes, sob o ponto de vista da religião, com excepção das respostas a *V2- Cristo é Deus*, como referimos acima.

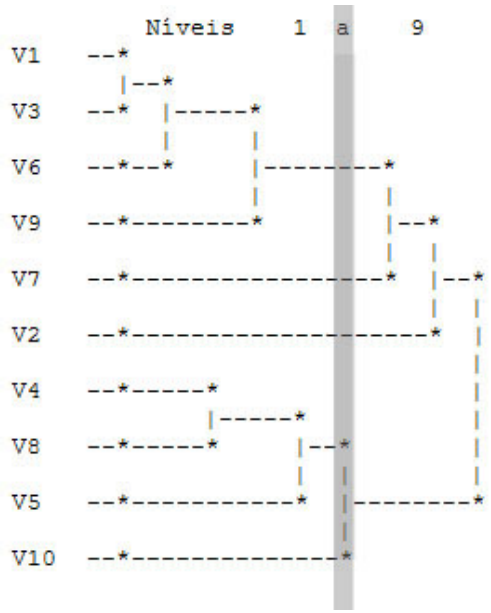

Figura 4.3.16. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+*AM*). O nível 6 é o mais importante, de acordo com o critério "estatística de níveis" (Lerman, 1970; Bacelar-Nicolau, 1972, 1980), STAT(6)= 4.5686.

Os resultados obtidos com o coeficiente P_L são análogos aos obtidos com o coeficiente s_{LC} (Anexo 5).

A estrutura destes dados não parece ser muito forte, pois obtemos partições que diferem de acordo com o algoritmo utilizado, embora as diferenças não sejam grandes.

Os resultados obtidos com os critérios de agregação sobre as semelhanças s_{LC} e P_L são, em parte, diferentes dos obtidos por (Sousa, 2005).

Queremos também chamar a atenção para um procedimento habitual, quer em Psicologia, quer em Medicina, quer noutras áreas, quando se estudam escalas constituídas por vários itens. Referimo-nos à quantificação das escalas pela soma das pontuações obtidas pelos seus itens. Nem sempre este procedimento é o mais adequado. O exemplo que tratamos, sucintamente, de seguida, sugere outro procedimento que consideramos ser mais correcto.

Exemplo 4.3.28. Comparação de variáveis modais – Escalas de Percepção do Questionário SERVQUAL Modificado (Doria *et al.*, 2007)[176]

O Bloco 1 do Questionário SERVQUAL Modificado é constituído por cinco escalas de percepções: "A. Elementos Tangíveis", "B. Fiabilidade dos Tratamentos e Cuidados", "C. Segurança/Garantia", "D. Interesse/Capacidade de Resposta" e "E. Empatia" (Secção 5.2,

176 Comunicação oral apresentada na XI Conferencia Española de Biometria e Primer Encuentro Iberoamericano de Biometria CEIB2007, Salamanca: *Comparison of Methodologies of Multivariate Analysis in a Palliative Care Context*.

Capítulo 5). Como vimos, " ... as respostas, aos itens daquelas escalas, são dadas sob a forma: 1-Totalmente em desacordo, 2-Desacordo, 3-Des/Acordo, 4-Acordo, 5-Totalmente de acordo, 6-Não se aplica, 9-Não sabe/Não responde. Os itens destas escalas são variáveis com modalidades parcialmente ordenadas, uma vez que entre as suas modalidades se verifica: "1<2<3<4<5, 6, 9." (Exemplo 4.3.6). Em Bacelar-Nicolau *et al.* (2005), para cada unidade estatística, calculou-se o resultado obtido em cada escala recorrendo à soma dos resultados obtidos nos itens da escala, em percentagem[177], eliminando as respostas "6-Não se aplica" e "9-Não sabe/Não responde". As escalas assim descritas são variáveis métricas e a análise subsequente entra em consideração com isso, tal como referimos na Subsecção 4.3.9.

Contudo, este procedimento leva à perda de informação. Por isso, sugerimos que as escalas sejam descritas, respectivamente, pelos perfis de respostas dados pelas unidades estatísticas. Isto é, em relação à escala/variável X, para cada unidade estatística i, regista-se o valor que a escala/variável toma para essa unidade estatística, $x_i = (p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}, p_{i6}, p_{i9})$, sendo p_{ij} a proporção de respostas dadas pela unidade estatística i a cada uma das pontuações j=1,2,3,4,5,6,9 da escala/variável X, $\sum p_{ij} = 1$. As escalas são, pois, variáveis simbólicas modais e podemos observar a sua distribuição na Figura 4.3.17. A matriz de dados é, neste caso, tridimensional.

[177] Em percentagem, porque as escalas são constituídas por diferente número de itens.

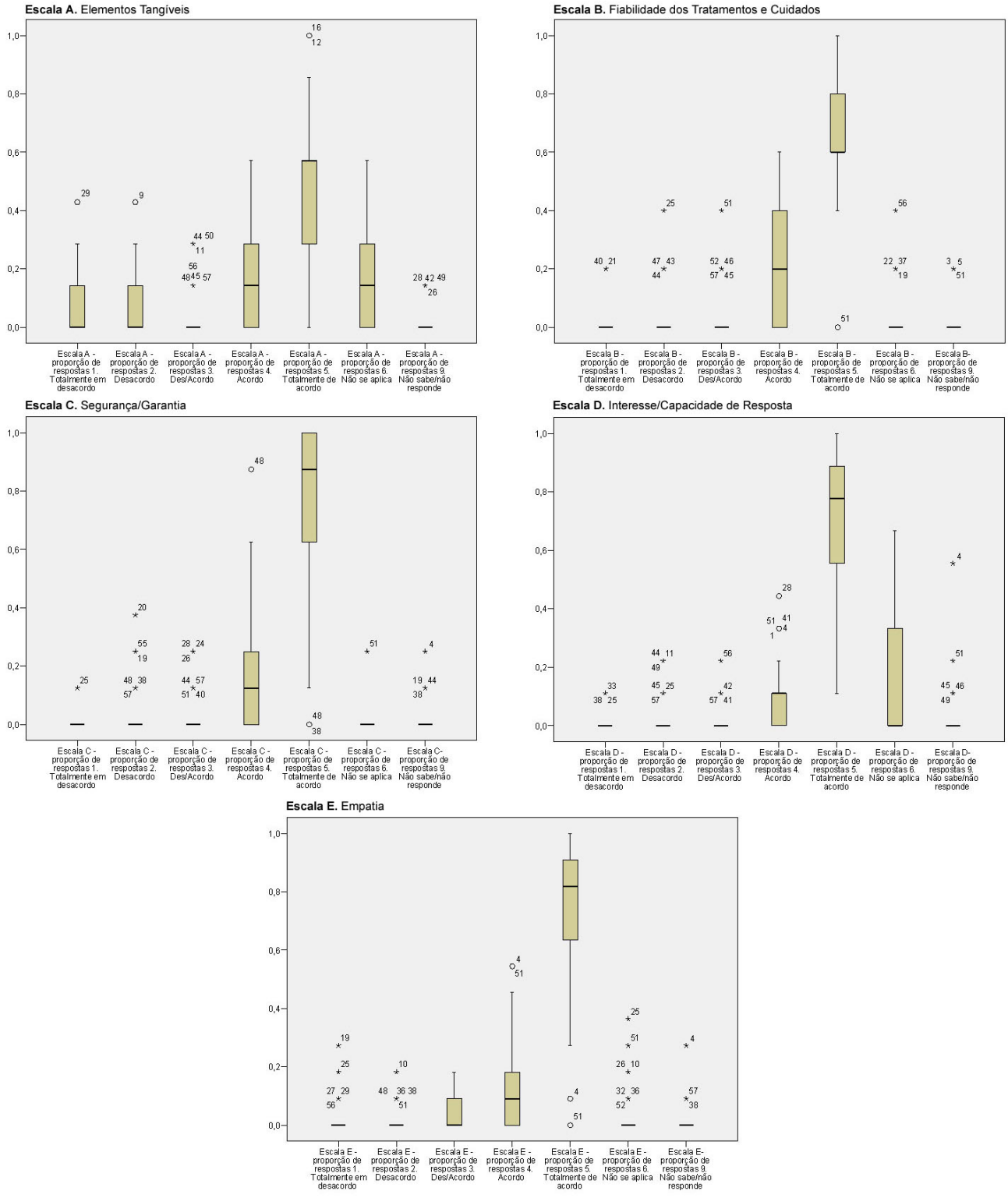

Figura 4.3.17. Diagramas em caixas de bigodes da distribuição de respostas "1-Totalmente em desacordo", "2-Desacordo", "3-Des/Acordo", "4-Acordo", "5-Totalmente de acordo", "6-Não se aplica", "9-Não sabe/Não responde", em proporção, às cinco escalas de percepção do Questionário SERVQUAL. Estão assinaladas as respostas consideradas *outlier* e os códigos *SPSS* das unidades estatísticas respectivas.

Com o objectivo de obter representações gráficas simples destas escalas e de encontrar tendências de resposta dos prestadores de cuidados paliativos, realizámos análises em componentes principais e análises classificatórias hierárquicas sobre as matrizes de semelhanças S_{LC} e P_L (Tabela 4.3.12 e Anexo 6, respectivamente).

Tabela 4.3.12. Matriz de semelhanças S_{LC} entre as cinco escalas de percepções do Questionário SERVQUAL Modificado

	A.Elementos Tangíveis	B.Fiabilidade	C.Segurança/ Garantia	D.Capacidade de Resposta	E.Empatia
A.Elementos Tangíveis	17.2098				
B.Fiabilidade	1.9168	11.9826			
C.Segurança/Garantia	2.2936	2.8741	10.7668		
D.Capacidade de Resposta	3.8001	2.6773	5.5724	16.8088	
E.Empatia	1.3216	5.6022	4.5600	5.0365	10.0032

Tabela 4.3.13. Resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre as cinco escalas de percepções do Questionário SERVQUAL Modificado

	Componente		
	1	*2*	*3*
Unidades de inércia (valores próprios)	28.22	15.26	11.67
Variabilidade Explicada (%)	42.27	22.85	17.48
Var. Exp. Acumulada (%)	42.27	65.12	82.60

A 1ª componente principal explica 42.27% da variabilidade total dos dados (Tabela 4.3.13), e nela sobressai a escala "D. Capacidade de Resposta", isto é, a capacidade de inspirar credibilidade e confiança, que apresenta uma maior proporção de respostas "Totalmente de acordo" e "Não se aplica" (Figura 4.3.18, Figura 4.3.17). A 2ª componente principal, que explica 22.85% da variabilidade total dos dados, caracteriza-se pela escala "A. Elementos Tangíveis", que se refere aos equipamentos e que revela uma maior insatisfação dos cuidadores (Figura 4.3.18, Figura 4.3.17). A 3ª componente principal caracteriza-se pela oposição das escalas "D. Capacidade de resposta" e "B. Fiabilidade" (Figura 4.3.19).

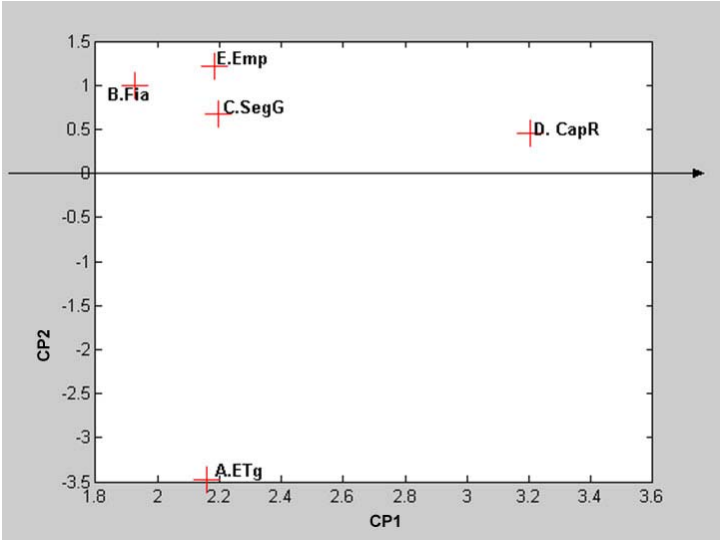

Figura 4.3.18. Representação das cinco escalas de percepções ("A.Elementos Tangíveis", "B.Fiabilidade", "C.Segurança/Garantia", "D.Capacidade de Resposta", "E.Empatia") do Questionário SERVQUAL Modificado no plano factorial (1,2), obtido com a ACP da matriz S_{LC}.

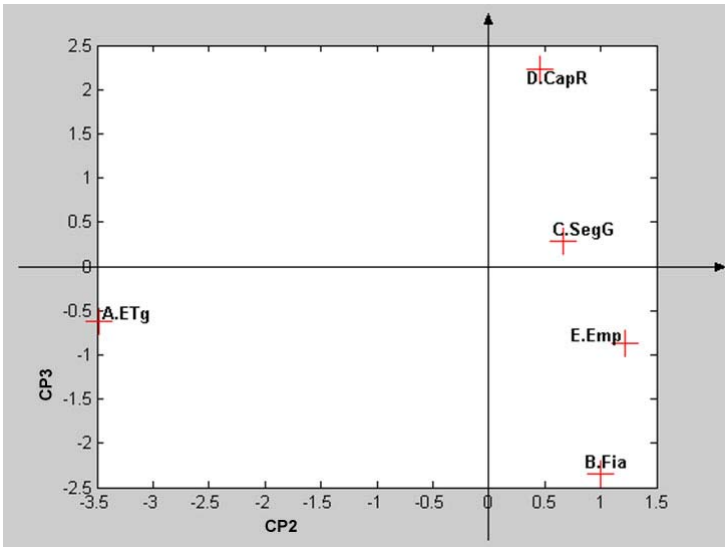

Figura 4.3.19. Representação das cinco escalas de percepções ("A.Elementos Tangíveis", "B.Fiabilidade", "C.Segurança/Garantia", "D.Capacidade de Resposta", "E.Empatia") do Questionário SERVQUAL Modificado no plano factorial (2,3), obtido com a ACP da matriz S_{LC}.

No nível 3 da hierarquia de partições (s_{LC}+Ligação única) (Figura 4.3.20), obtemos a partição constituída por duas classes:

- Classe 1 = {A. Elementos Tangíveis}. Classe que se refere ao equipamento e que apresenta um perfil de respostas que se destaca de todos os outros por revelar uma maior insatisfação dos cuidadores (Figura 4.3.17).
- Classe 2 = {"B.Fiabilidade", "E.Empatia", "C.Segurança/Garantia", "D.Capacidade de Resposta"}. Classe que se refere à capacidade de realizar os tratamentos com perfeição, de dar atenção pessoal, de fornecer prontamente os serviços e de inspirar

credibilidade e confiança, em que se destaca a predominância de respostas “4-Acordo” e “5-Totalmente de acordo”.

Nesta partição reconhecemos o 1º plano factorial (Figura 4.3.18).

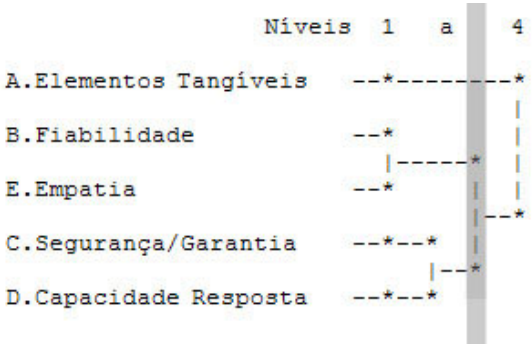

Figura 4.3.20. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+Ligação única). O nível 3 é o mais importante, seguido do nível 2, de acordo com o critério da “estatística de níveis” (Lerman, 1970; Bacelar-Nicolau, 1972, 1980), STAT(3)=2.13, STAT(2)=2.09.

Enquanto que, a partição obtida no nível 2 deste dendrograma, constituída por três classes, é bem complementada pela representação gráfica do plano (2,3) (Figura 4.3.19).

Podemos então, concluir que:

- Os dados apresentam uma estrutura classificatória hierárquica forte, pois diferentes critérios de agregação associados ao coeficiente s_{LC}, assim como ao coeficiente P_L, permitiram obter os mesmos dendrogramas.
- Os métodos de ACHA baseados na abordagem aqui apresentada e os obtidos por Bacelar-Nicolau *et al.* (2005) conduziram a resultados diferentes, enquanto que as ACP respectivas conduziram a resultados semelhantes.
- Foi detectada uma maior coerência entre os resultados obtidos com a ACP e com os algoritmos de ACHA aplicados directamente sobre a matriz S_{LC}.
- Há vantagem em recorrer à quantificação simbólica modal das escalas, que nos permitiu trabalhar com toda a informação fornecida pelos dados, aliada à utilização do coeficiente s_{LC}. Fica aqui a nossa sugestão para tratar este tipo de situações, muito frequentes na prática[178].

[178] Sabemos que as respostas “não sabe” são frequentemente consideradas como dados omissos, recorrendo-se a métodos de imputação. Preferimos usar a codificação simbólica modal que propomos.

4.4 Definição das matrizes score quando se comparam variáveis heterogéneas – O que representam os coeficientes s, s_{LC} e P_L

4.4.1 Introdução

Consideremos duas variáveis X e Y e as respectivas matrizes *score* geradas por elas. Como vimos, os *scores* associados às variáveis X e Y são matrizes reais, nxn, que são definidas tendo em conta os vectores respectivos, $(x_1, ..., x_n)$, $(y_1, ..., y_n)$, e a escala de medição que lhes é atribuída. A noção de estrutura intervém pois, quando queremos definir a relação entre estas variáveis.

O que é que representará o coeficiente bruto s, $s=\sum x_{ij}y_{ij}$= Tr XY^t= <X,Y>, quando as variáveis são heterogéneas, isto é, com estruturas diferentes?
Neste caso, as variáveis devem ser transformadas em matrizes *score*, de acordo, não só com o seu tipo, mas também com o tipo da variável com a qual se pretende relacionar. Com efeito, se não se toma cuidado com a definição dos *scores*, o coeficiente pode não ter sentido, como poderemos ver em seguida, para alguns casos.

Encontraremos também, em alguns dos casos analisados, coeficientes conhecidos.
Passemos, então, em revista os vários casos.

4.4.2 Atributo de descrição e qualquer outro tipo de variável

Como a variável atributo de descrição é um caso particular da variável nominal, aconselha-se a sua codificação como sendo nominal quando se pretende relacioná-la com variáveis de outro tipo. Nas próximas subsecções veremos que interpretação dar ao coeficiente *s*, nesses casos.

4.4.3 Variável nominal e variáveis ordinais

Como distinguimos vários tipos de variáveis ordinais, analisemos cada uma delas em seguida:

4.4.3.1 Variável nominal e variável com modalidades parcialmente ordenadas

Seja X uma variável nominal e Y uma variável com modalidades parcialmente ordenadas. Considerando as respectivas matrizes *score*, o coeficiente s corresponde ao número de indivíduos que estão em relação nas duas variáveis.

Exemplo 4.4.1. Semelhanças s_{LC} e P_L entre uma variável nominal e uma variável com modalidades parcialmente ordenadas

Consideremos as variáveis *Sexo do cuidador* e *Evolução dos sintomas – falta de ar* (variável com modalidades parcialmente ordenadas, 1<2<3, 4), que são apresentadas na Secção 5.3 do Capítulo 5. A informação obtida, na amostra de 57 cuidadores de doentes oncológicos, relativamente a estas duas variáveis está apresentada na Tabela 4.4.1. A semelhança s_{LC} calculada entre estas variáveis é forte positiva, s_{LC}=5.27, e o valor do coeficiente probabilístico está muito próximo de 1, $P_L \approx 1.00$. Observa-se, que os doentes da maioria dos cuidadores do sexo feminino nunca tiveram falta de ar (61.4% - 27), enquanto que, para 46.2% (6) dos cuidadores do sexo masculino, os doentes pioraram relativamente à falta de ar e, para 30.8% (4) destes cuidadores, a falta de ar dos doentes ao seu cuidado permaneceu igual.

Tabela 4.4.1. Tabela de contingência que cruza a informação sobre a *Evolução dos sintomas - Falta de ar* e o *Sexo do cuidador*, na amostra dos 57 cuidadores

			Sexo do cuidador		Total
			Masculino	Feminino	
Evolução dos sintomas - Falta de ar (1<2<3, 4)	1.Piorou	Frequência	6	6	12
		% por Sexo do cuidador	46,2%	13,6%	21,1%
	2.Permaneceu igual	Frequência	4	4	8
		% por Sexo do cuidador	30,8%	9,1%	14,0%
	3.Melhorou	Frequência	1	7	8
		% por Sexo do cuidador	7,7%	15,9%	14,0%
	4.Nunca teve	Frequência	2	27	29
		% por Sexo do cuidador	15,4%	61,4%	50,9%
	Total	Frequência	13	44	57
		% por Sexo do cuidador	100,0%	100,0%	100,0%

Este exemplo é meramente ilustrativo da relação entre variáveis destes tipos, que se destacou na análise multivariada realizada na Secção 5.3, mas não deixa de ser informativo.

4.4.3.2 Variável nominal e variável com modalidades estrita e totalmente ordenadas

Seja X uma variável nominal e Y uma variável com modalidades estrita e totalmente ordenadas (induz uma ordem sobre o conjunto das unidades estatísticas). Atribuindo os *scores* correspondentes ao tipo de cada uma das variáveis o coeficiente s é o produto escalar de uma matriz simétrica, X, e de uma matriz que é pseudo anti-simétrica, no sentido em que $y_{ij}+y_{ji}=1$[179]. Neste caso:

- $x_{ij}\, y_{ij} = 1$, se i e j apresentam a mesma modalidade de X, (i~j, para X), e i<j para Y.

Mas, se $y_{ij} = 1$ então $y_{ji} = 0$. Portanto, $x_{ij}\, y_{ij} + x_{ji}\, y_{ji} = x_{ij}$ ou $x_{ij}\, y_{ij} + x_{ji}\, y_{ji} = x_{ji}$, então $\Sigma_j\, x_{ij}\, y_{ij} + x_{ji}\, y_{ji}$ é independente de Y.

O coeficiente s representa o número de pares de unidades estatísticas que apresentam o mesmo valor da variável nominal, X, e que estão em relação de ordem (no mesmo sentido) em Y. Mas, como o valor do produto escalar é independente de Y, este caso não tem interesse.

No entanto, Le Calvé pensa que tem interesse codificar uma escala ordinal com modalidades estrita e totalmente ordenadas de maneira sequencial, quando queremos compará-la com uma nominal. Esta situação está contemplada na *Toolbox Coeficientes s, s_{LC} e P_L*. A sua interpretação encontra-se no ponto 4 desta subsecção.

4.4.3.3 Variável nominal e variável com modalidades totalmente ordenadas

Seja X uma variável nominal e Y uma variável com modalidades totalmente ordenadas (induz uma preordem sobre o conjunto das unidades estatísticas). Atribuindo os *scores* correspondentes a cada uma das variáveis, segundo o seu tipo, o coeficiente s é o produto escalar de uma matriz simétrica, X, (a que corresponde à variável nominal) e de uma matriz não simétrica, Y, (é anti-simétrica por blocos, isto é fora da diagonal).

Uma vez que, $y_{ij} + y_{ji} = 1$, se $i < j$ ou $j > i$, ou $y_{ij} + y_{ji} = 2$, se $i \leq j$ ou $j \geq i$, quer dizer, se i e j apresentam a mesma modalidade de Y, (i~j, para Y), temos então: s_{caso3}= s_{caso2}+ $s_{caso\ nominalxnominal}$. A estatística s é, neste caso, a soma das estatísticas do caso 2 (nominal, modalidades estrita e totalmente ordenadas) e do caso (nominal, nominal).

O coeficiente s é o número de pares de unidades estatísticas que apresentam o mesmo valor nas duas variáveis, ou que apresentam o mesmo valor da variável característica descritiva e que estão em relação de ordem (no mesmo sentido) em Y. Assim sendo, este caso não tem interesse.

---

[179] $y_{ij}+y_{ji}=0$, no sentido habitual de anti-simetria.

No entanto, Le Calvé pensa que tem interesse codificar esta variável ordinal de maneira sequencial, quando queremos compará-la com uma nominal. Esta situação está contemplada na *Toolbox Coeficientes s, s_{LC} e P_L*. A sua interpretação encontra-se na Subsecção 4.4.3.4 e está exemplificada na Subsecção 4.4.3.6 (Exemplo 4.4.2).

4.4.3.4 Variável nominal e variável de ordem sequencial

Seja X uma variável nominal e Y uma variável de ordem sequencial. O coeficiente s é o número de pares com o mesmo valor em X e que são consecutivos em Y. É igual a (n-1) menos o número de rupturas em Y, isto é, o número de mudanças de valor sobre duas categorias consecutivas em Y. A parcela (n-1) deve-se ao facto da diagonal principal ter zeros.

Depois da centragem e redução de s encontramos:

s_{LC}=(n-1)- teste das sequências homogéneas.

Quando o coeficiente s_{LC} toma valores elevados, $s_{LC}>>0$ ou $s_{LC}<<0$, significa que a repartição não é aleatória. Se a repartição fosse aleatória teríamos $s_{LC}=0$ e $P_L=0.5$.

Le Calvé considera este caso muito interessante e sugere que se codifique a variável ordinal como sequencial, quando se pretende compará-la com uma nominal.

Esta situação está exemplificada na Subsecção 4.4.3.6 (Exemplo 4.4.2).

4.4.3.5 Variável nominal e variável número de ordem

Seja X uma variável nominal e Y uma variável número de ordem. Neste caso, o coeficiente s é a soma dos números de ordem de Y, em que o número de ordem de cada unidade estatística i é ponderada pelo número, k_i, de elementos equivalentes a i em X: $s = \sum_{i=1}^{n} k_i \times$ nº de *ordem de i*. Isto é, o coeficiente s é a soma dos números de ordem dos pares de unidades estatísticas observadas em cada classe de X. O que faz sentido como ideia de semelhança.

O coeficiente s_{LC} é muito elevado se muitos elementos equivalentes em X são classificados com uma ordem elevada em Y, i.e., são os últimos em Y. Segundo Le Calvé, a estatística pareceria ser mais interessante se os seus valores fossem ordenados no outro sentido, i.e., se existissem muitos elementos iguais em X que fossem os primeiros em Y. Isto obtém-se codificando Y no outro sentido, ou seja, y_{ij} = n – (número de ordem de i). As duas opções apresentadas estão programadas na *Toolbox Coeficientes s, s_{LC} e P_L*.

Esta situação está exemplificada na subsecção seguinte (Exemplo 4.4.2).

4.4.3.6 *Exemplos de aplicação: semelhanças s_{LC} e P_L entre uma variável nominal e uma variável ordinal*

Exemplifica-se de seguida a comparação de variáveis nominais com variáveis do tipo ordinal utilizando os coeficientes s_{LC} e P_L.

Exemplo 4.4.2. Semelhanças s_{LC} e P_L entre uma variável nominal e uma variável com modalidades totalmente ordenadas e entre uma variável nominal e uma variável número de ordem

Consideremos as variáveis, *Sexo do cuidador*, *Estado físico e psíquico do cuidador – nervoso*, *Estado físico e psíquico do cuidador – calmo* e *Estado físico e psíquico do cuidador - capaz de cuidar do doente,* que são apresentadas na Secção 5.3 do Capítulo 5. A informação obtida, na amostra de 57 cuidadores de doentes oncológicos, relativamente a estas variáveis está apresentada, respectivamente, na Tabela 4.4.2, na Tabela 4.4.3 e na Tabela 4.4.4.

Tabela 4.4.2. Tabela de contingência que cruza a informação sobre o *Estado físico e psíquico do cuidador – nervoso* e o *Sexo do cuidador*, na amostra dos 57 cuidadores

			Sexo do cuidador		Total
			Masculino	Feminino	
Estado físico e psíquico do cuidador – nervoso	Nunca	Frequência	1	1	2
		% por Sexo do cuidador	7,7%	2,3%	3,5%
	Quase nunca	Frequência	2	1	3
		% por Sexo do cuidador	15,4%	2,3%	5,3%
	Às vezes	Frequência	6	13	19
		% por Sexo do cuidador	46,2%	29,5%	33,3%
	Quase sempre	Frequência	2	11	13
		% por Sexo do cuidador	15,4%	25,0%	22,8%
	Sempre	Frequência	2	18	20
		% por Sexo do cuidador	15,4%	40,9%	35,1%
Total		Frequência	13	44	57
		% por Sexo do cuidador	100,0%	100,0%	100,0%

Sendo estas variáveis ordinais codificadas como ordinais com modalidades totalmente ordenadas (preordem), quando se calcula o valor dos coeficientes de semelhança s_{LC} e P_L entre cada uma delas e a variável nominal *Sexo*, elas são recodificadas como variáveis ordem sequencial. Neste caso, as semelhanças s_{LC} calculadas são, nuns casos, fracas e noutro aproximadamente nula (Tabela 4.4.5).

Tabela 4.4.3. Tabela de contingência que cruza a informação sobre o *Estado físico e psíquico do cuidador – calmo* e o *Sexo do cuidador*, na amostra dos 57 cuidadores

			Sexo do cuidador		Total
			Masculino	Feminino	
Estado físico e psíquico do cuidador – calmo	Nunca	Frequência	1	6	7
		% por Sexo do cuidador	7,7%	13,6%	12,3%
	Quase nunca	Frequência	0	10	10
		% por Sexo do cuidador	,0%	22,7%	17,5%
	Às vezes	Frequência	6	20	26
		% por Sexo do cuidador	46,2%	45,5%	45,6%
	Quase sempre	Frequência	6	5	11
		% por Sexo do cuidador	46,2%	11,4%	19,3%
	Sempre	Frequência	0	3	3
		% por Sexo do cuidador	,0%	6,8%	5,3%
Total		Frequência	13	44	57
		% por Sexo do cuidador	100,0%	100,0%	100,0%

Neste último caso, s_{LC} = 0.0134 (P_L= 0.5055), as variáveis são independentes – podemos assim dizer que a repartição dos "valores" da variável *Estado físico e psíquico do cuidador – capaz de cuidar do doente* por cada uma das categorias da variável *Sexo* é aleatória (Tabela 4.4.4 e Tabela 4.4.5).

Tabela 4.4.4. Tabela de contingência que cruza a informação sobre o *Estado físico e psíquico do cuidador – capaz de cuidar do doente* e o *Sexo do cuidador*, na amostra dos 57 cuidadores

			Sexo do cuidador		Total
			Masculino	Feminino	
Estado físico e psíquico do cuidador - capaz de cuidar do doente	As vezes	Frequência	1	4	5
		% por Sexo do cuidador	7,7%	9,1%	8,8%
	Quase sempre	Frequência	5	15	20
		% por Sexo do cuidador	38,5%	34,1%	35,1%
	Sempre	Frequência	7	25	32
		% por Sexo do cuidador	53,8%	56,8%	56,1%
Total		Frequência	13	44	57
		% por Sexo do cuidador	100,0%	100,0%	100,0%

Quando estas variáveis ordinais são codificadas como variáveis número de ordem, o coeficiente s_{LC} detecta uma relação padronizada, entre aquelas variáveis, mais forte do que a encontrada anteriormente, nos casos, Sexo e *Estado físico e psíquico do cuidador –*

nervoso e Sexo e *Estado físico e psíquico do cuidador – calmo*. Enquanto que, tal como anteriormente, a semelhança s_{LC} entre o *Sexo* e o *Estado físico e psíquico do cuidador - capaz de cuidar do doente* é aproximadamente nula (Tabela 4.4.5), traduzindo a independência destas.

A classificação da variável *Estado físico e psíquico do cuidador - nervoso* em variável número de ordem permite, pois, realçar que 40.9% (18) dos cuidadores do sexo feminino se sentem sempre nervosos, enquanto que, 46.2% (6) dos cuidadores do sexo masculino, se sentem às vezes nervosos (Tabela 4.4.2).

A observação da Tabela 4.4.3 permitirá ver a interpretação a dar à relação encontrada entre o *Sexo* e o *Estado físico e psíquico do cuidador – calmo*. Neste caso, 46.2% (6) dos cuidadores do sexo masculino estão calmos "às vezes" e 22.7% (10) cuidadores do sexo feminino "quase nunca" estão calmos.

Tabela 4.4.5. Valores dos coeficientes de semelhança s_{LC} e P_L entre as variáveis apresentadas

Variáveis	Variáveis ordinais codificadas como variáveis com modalidades totalmente ordenadas	Variáveis ordinais codificadas como variável número de ordem
Sexo e *Estado físico e psíquico do cuidador - nervoso*	s_{LC} = 0.9594 e P_L = 0.8313	s_{LC} = 2.4051 e P_L = 0.9919
Sexo e *Estado físico e psíquico do cuidador – calmo*	s_{LC} = - 0.3657 e P_L = 0.3573	s_{LC} = - 2.0957 e P_L = 0.0181
Sexo e *Estado físico e psíquico do cuidador - capaz de cuidar do doente*	s_{LC} = 0.0134 e P_L = 0.5055	s_{LC} = 0.1293 e P_L = 0.5514

Quando estas variáveis ordinais são codificadas como ordinais com modalidades totalmente ordenadas (preordem), a relação s_{LC} detectada pode ser mais fraca, como vimos.

Neste contexto, apresentamos ainda o seguinte exemplo, que consideramos ser bem elucidativo para a interpretação destes coeficientes.

Exemplo 4.4.3. Semelhanças s_{LC} e P_L entre uma variável nominal e uma variável com modalidades totalmente ordenadas

Num inquérito a 100 puérperas, que recorreram à consulta de Pediatria (HSM), sobre o perigo de morte súbita e a posição de dormir do bebé constatámos que existe evidência de relação entre o *Local de vigilância da gravidez* (*Centro de saúde, Hospital, Privado*) e o *Nível socioeconómico (Índice de Graffar)* das puérperas inquiridas: s_{LC}=5.10 ($P_L \approx 1.00$) (Doria *et al*., 2006). Como se pode verificar, há uma tendência para que as puérperas com níveis socioeconómicos mais baixos (1.V, 2.IV, 3.III) tenham vigiado a gravidez em centros de saúde e as que têm níveis socioeconómicos mais elevados (predominantemente 4.II e 5.I)

tenham vigiado a gravidez no sistema privado, enquanto que as que vigiaram a gravidez no hospital têm níveis socioeconómicos, predominantemente, 3.III e 4.II (Figura 4.4.1, Figura 4.4.2).

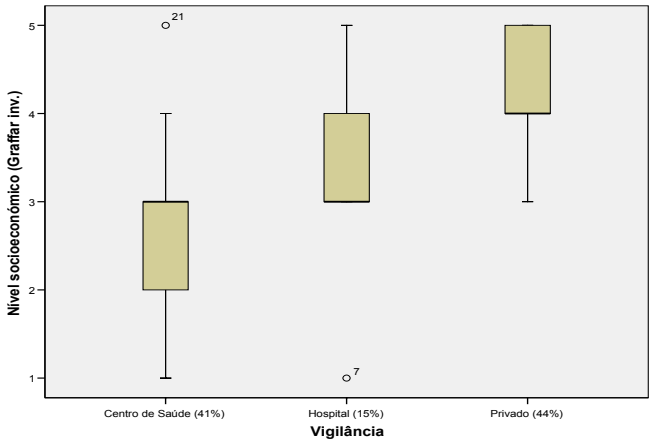

Figura 4.4.1. Diagramas em caixas de bigodes da distribuição do *Nível socioeconómico* das puérperas pelo *Local de vigilância da gravidez* (*Centro de saúde, Hospital, Privado*).

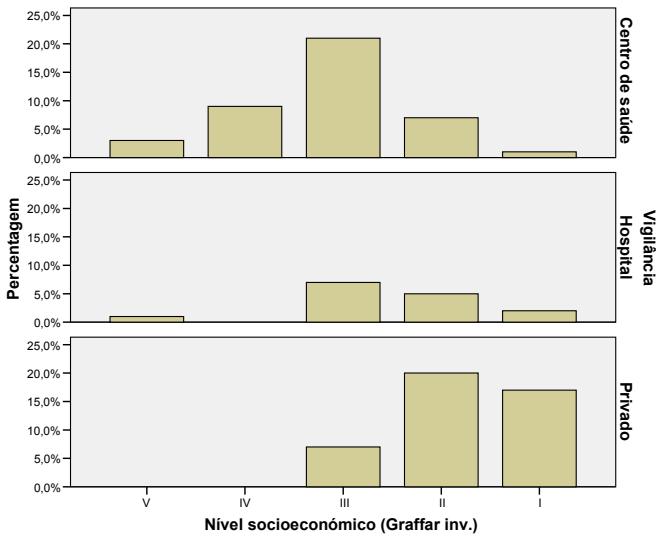

Figura 4.4.2. Gráficos de barras da distribuição do *Nível socioeconómico* das puérperas pelo *Local de vigilância da gravidez* (1.*Centro de saúde, 2.Hospital, 3.Privado*).

4.4.4 Variável nominal e variável métrica

Seja X uma variável nominal e Y uma variável métrica. Tal como no último caso apresentado, o coeficiente s é a soma dos valores de Y observados nos pares de unidades estatísticas que estão em cada classe de X, i. e., que são equivalentes em X.

No caso particular da variável nominal ser binária o coeficiente s_{LC} coincide com o coeficiente de correlação bisserial por pontos, r_{bp}, a menos de um factor, abstraindo-nos do sinal: $s_{LC} = \sqrt{n-1}\, r_{bp}$, sendo $r_{bp} = \frac{\overline{y}_p - \overline{y}_q}{s_Y}\sqrt{pq}$, p e q designam a proporção de "uns" e de "zeros" de X, respectivamente, s_Y é o desvio padrão de Y, e $\overline{y}_p$ e $\overline{y}_q$ são as médias dos valores de Y aos quais correspondem, respectivamente, "uns" e "zeros" em X (i.e., médias dos valores de Y, das unidades estatísticas que são equivalentes em X). A interpretação é a habitual: - Há uma tendência para emparelhar valores elevados/baixos da variável métrica com os uns/zeros da variável binária, abstraindo-nos do sinal, isto é, numa classe estão os valores mais altos e na outra os mais baixos.

Exemplo 4.4.4. Semelhanças s_{LC} e P_L entre uma variável binária e uma variável métrica

Consideremos a informação recolhida em 50 jovens diabéticos que recorrem à consulta de diabetes do Hospital A[180], relativamente à sua *Idade* (em anos) e ao *Consumo de álcool*[181] (não, sim). Verifica-se que existe uma relação forte negativa entre estas duas variáveis: s_{LC} = -4.42, $P_L \approx 0.0000$. Como seria de esperar, os jovens que consomem álcool (20 – 40%) são os mais velhos (idade média=17.8 anos, dp=1.64 anos) e os que não consomem álcool (30 – 60%) são os mais novos (idade média=14.3 anos, dp=2.5 anos)[182] (Figura 4.4.3).

[180] O hospital é conhecido e estes dados são reais.
[181] Sabe-se que o consumo de álcool nesta população não atinge níveis preocupantes.
[182] Valores certamente influenciados pelas duas observações *outlier* em evidência no gráfico.

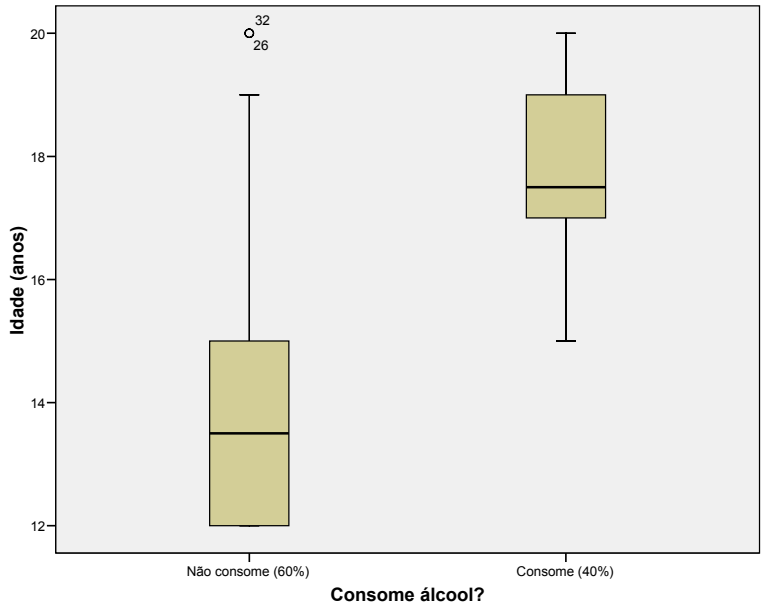

Figura 4.4.3. Diagramas em caixas de bigodes da distribuição da *idade*, dos 50 jovens diabéticos, por *consumo de álcool*. Os dois *outliers* assinalados referem-se à idade de 20 anos de dois jovens que não consomem álcool.

Também podemos verificar que o coeficiente s_{LC} coincide com o coeficiente de correlação bisserial por pontos, r_{bp}, a menos de um factor, abstraindo-nos do sinal (r_{bp}= 0.632, valor-p≈0.000). O valor-p do teste de significância apresentado, relaciona-se com o coeficiente P_L, valor-p= P_L (no caso do teste ser unilateral), como seria de esperar (ver Secção 4.8).

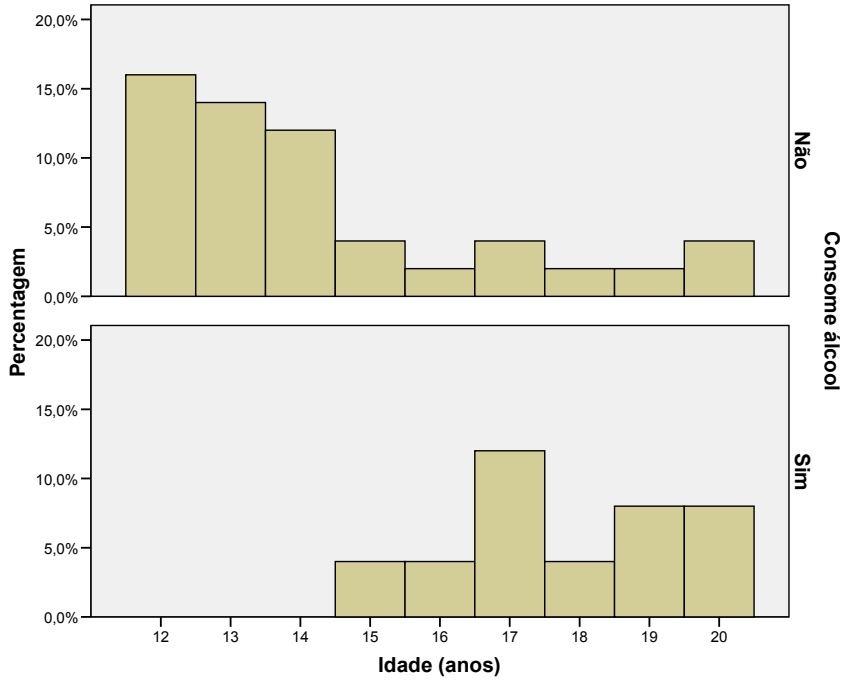

Figura 4.4.4. Gráficos de barras da distribuição da *idade* (em anos) dos diabéticos pelo *consumo de álcool* (0.*Não consome, 1.Consome*).

4.4.5 Variável nominal e variável simbólica/complexa

Seja X uma variável nominal e Y uma variável simbólica/complexa. Quando se pretende comparar variáveis destes tipos, as matrizes *score* são as que correspondem às variáveis referidas e já definidas anteriormente, com excepção da variável simbólica intervalar. A matriz *score*, neste caso, é a matriz assimétrica de semelhanças que se obtém da matriz assimétrica das distâncias de Haussdorf a partir da transformação max (H(i,i')) - H(i,i').

Este caso é análogo ao caso anterior, s é a soma dos valores de Y (semelhanças, entre as unidades estatísticas, relativas a Y) em cada classe de X.

Se a semelhança s_{LC} for muito forte, $s_{LC}>>0$ ($P_L \approx 1$), significa que em cada classe da variável nominal as unidades estatísticas são muito semelhantes, i.e., as classes são "homogéneas". Se $s_{LC}<<0$ ($P_L \approx 0$), as unidades estatísticas que se assemelham distribuem-se por classes diferentes, i. e., as classes não são homogéneas. Quando $s_{LC} = 0$ não há uma relação particular, é a relação média.

Exemplo 4.4.5. Semelhanças s_{LC} e P_L entre uma variável nominal e uma variável simbólica

Consideremos as variáveis, *Se pudesse escolher que preferia*? (1.Que o doente estivesse no hospital, 2.Que o doente estivesse em casa, 3.Outra) e a escala de percepções *"E.Empatia"* do Questionário SERVQUAL Modificado, que são apresentadas na Secção 5.3 do Capítulo 5. A escala de percepções *"E.Empatia"* é uma variável simbólica modal, já estudada na Subsecção 4.3.10.8. Dos 57 cuidadores inquiridos, a maioria (44 – 77.2%) preferia que *o doente estivesse em casa*, enquanto que 8 (14.0%) preferiam que *o doente estivesse no hospital* e 5 (8.8%) preferiam *outra* possibilidade. Verifica-se que existe uma relação moderada entre aquelas duas variáveis: s_{LC}= 2.0731 e P_L=0.9809. O perfil de respostas dadas à escala *Empatia* é peculiar a cada uma das categorias da variável nominal (Figura 4.4.5). Sobressai o elevado consenso de respostas positivas e, principalmente "totalmente de acordo", dadas pela maioria dos cuidadores, que preferiam que o doente estivesse em casa (44 – 77.2%).

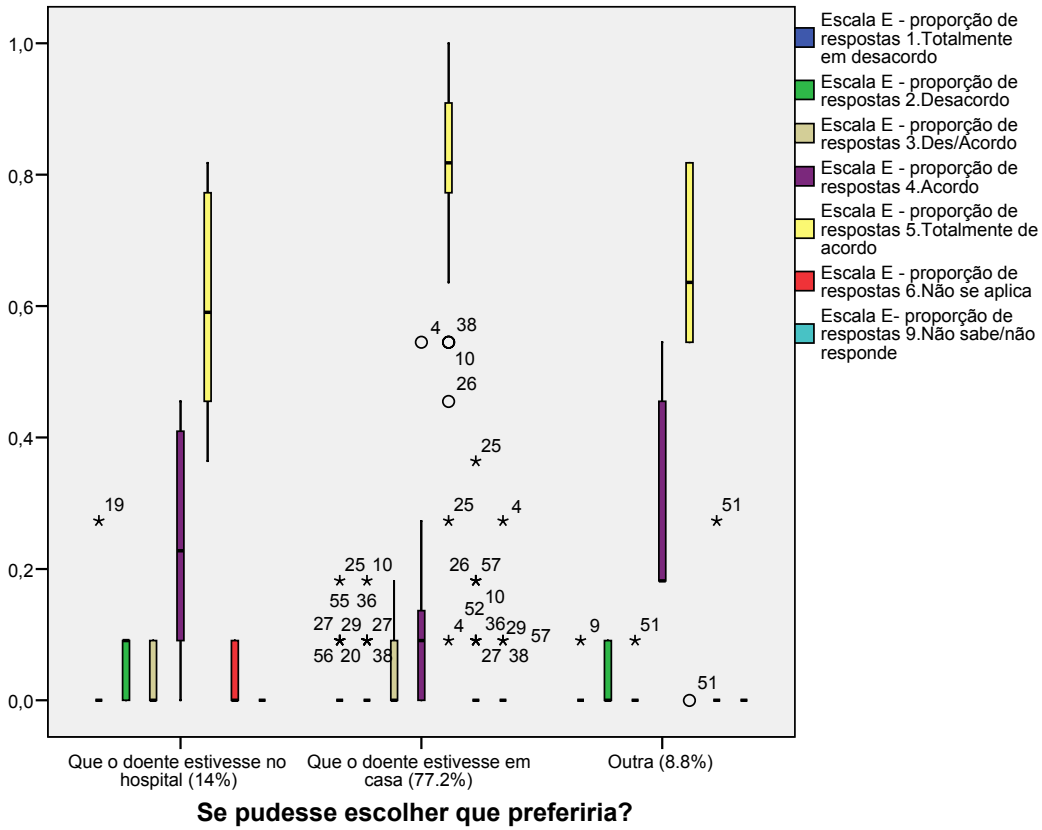

Figura 4.4.5. Diagramas em caixas de bigodes da distribuição de frequências relativas das categorias da escala *E. Empatia*, por cada uma das categorias/possibilidades de resposta a *Se pudesse escolher que preferia?* Estão assinaladas as respostas *outlier* com o código de identificação dos respectivos cuidadores.

4.4.6 Variáveis ordinais e variáveis ordinais

De forma geral, o coeficiente de semelhança s entre duas variáveis ordinais, representa o número de pares de unidades estatísticas que estão na mesma "ordem" nas duas variáveis. Em particular, destacaremos os casos que se seguem:

- **Seja X uma variável com modalidades estrita e totalmente ordenadas (ordem) e Y uma variável com modalidades totalmente ordenadas (preordem).** A estatística s é, neste caso, a estatística s do caso (ordem, ordem) apresentado na Subsecção 4.3.6: $s = s_{ordem,ordem}$.

- **Seja X uma variável com modalidades estrita e totalmente ordenadas e Y uma variável ordem sequencial.** O coeficiente de semelhança s é o número de pares consecutivos em Y e que estão dispostos na mesma ordem em X.

- **Seja X uma variável com modalidades totalmente ordenadas e Y uma variável ordem sequencial.** Caso análogo ao anterior.

- **Seja X uma variável com modalidades estrita e totalmente ordenadas e Y uma variável número de ordem.** O coeficiente de semelhança s é a soma dos números de ordem de Y correspondentes aos pares de unidades estatísticas que estão dispostos na mesma ordem em X. Neste caso, o coeficiente s é a soma das ordens de Y, em que a ordem de cada unidade estatística i é ponderada pelo número, k_i, de elementos equivalentes a i em X: $s = \sum_{i=1}^{n} k_i \times n.^{o}\ de\ ordem\ de\ i$ (4.4.1). Isto é, o coeficiente s é a soma dos números de ordem de Y correspondentes aos pares de unidades estatísticas que estão dispostas na mesma ordem em X.

- **Seja X uma variável com modalidades totalmente ordenadas e Y uma variável número de ordem.** Caso análogo ao anterior, com possibilidade de empates em X.

- **Seja X uma variável ordem sequencial e Y uma variável número de ordem.** O coeficiente de semelhança s é a soma dos números de ordem das unidades estatísticas de Y correspondentes aos pares de unidades estatísticas consecutivas de X.

4.4.7 Variáveis ordinais e variável métrica

Destacamos os casos que se seguem:

- **Seja X uma variável com modalidades totalmente ordenadas e Y uma variável métrica.** O coeficiente de semelhança s é a soma dos valores de Y correspondentes aos pares de unidades estatísticas que estão em relação em X, i. e., que estão dispostos na mesma ordem em X, ou que estão na mesma categoria de X . Neste caso, o coeficiente s é a soma dos valores de Y, em que o valor correspondente a cada unidade estatística i é ponderado pelo número, k_i, de elementos equivalentes a i em X: $s = \sum_{i=1}^{n} k_i y_i$ (4.4.2).

Este caso corresponde a uma classificação por ordem de grandeza.

Exemplo 4.4.6. Semelhanças s_{LC} e P_L entre uma variável com modalidades totalmente ordenadas e uma variável métrica

Foi realizado um inquérito a 100 puérperas, que recorreram à consulta de Pediatria (HSM), sobre o perigo de morte súbita e a posição de dormir do bebé. Constatámos que os dados apontam para uma relação negativa entre o *Nível socioeconómico (Índice de Graffar)* e a *Idade* das puérperas inquiridas: s_{LC}=-3.98 ($P_L \approx 0.0000$) (Doria *et al*., 2006).

Observamos que, de forma geral, há uma tendência para emparelhar gradualmente puérperas mais novas com os níveis socioeconómicos mais desfavorecidos, e puérperas, progressivamente mais velhas, com níveis socioeconómicos, progressivamente, mais favorecidos (Tabela 4.4.6, Figura 4.4.6, Figura 4.4.7).

Tabela 4.4.6. Estatísticas da *Idade* (em anos) por *Nível socioeconómico* das 100 puérperas

Variável		n_i	Idade média ± desvio padrão	Idade mediana
Índice de Graffar: Nível socioeconómico	1.V	4	23.2 ± 3.8	23.0
	2.IV	9	23.0 ± 3.4	23.0
	3.III	35	27.6 ± 6.7	27.0
	4.II	32	30.5± 5.0	30.5
	5.I	20	31.6± 4.6	30.5

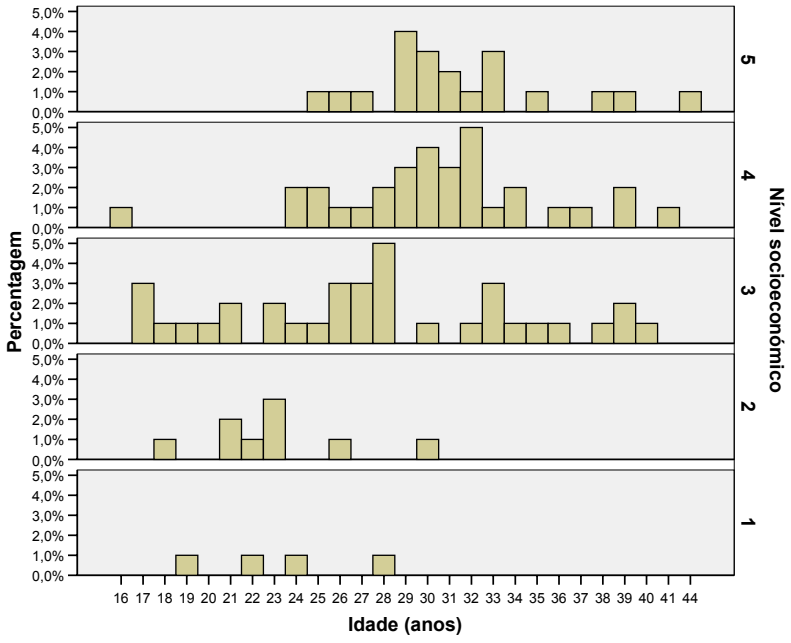

Figura 4.4.6. Histogramas da distribuição da *Idade* (em anos) por cada uma das categorias do *Nível socioeconómico* (Índice de Graffar: 1.V, 2.IV, 3:III, 4.II, 5.I) das 100 puérperas.

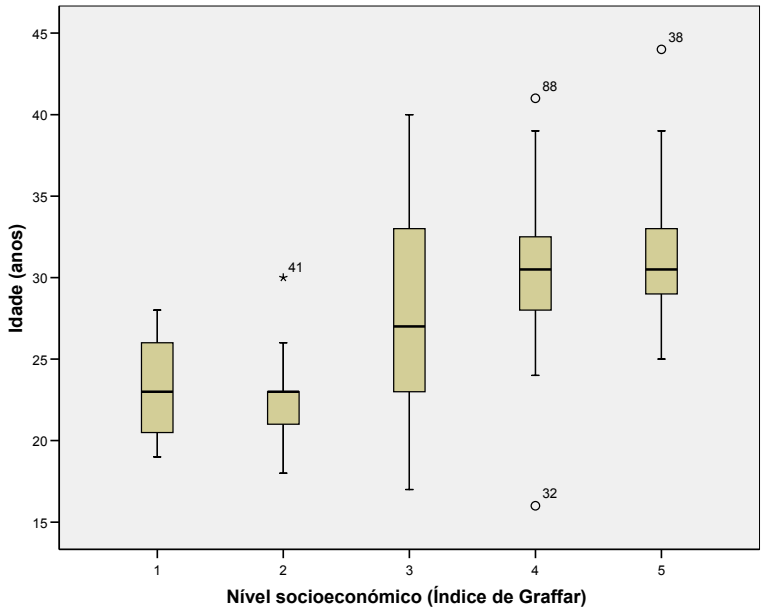

Figura 4.4.7. Diagramas em caixas de bigodes da distribuição da *Idade* (em anos) por cada uma das categorias do *Nível socioeconómico* (Índice de Graffar) das 100 puérperas.

- **Seja X uma variável número de ordem e Y uma variável métrica.** Neste caso, sugerimos que a matriz *score* da variável Y seja codificada como a da variável número de ordem. O coeficiente s_{LC} coincidirá, como vimos, com o coeficiente de correlação de Spearman.

- **Seja X uma variável de ordem sequencial e Y uma variável métrica.** O coeficiente de semelhança s é a soma dos valores de Y correspondentes aos pares de unidades estatísticas consecutivas de X.

4.4.8 Variáveis ordinais e variável simbólica/complexa

Quando se pretende comparar variáveis ordinais com variáveis simbólicas/complexas, as matrizes *score* são as que correspondem às variáveis referidas e já definidas anteriormente, com excepção da variável simbólica intervalar. A matriz *score*, neste caso, é a matriz assimétrica de semelhanças que se obtém da matriz assimétrica das distâncias de Haussdorf a partir da transformação max(H(i,i')) - H(i,i').

Aqui destacaremos os seguintes casos:

1. Variável com modalidades parcialmente ordenadas e variável simbólica/complexa. Um exemplo ajudar-nos-á a perceber melhor a relação entre estes dois tipos de variáveis.

Exemplo 4.4.7. Semelhanças s_{LC} e P_L entre uma variável com modalidades parcialmente ordenadas e uma variável categórica com valores múltiplos

Consideremos as variáveis, *Tem a quem recorrer quando necessita de companhia ou apoio de alguém?* (1.Não, 2.Sim, algumas vezes, 3.Sim, sempre, 4.Não responde) e *No apoio ao doente no domicílio, indique dois aspectos mais importantes …*(1.Material e equipamento, 2.Capacidade de apoio ao doente da equipa de saúde (de forma segura e precisa), 3.Confiança que inspiram os profissionais, 4.Capacidade de resposta adequada da equipa de saúde face às complicações, 5.Atenção que a equipa de saúde dá ao doente e à família), do Questionário SERVQUAL, que são apresentadas na Secção 5.3 do Capítulo 5.

A relação observada entre aquelas duas variáveis, com modalidades parcialmente ordenadas e categórica com valores múltiplos, respectivamente (s_{LC}= 3.0808 e P_L=0.9990), permite-nos aperceber que, embora as modalidades, "Atenção que a equipa de saúde dá ao doente e à família" (59.7%) e "Capacidade de apoio da equipa de saúde ao doente (de forma segura e precisa)" (52.6%) sejam as mais escolhidas, a sua preferência difere conforme os cuidadores tenham a quem recorrer, quando precisam de ajuda (Figura 4.4.8).

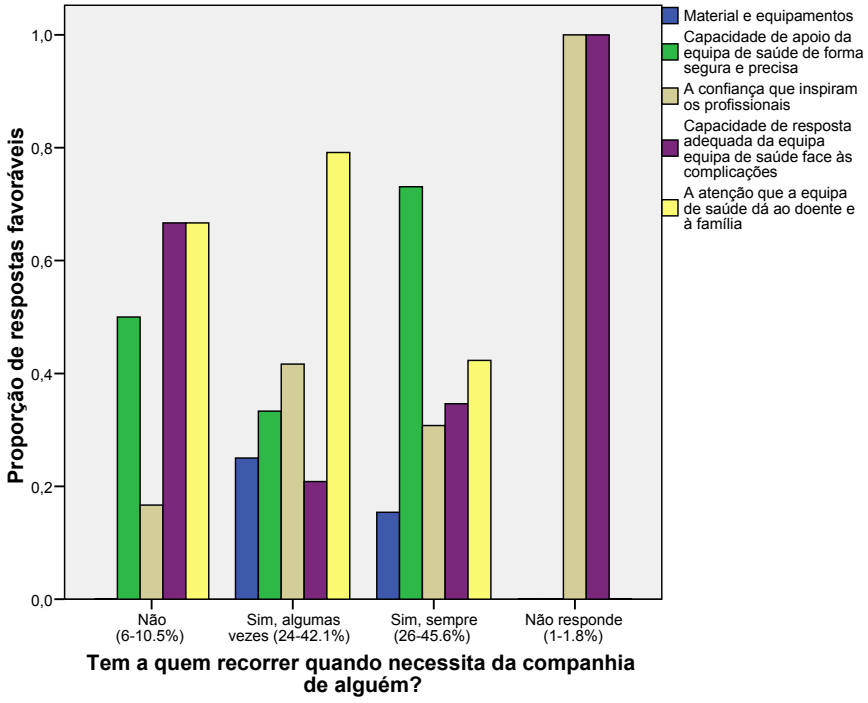

Figura 4.4.8. Gráfico de barras da distribuição das respostas favoráveis a *No apoio ao doente no domicílio, indique dois aspectos mais importantes …* por cada uma das categorias de resposta a *Tem a quem recorrer quando necessita de companhia ou apoio de alguém?*

2. Variável com modalidades totalmente ordenadas e variável simbólica/complexa. Seja X uma variável com modalidades totalmente ordenadas e Y uma variável simbólica/complexa. O coeficiente de semelhança s é a soma das semelhanças entre as

unidades estatísticas, em relação a Y, correspondentes aos pares de unidades estatísticas que estão em relação em X, i. e., que estão dispostas na mesma ordem em X com possibilidade de empates.

Exemplo 4.4.8. Semelhanças s_{LC} e P_L entre uma variável com modalidades totalmente ordenadas e uma variável categórica com valores múltiplos

Consideremos as variáveis, *Estado físico e psíquico do cuidador – trabalho acima das suas possibilidades* (1.Nunca, 2.Quase nunca, 3.Ás vezes, 4.Quase sempre, 5.Sempre) e *No apoio ao doente no domicílio, indique dois aspectos mais importantes …*(1.Material e equipamento, 2.Capacidade de apoio ao doente da equipa de saúde (de forma segura e precisa), 3.Confiança que inspiram os profissionais, 4.Capacidade de resposta adequada da equipa de saúde face às complicações, 5.Atenção que a equipa de saúde dá ao doente e à família), do Questionário SERVQUAL, que são apresentadas na Secção 5.3 do Capítulo 5.

Analisemos a relação encontrada entre estas duas variáveis: s_{LC}= 2.1412 e P_L=0.9839.

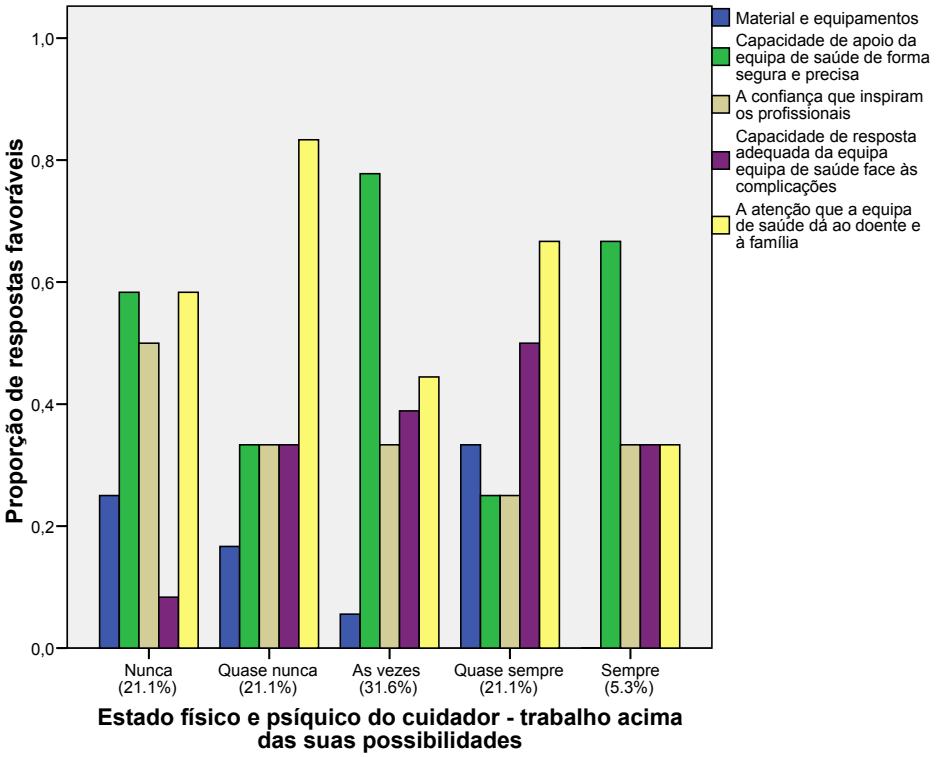

Figura 4.4.9. Gráfico de barras da distribuição das respostas favoráveis da proporção de respostas favoráveis *No apoio ao doente no domicílio, indique dois aspectos mais importantes …* por cada uma das modalidades do *Estado físico e psíquico do cuidador – trabalho acima das suas possibilidades*.

Embora a "*Atenção que a equipa de saúde dá ao doente e à família*" e a "*Capacidade de apoio da equipa de saúde ao doente (de forma segura e precisa)*" sejam as mais escolhidas, como vimos, a sua distribuição é diferente, consoante o *Estado físico e psíquico do cuidador*

– *trabalho acima das suas possibilidades*. Verifica-se que as maiores percentagens de respostas favoráveis à "Atenção que a equipa de saúde dá ao doente e à família" são de cuidadores que, nunca, quase nunca, ou quase sempre sentem o trabalho acima das suas possibilidades; enquanto que, as maiores percentagens de respostas favoráveis à "Capacidade de apoio da equipa de saúde ao doente (de forma segura e precisa)" são de cuidadores que, às vezes, nunca, ou sempre sentem o trabalho acima das suas possibilidades (Figura 4.4.9).

Neste exemplo, a interpretação não é facilitada pela "leitura" do gráfico (Figura 4.4.9). Contudo, os coeficientes s_{LC} e P_L detectam evidência de relação positiva entre as duas variáveis.

Exemplo 4.4.9. Semelhanças s_{LC} e P_L entre uma variável com modalidades totalmente ordenadas e uma variável categórica com valores múltiplos

Consideremos as variáveis, *Periodicidade das visitas* (1.Uma vez por mês, 2.Uma vez por semana, 3.Mais do que uma vez por semana) e *No apoio ao doente no domicílio, indique dois aspectos mais importantes* (1.Material e equipamento, 2.Capacidade de apoio ao doente da equipa de saúde (de forma segura e precisa), 3.Confiança que inspiram os profissionais, 4.Capacidade de resposta adequada da equipa de saúde face às complicações, 5.Atenção que a equipa de saúde dá ao doente e à família), do Questionário SERVQUAL Modificado, que são apresentadas na Secção 5.3 do Capítulo 5.

Analisemos a relação encontrada entre estas duas variáveis: s_{LC}= -1.4824 e P_L=0.0691.

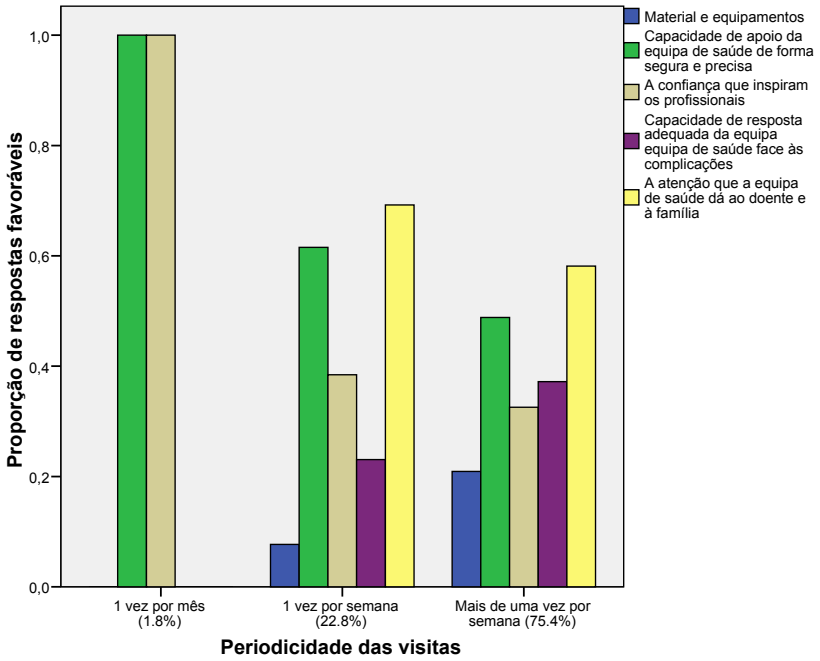

Figura 4.4.10. Gráfico de barras da proporção de respostas favoráveis a *No apoio ao doente no domicílio, indique dois aspectos mais importantes …* por cada uma das modalidades da *Periodicidade das visitas*.

Como se pode observar, há uma tendência para que as opções de resposta a *No apoio ao doente no domicílio, indique dois aspectos mais importantes …* mais frequentes decresçam com o aumento da *Periodicidade das visitas* (Figura 4.4.10).

3. Variável com modalidades estrita e totalmente ordenadas e variável simbólica/complexa. Seja X uma variável com modalidades totalmente ordenadas e Y uma variável simbólica/complexa. Caso análogo ao anterior, sem possibilidade de empates na variável X.

4. Variável número de ordem e variável simbólica/complexa. Seja X uma variável número de ordem e Y uma variável simbólica/complexa. O coeficiente de semelhança s é a soma das semelhanças entre cada unidade estatística i e as outras, i', $s_{ii'}$, em relação a Y, ponderadas pelo seu número de ordem em X, R_i: $s = \sum_{i=1}^{n} \sum_{i'=1}^{n} R_i s_{ii'}$ (4.4.3).

5. Variável ordem sequencial e variável simbólica/complexa. Seja X uma variável ordem sequencial e Y uma variável simbólica/complexa. O coeficiente de semelhança s é a soma das semelhanças, entre as unidades estatísticas em relação a Y, correspondentes aos pares de unidades estatísticas consecutivas de X.

4.4.9 Variável métrica e variável simbólica/complexa

Seja X uma variável métrica e Y uma variável simbólica/complexa. O coeficiente de semelhança s é a soma das semelhanças entre cada unidade estatística i e as outras, i', $s_{ii'}$, em relação a Y, ponderadas pelo seu valor em X, X(i)=x_i: $s = \sum_{i=1}^{n} \sum_{i'=1}^{n} x_i s_{ii'}$ (4.4.4).

4.5 Definição da matriz score no caso de dados sob a forma matricial

Os dados a analisar podem ser apresentados sob a forma matricial. Isto porque eles foram obtidos directamente – é o caso, por exemplo, das matrizes de preferências, da matriz de semelhanças entre nove sintomas de depressão (Streiner e Norman, 2003), das matrizes de

distâncias entre localidades ou das matrizes de distâncias filogenéticas[183] obtidas a partir dos dados originais, por exemplo, por alinhamento das sequências (Desdevises *et al*., 2003), como vimos no Capítulo 1 – ou a partir do cálculo de semelhanças/dissemelhanças entre variáveis – é o caso das matrizes de distâncias genéticas (por exemplo, Lefort-Buson e Vienne, 1985; Perrier, 1998) e de todas as outras que se podem obter através de coeficientes de proximidade.

As matrizes de dados também se podem obter como resultado de outras análises. Podemos pensar, em particular, nas matrizes de ultramétricas/hierarquias de partições que se obtêm em Análise Classificatória Hierárquica Ascendente (ACHA), e nas matrizes de distâncias filogenéticas[184] obtidas a partir das distâncias entre as espécies, calculadas a partir das árvores filogenéticas (Desdevises *et al*., 2003; Legendre *et al.*, 1994).

Definição 4.5.1. *Score* de dados sob a forma matricial

As matrizes *score* de dados de proximidade, que se apresentam sob a forma matricial, são as próprias matrizes, quer sejam de semelhanças, quer sejam de distâncias. Por convenção, a diagonal principal destas matrizes é nula.

Quando se comparam dados sob a forma de matriz convém que elas sejam do mesmo tipo, isto é, sejam ambas matrizes de semelhanças ou ambas matrizes de distâncias. No caso das matrizes não serem do mesmo tipo, podem-se transformar utilizando as operações habituais, que permitem transformar distâncias em semelhanças e vice-versa (Subsecção1.3.4).

4.5.1 O que representam os coeficientes s, s_{LC} e P_L quando se comparam matrizes *score* que são matrizes de proximidades

O coeficiente s, produto escalar entre matrizes de proximidades, representa a semelhança "bruta" e o coeficiente s_{LC} representa a semelhança padronizada. Enquanto que, o coeficiente P_L representa a função de distribuição dessa semelhança padronizada ser observada, isto é, a semelhança probabilística, como tem sido referido.

---

[183] Uma definição habitual de filogenia é a seguinte: Tradução, sob a forma de uma árvore, das relações entre as diferentes espécies fazendo aparecer os seus graus de parentesco, os seus antepassados comuns e traçando assim a história da descendência dos seres vivos.

[184] *Patristic distance matrix*, em inglês.

Tal como Escoufier (1973), podemos interpretar o coeficiente s como a covariância entre duas matrizes quadradas centradas, A e B: $CovV\left(A,B\right)=Tr\left(B^{\mathrm{T}}A\right)$. No caso de A e B serem matrizes quadradas simétricas e s.d.p. com a mesma dimensão, o coeficiente s coincide com o numerador do coeficiente R_V introduzido por Escoufier (1973): $Tr\left(B^{\mathrm{T}}A\right)$.

Verificamos também que o coeficiente s, como produto escalar entre matrizes de proximidades, coincide com a estatística Z de Mantel (1967), e é o dobro da estatística $\sum_{i<j}\sum x_{ij}y_{ij}$ (4.5.1) (em que, x_{ij} e y_{ij} são medidas espaciais e temporais entre os pontos i e j, respectivamente), também proposta por Mantel (1967) para relacionar a informação espaço-temporal da leucemia, numa abordagem designada como regressão generalizada, pois foi realizada entre duas matrizes de distâncias, para detectar esta doença.

O teste de Mantel é um teste permutacional. Embora, a sua estatística tenha um aspecto análogo ao das estatísticas que calculamos, ele não coincide com a abordagem aqui apresentada, pois a forma de obter a permutação é diferente da realizada por Le Calvé.

Já referimos que o coeficiente s é o índice aleatório S_1 (Lerman, 1973, 1977, 1981, 1992a). Sobre o índice aleatório S_1, Lerman (1992a) refere Le Calvé (1977)[185], e constata a existência da estatística de Mantel (1967), que também coincide com S_1. No entanto, a natureza formal da expressão da variância de S_1, var(S_1), obtida por Lerman é diferente da obtida por Mantel.

Este coeficiente, considerado como produto escalar de matrizes de proximidades entre os mesmos indivíduos, com a referência a Mantel (1967), é um coeficiente de semelhança geral utilizado frequentemente por investigadores, em várias situações e contextos, em particular nas áreas da biologia, antropologia, psicometria, sociologia, genética (e.g., Shannon *et al.*, 2002), do ambiente (e.g., Desdevises *et al*., 2003) e pela comunidade ecológica (pela primeira vez por Legendre e Troussellier (1988); posteriormente em Manly, 1992, 1994; Legendre e Legendre, 2000; e outros). O teste de Mantel tem sido alvo de diversos estudos comparativos, dos quais destacamos Dutilleul *et al*. (2000) e Legendre (2000). Dutilleul *et al*. (2000) comparam o teste de Mantel com a utilização do coeficiente de correlação de Pearson sobre os mesmos dados em bruto, uma vez que a distância Euclidiana, ou qualquer medida de proximidade entre duas unidades observacionais, podem ser calculadas; concluem que, sob o modelo de distribuição normal multivariado, e na maior parte dos casos, as duas técnicas conduzem a decisões semelhantes (situações de acordo), mas fora deste modelo encontram situações em que os valores da estatística de Mantel são

[185] Aparece referenciado como Le Calvé (1976).

significativos, enquanto que os valores do coeficiente de correlação de Pearson, r, não o são.

No nosso caso esta questão não se põe, pois como vimos, o coeficiente de semelhança s_{LC} coincide com o coeficiente de correlação de Pearson, a menos de um factor, $s_{LC} = \sqrt{n-1}\, r$. Ora, sabe-se que, no caso do par de variáveis métricas (X, Y) ter distribuição normal e, sob a hipótese nula de independência das variáveis, H_0: ρ=0, a distribuição da variável aleatória R é aproximada por uma distribuição normal, $R \dot{\frown} N\left(0, \frac{1}{\sqrt{n-1}}\right)$, para n >100 (Saporta, 1990).

Centrando e reduzindo R, obtemos a variável aleatória Z, $Z = \sqrt{n-1}\, R \dot{\frown} N(0,1)$, em que reconhecemos a expressão da variável aleatória S_{LC} para este tipo de variáveis. Quando o par de variáveis (X, Y) não tem distribuição normal, aqueles resultados permanecem com a condição de n ser grande (na prática, n>30), mas o facto de encontrar um valor de r que não seja significativamente diferente de 0, não permitirá concluir sobre a independência das variáveis (Saporta, 1990).

A comparação de várias classificações hierárquicas/hierarquias de partições sobre as mesmas unidades estatísticas tem constituído objecto de estudo por parte de vários investigadores (e.g., Nicolau, 1980; Lapointe e Legendre 1992; Lerman, 1999; Sousa e Nicolau, 2002; Youness e Saporta, 2004), devido à sua importância, interesse prático e dificuldade de concretização. Matematicamente, não é correcto utilizar o coeficiente de correlação linear de Pearson, r, para relacionar distâncias ultramétricas. Na literatura encontramos, entre outras, propostas que se baseiam no coeficiente de correlação linear (e.g., Rohlf, 1982; Lapointe e Legendre, 1992). Nós podemos fazê-lo com os coeficientes *s*, s_{LC} e P_L. Exemplificamos, na Secção 5.4 do Capítulo 5, a nossa proposta de utilização daqueles coeficientes para comparar várias hierarquias de partições sobre os mesmos objectos.

Perspectivamos pois um vasto domínio de aplicação dos coeficientes *s*, s_{LC} e P_L, no âmbito da ACP ou da ACHA.

4.6 Os dados omissos e os coeficientes s, s_{LC} e P_L

Os dados podem ser omissos porque se perderam ou estão ilegíveis ou ainda – no caso de questionários[186] – porque os inquiridos se esqueceram ou recusaram a responder. Nestes casos, dependendo da percentagem de dados omissos, o investigador poderá seguir várias estratégias: retirar os sujeitos da amostra/população, recuperá-los inquirindo de novo os sujeitos, utilizar métodos de imputação de dados omissos no caso dos dados serem irrecuperáveis, ou substituí-los por valores estimados a partir de outros dados.

Mas, os dados também podem ser omissos porque os inquiridos se recusam a responder (e.g. porque não sabem responder à questão ou não se identificam com as possibilidades de resposta ou a questão não se aplica à sua situação). Neste caso de dados omissos, a sua omissão dá informação que deverá ser tida em conta.

Por isso, é normalmente considerada boa prática, no caso das variáveis qualitativas e das variáveis simbólicas modais em questionários, a inclusão da(s) categoria(s) "Não sabe"/"Não responde"/"Não se aplica". No caso da(s) variável(eis) em questão serem variáveis ordinais, com a introdução da(s) categoria(s) "Não sabe"/"Não responde"/"Não se aplica" passam a ser variáveis com modalidades parcialmente ordenadas. No caso das variáveis serem nominais, a introdução da(s) categoria(s) "Não sabe"/"Não responde"/"Não se aplica" não altera a sua classificação, assim como no caso das variáveis simbólicas modais. Em qualquer destes casos, podemos usar os coeficientes s, s_{LC} e P_L para comparar as variáveis com outras do mesmo tipo ou de tipo diferente, como vimos. Vemos neste procedimento uma alternativa vantajosa à substituição dos valores omissos por métodos de imputação ou por valores estimados a partir de outros dados.

4.7 Conclusão sobre a interpretação geral dos coeficientes s_{LC} e P_L

Em resumo, podemos dizer que:

- Se o coeficiente de semelhança centrado e reduzido, s_{LC}, é muito elevado positivo ($s_{LC}>>0$ e, correspondentemente, $P_L\approx1$), os pares de indivíduos que estão em relação numa das variáveis, também o estão na outra. Estar em relação não quer dizer sempre a mesma coisa, pois depende do tipo de variável. No caso de variáveis nominais, significa estarem na mesma classe nas duas variáveis[187]; no caso de variáveis ordinais, significa "estarem no mesmo sentido"; nos casos métrico e

[186] Também podemos pensar nos protocolos médicos e fichas de doentes.

[187] Classes essas que não são forçosamente iguais nas duas variáveis. Por exemplo, presença na primeira variável e ausência na segunda, no caso das duas variáveis serem binárias.

simbólico, significa uma correspondência de "valores elevados" ou de "valores baixos", entrando em consideração com a interpretação da semelhança/distância usada no caso simbólico. No caso das variáveis serem heterogéneas, a interpretação é uma conjugação das ideias anteriores. Por exemplo, se uma variável é ordinal e a outra dicotómica encontramos os uns e os zeros concentrados ao longo da ordem.

- Se o coeficiente de semelhança centrado e reduzido, s_{LC}, é fortemente negativo ($s_{LC}<<0$ e, correspondentemente, $P_L \approx 0$), os pares de indivíduos que estão em relação numa das variáveis não o estão na outra. No caso de variáveis nominais, significa que os que estão na mesma classe numa das variáveis, estão em classes separadas na outra. No caso de variáveis ordinais, significa "estarem em sentidos contrários". Nos casos métrico e simbólico, significa uma correspondência de valores contrários ("valores elevados" com "valores baixos" e vice-versa), entrando em consideração com a interpretação da semelhança/distância usada no caso simbólico. No caso das variáveis serem heterogéneas, a interpretação é uma conjugação das ideias anteriores. Por exemplo, se uma variável é ordinal e a outra dicotómica encontramos os uns e os zeros distribuídos ao longo da ordem.
- Se o coeficiente de semelhança centrado e reduzido, s_{LC}, é nulo ($s_{LC}=0$ e, correspondentemente, $P_L=0.5$), uns pares de indivíduos estão em relação nas duas variáveis e outros não.

4.8 Os coeficientes s, s_{LC} e P_L e a inferência estatística

Sempre que estiverem satisfeitas as condições de aplicação da inferência estatística, terá interesse estudar a significância das semelhanças entre as variáveis calculadas a partir dos coeficientes s, s_{LC} e P_L.

A hipótese nula do teste, H_0: as variáveis são independentes (equivale a dizer que os pares de permutações são uniformemente distribuídas), corresponde a afirmar que a semelhança padronizada entre as variáveis é nula ou, de forma equivalente, que o valor do coeficiente s coincide com o valor médio da distribuição de S. O próprio coeficiente S_{LC} é a estatística de teste, que como se sabe, sob H_0, tem uma distribuição assintótica normal centrada e reduzida (Le Calvé, 1977). Enquanto que, o coeficiente P_L nos permite calcular o valor-p do teste:

- Valor-p=1-P_L (quando $P_L>0.5$) ou valor-p = P_L (quando $P_L<0.5$), no caso do teste ser unilateral.

- Valor-p=2(1-P_L) (quando P_L>0.5) ou valor-p = 2P_L (quando P_L<0.5), no caso do teste ser bilateral.

A hipótese nula é rejeitada sempre que o valor-p for muito pequeno. O critério de "pequeno" é-nos dado pelos valores do nível de significância, α, habitualmente mais utilizados na literatura científica por convenção (e.g., 0.05, 0.01); quando o nível de significância α "mede" a probabilidade de um erro que custa "caro", como por exemplo em Medicina, utilizam-se valores menores (e.g., 0.001, 0.0001). Observando os valores de P_L podemos dizer se são significativos, conforme o teste seja unilateral ou bilateral: $P_L \geq 0.95$, $P_L \leq 0.05$, $P_L \geq 0.975$, $P_L \leq 0.025$, entre outros.

Na prática, a aproximação da distribuição da estatística S_{LC} à distribuição normal centrada e reduzida, considerando a hipótese nula verdadeira, depende do tipo de variáveis que se estão a relacionar. Por analogia, como vimos, é válida para amostras com dimensão superior ou igual a oito, n≥8, no caso de se relacionarem variáveis com modalidades estrita e totalmente ordenadas (Subsecção 4.3.6). No caso das variáveis métricas e do par (X,Y) não ser gaussiano, a aproximação é considerada boa, quando as amostras têm dimensão superior ou igual a 30, n≥30 (Saporta, 1990). Quando as variáveis são métricas e o par (X,Y) é gaussiano, a aproximação é considerada boa quando as amostras têm dimensão superior a 100, n>100 (Saporta, 1990). No caso das variáveis ordinais em que se opte pelo número de ordem, aquela aproximação pode ser considerada boa para n>100 (Saporta, 1990), ou para n>30 (Conover, 1999), enquanto Siegel e Castellan (1989) referem n>20 ou n>25. Também têm sido realizados estudos sobre coeficientes probabilísticos deste tipo, que mostram uma convergência rápida quando as amostras têm pequena dimensão (Sousa *et al*., 2005; Secção 3.3). Convém, no entanto, num futuro próximo fazer um estudo mais detalhado sobre este assunto.

5 APLICAÇÕES

5.1 Introdução geral

Este capítulo pretende analisar a aplicação prática dos coeficientes *s*, s_{LC} e P_L apresentados no Capítulo 4. Em particular, pretende-se analisar como estes coeficientes dão resposta a situações concretas, especialmente na presença de variáveis de vários tipos. As variáveis utilizadas para este efeito são baseadas em dados recolhidos num âmbito multidisciplinar relacionado com as ciências médicas e a biologia. Excluindo os casos pontuais em que se exemplificou a utilização destes coeficientes, falta ainda compreender de forma clara como é que eles se comportam perante dados complexos de características variadas, como os que são frequentemente obtidos em diversas áreas científicas. Pretende-se também comparar, quando possível, os resultados aqui obtidos com os já analisados pela aplicação do coeficiente de afinidade sob as suas diversas formas (Capítulo 3) ou por alguma abordagem tradicional.

Para atingir os objectivos deste capítulo, foram realizadas análises com três tipos de dados. No primeiro caso, baseado no Questionário SERVQUAL Modificado, pretende-se aplicar os coeficientes *s*, s_{LC} e P_L à comparação de variáveis com modalidades parcialmente ordenadas (Secção 5.2). No segundo caso, baseado no mesmo questionário, os coeficientes foram aplicados a variáveis heterogéneas (Secção 5.3). No terceiro caso, os coeficientes foram aplicados a matrizes de distâncias ultramétricas (Secção 5.4). Neste terceiro caso foi também aplicado o coeficiente de afinidade, de forma a comparar os resultados dos vários coeficientes. No final apresentaremos as conclusões deste capítulo (Secção 5.5).

5.2 Variáveis de ordem parcial: O Questionário SERVQUAL Modificado

5.2.1 Introdução

O objectivo desta secção é o de analisar os itens de um questionário de qualidade e satisfação na saúde que apresentam maior variabilidade de respostas. Ou seja, pretende-se comparar os itens das escalas “A. Elementos Tangíveis” e “D. Interesse/Capacidade de

Resposta" (Tabela 5.2.1) do Bloco 1 do Questionário SERVQUAL Modificado, utilizando os coeficientes *s*, s_{LC} e P_L, uma vez que estes itens são variáveis com modalidades parcialmente ordenadas.

O Questionário SERVQUAL Modificado foi desenvolvido no âmbito do Projecto de Humanização dos Cuidados Paliativos em Contexto Domiciliário[188], cujo primeiro objectivo é o de melhorar e humanizar os cuidados paliativos prestados aos doentes terminais, no seu domicílio, por equipas de saúde ou voluntários. Para diagnosticar a qualidade e satisfação com os cuidados prestados aos doentes com apoio domiciliário, utilizou-se esta versão do Questionário SERVQUAL (*Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality*) (Anexo 7) traduzido e adaptado para este projecto a partir de um questionário já aplicado em Espanha[189] para um fim semelhante. Este questionário é constituído por quatro partes:

- Bloco 1 - Escala de Percepções ("A. Elementos Tangíveis", "B. Fiabilidade dos Tratamentos e Cuidados", "C. Segurança/Garantia", "D. Interesse/Capacidade de Resposta" e "E. Empatia")
- Bloco 2 - Escala de Preferências
- Dados do doente
- Dados do cuidador

O questionário foi respondido por 58 cuidadores de utentes da região de Lisboa (IPOFG e Centro de Saúde de Odivelas) que receberam cuidados paliativos em contexto domiciliário.

Estes dados já foram objecto de uma análise preliminar (Sousa Ferreira *et al.,* 2003; Bacelar-Nicolau *et al.,* 2005), onde se aplicaram técnicas de análise multivariada para estudar as pontuações globais, em percentagem, obtidas nas cinco escalas de percepções acima referidas. No último trabalho apresentado sobre este tema, por Dias *et al.* (2006), foi realizada a análise dos itens das referidas escalas de percepções, que já estava prevista no âmbito do projecto.

Aqui exploramos as respostas dadas aos itens das duas escalas do Bloco 1, que apresentam maior variabilidade (escalas "A. Elementos Tangíveis" e "D. Interesse/Capacidade de Resposta"), a partir dos coeficientes *s*, s_{LC} e P_L (Doria *et al.*,

[188] Projecto SDH. MD/P.I.01.13, subsidiado pela Fundação Calouste Gulbenkian e coordenado pelo Professor Doutor Manuel Silvério Marques do IPOFG.
[189] Projecto coordenado pelo Professor Óscar Lozano, da Universidade de Sevilha.

2006b[190]) e temos a vantagem de poder comparar estes resultados com os obtidos por aplicação de uma análise mais tradicional (Dias *et al.,* 2006[191]).

Tabela 5.2.1. Itens das escalas de percepções do Bloco 1, "A. Elementos Tangíveis" e "D.Interesse/Capacidade de Resposta", do Questionário SERVQUAL Modificado. As respostas a estes itens são dadas sob a forma: 1-Totalmente em desacordo, 2- Desacordo, 3-Des/Acordo, 4-Acordo, 5-Totalmente de acordo, 6-Não se aplica, 9-Não sabe/Não responde

Escala A. Elementos Tangíveis	***Escala D. Interesse/Capacidade de Resposta***
Equipamento, material médico, medicamentos, informação clínica e terapêutica no domicílio.	*Acessibilidade, desejo de ajudar, capacidade de responder às dúvidas, capacidade para formar o cuidador.*
a1. Os equipamentos médicos são adequados …	*d1p. A equipa de saúde não tem demasiada pressa …*
a2. Existe uma cópia do relatório clínico no seu domicílio ...	*d2. A equipa de saúde mostra sincero interesse …*
a3. Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento ...	*d3. É fácil contactar a equipa de saúde …*
a4. A equipa de saúde fornece as receitas necessárias ...	*d4. Quando chamada, a equipa de saúde dá resposta rapidamente …*
a5. A equipa de saúde fornece outra medicação ...	*d5p. Quando chamada, a equipa de saúde não dá a sensação de vir contrariada …*
a6. A equipa de saúde proporciona o material clínico necessário …	*d6p. Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir ...*
a7. A equipa de saúde facilita o acesso a outro equipamento de acesso …	*d7p. A maior parte das vezes não lhe parece que a equipa de saúde preferia mandar o doente para o hospital ...*
	d8. A equipa de saúde ensinou-lhe a cuidar do doente …
	d9. A equipa de saúde facilita o acesso ao hospital …

Os itens destas escalas são variáveis de ordem parcial (1<2<3<4<5, 6, 9), devido à existência das categorias "6-Não se aplica" e "9-Não sabe/Não responde". Os coeficientes *s*, s_{LC} e P_L têm em conta esta natureza dos itens e permitem relacioná-los.

Para analisarmos as respostas dadas aos itens destas escalas, estudámos as relações entre os itens de cada escala (Subsecções 5.2.2 e 5.2.3) e entre os itens das duas escalas simultaneamente (Subsecção 5.2.4), utilizando os coeficientes *s*, s_{LC} e P_L. Com o objectivo de visualizar as relações entre as respostas dadas aos itens, sintetizar a informação e de ter a possibilidade reconhecer as escalas a partir da proximidade dos seus itens, foram

[190] Apenas uma parte da análise realizada nesta Secção foi tema de uma apresentação oral nas JOCLAD 2006.

[191] Os resultados obtidos pelos autores não foram publicados e sou-lhes grata pela sua disponibilização.

realizadas análises em componentes principais das matrizes de semelhanças *S*, S_{LC} e P_L obtidas nas análises referidas, assim como análises classificatórias hierárquicas ascendentes, ACHA, daquelas matrizes. Sempre que for utilizada a análise classificatória hierárquica ascendente recorreremos ao critério da "estatística de níveis" (Lerman, 1970; Bacelar-Nicolau, 1972, 1980), para escolher a melhor partição da hierarquia obtida. Quando a análise dos itens é realizada dentro de cada escala, a sua interpretação tem que se basear mais no tipo de respostas do que no significado dos itens, ou seja, nos sentimentos/percepções afins que eles pretendem medir.
No final faremos a discussão dos resultados obtidos e apresentaremos algumas conclusões (Subsecção 5.2.5).

5.2.2 Resultados: análise dos itens da "Escala A. Elementos Tangíveis"

Em todos os itens da "Escala A. Elementos Tangíveis" encontramos respostas nas categorias "6. Não se aplica" e/ou "9. Não sabe/Não responde". Nesta escala, os itens "a1", "a3", "a4", "a5" e "a6" apresentam maioria de respostas fortemente favoráveis (5.Totalmente de acordo). Enquanto que "a2. Existe uma cópia do relatório clínico no seu domicílio..." recebe a maior percentagem (39.7%) de respostas fortemente desfavoráveis (1. Totalmente em desacordo) e a maioria de respostas "6. Não se aplica" é dada a "a7. A equipa de saúde facilita o acesso a outro equipamento de acesso " (55.2%) (Anexo 1).

As semelhanças *s* (Tabela 5.2.2), s_{LC} (Tabela 5.2.3) e P_L (Tabela 5.2.4) entre as respostas dadas aos itens da "Escala A. Elementos Tangíveis" permitem-nos distinguir vários itens que estão fortemente associados, quer "positiva", quer "negativamente", e outros para os quais a semelhança padronizada é nula [ver também o Exemplo 4.3.6, Subsecção 4.3.4, Capítulo 4].

Tabela 5.2.2. Matriz de semelhanças *S* entre as respostas dadas aos itens da "Escala A. Elementos Tangíveis"

	a1	a2	a3	a4	a5	a6	a7
a1	1624						
a2	1009	2017					
a3	1130	1465	2301				
a4	1100	1321	1564	2173			
a5	702	990	1234	1172	1543		
a6	1044	1180	1311	1326	969	1969	
a7	690	831	883	923	635	793	1355

Tabela 5.2.3. Matriz de semelhanças S_{LC} entre as respostas dadas aos itens da "Escala A. Elementos Tangíveis". Destacam-se a negrito as semelhanças mais fortes e as mais fracas em itálico

	a1	a2	a3	a4	a5	a6	a7
a1	23.4522						
a2	0.4827	12.8988					
a3	*-0.0068*	**1.3056**	10.7336				
a4	0.7054	*-0.0948*	0.8105	11.8124			
a5	**-1.4517**	0.9323	**3.1880**	**3.0620**	17.1872		
a6	**1.7699**	-0.5021	-0.9821	0.5431	1.0876	13.9794	
a7	0.8791	0.1663	**-1.6665**	0.9394	*0.0922*	-0.3964	31.5682

Tabela 5.2.4. Matriz de semelhanças P_L entre as respostas dadas aos itens da "Escala A. Elementos Tangíveis". Destacam-se a negrito as semelhanças mais fortes e as mais fracas em itálico

	a1	a2	a3	a4	a5	a6	a7
a1	1.0000						
a2	0.6854	1.0000					
a3	*0.4973*	**0.9042**	1.0000				
a4	0.7597	*0.4622*	0.7912	1.0000			
a5	**0.0733**	0.8244	**0.9993**	**0.9989**	1.0000		
a6	**0.9616**	0.3078	0.1630	0.7065	0.8616	1.0000	
a7	0.8103	0.5660	**0.0478**	0.8262	*0.5367*	0.3459	1.0000

5.2.2.1 Análise em componentes principais e ACHA da matriz de semelhanças S entre os itens da Escala A

A matriz de semelhanças *S* é definida positiva (Anexo 8). A 1ª componente principal explica 64.97% da variabilidade total e o 1º plano factorial explica 72.41% desta variabilidade (Tabela 5.2.5).

Tabela 5.2.5. Resultados obtidos com a ACP da matriz de semelhanças *S* entre os itens da "Escala A. Elementos Tangíveis"

	Componente	
	1	**2**
Unidades de inércia (valores próprios)	*8434.7*	*965.6*
Variabilidade Explicada (%)	*64.97*	*7.44*
Var. Exp. Acumulada (%)	*64.97*	*72.41*
a1. Os equipamentos médicos são adequados	*30.01*	***-18.60***
a2. Existe uma cópia do relatório clínico no seu domicílio...	*36.97*	*2.32*
a3. Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento..	***41.87***	*11.49*
a4. A equipa de saúde fornece as receitas necessárias...	***40.35***	*3.77*
a5. A equipa de saúde fornece outra medicação...	*30.02*	***15.87***
a6. A equipa de saúde proporciona o material clínico necessário	*35.86*	***-11.95***
a7. A equipa de saúde facilita o acesso a outro equipamento de acesso	*24.46*	*-8.56*

Todos os itens têm coordenadas positivas na 1ª componente principal, destacando-se os itens “a3” e “a4” sobre a informação clínica e terapêutica. A 2ª componente principal opõe o item “a1” ao item “a5” (Tabela 5.2.5, Figura 5.2.1); esta componente mede a oposição entre o “equipamento médico” e “outra medicação”.

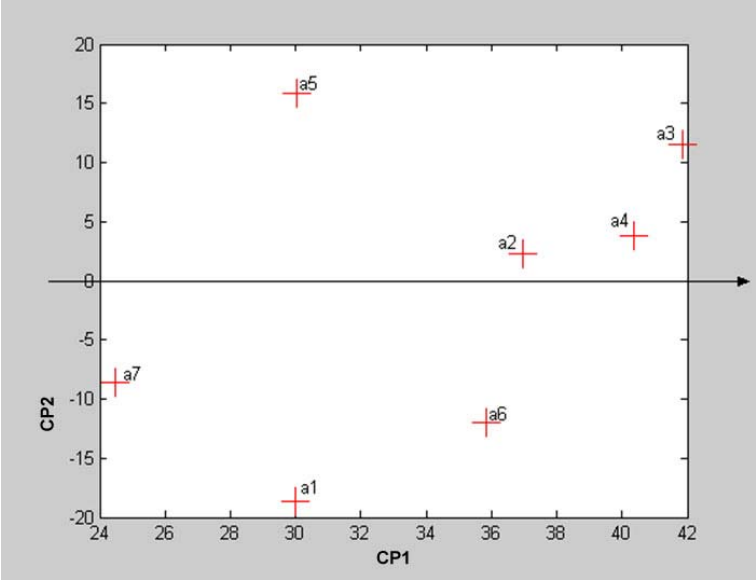

Figura 5.2.1. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças *S* entre os itens da “Escala A. Elementos Tangíveis”.

A análise em componentes principais das semelhanças *s* permitiu visualizar, no 1º plano factorial, relações de proximidade e de afastamento entre os itens da Escala A (Figura 5.2.1). Podemos assim considerar:

- (“a3“, “a4“, “a2“, “a6”), (“a1”), (“a5”) e (“a7”)

Nestas relações de proximidade e afastamento reconhecemos, precisamente, as classes da melhor partição obtida pela ACHA (*s* + *Ligação completa*), segundo o critério da ”estatística de níveis”[192], no nível 3: {a1}, {a2, a3, a4, a6}, {a5} e {a7} (Figura 5.2.2).

[192] Lerman (1970); Bacelar-Nicolau (1972, 1980).

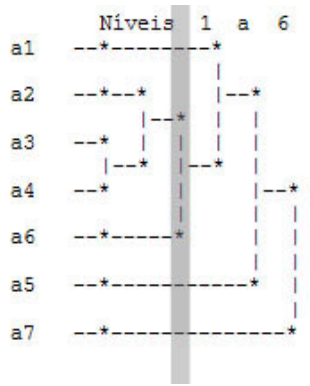

Figura 5.2.2. Dendrograma obtido com a análise classificatória hierárquica ascendente (s+Ligação completa). Está assinalado o melhor nível, segundo o critério da "estatística de níveis", STAT(3)=3.42.

Podemos então caracterizar os itens pelos seus perfis de resposta:

- O item "a1. Os equipamentos médicos são adequados" apresenta predominância da resposta "5-Totalmente de acordo", não sendo referidas respostas negativas e/ou do tipo negativo, mas sobressaindo os cerca de 26% de respostas "6-Não se aplica".
- Informação clínica/terapêutica e receitas – "a3", "a4" – e "a6. A equipa de saúde proporciona o material clínico necessário", que se caracterizam pela maioria de respostas "5-Totalmente de acordo" e "a2" que apresenta como resposta mais frequente "1-Totalmente em desacordo" (39.7%).
- No item "a5"/"Outra medicação" são contempladas todas as possibilidades de resposta com predominância, igualmente distribuída pelas categorias positivas, "4-Acordo" e "5-Totalmente de acordo" (58.6%).
- No item "a7. A equipa de saúde facilita o acesso a outro equipamento de acesso" são contempladas todas as possibilidades de resposta com predominância de respostas "6-Não se aplica" (55.2%).

5.2.2.2 Análise em componentes principais e ACHA da matriz de semelhanças S_{LC} entre os itens da Escala A

A matriz de semelhanças S_{LC} é definida positiva (Anexo 9).

A 1ª componente principal só explica 31.84% da variabilidade total e o 1º plano factorial explica 45.87% desta variabilidade (Tabela 5.2.6). Os 74.21% de variabilidade total explicada só é atingida no espaço a quatro dimensões.

Tabela 5.2.6. Resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre os itens da "Escala A. Elementos Tangíveis"

	Componente		
	1	*2*	*3*
Unidades de inércia (valores próprios)	*31.84*	*23.95*	*19.92*
Variabilidade Explicada (%)	*26.18*	*19.69*	*16.38*
Var. Exp. Acumulada (%)	*26.18*	*45.87*	*62.25*
a1. Os equipamentos médicos são adequados	***0.613***	***-4.694***	*-0.718*
a2. Existe uma cópia do relatório clínico no seu domicílio...	*0.031*	*-0.052*	*-0.732*
a3. Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento..	*-0.438*	*0.225*	*-1.446*
a4. A equipa de saúde fornece as receitas necessárias...	*0.254*	*0.000*	*-1.638*
a5. A equipa de saúde fornece outra medicação...	*-0.069*	*0.990*	***-3.681***
a6. A equipa de saúde proporciona o material clínico necessário	*-0.036*	*-0.766*	*-0.736*
a7. A equipa de saúde facilita o acesso a outro equipamento de acesso	***5.586***	*0.540*	*-0.006*

A ACP da matriz de semelhanças S_{LC} permitiu evidenciar alguns itens da "Escala A. Elementos Tangíveis", que se distinguem pelas suas características únicas:

- Na 1ª componente principal destaca-se o item "a7" (Figura 5.2.3). Este item ao afastar-se de todos os outros revela a predominância de respostas na categoria "6. Não se aplica" (55.2%), entre todas as possibilidades de resposta observadas neste item, que também foram respondidas. Não é de estranhar a ocorrência de grande frequência deste tipo de resposta neste item, dado que trata de acesso a outros equipamentos que muitos doentes não necessitam.

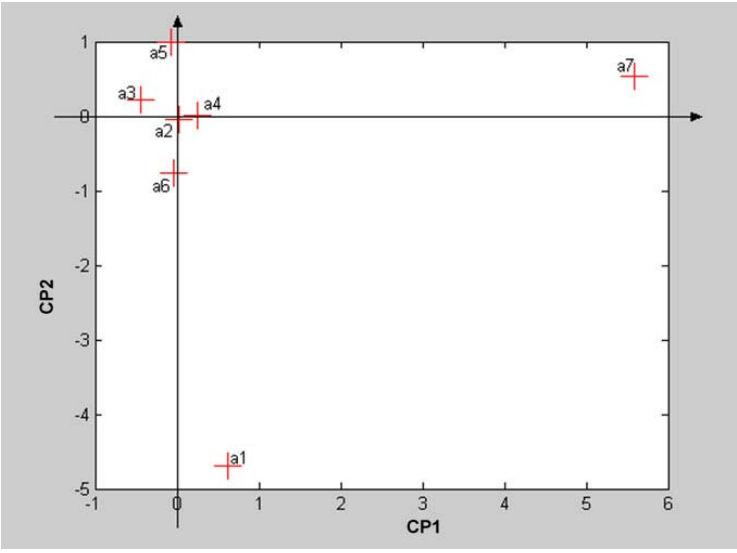

Figura 5.2.3. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças S_{LC} entre os itens da "Escala A. Elementos Tangíveis".

- Na 2ª componente destaca-se “a1.Os equipamentos médicos são adequados” (Figura 5.2.3), que se caracteriza por não ter respostas do tipo negativo (“1- Desacordo total”, “2- Desacordo” e “9- Não sabe/Não responde”). Aquele item opõe-se, eventualmente, a “a5”.

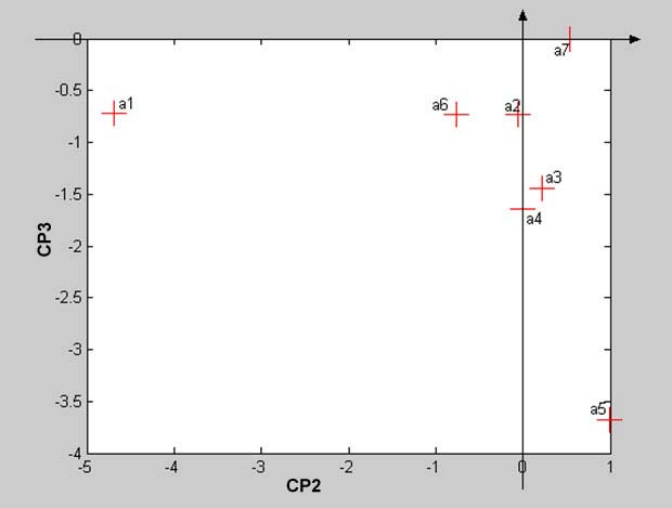

Figura 5.2.4. Representação gráfica do plano factorial (2,3) obtido com a ACP da matriz de semelhanças S_{LC} entre os itens da “Escala A. Elementos Tangíveis”.

- A 3ª componente principal (Figura 5.2.4) faz sobressair “a5.A equipa de saúde fornece outra medicação...” que, embora apresentando respostas em todas as categorias, revela predominância nas categorias positivas (“4-Acordo“ e “5-Totalmente de acordo“) – 58.6% (Anexo 1).

Em vários algoritmos de ACHA, (s_{LC}+ Ligação única), (s_{LC}+ Ligação pela média), (s_{LC}+ AVM), (s_{LC}+ AVL), (s_{LC}+ AVB) (Anexo 9), obtém-se, no nível 4, como melhor partição: {{a_1, a_6}, {a_3, a_5, a_4, a_2}, {a_7}} (Figura 5.2.5, Figura 5.2.6).

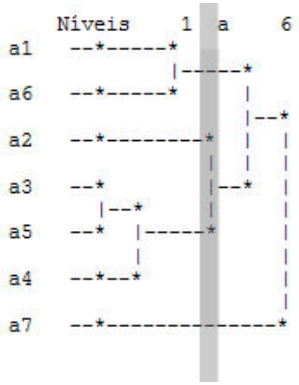

Figura 5.2.5. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+Ligação única). Está assinalado o melhor nível, segundo o critério da ”estatística de níveis”, STAT(4)=2.6112.

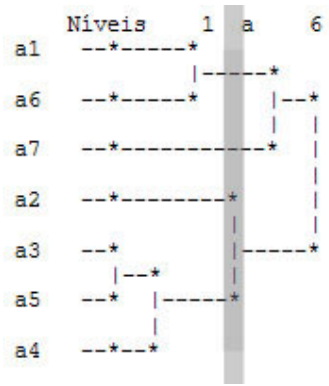

Figura 5.2.6. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+ Ligação pela média). Está assinalado o melhor nível, segundo o critério da "estatística de níveis", STAT(4)=2.6112.

Reconhecemos, nesta partição, as semelhanças observadas entre os itens (Tabela 5.2.3) e projectadas no 1º plano factorial (Figura 5.2.3).

5.2.2.3 Análise em componentes principais e ACHA da matriz de semelhanças P_L entre os itens da Escala A

Alguns resultados obtidos com a ACP da matriz de semelhanças P_L estão apresentados na Tabela 5.2.7 e na Figura 5.2.7

A matriz de semelhanças P_L não é semidefinida positiva (Anexo 10).

Tabela 5.2.7. Resultados obtidos com a ACP da matriz de semelhanças P_L entre os itens da "Escala A. Elementos Tangíveis"

	Componente	
	1	**2**
Unidades de inércia (valores próprios)	*4.7988*	*1.4476*
Variabilidade Explicada (%)	*68.55*	*20 .68*
Var. Exp. Acumulada (%)	*68.55*	*89.23*
a1.Os equipamentos médicos são adequados	*0.805*	*0.577*
a2.Existe uma cópia do relatório clínico no seu domicílio...	*0.818*	*-0.303*
a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...	*0.786*	*-0.655*
a4.A equipa de saúde fornece as receitas necessárias...	*0.963*	*0.057*
a5.A equipa de saúde fornece outra medicação...	*0.925*	*-0.440*
a6.A equipa de saúde proporciona o material clínico necessário	*0.758*	*0.406*
a7.A equipa de saúde facilita o acesso a outro equipamento de acesso	*0.713*	*0.482*

A 1ª componente principal explica grande parte da variabilidade (68.55%) e o 1º plano factorial explica a maior parte da variabilidade total (89.23% - Tabela 5.2.7).

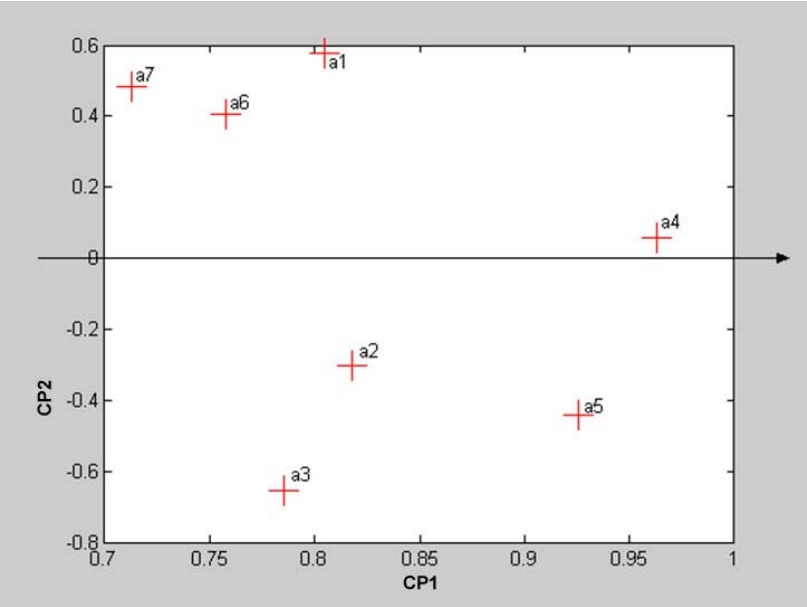

Figura 5.2.7. Representação dos itens da "Escala A. Elementos Tangíveis" no 1º plano factorial, obtido com a ACP da matriz P_L.

Como se pode verificar:

- Todas as variáveis têm coordenadas positivas na 1ª componente principal, destacando-se a relação/proximidade entre "a4. A equipa de saúde fornece as receitas necessárias..." e "a5. A equipa de saúde fornece outra medicação..." (P_L(a4, a5)= 0.9989). Aqui a 1ª componente caracteriza-se pela relação entre "receitas" e "outra medicação". Verificamos pois, que todas as possibilidades de resposta do item "a5" (com predominância das opções positivas, 4 e 5) são respondidas pelos cuidadores que dão a maioria de respostas "5-Totalmente de acordo" ao item "a4.A equipa de saúde fornece as receitas necessárias..." (Anexo 1).

- A 2ª componente caracteriza-se pela oposição entre "a1"/"equipamentos médicos" e "a3"/"informação terapêutica". Aos dois itens é dada uma maioria de respostas "5-Totalmente de acordo". No entanto, o item "a1" obtém respostas mais consensuais, pois não lhe são dadas respostas do tipo negativo ("1-Totalmente em desacordo","2-Desacordo" e "9-Não sabe/Não responde").

No 1º plano factorial observa-se a oposição entre “a7”, “a6”, “a1” e os itens “a3“, “a2“, “a5“ e “a4“. A ACP das semelhanças P_L permitiu destacar relações entre alguns itens da Escala A, podendo visualizar-se, na Figura 5.2.7, dois conjuntos:

- Os itens com respostas extremas: “a7”, “a1” e “a6”.

  Os itens “a1. Os equipamentos médicos são adequados” e “a6. A equipa de saúde proporciona o material clínico necessário” apresentam predominância da resposta “5-Totalmente de acordo” seguida da “6-Não se aplica”, não sendo referidas respostas do tipo negativo.

  Enquanto que, em “a7. A equipa de saúde facilita o acesso a outro equipamento de acesso” são contempladas todas as possibilidades de resposta com predominância de respostas “6-Não se aplica” (32 – 55.2%).

  O item “a7” está próximo dos itens “a1” e “a6” porque se opõe aos restantes itens.

- Os itens “a2”, “a3” e “a4”, “a5”

  O item “a2. Existe uma cópia do relatório clínico no seu domicílio...” - Apresenta predominância da resposta “1-Totalmente em desacordo” (23 - 39.7%) seguida da resposta “5-Totalmente de acordo” (20 – 34.5%), (P_L(a2,a3)=0.9042).

  O item “a3. Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...”, “a4. A equipa de saúde fornece as receitas necessárias...” e “a5.A equipa de saúde fornece outra medicação...”. São itens que se caracterizam por estarem contempladas todas as possibilidades de resposta (“a4” não tem respostas totalmente em desacordo) com predominância da(s) resposta(s) positivas “5-Totalmente de acordo” e “4-Acordo” (P_L(a3,a5)=0.9993, P_L(a4,a5)=0.9989).

Os algoritmos de ACHA – (P_L+ Ligação única) e (P_L+ Ligação pela média) – permitiram obter como melhor partição, no nível 4: {{a_1, a_6}, {a_7}, {a_3, a_5, a_4, a_2}} (Figura 5.2.8). Esta partição está de acordo com o resultado da ACP apresentado.

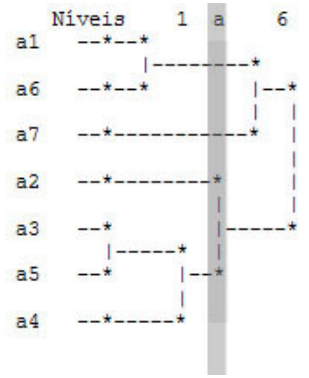

Figura 5.2.8. Dendrograma obtido com a análise classificatória hierárquica ascendente (P_L+A.M.). Está assinalado o melhor nível, segundo o critério da ”estatística de níveis”, STAT(4)=2.6112.

5.2.2.4 Discussão dos resultados obtidos com as ACP e ACHA das matrizes de semelhanças entre os itens da Escala A

Cada uma das ACP realizadas, permitiu realçar aspectos diferentes das relações entre os itens da Escala A que passamos a apresentar na seguinte tabela:

Tabela 5.2.8. Comparação dos resultados obtidos com as ACP das matrizes de semelhanças *S*, S_{LC}, P_L e a análise das ordens entre os itens da "Escala A. Elementos Tangíveis", no 1º plano factorial

ACP da matriz S (n=58)	ACP da matriz S_{LC} (n=58)	ACP da matriz P_L (n=58)	Análise das ordens (n=14)
A5			*a3*
a3, a2, a4, a6	*A5, a3, a2, a4, a6* [193]	*a3, a4, a5, a2*	*a2, a4, a5, a6*
a1	***a1***	*a1, a6, a7*	*a1, a7*
a7	***a7***		

Na abordagem mais utilizada por Dias *et al*. (2006) foram retirados da amostra todos os cuidadores que tivessem dado respostas "6-Não se aplica" e/ou "9-Não sabe/Não responde", ficando a amostra reduzida a 14 indivíduos. Os resultados obtidos com a análise das ordens[194], no 1º plano factorial, permitem detectar os itens "a2", "a4", "a5" e "a6" como importantes na formação da 1ª componente principal. Enquanto que os itens "a7" e "a3" são importantes na formação da 2ª componente principal. Estes resultados só apresentam analogia com os obtidos com os coeficientes *s*, s_{LC}, P_L, no que se refere ao item "a7" e ao facto deste se afastar do item "a3". Contudo, o item "a7" destaca-se nos dois tipos de análise por motivos diferentes. Numas é a predominância de respostas "6-Não se aplica" que o destaca, noutra (na abordagem clássica) é precisamente o facto de passar a ser um item incaracterístico, ao ser-lhe retirada a possibilidade de respostas "6-Não se aplica" (Anexo 1).

Em Dias *et al*. (2006) também se recorreu à análise classificatória hierárquica ascendente sobre a matriz das correlações de Spearman, R_S. Estes resultados (Tabela 5.2.9) são diferentes dos obtidos com os coeficientes s, s_{LC} e P_L.

[193] Centro de gravidade da distribuição; *ventre mou* (em francês).
[194] Em francês, *analyse des rangs* (Lebart *et al*., 1995)

Tabela 5.2.9. Comparação dos resultados obtidos com as ACHA das matrizes de semelhanças *S*, S_{LC}, P_L e R_S entre os itens da "Escala A. Elementos Tangíveis"

Algoritmo: (s+ Lig compl)	**Algoritmo: (s_{LC}+Lig média) ou (s_{LC}+Lig única)**	**Algoritmo: (P_L+Lig média) ou (P_L+Lig única)**	**Algoritmo: (r_S+Lig compl)**
Partição obtida no nível k=3	**Partição obtida no nível k=4**	**Partição obtida no nível k=4**	**Partição obtida no nível k=3**
STAT(3)=3.42	**STAT(4)=2.611**	**STAT(4)=2.611**	
(n=58)	**(n=58)**	**(n=58)**	**(n=14)**
{a5}			*{a3}*
{a3, a4, a2, a6}	*{a3, a5, a4, a2}*	*{a3, a5, a4, a2}*	*{a2, a5, a6, a4}*
{a1}	*{a1, a6}*	*{a1, a6}*	*{a1, a7}*
{a7}	*{a7}*	*{a7}*	

Podemos observar que os resultados obtidos com a ACHA sobre as matrizes de semelhanças S_{LC} e P_L quando se utilizam os critérios de agregação "Ligação única" e "Ligação pela média" são os mesmos. Enquanto que, a projecção daquelas matrizes de semelhanças afecta mais a visualização dos resultados, como vimos.

Pensamos que a partição constituída pelas classes, $\{a_1, a_6\}$, $\{a_7\}$ e $\{a_3, a_5, a_4, a_2\}$, já caracterizadas anteriormente, traduz bem a tipologia de respostas dadas pelos cuidadores aos itens da "Escala A. Elementos Tangíveis". Os coeficientes s_{LC} e P_L permitem pois, obter os resultados que melhor traduzem as respostas dos cuidadores.

5.2.3 Resultados: análise dos itens da "Escala D. Interesse/Capacidade de Resposta"

Na "Escala D. Interesse/Capacidade de Resposta", todos os itens (Tabela 5.2.1) apresentam uma maior percentagem de respostas na categoria "5-Totalmente de acordo" (Anexo 1). Com excepção do item "d1p. A equipa de saúde não tem demasiada pressa", todos eles apresentam respostas "6-Não se aplica" e/ou "9-Não sabe/Não responde". As respostas "6-Não se aplica" tornam-se expressivas nos itens "d3", "d4", "d5p" e "d6p" pois atingem cerca de 25% das respostas dadas a cada um deles (Anexo 1).

Os valores das semelhanças *s*, s_{LC} e P_L, entre as respostas dadas aos itens da "Escala D. Interesse/Capacidade de Resposta", são apresentados na Tabela 5.2.10, na Tabela 5.2.11 e na Tabela 5.2.12, respectivamente.

Tabela 5.2.10. Matriz de semelhanças *S* entre as respostas dadas aos itens da "Escala D. Interesse/Capacidade de Resposta"

	d1p	d2	d3	d4	d5p	d6p	d7p	d8	d9
d1p	*2596*								
d2	*2406*	*2977*							
d3	*1456*	*1719*	*1828*						
d4	*1375*	*1656*	*1667*	*1766*					
d5p	*1465*	*1743*	*1744*	*1696*	*1891*				
d6p	*1117*	*1245*	*1250*	*1309*	*1271*	*1319*			
d7p	*2264*	*2659*	*1658*	*1594*	*1720*	*1170*	*2872*		
d8	*1767*	*1997*	*1337*	*1232*	*1317*	*965*	*1921*	*2165*	
d9	*1239*	*1438*	*903*	*855*	*884*	*640*	*1374*	*1101*	*1467*

Tabela 5.2.11. Matriz de semelhanças S_{LC} entre as respostas dadas aos itens da "Escala D. Interesse/Capacidade de Resposta"

	d1p	d2	d3	d4	d5p	d6p	d7p	d8	d9
d1p	9.0562								
d2	1.6451	9.0484							
d3	0.5334	2.0278	17.9590						
d4	-0.3466	2.0415	**16.5566**	21.3944					
d5p	-0.6387	1.2172	**15.7462**	**16.7658**	18.3546				
d6p	2.0369	1.9719	**16.1615**	**20.5080**	**17.6031**	26.8328			
d7p	0.2050	2.0281	1.6893	1.6042	1.9801	0.7705	9.3897		
d8	1.1891	1.0665	2.8116	1.6866	1.7080	2.5066	0.8202	12.2685	
d9	1.8107	3.0317	2.0577	1.7637	1.0715	1.5198	2.3193	2.6310	17.2464

Facilmente observamos:

- A relação padronizada forte positiva que existe entre os itens "d3","d4","d5p" e "d6p".
- A relação padronizada negativa entre os itens "d1p" e "d5p" (s_{LC}(d1p,d5p)=-0.6387, P_L(d1p,d5p)=0.2615).

Tabela 5.2.12. Matriz de semelhanças P_L entre as respostas dadas aos itens da "Escala D. Interesse/Capacidade de Resposta"

	d1p	d2	d3	d4	d5p	d6p	d7p	d8	d9
d1p	1.0000								
d2	0.9500	1.0000							
d3	0.7031	0.9787	1.0000						
d4	0.3644	0.9794	**1.0000**	1.0000					
d5p	0.2615	0.8882	**1.0000**	**1.0000**	1.0000				
d6p	0.9792	0.9757	**1.0000**	**1.0000**	**1.0000**	1.0000			
d7p	0.5812	0.9787	0.9544	0.9457	0.9762	0.7795	1.0000		
d8	0.8828	0.8569	0.9975	0.9542	0.9562	0.9939	0.7940	1.0000	
d9	0.9649	0.9988	0.9802	0.9611	0.8580	0.9357	0.9898	0.9957	1.0000

5.2.3.1 Análise em componentes principais e ACHA da matriz de semelhanças S entre os itens da Escala D

A matriz de semelhanças *S* é definida positiva (Anexo 11).

A percentagem de variabilidade total explicada pelo 1º factor é de 76.36%, enquanto que o 2º factor só permite explicar 9.82% da variabilidade total (Tabela 5.2.13). O 1º plano factorial explica, pois, grande parte da variabilidade total, 86.18%.

Tabela 5.2.13. Resultados obtidos com a ACP da matriz de semelhanças S entre os itens da "Escala D. Interesse/Capacidade de Resposta"

	Componente	
	1	**2**
Unidades de inércia (valores próprios)	*14419*	*1853*
Variabilidade Explicada (%)	*76.36*	*9.82*
Var. Exp. Acumulada (%)	*76.36*	*86.18*
d1p. A equipa de saúde não tem demasiada pressa	***44.96***	***-13.99***
d2. A equipa de saúde mostra sincero interesse	***51.15***	***-11.93***
d3. É fácil contactar a equipa de saúde	*37.80*	***16.88***
d4. Quando chamada, a equipa de saúde dá resposta rapidamente	*36.54*	***18.67***
d5p. Quando chamada, a equipa de saúde não dá a sensação de vir contrariada	*38.32*	***17.58***
d6p. Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir...	*28.33*	***16.04***
d7p. A maior parte das vezes não lhe parece que a equipa de saúde preferia mandar o doente para o hospital	***49.46***	***-11.45***
d8. A equipa de saúde ensinou-lhe a cuidar do doente	*39.20*	*-9.56*
d9. A equipa de saúde facilita o acesso ao hospital	*27.62*	*-9.69*

Todos os itens desta escala têm coordenadas positivas na 1ª componente principal. No entanto, nesta componente destaca-se o interesse da equipa de saúde: destacam-se bem os itens "d2" e "d7p". Estes itens caracterizam-se pelo maior consenso de respostas "5-Totalmente de acordo" (91.4%, 89.7%, respectivamente) e 84.5% dos cuidadores estão totalmente de acordo com os dois itens (Anexo 1).

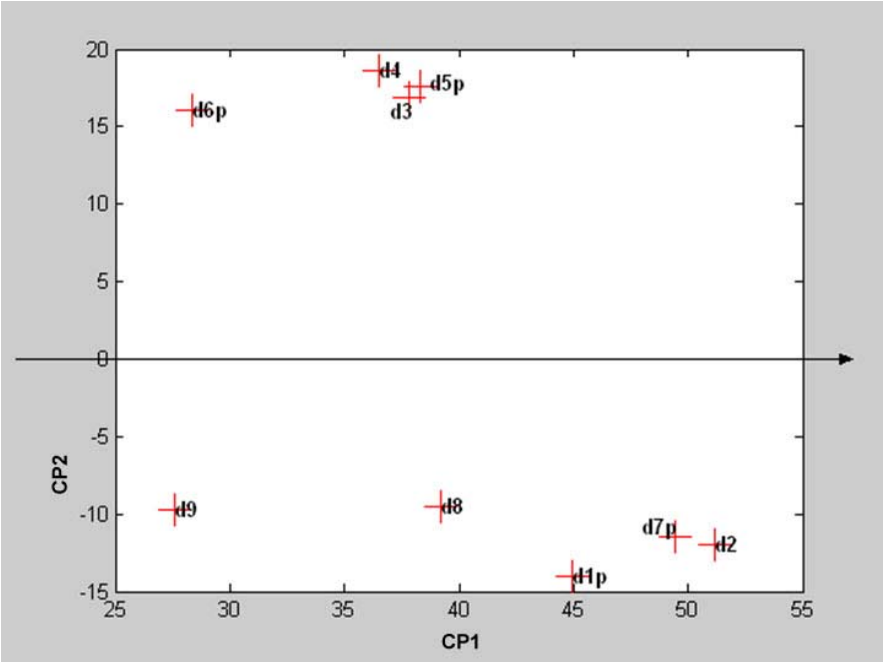

Figura 5.2.9. Representação dos itens da "Escala D. Interesse/Capacidade de Resposta" no 1º plano factorial obtido com a ACP da matriz *S*.

A 2ª componente principal destaca os itens que se referem à "acessibilidade" da equipa de saúde, "d4", "d5p", "d3" e "d6p". Como referimos acima, eles apresentam cerca de ¼ de respostas "6-Não se aplica". Observamos também a oposição deles ao item "d1p" que não apresenta respostas "6-Não se aplica" e "9-Não sabe/Não responde". Esta componente opõe os itens que apresentam as maiores percentagens de respostas "6-Não se aplica" aos restantes itens.

A ACP dos coeficientes de semelhança *s* permitiu visualizar, no 1º plano factorial, relações de proximidade e de afastamento entre os itens da Escala D (Figura 5.2.9): [("d6p"), ("d3","d4","d5p")], [("d2", "d7p", "d1p"), "d8", "d9"]. Podemos caracterizá-los pelos seus perfis de resposta:

- Itens que se referem à "acessibilidade" da equipa de saúde e que apresentam cerca de 25% de respostas "6-Não se aplica":

  - O item "d6p" é o único que apresenta contempladas todas as possibilidades de resposta. A resposta "5-Totalmente de acordo" embora seja a mais frequente (39,7%), é a que traduz menos adesão total entre todos os itens desta escala. Neste item também sobressai a percentagem de respostas "6-Não se aplica" (27.6%).

  - Nos itens "d3", "d4", "d5p" não estão contempladas todas as possibilidades de resposta e, embora apresentem maior percentagem de respostas "5-Totalmente de acordo", também sobressaem as respostas "6-Não se aplica" (24.1%, 27.6% e 25.9%, respectivamente).

- Os itens sobre o interesse da equipa de saúde “d2”, “d7p” “concentram” as respostas em “5-Totalmente de acordo” (91.4% e 89.7%, respectivamente).

  No item “d1p. A equipa de saúde não tem demasiada pressa” as respostas “5-Totalmente de acordo” também constituem a maioria (74.1%). No entanto, é o único item que só apresenta respostas na escala ordinal, não apresentando respostas às opções 6 e 9.

  O item “d8. A equipa de saúde ensinou-lhe a cuidar do doente” apresenta grande variabilidade de respostas (só não é respondida a opção “9-Não sabe/Não responde”) com maioria de respostas “5-Totalmente de acordo” (69.0%).

  O item “d9. A equipa de saúde facilita o acesso ao hospital” apresenta grande variabilidade de respostas (só não é respondida a opção “1-Totalmente em desacordo”), sendo “5-Totalmente de acordo” a mais frequente (48.3%) e que se caracteriza por ser o item com a maior percentagem de respostas na categoria “9-Não sabe/Não responde” (12.1%).

Os resultados obtidos, sobre os itens da “Escala D”, com o algoritmo (s+Ligação completa) da ACHA (Figura 5.2.10) complementam bem o que acabámos de ver sobre a ACP da matriz S.

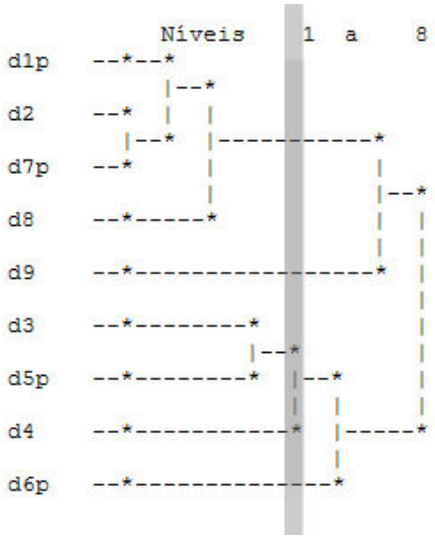

Figura 5.2.10. Dendrograma obtido com a análise classificatória hierárquica ascendente (s+Ligação completa). Está assinalado o melhor nível, segundo o critério da ”estatística de níveis”, STAT(5)=4.22.

A partição obtida no nível 5 é a melhor, segundo o critério da ”estatística de níveis”: {{d2, d7p, d1p, d8}, {d9}, {d3, d5p, d4}, {d6p}}. Nela reconhecemos as relações entre os itens, visualizadas no 1º plano factorial da projecção da matriz de semelhanças S (Figura 5.2.9).

5.2.3.2 Análise em componentes principais e ACHA da matriz de semelhanças S_{LC} entre os itens da Escala D

A matriz de semelhanças S_{LC} é definida positiva (Anexo 12).
A percentagem de variabilidade total explicada pelo 1º factor é de 52.44%, enquanto que o 2º factor só permite explicar 14.21% da variabilidade total (Tabela 5.2.14).

Na 1ª componente principal visualiza-se bem a relação entre os itens ligados à acessibilidade/contacto da equipa de saúde, "d3", "d4", "d5p" e "d6p" (Tabela 5.2.14, Figura 5.2.11) que apresentam as maiores percentagens de respostas na categoria "6-Não se aplica", como já foi referido. Enquanto na 2ª componente destaca-se o item "d9. A equipa de saúde facilita o acesso ao hospital" que se caracteriza por ser o que tem a maior percentagem de respostas na categoria "9-Não sabe/Não responde" (12.1%) (Anexo 1). Na 3ª componente principal é o item "d8" que sobressai (Anexo 12).

Tabela 5.2.14. Resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre os itens da "Escala D. Interesse/Capacidade de Resposta"

	Componente	
	1	**2**
Unidades de inércia (valores próprios)	*74.24*	*20.11*
Variabilidade Explicada (%)	*52.44*	*14.21*
Var. Exp. Acumulada (%)	*52.44*	*66.65*
d1p. A equipa de saúde não tem demasiada pressa	*0.16*	*0.96*
d2. A equipa de saúde mostra sincero interesse	*0.53*	*1.34*
d3. É fácil contactar a equipa de saúde	***3.87***	*0.04*
d4. Quando chamada, a equipa de saúde dá resposta rapidamente	***4.40***	*-0.30*
d5p.Quando chamada, a equipa de saúde dá a sensação de não vir contrariada	***3.97***	*-0.39*
d6p.Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir...	***4.79***	*-0.33*
d7p. A maior parte das vezes parece-lhe que a equipa de saúde não preferia mandar o doente para o hospital...	*0.43*	*1.05*
d8.A equipa de saúde ensinou-lhe a cuidar do doente	*0.64*	*1.44*
d9. A equipa de saúde facilita o acesso ao hospital	*0.56*	***3.72***

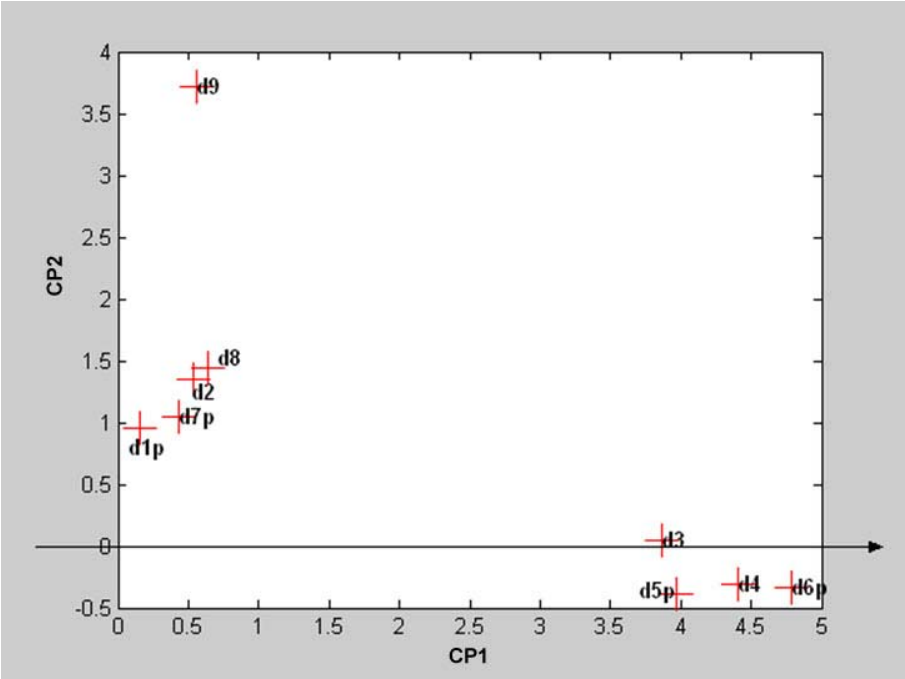

Figura 5.2.11. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças S_{LC} entre os itens da "Escala D. Interesse/Capacidade de Resposta".

A ACP das semelhanças s_{LC} permitiu encontrar proximidade/afastamento entre alguns itens desta escala, bem visualizados no 1º plano factorial e aos quais correspondem diferentes perfis de resposta:

- O item "d9". A equipa de saúde facilita o acesso ao hospital" é o que apresenta a maior percentagem de respostas na categoria "9-Não sabe/Não responde" (7-12.1%), entre os itens desta escala;
- Os itens de respostas ligadas à acessibilidade/contacto com a equipa de saúde, "d3", "d4", "d5p" e "d6p", que apresentam maior percentagem de respostas "6-Não se aplica" (respectivamente, 24.1%, 27.6%, 25.9%, 27,6%);
- Os itens "d1p", "d2", "d7p" e "d8" que se caracterizam pela maioria de respostas totalmente em acordo, tal como foi visto.

No entanto, o algoritmo (s_{LC}+Ligação completa) da ACHA (Figura 5.2.12) também faz sobressair relações entre alguns dos itens desta escala em que reconhecemos a 3ª e a 4ª componentes da ACP (Anexo 12). É o caso dos itens "d1p", "d8" e "d9". No nível 5 obtém-se a melhor partição, segundo o critério da "estatística dos níveis", cujas classes são:

- Classe 1 = {d1p} – Este item caracteriza-se por ser o único que não teve respostas nas categorias "6-Não se aplica" e "9-Não sabe/Não responde", apresentando maioria de respostas totalmente de acordo (43 - 74.1%).

- Classe 2 = {d2, d9, d7p} – os itens “d2” e “d7p” têm respostas totalmente em acordo que rondam os 90%, que se distribuem pelas várias possibilidades de resposta de “d9” (no qual o “1-Totalmente em desacordo” não tem respostas).
- Classe 3 = {d3, d4, d5p, d6p} – itens de respostas ligadas à acessibilidade/contacto com a equipa de saúde, que apresentam maior percentagem de respostas “6.Não se aplica” (respectivamente, 24.1%, 27.6%, 25.9%, 27.6%).
- Classe 4 = {d8} – item com maioria de respostas totalmente em acordo (40 - 69%), em que a possibilidade “9-Não sabe/Não responde” não é contemplada, apresentando contudo 6 (10.3%) respostas “6-Não se aplica”.

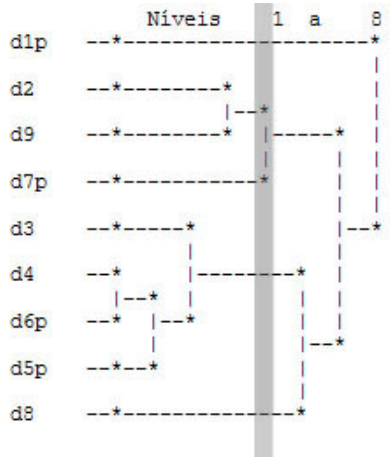

Figura 5.2.12. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+Ligação completa). Está assinalado o melhor nível, segundo o critério da ”estatística de níveis”, STAT(5)=4.11.

5.2.3.3 Análise em componentes principais e ACHA da matriz de semelhanças P_L entre os itens da Escala D

A matriz de semelhanças P_L entre as respostas dadas aos itens da “Escala D. Interesse/Capacidade de Resposta" não é semidefinida positiva (Anexo 13). O 1º factor explica grande parte da variabilidade total, ou seja, 91.63%, enquanto o 1º plano factorial explica mais do que 100% da variabilidade total, devido ao facto da matriz P_L não ser s.d.p. (Tabela 5.2.15).

A 1ª componente principal é bastante geral pois todos os itens têm coordenadas positivas e bastante próximas. Na 2ª componente encontramos a oposição entre o item “d1p” e os itens “d5p”, “d4” e “d7p”. Parece-nos que esta é uma componente ligada principalmente à satisfação com o atendimento do doente.

Tabela 5.2.15. Resultados obtidos com a ACP da matriz de semelhanças P_L entre os itens da "Escala D. Interesse/Capacidade de Resposta"

	Componente	
	1	**2**
Unidades de inércia (valores próprios)	*8.25*	*1.03*
Variabilidade Explicada (%)	*91.63*	*11.45*
Var. Exp. Acumulada (%)	*91.63*	*103.08*
d1p. A equipa de saúde não tem demasiada pressa	*0.77*	***0.75***
d2. A equipa de saúde mostra sincero interesse	***0.997***	*0.12*
d3. É fácil contactar a equipa de saúde	***1.00***	*-0.10*
d4. Quando chamada, a equipa de saúde dá resposta rapidamente	*0.96*	*-0.38*
d5p. Quando chamada, a equipa de saúde dá a sensação de não vir contrariada	*0.93*	***-0.46***
d6p. Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir...	***1.00***	*0.13*
d7p. A maior parte das vezes parece-lhe que a equipa de saúde não preferia mandar o doente para o hospital...	*0.93*	*-0.21*
d8. A equipa de saúde ensinou-lhe a cuidar do doente	***0.98***	*0.08*
d9. A equipa de saúde facilita o acesso ao hospital	***1.01***	*0.16*

No 1º plano factorial (Tabela 5.2.15, Figura 5.2.13), o coeficiente P_L faz sobressair a oposição do item "d1p" aos restantes itens:

- O item "d1p. A equipa de saúde não tem demasiada pressa", que só apresenta respostas na escala ordinal, destaca-se dos restantes itens desta escala.
- "d2", "d6p", "d8", "d9" e "d3" são os itens que apresentam maior variabilidade de respostas (com excepção dos itens "d2" e "d3") e "d4","d5p" e "d7p" são itens com respostas predominantemente positivas além das respostas "6-Não se aplica" e "9-Não sabe/Não responde".

Aqueles dois conjuntos de itens são as classes da partição, obtida no nível 5, pelos algoritmos (P_L+Ligação completa) e (P_L+Ligação pela média) da ACHA: {{d1p}, {d2, d6p, d8, d9, d3, d4, d5p, d7p}} (Figura 5.2.14).

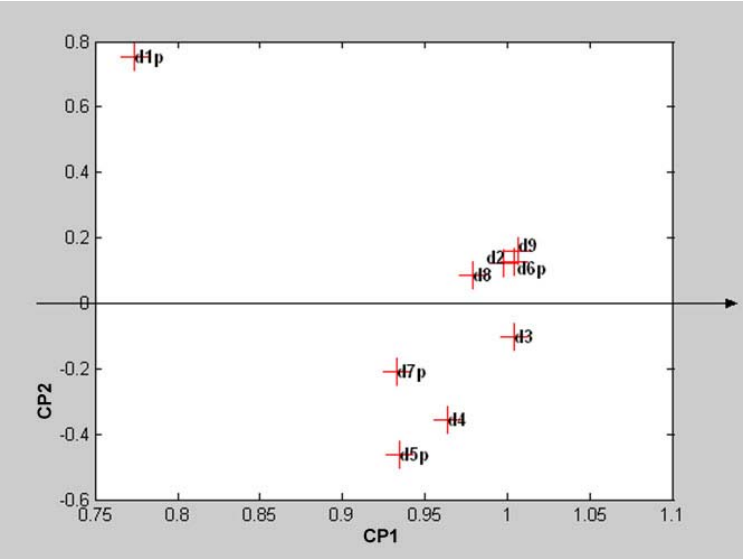

Figura 5.2.13. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças P_L entre os itens da "Escala D. Interesse/Capacidade de Resposta".

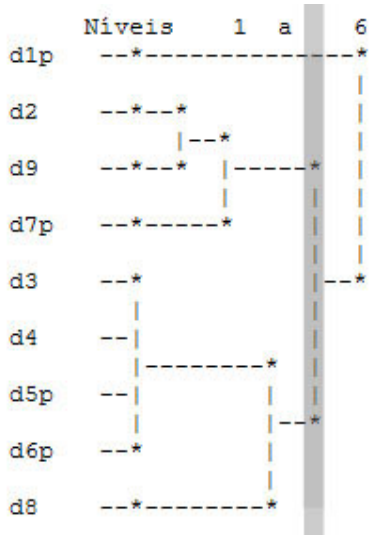

Figura 5.2.14. Dendrograma obtido com a análise classificatória hierárquica ascendente (P_L+Ligação completa). Está assinalado o melhor nível, segundo o critério da "estatística de níveis", STAT(5)=4.11.

5.2.3.4 Discussão dos resultados obtidos com as ACP e ACHA das matrizes de semelhanças entre os itens da Escala D

As análises em componentes principais realizadas permitiram realçar aspectos diferentes das relações entre os itens da "Escala D", que passamos a apresentar na seguinte tabela:

Tabela 5.2.16. Comparação dos resultados obtidos com as ACP das matrizes de semelhanças *S*, S_{LC}, P_L e a análise das ordens entre os itens da "Escala D. Interesse/Capacidade de Resposta", no 1º plano factorial

ACP da matriz S (n=58)	ACP da matriz S_{LC} (n=58)	ACP da matriz P_L (n=58)	Análise das ordens (n=31) excluindo d3 e d4
	"d9"	*"d1p"*	*"d1p", "d2","d6p"*
"d2", "d7p"," d1p", "d8", "d9"	*"d1p", "d2", "d7p" e "d8*	*"d2", "d6p", "d8", "d9", "d3", "d4", "d5p", "d7p"*	*"d7p","d8","d9"*
"d3", "d4","d5p", "d6p"	*"d3", "d4", "d5p" e "d6p"*		*"d5p"*

Os resultados das análises com as semelhanças *s*, s_{LC} e P_L são comparáveis. A diferença entre elas reside essencialmente nos itens "d9" e "d1p".

Na abordagem mais utilizada em Dias *et al*. (2006) foram retirados do estudo todos os cuidadores que tivessem dado respostas "6-Não se aplica" e/ou "9-Não sabe/Não responde", ficando a amostra reduzida a 31 indivíduos. Só tivemos acesso aos resultados obtidos excluindo os itens "d3" e "d4" da análise, por isso é difícil compará-los com os que obtivemos.

No entanto, observamos que nas análises com os coeficientes s, s_{LC} e P_L também é detectado o afastamento dos itens "d1p" e "d5p". No 2º factor das ACP obtidas com os coeficientes s_{LC} e P_L, encontra-se bem visualizada a relação padronizada negativa, já referida acima, que existe entre eles (s_{LC}(d1p,d5p)=-0.64 e P_L(d1p,d5p)=0.26).

Tabela 5.2.17. Comparação dos resultados obtidos com as ACHA das matrizes de semelhanças S, S_{LC}, P_L e R_S entre os itens da "Escala D. Interesse/Capacidade de Resposta"

Algoritmo: (s+ Lig compl) Partição obtida no nível k=5 STAT(5)=4.22 (n=58)	Algoritmo: (s_{LC}+Lig. comp) Partição obtida no nível k=5 STAT(5)=4.11 (n=58)	Algoritmo: (P_L+Lig média) ou (P_L+Lig completa) Partição obtida no nível k=5 STAT(5)=4.11 (n=58)	Algoritmo: (r_S+Lig média) Partição obtida no nível k=4 (n=14)
{d9} *{d2, d7p, d1p, d8}* *{d3, d5p, d4}* *{d6p}*	*{d1p}* *{d2, d9, d7p}* *{d3, d4, d5p, d6p}* *{d8}*	*{d1p}* *{(d2, d9, d7p), (d3, d4, d5p, d6p), (d8)}*	*{d2, d6p, d1p}* *{d5p}* *{d3}* *{d4}*

A hierarquia de partições obtida com o algoritmo (P_L+Ligação completa) como que "cobre" a hierarquia de partições obtida com o algoritmo (s_{LC}+Ligação completa), fazendo sobressair as mesmas partições que esta última: {{d1p}, {{d3, d4, d5p, d6p}, {d2, d9, d7p}, {d8}} (Tabela 5.2.17).

A ACP da matriz de semelhanças P_L retém no 1º plano a maior parte da informação, distorcendo, em parte, o aspecto das proximidades entre os itens no pormenor. No entanto, de forma global, permite obter um resultado semelhante ao obtido com a ACP da matriz de semelhanças S_{LC}, ao aproximar os conjuntos de itens que se visualizaram na projecção da matriz S_{LC} no 1º plano factorial. As duas metodologias complementam-se bem.

5.2.4 Resultados: análise dos itens das escalas "A. Elementos Tangíveis" e "D. Interesse / Capacidade de Resposta"

As semelhanças s, s_{LC} e P_L (Anexo 14) entre as respostas dadas aos itens das escalas "A. Elementos Tangíveis" e "D. Interesse/Capacidade de Resposta", em conjunto, permitem-nos distinguir, por exemplo, a semelhança padronizada muito forte positiva entre as respostas aos itens "d3", "d4", "d5p" e "d6", já observada anteriormente quando se estudaram apenas os itens da "Escala D" (Secção 5.2.3).

5.2.4.1 Análise em componentes principais e ACHA da matriz de semelhanças S entre os itens das escalas A e D

A matriz de semelhanças S entre os itens das escalas "A. Elementos Tangíveis" e "D. Interesse/Capacidade de Resposta" é definida positiva (Anexo 15).
A percentagem de variabilidade total explicada pelo 1º plano factorial é de 75.25%, devendo-se a maior parte dela ao 1º factor (68.93%) (Tabela 5.2.18).

De forma geral, podemos dizer que a ACP realizada com o coeficiente s permite obter resultados análogos aos que foram obtidos com cada uma das escalas separadamente (Subsecções 5.2.2 e 5.2.3), com destaque para a "Escala D".

A 1ª componente principal caracteriza-se principalmente pelo interesse manifestado pela equipa de saúde: "d2", "d7p" e "d1p" (itens da "Escala D", que apresentam maior consenso da parte dos cuidadores em estarem totalmente de acordo e menos variabilidade de respostas) (Tabela 5.2.18; Figura 5.2.15 e Anexo 1).

Tabela 5.2.18. Resultados obtidos com a ACP da matriz de semelhanças *S* entre os itens das Escalas "A.Elementos Tangíveis" e "D.Interesse/Capacidade de Resposta"

	Componente	
	1	2
Unidades de inércia (valores próprios)	*21964*	*2013*
Variabilidade Explicada (%)	*68.93*	*6.32*
Var. Exp. Acumulada (%)	*68.93*	*75.25*
a1.Os equipamentos médicos são adequados	*29.37*	*-4.21*
a2.Existe uma cópia do relatório clínico no seu domicílio...	*34.91*	***-9.90***
a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...	***41.03***	*-3.23*
a4.A equipa de saúde fornece as receitas necessárias...	*39.30*	*-2.34*
a5.A equipa de saúde fornece outra medicação...	*28.19*	*-2.26*
a6.A equipa de saúde proporciona o material clínico necessário	*34.90*	*-3.02*
a7.A equipa de saúde facilita o acesso a outro equipamento de acesso	*23.31*	*-6.42*
d1p. A equipa de saúde não tem demasiada pressa	***44.93***	***-9.78***
d2.A equipa de saúde mostra sincero interesse	***51.45***	*-8.31*
d3. É fácil contactar a equipa de saúde	*36.52*	***19.37***
d4. Quando chamada, a equipa de saúde dá resposta rapidamente	*35.24*	***20.96***
d5p.Quando chamada, a equipa de saúde dá a sensação de não vir contrariada	*37.11*	***19.80***
d6p.Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir...	*27.13*	*17.91*
d7p. A maior parte das vezes parece-lhe que a equipa de saúde não preferia mandar o doente para o hospital...	***49.69***	*-8.08*
d8. A equipa de saúde ensinou-lhe a cuidar do doente	*38.99*	*-5.61*
d9. A equipa de saúde facilita o acesso ao hospital	*27.46*	*-6.08*

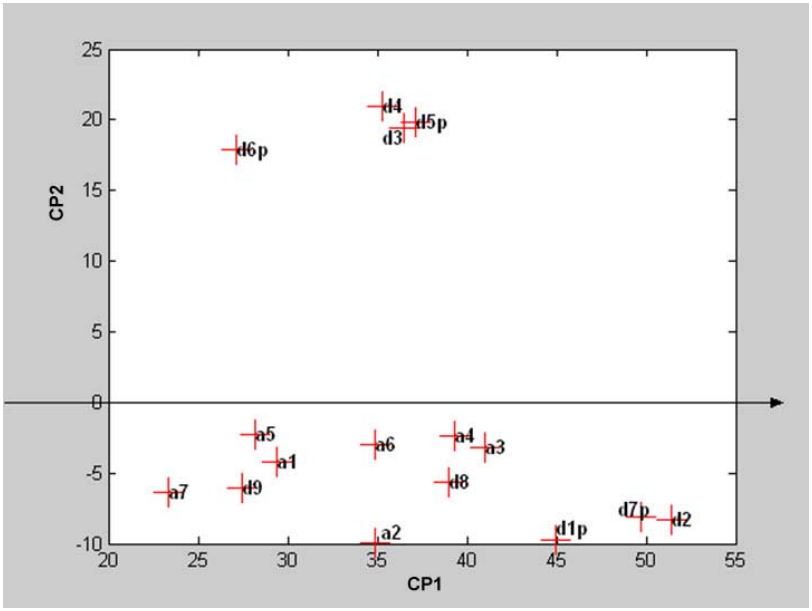

Figura 5.2.15. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças *S* entre os itens das escalas "A. Elementos Tangíveis" (*a1* a *a7*) e "D. Interesse/Capacidade de Resposta" (*d1p* a *d9*).

O 2º factor opõe os itens “d3”, “d4”, “d5p” e “d6p” (itens que se caracterizam pelas respostas “6-Não se aplica”, que rondam os 25%) principalmente aos itens “a2. Existe uma cópia do relatório clínico no seu domicílio...” (cuja moda é estar totalmente em desacordo) e “d1p” (único que não tem respostas 6 e 9).

No 1º plano factorial as escalas A e D aparecem praticamente separadas. Apenas os itens “d8” e “d9”, que apresentam maior variabilidade de respostas se aproximam dos itens da “Escala A”. Neste plano destacam-se pois vários itens com características muito próprias:

- Os itens “d2”, “d7p” e “d1p”.
- Todos os itens da “Escala A”, assim como “d8” e “d9”.
- Os itens “d3”, “d4”, “d5p” e “d6p”.

A hierarquia de partições obtida com o algoritmo (s+ Ligação completa) da ACHA, traduz de certa maneira o que observámos no 1º plano factorial: - a constituição da classe {d3, d4, d5p e d6p} e o “núcleo” {d2, d7p, d1p} ao qual se vão agregando vários itens. No nível 10 observamos a melhor partição (STAT(10)=7.0177): {{a1}, {a2, a3, a4, d1p, d2, d7p, d8, a6}, {d3, d4, d5p, d6p}, {a5}, {d9}, {a7}} (Figura 5.2.16).

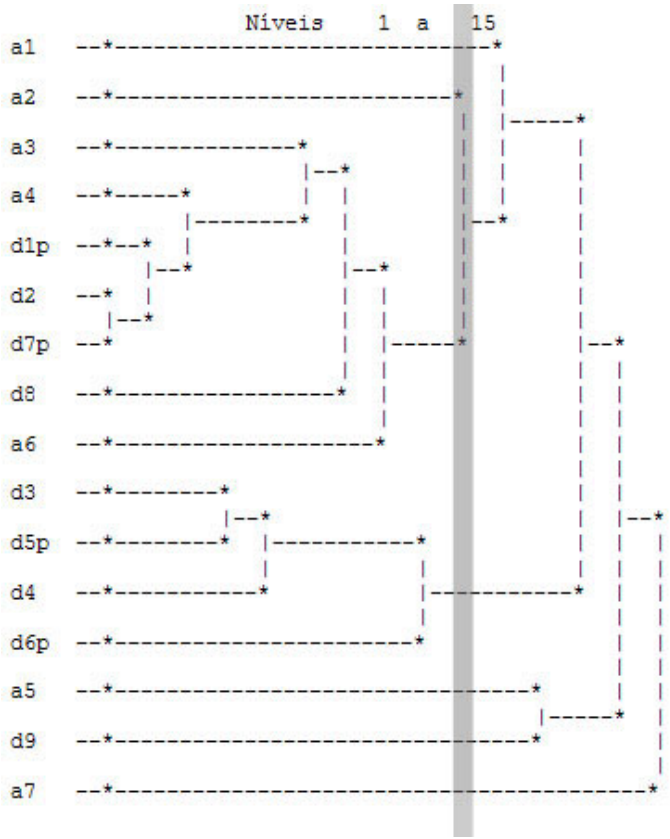

Figura 5.2.16. Dendrograma obtido com a análise classificatória hierárquica ascendente (s+Ligação completa). Está assinalado o melhor nível, segundo o critério da ”estatística de níveis”, STAT(10)=7.02.

5.2.4.2 Análise em componentes principais e ACHA da matriz de semelhanças S_{LC} entre os itens das escalas A e D

A matriz de semelhanças S_{LC} entre os itens das escalas "A. Elementos Tangíveis" e "D. Interesse/Capacidade de Resposta" é definida positiva (Anexo 16).

A percentagem de variabilidade total explicada pelo 1º factor é pequena (29.2%) e atinge apenas os 50.81%, no espaço definido pelos três primeiros factores (Tabela 5.2.19).

Tabela 5.2.19. Resultados obtidos com a ACP da matriz de semelhanças S_{LC} entre os itens das Escalas "A. Elementos Tangíveis" e "D. Interesse/Capacidade de Resposta"

	Componente		
	1	*2*	*3*
Unidades de inércia (valores próprios)	*76.85*	*31.99*	*24.88*
Variabilidade Explicada (%)	*29.20*	*12.15*	*9.45*
Var. Exp. Acumulada (%)	*29.20*	*41.36*	*50.81*
a1.Os equipamentos médicos são adequados	*0.86*	*0.69*	***4.38***
a2.Existe uma cópia do relatório clínico no seu domicílio...	*-0.12*	*0.01*	*-0.09*
a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...	*0.76*	*-0.46*	*-0.34*
a4.A equipa de saúde fornece as receitas necessárias...	*0.93*	*0.26*	*-0.03*
a5.A equipa de saúde fornece outra medicação...	*0.88*	*-0.08*	*-1.00*
a6.A equipa de saúde proporciona o material clínico necessário	*0.60*	*-0.03*	*0.56*
a7.A equipa de saúde facilita o acesso a outro equipamento de acesso	*-0.06*	***5.57***	*-0.61*
d1p. A equipa de saúde não tem demasiada pressa	*0.16*	*0.09*	*0.67*
d2.A equipa de saúde mostra sincero interesse	*0.52*	*-0.20*	*0.07*
d3. É fácil contactar a equipa de saúde	***3.86***	*-0.03*	*-0.17*
d4. Quando chamada, a equipa de saúde dá resposta rapidamente	***4.38***	*-0.12*	*-0.46*
d5p.Quando chamada, a equipa de saúde dá a sensação de não vir contrariada	***3.94***	*-0.14*	*-0.56*
d6p.Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir...	***4.75***	*0.26*	*0.0001*
d7p. A maior parte das vezes parece-lhe que a equipa de saúde não preferia mandar o doente para o hospital...	*0.43*	*-0.16*	*0.62*
d8. A equipa de saúde ensinou-lhe a cuidar do doente	*0.68*	*-0.07*	*0.99*
d9. A equipa de saúde facilita o acesso ao hospital	*0.66*	*0.12*	*1.22*

Este coeficiente, como já foi referido, faz sobressair as situações extremas. No 1º factor observa-se a oposição dos itens "d3", "d4", "d5p" e "d6p" a todos os outros. O 2º factor opõe o item "a7" (que se caracteriza por apresentar todas as possibilidades de resposta com maioria de respostas "6-Não se aplica") a todos os outros (Figura 5.2.17). No 3º factor destaca-se o item "a1", que não tem respostas negativas, opondo-se a todos os outros (Anexo 15).

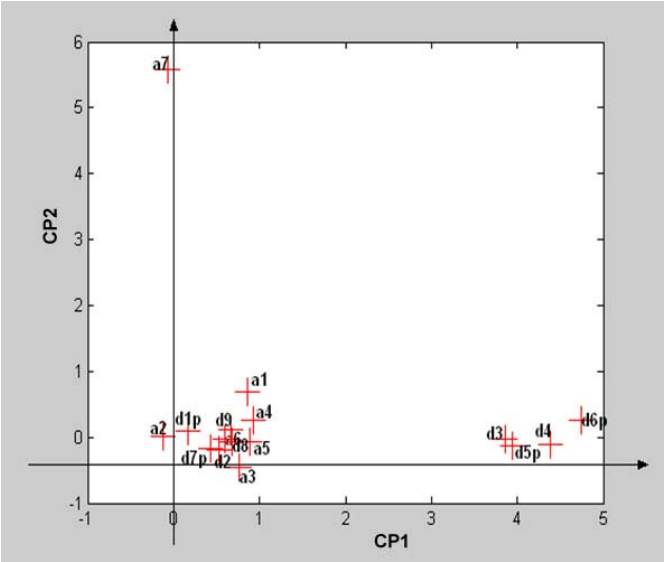

Figura 5.2.17. Representação gráfica do plano factorial (1,2) obtido com a ACP da matriz de semelhanças S_{LC} entre os itens das escalas "A. Elementos Tangíveis" (a1 a a7) e "D. Interesse/Capacidade de Resposta" (d1p a d9).

Todas as hierarquias de partições, obtidas com o coeficiente de semelhança s_{LC} associado a diversos critérios de agregação, evidenciam a semelhança entre os itens "d3", "d4", "d5p" e "d6p", aos quais se vão agregando sucessivamente outros itens, enquanto o item "a7" constitui uma classe singular (Anexo 16). Um exemplo disto é dado pela melhor partição das hierarquias de partições obtidas com os critérios de agregação da família AVL, AVB e AV1, no nível 10 (STAT(10)= 6.45): {{a1, d3, d4, d5p, d6p, a4, a5, a3, d2, d9, d8, a6, d7}, {d1p}, {a2}, {a7}} (Figura 5.2.18).

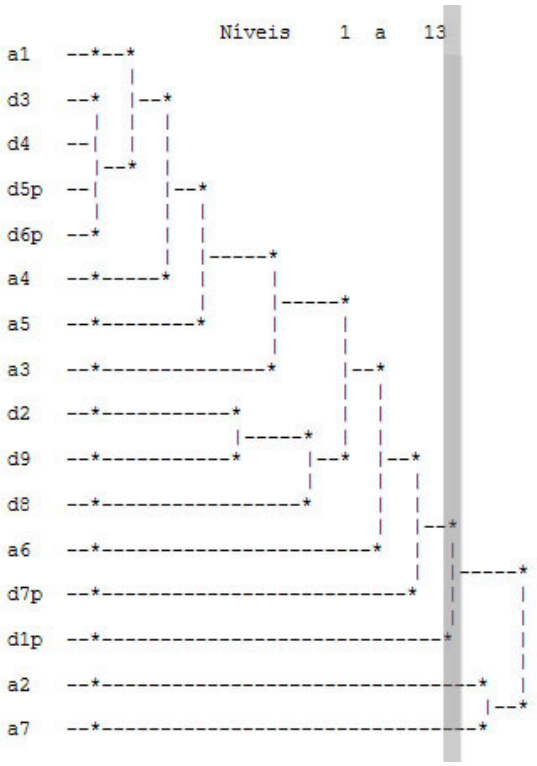

Figura 5.2.18. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+AVL, EPSILON = 1.00, qsi= 1.00). Está assinalado o melhor nível, segundo o critério da "estatística de níveis", STAT(11)=6.45.

5.2.4.3 Análise em componentes principais e ACHA da matriz de semelhanças P_L entre os itens das escalas A e D

A matriz de semelhanças P_L, entre as respostas dadas aos itens das escalas "A. Elementos Tangíveis" e "D. Interesse/Capacidade de Resposta" não é semidefinida positiva (Anexo 17). A 1ª componente principal explica grande parte da variabilidade (77.44%) e o 1º plano factorial explica a maior parte da variabilidade total (90.61%), (Tabela 5.2.20).

Tabela 5.2.20. Resultados obtidos com a ACP da matriz de semelhanças P_L entre os itens das escalas "A. Elementos Tangíveis" e "D. Interesse/Capacidade de Resposta"

	Componente	
	1	*2*
Unidades de inércia (valores próprios)	*12.39*	*2.11*
Variabilidade Explicada (%)	*77.44*	*13.17*
Var. Exp. Acumulada (%)	*77.44*	*90.61*
a1.Os equipamentos médicos são adequados	*0.90*	*-0.48*
a2.Existe uma cópia do relatório clínico no seu domicílio...	*0.54*	*0.45*
a3.Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...	*0.86*	***0.69***
a4.A equipa de saúde fornece as receitas necessárias...	*0.93*	*-0.15*
a5.A equipa de saúde fornece outra medicação...	*0.86*	*0.44*
a6.A equipa de saúde proporciona o material clínico necessário	*0.80*	*-0.29*
a7.A equipa de saúde facilita o acesso a outro equipamento de acesso	*0.53*	*-0.42*
d1p. A equipa de saúde não tem demasiada pressa	*0.71*	*-0.64*
d2.A equipa de saúde mostra sincero interesse	*0.87*	*0.26*
d3. É fácil contactar a equipa de saúde	*1.04*	*0.04*
d4. Quando chamada, a equipa de saúde dá resposta rapidamente	*0.9992*	*0.14*
d5p.Quando chamada, a equipa de saúde dá a sensação de não vir contrariada	*0.9905*	*0.27*
d6p.Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir...	*1.05*	*-0.18*
d7p. A maior parte das vezes parece-lhe que a equipa de saúde não preferia mandar o doente para o hospital...	*0.84*	*0.17*
d8. A equipa de saúde ensinou-lhe a cuidar do doente	*0.94*	*-0.24*
d9. A equipa de saúde facilita o acesso ao hospital	*1.004*	*-0.14*

Os itens "d6p", "d9", "d3", "d4", "d5p" da "Escala D" foram os que mais contribuíram para a formação do 1º factor. Enquanto que o 2º factor opõe os itens "a3" e "d1p" (Tabela 5.2.20, Figura 5.2.19).

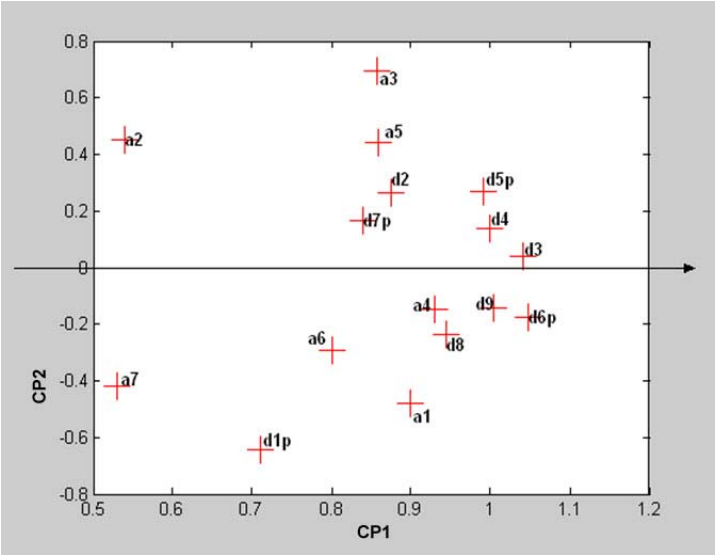

Figura 5.2.19. Representação gráfica plano factorial (1,2) obtido com a ACP da matriz de semelhanças P_L entre os itens das escalas "A. Elementos Tangíveis" (a1 a a7) e "D. Interesse/Capacidade de Resposta" (d1 a d9).

A ACP baseada no coeficiente P_L faz sobressair no 1º plano factorial:

- A visualização da singularidade dos itens "a2", "a7" e, eventualmente, "d1p", que se opõem aos restantes itens das duas escalas. A quase separação entre as escalas A e D, com excepção dos itens "d1p", "d2" e "d7p" da "Escala D", que estão mais próximos dos da "Escala A".
- A proximidade entre os itens da "Escala D", "d3", "d4", "d5p", "d6p", "d9" e o item "a4" da "Escala A", bem patente no 1º eixo factorial, já apercebida a partir dos valores das semelhanças P_L, próximos de 1, entre aqueles itens (Anexo 14 e Anexo 17).

Os algoritmos de análise classificatória hierárquica ascendente utilizados permitiram observar dois tipos de estruturas (Anexo 17) nestes itens:

- As melhores partições, segundo o critério da "estatística dos níveis", das hierarquias obtidas com os algoritmos (P_L+ Ligação única) e (P_L+ AVM[195]), traduzem bem as relações de semelhança visualizadas no 1º plano factorial (Figura 5.2.19), tal como podemos observar, por exemplo, na partição obtida, no nível 10 (STAT(10)=6.45), pelo primeiro algoritmo referido: {{a1, d3, d4, d5p, d6p, a4, a5, a3, d9, d2, d8, a6, d7p}, {d1p}, {a2}, {a7}} (Figura 5.2.20). No entanto, esta partição não mostra a quase separação das duas escalas.

[195] Critério de agregação da média aritmética com transformação e com f.d.d.

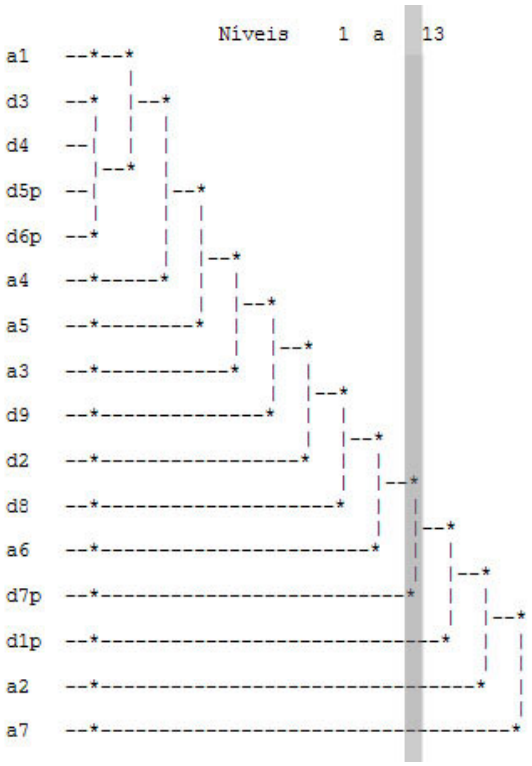

Figura 5.2.20. Dendrograma obtido com a análise classificatória hierárquica ascendente (P_L+Ligação única). Está assinalado o melhor nível, segundo o critério da "estatística de níveis", STAT(10)=6.45**.**

- A quase separação das escalas é mais visível na partição obtida no nível 9 (STAT(9)=6.87) pelo algoritmo (P_L+ Ligação pela média):

{{a1, d8, a6}, {a3, a5, a4, d3, d4, d5p, d6p}, {d2, d9, d7p}, {d1p}, {a2}, {a7}} (Figura 5.2.21).

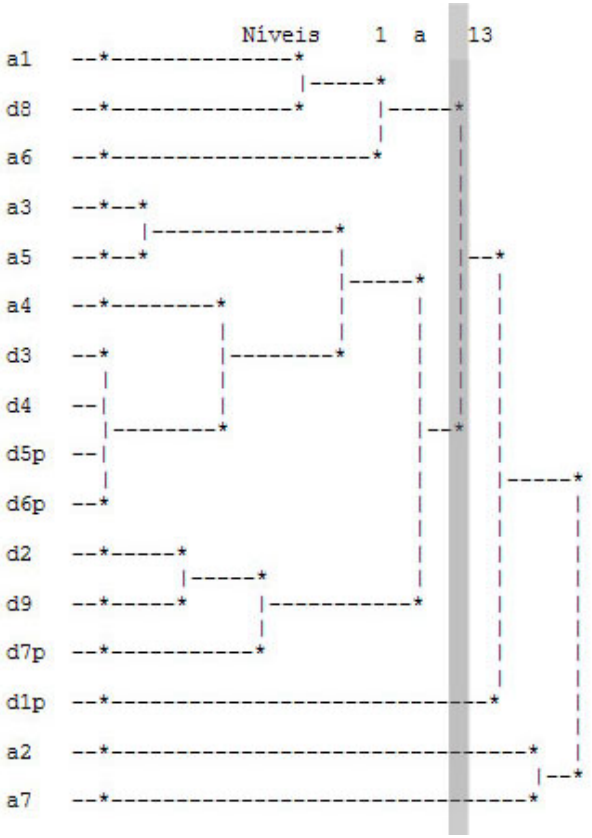

Figura 5.2.21. Dendrograma obtido com a análise classificatória hierárquica ascendente (P_L+Ligação pela média). Está assinalado o melhor nível, segundo o critério da "estatística de níveis", STAT(10)=6.87.

5.2.4.4 Discussão dos resultados obtidos com as ACP e ACHA das matrizes de semelhanças entre os itens das escalas A e D

Quando são tratados os itens das duas escalas A e D, em conjunto:

- As escalas A e D aparecem praticamente separadas, quando se estuda a relação entre os seus itens utilizando os coeficientes *s* e P_L e a ACP das respectivas matrizes de semelhanças.
- Apenas alguns algoritmos de ACHA em que foram utilizados os coeficientes s_{LC} e P_L puseram em evidência a quase separação das escalas: (s_{LC}+Ligação completa), (s_{LC}+Ligação pela média), (P_L+Ligação completa), (P_L+Ligação pela média), (P_L+ AVL), (P_L+ AVB), (P_L+ AV1).
- Na ACP baseada no coeficiente s_{LC} sobressaem itens que também se destacaram nas análises de cada escala, quando tratadas separadamente. É o caso, por exemplo, do item "a7. A equipa de saúde facilita o acesso a outro equipamento de acesso" e dos itens "d3. É fácil contactar a equipa de saúde ", "d4. Quando chamada,

a equipa de saúde dá resposta rapidamente ",“d5p. Quando chamada, a equipa de saúde dá a sensação de não vir contrariada“ e “d6p. Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir... “, que nas respectivas escalas apresentam as percentagens mais elevadas de respostas na modalidade “6-Não se aplica”. Estes resultados são comparáveis aos obtidos com os coeficientes s e P_L.

- Alguns algoritmos de ACHA conduziram a resultados que complementaram bem a visualização obtida com as ACP das matrizes de semelhanças S, S_{LC} e P_L.
- Os algoritmos de ACHA que se baseiam nos mesmos critérios de agregação clássicos (“ligação única”, “ligação completa” e “ligação pela média”) em conjugação com os coeficientes de semelhança s_{LC} e P_L conduzem a hierarquias de classificação análogas, em que classes das partições obtidas com o coeficiente de semelhança P_L contêm, frequentemente, classes das que foram obtidas com o coeficiente s_{LC}.

Não encontrámos esta análise em Dias *et al*. (2006).

5.2.5 Conclusões preliminares

Em todas as análises desta secção, as matrizes de semelhanças S e S_{LC} entre variáveis de ordem parcial são definidas positivas (d.p.), enquanto as matrizes de semelhanças P_L não o são. Os resultados obtidos com as análises em componentes principais baseadas nos coeficientes s, s_{LC} e P_L são comparáveis. No entanto, há aspectos particulares de cada um deles que convém referir. A matriz S é sensível ao efeito de dimensão e não entra em conta com a estrutura dos dados. O coeficiente s_{LC} faz sobressair as situações extremas, quando projectada a respectiva matriz de semelhanças. Geralmente, nas ACP das matrizes S_{LC}, a cada factor associa-se predominantemente um item ou um pequeno conjunto de itens das escalas estudadas. Os resultados obtidos com as ACP baseadas no coeficiente s_{LC} permitiram destacar/caracterizar alguns itens pelos tipos de respostas que lhes foram dadas, quando as escalas A e D foram tratadas, quer separadamente, quer em conjunto (caso dos itens “d3”, “d4”, “d5p” e “d6p” da Escala D e do item “a7” da “Escala A”). Enquanto que, as ACP das matrizes *S* e P_L permitem fazer uma análise de conjunto, mais geral. Observa-se ainda uma maior “contracção” do espaço obtido com as ACP das matrizes P_L.

Os algoritmos de ACHA compostos pelo mesmo critério de agregação “clássico” (“ligação única”, “ligação completa”, “ligação pela média”) associado aos coeficientes de semelhança s_{LC} e P_L conduzem a hierarquias de classificação análogas, em que as melhores partições obtidas com o coeficiente P_L são constituídas pelas mesmas classes ou por menos classes,

que resultam da agregação de classes já existentes nas melhores partições obtidas com o coeficiente de semelhança s_{LC}.

Não podemos dizer que as escalas A e D estão completamente separadas. Quando analisadas em conjunto verifica-se que as relações fortes (positivas ou negativas) já existentes entre alguns itens de cada uma das escalas se mantêm; mas, também são detectadas novas relações fortes entre alguns itens das duas escalas. Destacamos, por ter um perfil de respostas peculiar, o resultado obtido com o algoritmo de ACHA, (P_L+Ligação pela média) no nível 9, (STAT(9)= 6.87):
{a1, d8, a6}, {a3, a5, a4, d3, d4, d5p, d6p}, {d2, d9, d7p}, {d1p}, {a2}, {a7} (Figura 5.2.21).

- "d1p. A equipa de saúde não tem demasiada pressa ..." com maioria de respostas totalmente de acordo (43 – 74.1%) e único item, nas duas escalas, que não apresenta respostas "6-Não se aplica" e "9-Não sabe/Não responde".
- O item "a2. Existe uma cópia do relatório clínico no seu domicílio..." apresenta predominância da resposta "1-Totalmente em desacordo" (23 - 39.7%).
- No item "a7. A equipa de saúde facilita o acesso a outro equipamento de acesso ..." são contempladas todas as possibilidades de resposta com predominância de respostas "6-Não se aplica" (32 – 55.2%). Não é de estranhar a ocorrência de grande frequência deste tipo de resposta neste item, dado que trata de acesso a outros equipamentos que muitos doentes não necessitam.
- Os itens "d2. A equipa de saúde mostra sincero interesse ..." e "d7p. A maior parte das vezes parece-lhe que a equipa de saúde não preferia mandar o doente para o hospital..." são itens da "Escala D" que apresentam maior consenso da parte dos cuidadores em estarem totalmente de acordo, e menos variabilidade de respostas, pois têm respostas totalmente de acordo que rondam os 90%, que se distribuem pelas várias possibilidades de resposta de "d9. A equipa de saúde facilita o acesso ao hospital" (no qual o "1-Totalmente em desacordo" não tem respostas) concentrando cerca de 48% de respostas no totalmente de acordo.
- Os itens de respostas ligadas à acessibilidade/contacto com a equipa de saúde,- "d3. É fácil contactar a equipa de saúde ...", "d4. Quando chamada, a equipa de saúde dá resposta rapidamente ...", "d5p. Quando chamada, a equipa de saúde dá a sensação de não vir contrariada ...", "d6p. Quando contacta a equipa de saúde ela não tenta resolver o problema para não ter que ir..." – , com maioria de respostas totalmente de acordo, que se caracterizam pelas respostas "6-Não se aplica", que rondam os 25%. A estes itens vão-se agregando sucessivamente os itens da "Escala A" relativos à informação clínica/terapêutica e receitas, "a4. A equipa de saúde fornece as receitas

necessárias...", "a5. A equipa de saúde fornece outra medicação..." e "a3. Existe uma cópia da guia de tratamento no seu domicílio para que saiba o tratamento...". São itens que se caracterizam por estarem contempladas todas as possibilidades de resposta ("a4" não tem respostas totalmente em desacordo) com predominância da(s) resposta(s) positivas "5-Totalmente de acordo".

- Os itens "a1. Os equipamentos médicos são adequados ..." e "a6. A equipa de saúde proporciona o material clínico necessário ..." sem respostas "negativas" que apesar do consenso têm, respectivamente, 26% e 15,5% de respostas "6-Não se aplica" e "d8. A equipa de saúde ensinou-lhe a cuidar do doente". Cerca de 36% dos cuidadores estão totalmente de acordo com estes três itens.

Na análise realizada por Dias *et al*. (2006) foram retiradas as respostas "6-Não se aplica" e "9-Não sabe/Não responde", o que levou à redução da dimensão da amostra utilizada na análise, perdendo-se assim informação importante. Este facto justifica também a obtenção de resultados diferentes, quando se recorre à análise proposta por Dias *et al*. (2006).
Finalmente podemos dizer que se atingiu o objectivo proposto. Relacionaram-se variáveis de ordem parcial que não são, habitualmente, contempladas na literatura. Os resultados obtidos permitem-nos ter uma visão global e sintetizada, muito satisfatória, das respostas dadas aos itens das escalas A e D.

5.3 Variáveis heterogéneas: O Questionário SERVQUAL Modificado

5.3.1 Introdução

O Questionário SERVQUAL Modificado[196] sobre qualidade e satisfação dos doentes com apoio domiciliário é constituído por dois blocos (Bloco1, Bloco 2) e por questões sobre "Dados do Doente" e sobre "Dados do Cuidador". Tal como já foi referido anteriormente, o Bloco 1 consiste numa "Escala de Percepções" ("A. Elementos Tangíveis", "B. Fiabilidade dos Tratamentos e Cuidados", "C. Segurança/Garantia", "D. Interesse/Capacidade de Resposta" e "E. Empatia") e o Bloco 2 é uma "Escala de Preferências". Enquanto que as respostas dadas pelos cuidadores ao Bloco 1 já foram estudadas (Bacelar Nicolau *et al.*, 2005; Dias *et al*., 2006; Doria *et al*., 2006, Exemplo 4.3.28 e Secção 5.2), as restantes

[196] Projecto SDH. MD/P.I.01.13, subsidiado pela Fundação Calouste Gulbenkian e coordenado pelo Professor Doutor Manuel Silvério Marques (IPOGM).

respostas ainda não o foram. Nesta secção, explorámos de forma univariada todas as questões do Bloco 2, dos "Dados do Doente" e dos "Dados do Cuidador" do questionário, assim como, de forma bivariada, algumas delas (Anexo 18).

Aqui pretendemos explorar de forma multivariada as possíveis relações entre as cinco escalas de percepções do Bloco 1 (descritas através dos perfis de distribuição das frequências relativas no Exemplo 4.3.28), a "Escala de Preferências" (Bloco 2), alguns "Dados do doente" e alguns "Dados do cuidador", que foram considerados mais pertinentes para o estudo[197] sob o ponto de vista clínico (Tabela 5.3.1). Como se pode constatar, as variáveis deste estudo são heterogéneas. O objectivo desta secção é pois, o de tentar encontrar um perfil que permita caracterizar de forma mais sintética/resumida as respostas dos 57[198] cuidadores a estas 31 questões de natureza diversa.

Tabela 5.3.1. Descrição das variáveis, seus códigos de representação e respectiva escala de medição

Variáveis	Escala de medição
A. Elementos Tangíveis (**A.ETg**) *Distribuição das frequências relativas*	Simbólica modal
B. Fiabilidade (**B.Fia**) *Distribuição das frequências relativas*	Simbólica modal
C. Segurança/Garantia (**C.Seg**) *Distribuição das frequências relativas*	Simbólica modal
D. Capacidade de Resposta (**D.CapR**) *Distribuição das frequências relativas*	Simbólica modal
E. Empatia (**E.Emp**) *Distribuição das frequências relativas*	Simbólica modal
p1. Se pudesse escolher que preferia ...? (**Prefer**) *Que o doente estivesse: 1. no hospital, 2. em casa, 3. outra*	Nominal
p2. No apoio ao doente no domicílio, indique dois aspectos mais importantes ...(**Apoio**) *(p_{21}, p_{22}, p_{23}, p_{24}, p_{25})*	Simbólica categórica com valores múltiplos
d3. Nível de instrução do doente (**InstD**) *0.Não sabe ler nem escrever, 1.Não frequentou estudos mas sabe ler e escrever, 2.Ensino primário, 3.Preparatório/ Secundário, 4.Bacharelato/ Licenciatura, 5.Mestrado/ Doutoramento, 6.Outro*	Ordem total, $\leq$ (porque ninguém responde a "Outro")

[197] Agradeço ao Prof. Doutor Manuel Silvério Marques e à Profª. Doutora Ana Sousa Ferreira as conversas que tivemos sobre este assunto.

[198] O indivíduo 23 foi retirado do estudo por não responder a várias questões.

Variáveis (cont.)	Escala de medição
d.4. O doente conhece/conhecia a doença que tem/tinha? (**Conhecia**) *0.Não, 1.Sim,parcialmente, 2.Sim, totalmente, 9.Não sabe*	Ordem parcial 0<1<2, 9
c.2. Sexo do cuidador (**Sexo**) *1.Masculino, 2.Feminino*	Nominal
c.4. Nível de instrução do cuidador (**InstC**) *0.Não sabe ler nem escrever, 1.Não frequentou estudos mas sabe ler e escrever, 2.Ensino primário, 3.Preparatório/ Secundário, 4.Bacharelato/ Licenciatura, 5.Mestrado/ Doutoramento, 6.Outro*	Ordem total, ≤ (porque ninguém responde "Outro")
c.5.Situação profissional do cuidador (**Prof**) *1.Trabalha, 2.Desempregado, 3.Reformado, 4.Dona de casa, 5.Estudante, 6.Incapacitado ou invalidez permanente, 7.Outro*	Nominal com 7 categorias
c.11. Como se sente/sentiu durante o tempo que está/esteve a cuidar do doente? *1.Nunca, 2.Quase nunca, 3.Às vezes, 4.Quase sempre, 5.Sempre*	
c.11.1. Capaz de cuidar do doente (**+Capaz**)	Ordem total, ≤
c.11.2. Nervoso/a (**nerv**)	Ordem total, ≤
c.11.3. Com a moral tão em baixo que nada o/a animava (**mor-bx**)	Ordem total, ≤
c.11.4. Calmo/a e tranquilo/a **(+Calm**)	Ordem total, ≤
c.11.5. Desanimado/a e triste (**trist**)	Ordem total, ≤
c.11.6. A fazer coisas acima das suas possibilidades (**a-possib**)	Ordem total, ≤
c.11.7. Esgotado/a após os cuidados prestados ao doente (**esg**)	Ordem total, ≤
c.11.8. Bem porque estava a fazer tudo o que podia **(+Bem**)	Ordem total, ≤
c.11.9. Valeu/vale a pena o esforço que fez/faz para cuidar do doente (**+Vale**)	Ordem total, ≤
c.12. Tem a quem recorrer quando necessita de companhia ou apoio de alguém? (**Ajuda**) *0.Não, 1.Sim, algumas vezes, 2.Sim,sempre, 9.Não responde*	Ordem parcial 0<1<2, 9
c.19. Nos últimos seis meses, como evoluíram os seguintes sintomas do doente?: *1. Piorou, 2.Permaneceu igual, 3.Melhorou, 4.Nunca teve* Acrescenta-se a opção 9 quando não há resposta.	
c.19.1. Dores (**dores**)	Ordem parcial 1<2<3 , 4
c.19.2. Vómitos (**vómitos**)	Ordem parcial 1<2<3 , 4
c.19.3. Falta de ar (**f-ar**)	Ordem parcial 1<2<3 , 4
c.19.4. Insónias (**insónias**)	Ordem parcial 1<2<3, 4, 9
c.19.5. Prisão de ventre (**p-ventre**)	Ordem parcial 1<2<3 , 4
c.19.6. Alterações na boca (**alt-boca**)	Ordem parcial 1<2<3 , 4
c.19.7. Falta de apetite (**f-apet**)	Ordem parcial 1<2<3 , 4
c.19.8. Ansiedade/depressão (**ansiedade**)	Ordem parcial 1<2<3, 4, 9
c.21. Com que periodicidade vão/iam visitá-lo habitualmente: (**Period**) *1. 1 vez por mês, 2. Cada 15 dias, 3. 1 vez por semana, 4. Mais de uma vez por semana*	Ordem total, ≤

Para isso, propomos a aplicação da análise em componentes principais (ACP) e da análise classificatória hierárquica ascendente (ACHA) directamente sobre as matrizes de semelhanças S_{LC} e P_L (Subsecção 5.3.2). Neste caso, o estudo da matriz de semelhanças S

não tem interesse, pois as variáveis são heterogéneas. No final discutiremos os resultados obtidos (Subsecção 5.3.3) e tiraremos conclusões (Subsecção 5.3.4).

A análise exploratória univariada permitiu observar que os doentes desta amostra são, ligeiramente, em maior número do sexo masculino (52.6%), enquanto que os cuidadores são maioritariamente de sexo feminino (77.2%) e, de forma geral, mais novos (57.3±14.7 anos) do que os doentes (68.5±14.6 anos), sendo de assinalar a existência de um doente com 7 anos, idade considerada *outlier* nesta distribuição. Também permitiu verificar que existe variabilidade nas respostas dadas pelos cuidadores e identificar alguns dados omissos. Às variáveis com dados omissos foi acrescentada a categoria "9 - Não responde" (Tabela 5.3.1).

5.3.2 Comparação de variáveis heterogéneas: Resultados

As semelhanças s_{LC} e P_L (Anexo 19) entre as trinta e uma questões permitem distinguir, por exemplo, as semelhanças padronizadas positivas entre variáveis que se referem à evolução de sintomas dolorosos – "prisão de ventre", "ansiedade/depressão", "falta de apetite", "dores" – e exemplificadas para algumas delas na Tabela 5.3.2 e na Tabela 5.3.3. Nestes sintomas observam-se as maiores percentagens da situação "piorou" (Anexo 18).

Tabela 5.3.2. Valores das semelhanças s_{LC} (P_L) entre a evolução de alguns sintomas dolorosos

	Prisão de ventre	*Dores*	*Falta de apetite*
Ansiedade/depressão	*4.29 (≈ 1.000)*	*3.00 (0.9987)*	*2.63 (0.9958)*

Tabela 5.3.3. Tabela de contingência das variáveis "Evolução dos sintomas - Prisão de Ventre" e "Evolução dos sintomas - Ansiedade/Depressão" – (s_{LC}(p-ventre, ansiedade)=4.29 e P_L(p-ventre, ansiedade) ≈ 1.00)

		Evolução dos sintomas - Ansiedade/Depressão					Total
		Piorou	Permaneceu igual	Melhorou	Nunca teve	Não responde	
Evolução dos sintomas - Prisão de Ventre	Piorou	**19**	*7*	*4*	0	0	30
	Permaneceu igual	1	**5**	*3*	*2*	0	11
	Melhorou	0	1	**2**	0	1	4
	Nunca teve	2	**8**	0	**2**	0	12
Total		22	21	9	4	1	57

O valor negativo de s_{LC}(esg, Ajuda)= -2.71 (P_L(esg, Ajuda)=0.0033), traduz bem a relação entre o estado do cuidador "esgotado" e o facto de "ter a quem recorrer quando necessita da companhia de alguém" (Tabela 5.3.4).

Tabela 5.3.4. Tabela de contingência das variáveis "Estado físico e psíquico do cuidador – esgotado" e "Tem a quem recorrer quando necessita da companhia de alguém? " (s_{LC}(esg, Ajuda)= -2.71 e P_L(esg,Ajuda)=0.0033)

		Tem a quem recorrer quando necessita da companhia de alguém?				
		Não	Sim, algumas vezes	Sim, sempre	Não responde	Total
Estado físico e psíquico do cuidador - esgotado	Nunca	1	2	**6**	1	10
	Quase nunca	1	0	**3**	0	4
	As vezes	0	5	**13**	0	18
	Quase sempre	**4**	**10**	3	0	17
	Sempre	0	**7**	1	0	8
Total		6	24	26	1	57

A matriz de semelhanças S_{LC} é definida positiva (Anexo 20). No entanto, os resultados obtidos com a ACP desta matriz são pobres. A percentagem de variabilidade explicada por cada componente principal é muito pequena, decrescendo muito gradualmente, e o 1º plano factorial só consegue explicar 17.14% da variabilidade total (Tabela 5.3.5). A 1ª componente associa-se principalmente à evolução dos sintomas dos doentes no que se refere a "vómitos", "falta de ar" e "alterações na boca" (são os que apresentam maior percentagem de doentes que "nunca tiveram" – respectivamente, 43.9%, 50.9% e 54.4% –, mas, com grande percentagem de doentes que pioraram) opondo-os ligeiramente a "aspectos que considera importantes no apoio domiciliário".

É de notar, que nesta análise, as variáveis que acabámos de referir – as evoluções dos sintomas "vómitos", "falta de ar" e "alterações na boca", e "aspectos que considera importantes no apoio domiciliário" – são as que se destacam sempre na formação da primeira, segunda, terceira e quarta componentes principais (Anexo 20).

Tabela 5.3.5. Resultados obtidos com a ACP da matriz de semelhanças S_{LC}

	Componente	
	1	*2*
Unidades de inércia (valores próprios)	*47.00*	*41.25*
Variabilidade Explicada (%)	*9.13*	*8.01*
Var. Exp. Acumulada (%)	*9.13*	*17.14*

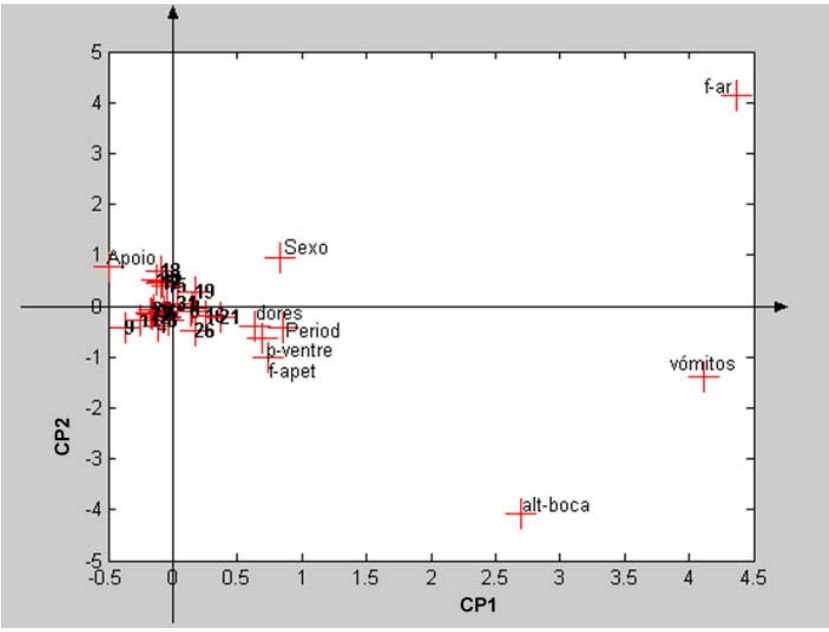

Figura 5.3.1. Representação gráfica do 1º plano obtido com a ACP da matriz de semelhanças S_{LC}.

A matriz de semelhanças P_L não é s.d.p. (Anexo 21). O 1º plano factorial explica 70.57% da variabilidade total (Tabela 5.3.6).

Tabela 5.3.6. Resultados obtidos com a ACP da matriz de semelhanças P_L

	Componente	
	1	*2*
Unidades de inércia (valores próprios)	*16.76*	*5.12*
Variabilidade Explicada (%)	*54.06*	*16.51*
Var. Exp. Acumulada (%)	*54.06*	*70.57*

Todas as variáveis têm coordenadas positivas na 1ª componente principal (Figura 5.3.2). Grande número de variáveis foi importante para a formação desta componente. A 1ª componente principal está associada, principalmente, à evolução dos sintomas dolorosos dos doentes – "dores" (52.6% pioraram), "vómitos" (43.9% nunca tiveram), "ansiedade/depressão" (38.6% pioraram) – ao modo como os cuidadores percepcionam a "Capacidade de Resposta da equipa de saúde" (Escala D), à "Fiabilidade dos cuidados e tratamentos" (Escala B), a "Empatia" (Escala E), os "Elementos Tangíveis" (Escala A) e aos sentimentos positivos dos cuidadores face aos doentes – na sua maioria sentem que são "capazes de cuidar do doente" ('Sempre' 56.1%; 'Quase sempre' 35.1%) e que "vale a pena o esforço" ('Sempre' 78.9%; 'Quase sempre' 10.5%).

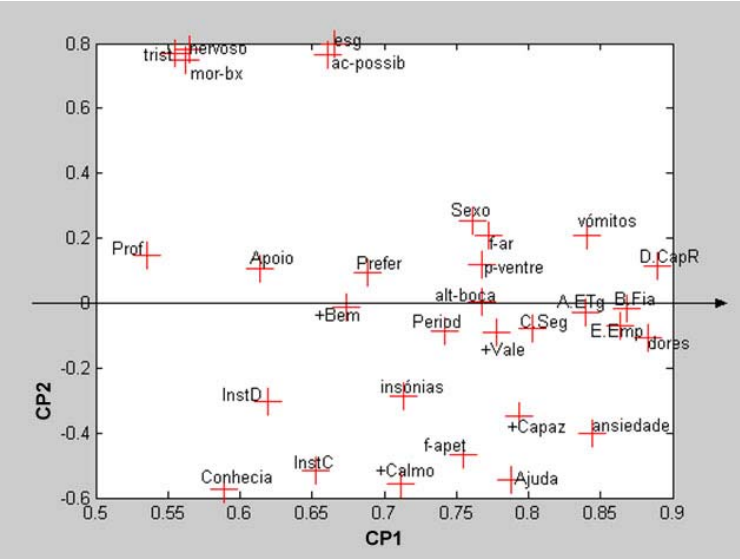

Figura 5.3.2. Representação gráfica do 1º plano obtido com a ACP da matriz de semelhanças P_L.

A 2ª componente principal refere-se principalmente aos aspectos negativos e positivos dos sentimentos dos cuidadores face aos doentes.

A utilização da ACHA directamente sobre as matrizes de semelhanças S_{LC} e P_L vai complementar a informação que obtivemos e permitir encontrar os grupos de variáveis que nos dão o perfil de respostas dos cuidadores.

De entre os vários algoritmos de ACHA utilizados destacamos as hierarquias de partições obtidas com os algoritmos (s_{LC}+AVM) e (P_L+AVM) (Anexo 20, Anexo 21 e Figura 5.3.3). A hierarquia de partições obtida com o coeficiente de semelhança P_L, embora apresente um efeito de cadeia assinalável, como que "contém" a hierarquia de partições obtida com o coeficiente s_{LC}. As duas hierarquias traduzem, de certa maneira, o que observamos na representação gráfica do 1º plano factorial da ACP da matriz de semelhanças P_L.

Considerando a melhor partição do algoritmo (s_{LC}+AVM) obtida no nível 21 (STAT(21)=10.89), observamos que é constituída pelas seguintes 7 classes:

- Classe 1={A.Etg, B.Fia, E.Emp, C.Seg, D.CapR, Prefer, Sexo, f-ar, vómitos, +Vale, dores, p-ventre, ansiedade, f-apet, insónia, alt-boca},
- Classe 2={Period},
- Classe 3={InstD, InstC, Ajuda, Conhecia, +Capaz, +Calm},
- Classe 4={+Bem},
- Classe 5={Prof},
- Classe 6={Apoio},
- Classe 7={nerv, mor-bx, trist, esg, a-possib}.

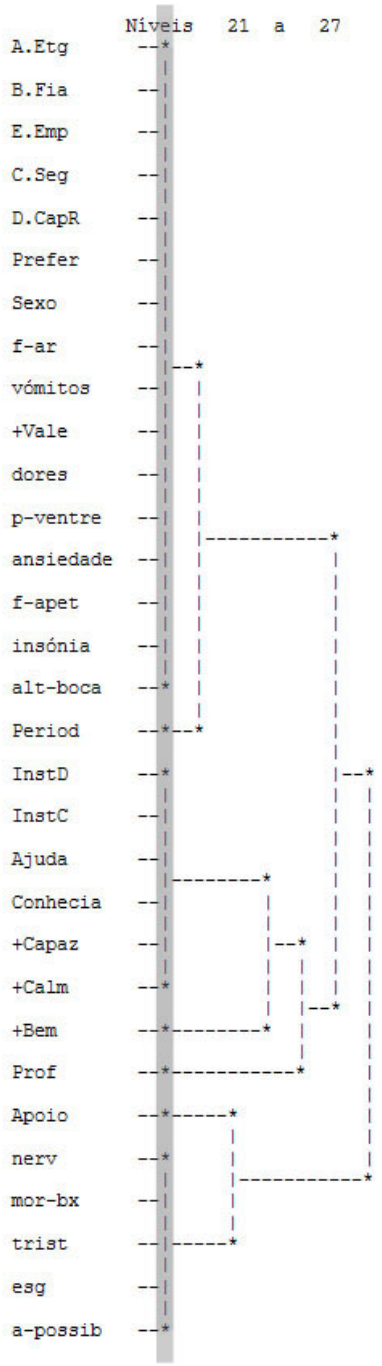

Figura 5.3.3. Dendrograma obtido com a análise classificatória hierárquica ascendente (s_{LC}+AVM) . Está assinalado o melhor nível, segundo o critério da "estatística de níveis", STAT(21)=10.89 e são apresentados apenas os últimos níveis (21 a 27).

5.3.3 Discussão e interpretação dos resultados

Neste estudo com variáveis heterogéneas, a ACP da matriz de semelhanças P_L permitiu obter mais informação do que a ACP da matriz de semelhanças S_{LC}, embora esta seja definida positiva e a matriz P_L não o seja.

A matriz de semelhanças S_{LC} faz sobressair em cada componente principal uma variável ou um número reduzido delas, enquanto as restantes se concentram no "centro de gravidade" da distribuição ou próximo dele. A percentagem de variabilidade explicada pelas componentes principais é muito baixa.

A informação obtida com a ACP da matriz de semelhanças P_L é bem complementada pela ACHA desta matriz de semelhanças e da matriz de semelhanças S_{LC}, quando associadas ao critério de agregação AVM. As hierarquias de partições obtidas desenrolam-se aos nossos olhos como um bom complemento do que vemos na representação do 1º plano factorial da ACP da matriz de semelhanças P_L, (Anexo 21 e Figura 5.3.2). A hierarquia de partições (P_L+AVM) como que "cobre" a hierarquia de partições obtidas pelo algoritmo (s_{LC}+AVM). Interpretando a melhor partição obtida, segundo o "critério da estatística de níveis", no nível 21 (STAT(21)=10.89) e já apresentada acima, obtemos um perfil destes cuidadores e doentes relativo às variáveis escolhidas:

- Classe 1={A.Etg, B.Fia, E.Emp, C.Seg, D.CapR, Prefer, Sexo, f-ar, vómitos, +Vale, dores, p-ventre, ansiedade, f-apet, insónia, alt-boca}

A percepção que os cuidadores têm sobre a qualidade dos cuidados paliativos prestados pelo apoio domiciliário traduz, de forma geral, uma satisfação que é elevada, com algum destaque para as escalas "C. Segurança/Garantia" e "E. Empatia", mas apresentando algumas respostas "Não se aplica" e "Não sabe/Não responde" relevantes nas escalas "A. Elementos Tangíveis" e "D. Capacidade de Resposta". Estes cuidadores apresentam uma preferência elevada (44 – 77.2%) por ter o doente em casa, referindo sobre a evolução dos sintomas que estes pioraram na maior parte dos doentes que têm dores (60.0%), dos que têm prisão de ventre (66.7%), dos que têm alterações da boca (65.4%), dos que têm falta de apetite (61.4%), pioraram um pouco menos nos doentes que têm insónias (45.4%), nos que têm falta de ar (42.9%), nos que têm ansiedade/depressão (42.3%) e são em menor número nos que têm vómitos (31.3%).

A relação entre as variáveis que se referem à evolução de sintomas dolorosos, "prisão de ventre" e "ansiedade/depressão", observa-se na Tabela 5.3.3, em que se destacam as elevadas percentagens da situação "piorou".

- Classe 2={Period}

As visitas dos profissionais de saúde são realizadas predominantemente "mais do que uma vez por semana" (43 - 75.4%).

- Classe 3={InstD, InstC, Ajuda, Conhecia, +Capaz, +Calm}

Quer os doentes, quer os seus cuidadores têm, na sua maior parte, nível de instrução que se reporta ao Ensino Primário concluído ou ao Ensino Preparatório/Secundário. Os cuidadores que "têm a quem recorrer quando necessitam de companhia ou de apoio"

sempre, são cerca de 46% e sentem-se “capazes de cuidar do doente ...” (‘Sempre‘ 56.1%), ‘Quase sempre‘ 35.1%), mas só 5.3% se dizem sentir “sempre calmos”, enquanto que 75.4% deles se dizem estar, no máximo, “calmos às vezes”.

A maior parte dos doentes “conhecia a sua doença”: (‘Totalmente’ 50.9%; ‘Parcialmente’ 33.3%).

A relação da variável “Conhecia a sua doença” com as variáveis “Nível de instrução do cuidador” e “Nível de instrução do doente” é forte:

- s_{LC}(Conhecia, InstC)=2.6748 (P_L(Conhecia, InstC)=0.9963),
- s_{LC}(Conhecia, InstD)=2.1493 (P_L(Conhecia, InstD)=0.9842)

O conhecimento total da doença aumenta progressivamente com o nível de instrução do cuidador e do doente (Anexo 18).

- Classe 4={+Bem}

80.7% dos cuidadores “sente-se bem porque está/estava a fazer tudo ...”, embora um diga que ‘nunca se sente bem’ e outro responda ‘às vezes’ ...

- Classe 5={Prof}

A situação profissional dos cuidadores mais frequente é a de reformado (49.1%).

- Classe 6={Apoio}

No apoio ao doente no domicílio os dois aspectos mais importantes referidos pelos cuidadores foram: “A atenção que a equipa de saúde dá ao doente e à família” (59.6%) e “Capacidade de apoio da equipa de saúde de forma segura e precisa” (52.6%).

- Classe 7={nerv, mor-bx, trist, esg, a-possib}

Esta classe, muito coesa, é constituída pelos sentimentos negativos dos cuidadores face ao sofrimento dos doentes a seu cargo:

- sente-se/sentiu “nervoso/a ...” (‘Quase sempre’ e ‘Sempre’ , 57.9%),
- “com a moral tão em baixo ...” (‘Às vezes’ e ‘Quase sempre’ , 75.4%),
- “desanimado/a e triste ...” (‘Às vezes’ e ‘Quase sempre’, 71.9%),
- “esgotado/a ...” (‘Às vezes’ e ‘Quase sempre’, 61.4%), e
- “a fazer coisas acima das suas possibilidades ...” (‘Às vezes’ e ‘Quase sempre’, 52.7%), (Anexo 18).

Os cuidadores são na sua maioria do sexo feminino (77.2%). Embora a variável “Sexo do cuidador” se encontre na 1ª classe da partição apresentada, não significa que exista uma relação importante entre ela e todas as outras variáveis desta classe. O que se detecta, é a relação forte que esta variável apresenta com a variável “Evolução dos sintomas - falta de ar”: (s_{LC}(Sexo, f-ar)=5.27, P_L(Sexo, f-ar)$\approx$1.00), que não tem significado clínico, mas apenas informativo, como vimos (Exemplo 4.4.1, Subsecção 4.4.3, Capítulo 4).

Sob o ponto de vista da interpretação geral dos resultados, o *Sexo do cuidador* apresenta uma relação moderada:

- com o facto de se *sentir nervoso* (s_{LC}(Sexo, nerv)=0.9594 (P_L(Sexo, nerv)=0.8313), manifestando que a situação mais frequente, para as mulheres, é "estar sempre" nervosa (40.9%) e, para os homens, é a de estar "às vezes" nervoso (46.2%). Observa-se, também, uma tendência para as mulheres serem mais nervosas do que os homens (P_{50Fem}='Quase sempre', P_{75Fem}='Sempre'; P_{50Masc}='Às vezes', P_{75Masc}='Quase sempre').
- com o facto de se *sentir calmo* (s_{LC}(Sexo, calm)=-0.3657 (P_L(Sexo, calm)=0.3573), manifestando que, as situações mais frequentes, para as mulheres, são as de estarem "às vezes" (45.5%) ou "quase nunca" (22.7%) calmas e, para os homens, são as de estarem "às vezes" (46.2%) ou "quase sempre" (46.2%) calmos. Observa-se, também, uma tendência para as mulheres serem "menos calmas" do que os homens (P_{75Fem}='Às vezes', P_{75Masc}='Quase sempre').
- com o facto de se sentir esgotado (s_{LC}(Sexo, esg)= -1.20 (P_L(Sexo, esg)=0.1151), manifestando que, a situação mais frequente, para as mulheres, é a de estar "às vezes" esgotada (38.6%) e, para os homens, é a de estar "quase sempre" esgotado (46.2%).

Aquelas relações tornam-se mais "visíveis" quando as variáveis são codificadas como variáveis número de ordem: s_{LC}(Sexo, nerv)=2.405, s_{LC}(Sexo, calm)= -2.0957, s_{LC}(Sexo, esg)= 1.1907.

5.3.4 Conclusões preliminares

A matriz de semelhanças S_{LC} entre as variáveis heterogéneas estudadas é definida positiva, enquanto a matriz P_L não o é. As ACP baseadas nas matrizes de semelhanças S_{LC} e P_L destacaram informações diferentes, que se complementam. Contudo a ACP da matriz de semelhanças P_L é mais informativa.

A 1ª componente principal da matriz de semelhanças P_L caracteriza-se principalmente pela "Percepção" dos cuidadores, pela dor dos doentes e pela ideia de que os cuidadores sentem que vale a pena "baterem-se" pelos doentes, isto é, cuidarem deles. A 2ª componente principal é a componente dos sentimentos dos cuidadores.

No 1º plano factorial apercebemo-nos bem da oposição entre os aspectos negativos dos sentimentos dos cuidadores face aos doentes e os aspectos positivos destes que estão mais próximos das restantes variáveis.

Mais uma vez podemos dizer que os algoritmos de ACHA compostos pelo mesmo critério de agregação "clássico" ("ligação única", "ligação completa", "ligação pela média" e AVM) associado aos coeficientes de semelhança s_{LC} e P_L conduzem a hierarquias de classificação análogas, em que as melhores partições obtidas com o coeficiente P_L são constituídas pelas mesmas classes ou por menos classes, que resultam da agregação de classes já existentes nas melhores partições obtidas com o coeficiente de semelhança s_{LC}. A ACHA permitiu-nos comparar as respostas dadas pelos cuidadores e obtermos um perfil dos cuidadores e dos doentes.

As duas metodologias de análise multivariada utilizadas, ACP e ACHA, complementam-se bem.

Um outro aspecto, que é de realçar, refere-se à possibilidade de estudar a significância estatística das relações encontradas com os coeficientes s_{LC} e P_L, caso seja aplicável esse estudo. Em particular, quando são violadas as condições de aplicação do teste do qui-quadrado de independência (tabelas de contingência com zeros e/ou pequenas frequências esperadas) – situação muito frequente neste estudo –, consideramos que estes coeficientes de semelhança, s_{LC} e P_L, se apresentam como uma boa alternativa.

5.4 Matrizes ultramétricas: o exemplo utilizado por Hubálek (1982)

5.4.1 Introdução

No estudo que se segue estamos interessados em comparar várias classificações/ultramétricas dos mesmos objectos, obtidas a partir de diversos algoritmos de análise classificatória hierárquica ascendente (ACHA), que se diferenciam uns dos outros apenas pelos diferentes coeficientes de semelhança utilizados. Ou seja, que são obtidas a partir de diferentes coeficientes de semelhança entre objectos e o mesmo critério de agregação entre classes. Com esse objectivo usamos os coeficientes s, s_{LC}, P_L e o coeficiente VAL_{Aw}.

Decidimos fazer esta comparação utilizando o exemplo que Hubálek (1982) usou e que resulta de um problema concreto: - Comparar 20 coeficientes de semelhança para dados binários (Tabela 5.4.1).

Tabela 5.4.1. Coeficientes de semelhança para dados binários propostos por diversos autores, em que *a, b, c* e *d* referem-se às quatro células da tabela de contingência 2x2, designando respectivamente, o número de co-presenças, presenças/ausências, ausências/presenças e co-ausências dos atributos. Excerto da tabela apresentada por Hubálek (1982) (Tabela 1.3.4)

Coeficiente de semelhança	*Autor*
$A_3 = a/(b+c)$	*Kulczynski (1927)*
$A_4 = a/(a+b+c)$	*Jaccard (1901), Sneath (1957)*
$A_5 = a/[a+ \frac{1}{2}(b+c)]$	*Dice (1945), Sørensen (1948)*
$A_6 = a/[a+2(b+c)]$	*Sokal & Sneath (1963)*
$A_7 = \frac{1}{2}[a/(a+b)+a/(a+c)]$	*Kulczynski (1927)*
$A_{11} = a/[(a+b).(a+c)]^{\frac{1}{2}}$	*Driver & Kroeber (1932)*
$A_{14} = a/(a+b+c+d) = a/n$	*Russell & Rao (1940)*
$A_{15} = a/[\frac{1}{2}(ab+ac)+bc]$	*Mountford (1962)*
$A_{18} = \frac{1}{4}[a/(a+b)+a/(a+c)+d/(c+d)+d/(b+d)]$	*Sokal & Sneath (1963)*
$A_{25} = ad/[(a+b).(c+d).(a+c).(b+d)]^{\frac{1}{2}}$	*Sokal & Sneath (1963)*
$A_{30} = (ad-bc)/[(a+b).(c+d).(a+c).(b+d)]^{\frac{1}{2}}$	*Yule (1912), Pearson & Heron (1913)*
$A_{32} = (\sqrt{(ad)}+a-b-c)/(\sqrt{(ad)}+a+b+c)$	*Baroni-Urban & Buser (1976)*
$A_{34} = \pm(\chi^2/\chi^2_{max})^{\frac{1}{2}}$ com o sinal de (ad-bc)	*Cole (1949): 'C_7'; Hurlbert (1969)*
$A_{35} = \pm[(\chi^2-\chi^2_{min})/(\chi^2_{max}-\chi^2_{min})]^{\frac{1}{2}}$	*Hurlbert (1969)*
$A_{36} = (ad-bc)/(ad+bc)$	*Yule (1900)*
$A_{37} = (\sqrt{(ad)}-\sqrt{(bc)})/(\sqrt{(ad)}+\sqrt{(bc)})$	*Yule (1912)*
$A_{38} = \cos[180\sqrt{(bc)}/(\sqrt{(ad)}+\sqrt{(bc)})]$	*Pearson & Heron (1913)*
$A_{39} = 4(ad-bc)/[(a+d)^2+(b+c)^2]$	*Michael (1920)*
$A_{40} = n.a/[(a+b).(a+c)]$	*Forbes (1907)*
$A_{43} = (a-a')/(a+a') = [n.a-(a+b).(a+c)]/[n.a+(a+b).(a+c)]$	*Tarwid (1960)*

Foram calculados os coeficientes de semelhança s, s_{LC}, P_L e VAL_{Aw} entre as 20 ultramétricas geradas por 20 coeficientes para dados binárias e pelo mesmo critério de agregação entre classes (critério de agregação da ligação pela média[199]). Obtemos assim, quatro matrizes de semelhanças entre aqueles coeficientes/ultramétricas. Sendo, então, estas matrizes assimiladas a produtos escalares, a sua diagonalização é usada como uma análise em componentes principais, ACP, como já foi referido. Também utilizamos análise classificatória hierárquica ascendente, ACHA, directamente sobre estas matrizes.

Realizámos as mesmas análises com os coeficientes s, s_{LC}, P_L e VAL_{Aw} sobre as 20 ultramétricas de forma a poder comparar o comportamento destes coeficientes quando aplicados a dados reais, assim como, com os resultados obtidos por Hubálek.

[199] *Average linkage*, em inglês.

Finalmente comparamos o coeficiente *simple matching* (Sokal e Michener, 1958), com os outros coeficientes de semelhança estudados.

5.4.2 Comparação de ultramétricas e coeficientes: Resultados

Os dados que analisámos referem-se à ocorrência simultânea de 7 tipos de cogumelos *Chaetomium (Ascomycetes)* em 869 ninhos (Hubálek, 1982) (Anexo 22).
Realizámos 20 classificações dos sete cogumelos recorrendo à análise classificatória hierárquica ascendente. Utilizámos, tal como Hubálek, o critério de agregação da ligação pela média com cada um dos 20 índices de semelhança, obtendo assim 20 classificações dos mesmos objectos que diferem apenas na forma como a semelhança entre eles foi observada. Obtivemos então 20 ultramétricas/dendrogramas (Anexo 22). Elas mostram claramente que resultados diferentes foram obtidos com coeficientes diferentes e pretendemos compará-las.

Para fazer isso calculámos os valores dos coeficientes s, s_{LC} e P_L, entre as ultramétricas que, neste caso, são as matrizes *score*. Uma vez calculados os valores dos três coeficientes entre as ultramétricas, obtivemos as respectivas matrizes de semelhanças: *S*(20x20), S_{LC}(20x20) e P_L(20x20) (Anexo 22).
Também foram calculados os valores do coeficiente VAL_{Aw} entre as 20 ultramétricas que, neste caso, são tratados como vectores coluna (as matrizes *score* diagonais inferiores/superiores desdobradas[200]). Obtivemos assim a matriz de semelhanças VAL_{Aw}(20x20).

Para representar estas quatro matrizes utilizámos técnicas de análise de dados: análises classificatórias hierárquicas ascendentes (ACHA) com o critério de agregação da ligação pela média e análises em componentes principais destas matrizes de semelhanças (Anexo 22).

[200] *Unfolded*, em inglês.

5.4.2.1 Resultados obtidos com os coeficientes s, s_{LC}, P_L e VAL_{Aw}

Os resultados obtidos com os coeficientes s, s_{LC} e P_L, estão apresentados no Anexo 21. Verificamos que eles são praticamente os mesmos, quando se utilizam os coeficientes s_{LC} e P_L. Estes, por sua vez, são semelhantes aos obtidos com o coeficiente s.

Para não sobrecarregar esta apresentação, limitar-nos-emos a analisar os resultados obtidos com os coeficientes probabilísticos, P_L e VAL_{Aw}, respectivamente:

- os da análise classificatória hierárquica ascendente (Tabela 5.4.2, Figura 5.4.1) e da análise em componentes principais da matriz P_L (Tabela 5.4.3, Figura 5.4.2).
- os da análise classificatória hierárquica ascendente (Tabela 5.4.4 e) e da análise em componentes principais da matriz VAL_{Aw} (Tabela 5.4.5 e Figura 5.4.4).

Tabela 5.4.2. Resultados obtidos com a análise classificatória hierárquica ascendente (P_L+Ligação pela média): as melhores partições foram obtidas com o critério da "estatística de níveis" (Lerman (1970); Bacelar-Nicolau (1972, 1980)), nos dois melhores níveis

Algoritmo	*Resultados:* *Partições obtidas no nível k*	*Estatística* *STAT(k)*
Coeficiente P_L *Ligação pela média*	*{A_3,A_{39},A_6,A_4, A_{14},A_5,A_{11},A_{25},A_{32}}* *{A_7, A_{30}, A_{15}, A_{37}, A_{38}, A_{36}, A_{40},A_{43},A_{18}}* *{A_{34}, A_{35}}*	*STAT(17)* *11.4820*
Coeficiente P_L *Ligação pela média*	*{A_3,A_{39},A_6,A_4, A_{14},A_5,A_{11},A_{25}}* *{A_{32}}* *{A_7, A_{30}, A_{15}, A_{37}, A_{38}, A_{36}, A_{40},A_{43},A_{18}}* *{A_{34}, A_{35}}*	*STAT(15)* *11.1466*

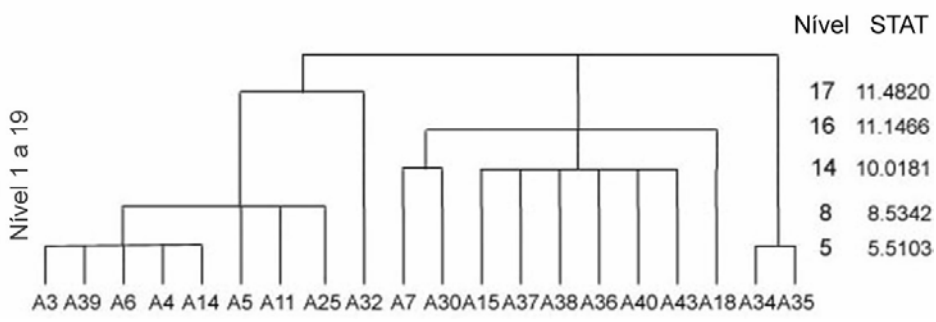

Figura 5.4.1. Dendrograma condensado aos níveis mais importantes, de acordo com o critério "estatística de níveis", obtida com a análise classificatória hierárquica ascendente (P_L+Ligação pela média).

Tabela 5.4.3. Resultados obtidos com a ACP da matriz de semelhanças P_L

	Componente	
	1	*2*
Unidades de inércia (valores próprios)	*18.75*	*1.39*
Variabilidade Explicada (%)	*94.06*	*6.96*
Var. Exp. Acumulada (%)	*94.06*	*101.02*

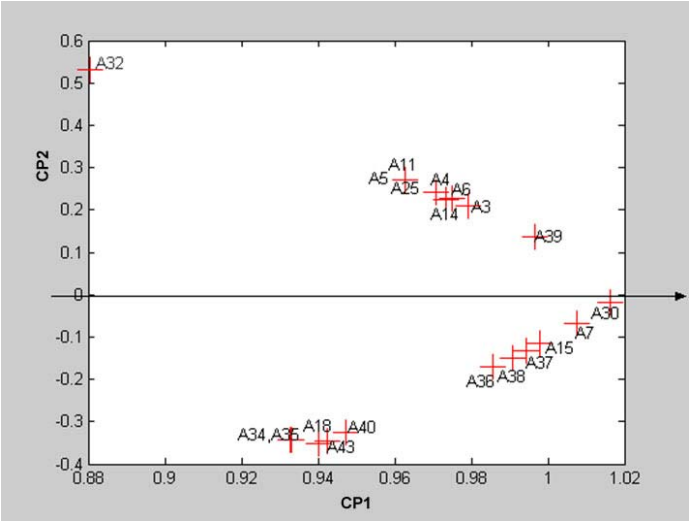

Figura 5.4.2. Representação gráfica do 1º plano obtido com a ACP da matriz de semelhanças P_L.

Tabela 5.4.4. Resultados obtidos com a análise classificatória hierárquica ascendente (VAL_{Aw}+ Ligação pela média): as melhores partições foram obtidas com o critério da "estatística de níveis" (Lerman, 1970; Bacelar-Nicolau, 1972, 1980), nos dois melhores níveis

Algoritmo	*Resultados:* *Partições obtidas no nível k*	*Estatística* *STAT(k)*
Coeficiente VAL_{Aw} *Ligação pela média*	*{A_3, A_6, A_4, A_5, A_{11}, A_{25}, A_{14}, A_{39}, A_{32} }* *{A_7, A_{30}, A_{15}, A_{37}, A_{36}, A_{38}, A_{18}, A_{40}, A_{43}, A_{34}, A_{35}}*	*STAT(9)* *11.8955*
Coeficiente VAL_{Aw} *Ligação pela média*	*{A_3, A_6, A_4, A_5, A_{11}, A_{25}, A_{14}, A_{39}}* *{ A_{32} }* *{A_7, A_{30}, A_{15}, A_{37}, A_{36}, A_{38}, A_{18}, A_{40}, A_{43} }* *{A_{34}, A_{35}}*	*STAT(7)* *11.2968*

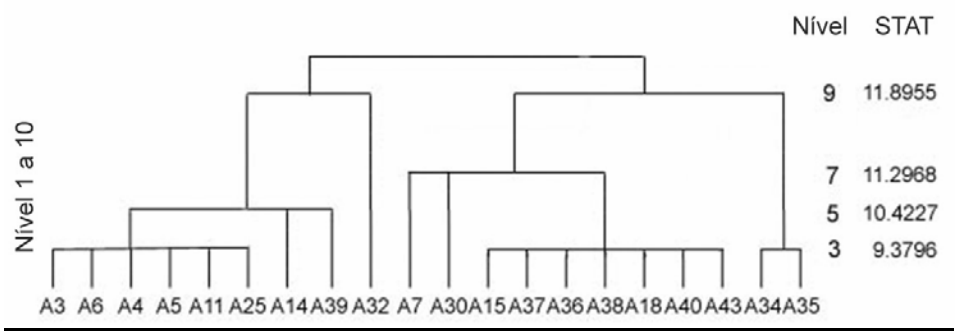

Figura 5.4.3. Dendrograma condensado aos níveis mais importantes, de acordo com o critério "estatística de níveis", obtida com a análise classificatória hierárquica ascendente (VAL_{Aw}+Ligação pela média).

Tal como podemos ver os resultados obtidos com estes coeficientes, P_L e VAL_{Aw} , são muito semelhantes.

Tabela 5.4.5. Resultados obtidos com a ACP da matriz de semelhanças VAL_{Aw}

	Componente	
	1	*2*
Unidades de inércia (valores próprios)	18.23	2.11
Variabilidade Explicada (%)	91.14	10.54
Var. Exp. Acumulada (%)	91.14	101.68

As matrizes de semelhanças P_L e VAL_{Aw} não são semidefinidas positivas. A 1ª componente explica a maior parte da variabilidade total, em ambas as análises (ACP+P_L: 93.77%; ACP+VAL_{Aw}: 91.14%). A 2ª componente tem associada a ela uma fraca percentagem de inércia explicada.

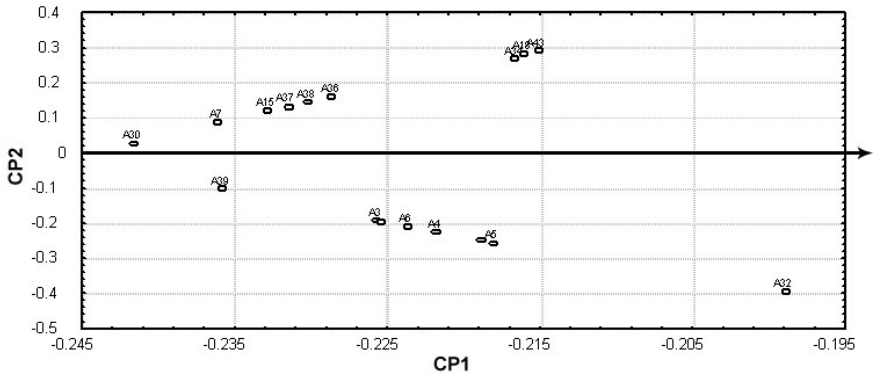

Figura 5.4.4. Representação gráfica do 1º plano obtido com a ACP da matriz de semelhanças VAL_{Aw}.

Nas duas análises em componentes principais vemos que:

- Todas as variáveis têm coordenadas positivas (negativas) no 1º eixo.
- O 2º eixo permite separar as variáveis.

Como podemos observar, os resultados obtidos com estes coeficientes são muito semelhantes. Além disso, para cada coeficiente, as melhores partições (Tabela 5.4.2, Tabela 5.4.4) coincidem com os resultados obtidos no 1º plano factorial (Figura 5.4.2, Figura 5.4.4).

5.4.2.2 Discussão e interpretação dos resultados

Considerando a partição obtida com os algoritmos (P_L+Ligação pela média) e (VAL_{Aw}+Ligação pela média), – Classe 1={A_3, A_{39}, A_6, A_4, A_{14}, A_5, A_{11}, A_{25}}, Classe 2={A_{32}}, Classe 3={A_7, A_{30}, A_{15}, A_{37}, A_{38}, A_{36}, A_{40}, A_{43}, A_{18}} e Classe 4={A_{34}, A_{35}} –, podemos dizer que as classes destes coeficientes estão separadas mais pelas suas propriedades matemáticas

do que pelo seu aspecto matemático (Tabela 5.4.6), com excepção da classe {A_{32}} para a qual não encontramos explicação.

Tabela 5.4.6. Descrição das classes

Classe 1 {A_3, A_{39}, A_6, A_4, A_{14}, A_5, A_{11}, A_{25}}	- Todos os índices apresentam associação nula quando a = 0, excluindo A_{39}. - Os índices A_3, A_4, A_6, A_{25}, A_{39} não são lineares. - d não faz parte dos índices A_3, A_4, A_5, A_6, A_{11}, A_{14}.
Classe 2 {A_{32}}	- A_{32} é igual a –1, quando a= 0. - Não é linear. - d faz parte da sua expressão.
Classe 3 {A_7, A_{30}, A_{15}, A_{37}, A_{38}, A_{36}, A_{40}, A_{43}, A_{18}}	- Alguns destes índices apresentam associação nula quando a = 0: A_7,A_{15}, A_{40}. - Os índices A_{15}, A_{36}, A_{37}, A_{38}, A_{43} não são lineares. - d faz parte da maior parte dos índices desta classe: A_{18},A_{30},A_{36},A_{37},A_{38}, A_{40},A_{43}.
Classe 4 {A_{34}, A_{35}}	Estes índices apresentam as duas propriedades: linearidade, associação nula quando a = 0.

Quando comparamos estes resultados com as melhores partições obtidas por Hubálek (r de Pearson + Ligação pela média*)* (Tabela 5.4.7), observamos que a maior diferença existente entre elas refere-se aos coeficientes A_7, A_{18} e A_{30}.

Tabela 5.4.7. As melhores partições obtidas por Hubálek (r de Pearson + Ligação pela média)

Nível (r=0.96)	Nível (r=0.95)
Classe 1: {A_{14},A_{39}}	
Classe 2:	Classe 1:
{A_3,A_4,A_5,A_6,A_7,A_{11},A_{25},A_{18},A_{30}}	{A_{14},A_{39},A_3,A_4,A_5,A_6,A_7,A_{11},A_{25},A_{18},A_{30}}
Classe 3: {A_{32}}	Classe 2: {A_{32}}
Classe 4: {A_{34},A_{35}}	Classe 3: {A_{34},A_{35}}
Classe 5: {A_{36},A_{37},A_{38},A_{42},A_{15},A_{40}}	Classe 4:{A_{36},A_{37},A_{38},A_{42},A_{15},A_{40}}

Considerando os resultados obtidos com os coeficientes binários, quando se classificaram os cogumelos, pensamos que A_{18}, A_{30} e A_7 estariam melhor colocados na Classe 4 – r=0.95 (Tabela 5.4.7) –, tal como obtivemos com os coeficientes P_L e VAL_{Aw} (Classe 3). Isto verifica-se igualmente na partição mediana obtida por Hubálek.

5.4.3 O que é que acontece ao coeficiente *simple matching* quando comparado com os 20 coeficientes?: Resultados

Será interessante estudar o coeficiente *simple matching* (Sokal e Michener, 1958), uma vez que ele tem um aspecto matemático simples, A_{20}=(a+d)/n, parecido com o do coeficiente A_{14}= a/n, apenas aparentemente.

Quando Hubálek classifica os cogumelos observa que, com este coeficiente se obtêm resultados muito diferentes dos obtidos com os outros coeficientes. Este facto especial é confirmado, quando introduzimos este novo coeficiente/ultramétrica na ACP (Figura 5.4.5 e Figura 5.4.6). Isto pode acontecer por, neste exemplo, ***d*** ser demasiado grande em relação a ***a***. Por isso este coeficiente pode conduzir a resultados muito diferentes dos outros no caso de fenómenos raros.

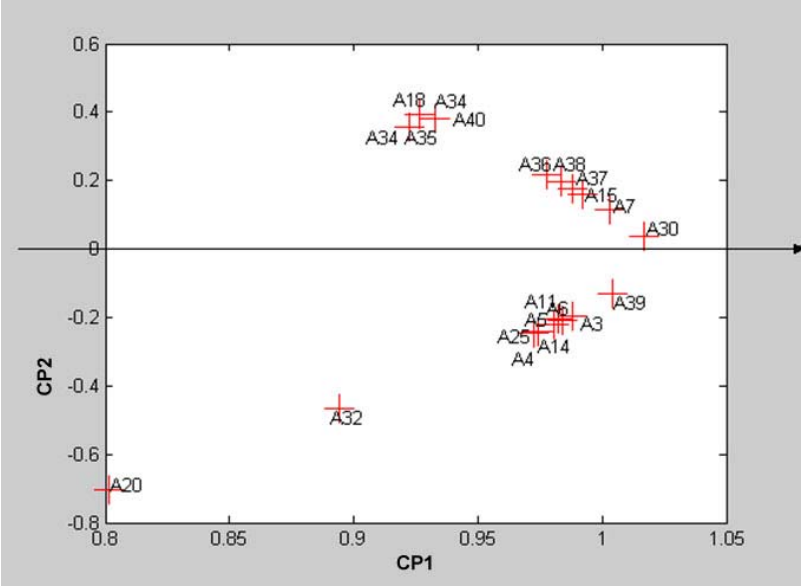

Figura 5.4.5. Gráfico do 1º plano obtido com a ACP da matriz de semelhanças P_L(21x21).

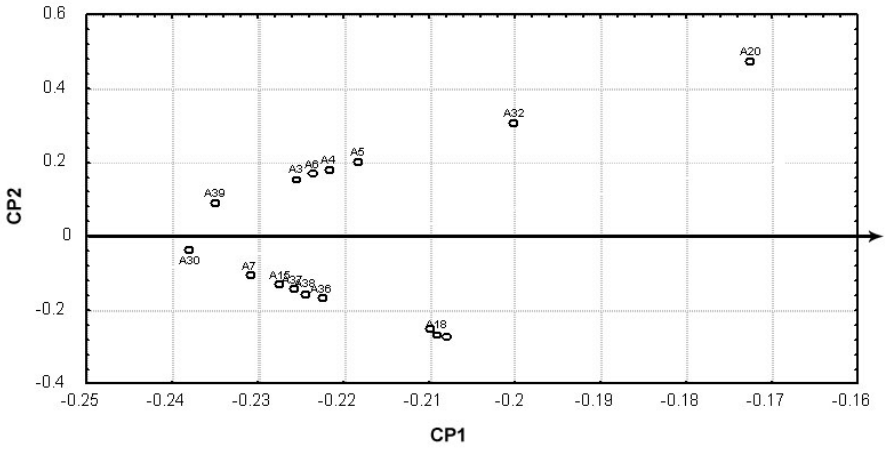

Figura 5.4.6. Gráfico do 1º plano obtido com a ACP da matriz de semelhanças VAL_{Aw}(21x21).

5.4.4 Conclusões preliminares

Os coeficientes – s, s_{LC}, P_L, VAL_{Aw} e r – quando aplicados a ultramétricas obtidas com estes vinte coeficientes para variáveis dicotómicas sobre os mesmos dados, conduziram a resultados/partições semelhantes. Este facto mostra-nos uma estrutura classificatória hierárquica forte. Contudo, é de realçar a diferença encontrada em relação aos coeficientes A_{18}, A_{30} e A_7, pois consideramos que os coeficientes de semelhança s, s_{LC}, P_L, VAL_{Aw} são mais apropriados do que o coeficiente r de Pearson, para relacionar ultramétricas.

Quando observamos as partições verificamos que as classes estão separadas mais pelas suas propriedades matemáticas do que pelo seu aspecto matemático.

Na nossa análise utilizámos duas técnicas de análise de dados, ACP e ACHA, que se enriquecem mutuamente. Enquanto que Hubálek só utilizou a ACHA.

O coeficiente *simple matching* (Sokal e Michener, 1958), quando comparado com os vinte coeficientes estudados, utilizando os coeficientes s, s_{LC}, P_L ou VAL_{Aw} com a ACP, comporta-se de maneira diferente, o que está de acordo com Hubálek. Talvez isto aconteça por estarmos em presença de um fenómeno raro.

5.5 Conclusões do capítulo

Em todas as análises realizadas neste capítulo, as matrizes de semelhanças S e S_{LC}, entre variáveis de ordem parcial ou entre variáveis heterogéneas são definidas positivas, entre ultramétricas são semidefinidas positivas, enquanto as matrizes de semelhanças P_L não o são. As ACP das matrizes S e S_{LC} permitiram pois, obter representações euclidianas exactas das variáveis, quando se compararam variáveis de ordem parcial. No caso das variáveis heterogéneas, a ACP da matriz S_{LC} também permitiu obter uma representação exacta das variáveis. As ACP das matrizes de semelhanças S_{LC}, em certos casos, fizeram sobressair, apenas algumas variáveis como responsáveis pela formação das primeiras componentes principais. Além disso, a variabilidade explicada pelas primeiras componentes principais nem sempre foi elevada, distribuindo-se muito gradualmente, por todas as componentes. Contudo, segundo Legendre e Legendre (2000), a projecção num espaço reduzido pode ser informativa mesmo que esse espaço contribua para uma pequena fracção da variância. A ACP da matriz P_L só permitiu obter representações euclidianas aproximadas,

mas bem interpretáveis e, em alguns casos, mais informativas do que as obtidas com a matriz S_{LC}.

Os algoritmos de ACHA compostos pelo mesmo critério de agregação "clássico" ("ligação única", "ligação completa", "ligação pela média") associado aos coeficientes de semelhança s_{LC} e P_L conduzem a hierarquias de classificação análogas, em que as melhores partições obtidas com o coeficiente P_L são constituídas pelas mesmas classes ou por menos classes, que resultam da agregação de classes já existentes nas melhores partições obtidas com o coeficiente de semelhança s_{LC}, como seria de esperar.

6 CONCLUSÕES E PERSPECTIVAS

O objectivo principal deste trabalho relaciona-se com os problemas de representação de dados em análise multivariada, e em particular, com as representações euclidianas de variáveis heterogéneas, no âmbito da Biomatemática. A necessidade de representar variáveis heterogéneas ocorre frequentemente quando se analisam, por exemplo, os resultados de questionários ou inquéritos, os quais são habitualmente apresentados em matrizes que cruzam a informação dos indivíduos e das variáveis. A visualização das relações entre as variáveis permitirá conhecer a estrutura dos dados, caso esta exista, com um impacto evidente.

Este problema de representação de variáveis com natureza matemática diversa, em análise multivariada, recorrendo a coeficientes de semelhança tem sido pouco tratado e por isso consideramos esta abordagem inovadora.

As análises em componentes principais de matrizes de semelhanças, que resultam de comparar variáveis de natureza diversa, permitiram-nos alcançar o objectivo principal – isto é, o de visualizar as variáveis e implicitamente as suas relações – quando recorremos aos coeficientes s, s_{LC} e P_L.

Os coeficientes s (semelhança observada), s_{LC} (semelhança padronizada) e P_L (coeficiente probabilístico) propostos por Le Calvé (1977) generalizam uma ideia de Daniels (1944) retomada por Lerman (1973). Estes coeficientes permitem relacionar variáveis atributos de descrição, nominais, ordinais (com modalidades parcialmente ordenadas, com modalidades totalmente ordenadas, com modalidades estrita e totalmente ordenadas e número de ordem) e variáveis métricas, assim como matrizes de semelhanças ou de dissemelhanças. No caso particular de se compararem algumas das variáveis do mesmo tipo, o coeficiente s_{LC} coincide, a menos de um factor, com coeficientes de correlação conhecidos (e.g., coeficiente Φ de Pearson, coeficiente de correlação tau de Kendall, coeficiente de correlação de Spearman, coeficiente de correlação linear de Pearson). Quando se compara uma variável dicotómica com uma variável métrica, o coeficiente s_{LC} coincide com o coeficiente de correlação bisserial por pontos, a menos de um factor, como seria de esperar. No caso dos atributos de descrição, o coeficiente P_L coincide com o coeficiente geral de semelhança do tipo VL para variáveis binárias p_{xy} proposto por Bacelar-Nicolau (1980).

Neste trabalho, estendemos a possibilidade destes coeficientes também compararem variáveis de ordem sequencial e alguns tipos de variáveis simbólicas: variáveis categóricas

com valores múltiplos, variáveis categóricas com todas as modalidades ordenadas pelas unidades estatísticas, variáveis intervalares e variáveis modais. No caso das variáveis simbólicas recorremos a coeficientes de semelhança e de dissemelhança conhecidos – como por exemplo, o coeficiente de afinidade quando se comparam variáveis simbólicas modais (distribuições de frequências, histogramas, barras) e a distância de Hausdorff no caso das variáveis intervalares – para definir as matrizes *score* dessas variáveis, o que resultou num enlace feliz com os coeficientes s, s_{LC} e P_L.

Os coeficientes s, s_{LC} e P_L permitem relacionar variáveis dos mais diversos tipos sempre que exista variabilidade nelas. Por isso, as variáveis que têm sempre o mesmo valor devem ser retiradas do estudo.

Pelo facto do coeficiente s não entrar em conta com a estrutura dos dados no seu cálculo, existe vantagem em usar os coeficientes s_{LC} e P_L, quer no caso das variáveis serem heterogéneas, quer nos casos das variáveis métricas e das variáveis simbólicas intervalares com unidades de medida diferentes, para as comparar.

Em todos os exemplos tratados com dados verídicos – comparação de variáveis de ordem parcial, variáveis simbólicas intervalares, variáveis simbólicas modais, variáveis heterogéneas – verificou-se que as matrizes de semelhanças S e S_{LC} são sempre definidas positivas. Quando se compararam ultramétricas, as matrizes de semelhanças S e S_{LC} são semidefinidas positivas. As ACP das matrizes S e S_{LC} permitiram pois, obter representações euclidianas exactas das variáveis, naqueles casos. As matrizes S são sensíveis ao efeito de dimensão. Por outro lado, nas ACP das matrizes S_{LC}, quando se comparam variáveis heterogéneas, a variabilidade explicada pelas primeiras componentes principais é muito baixa e sobressaem apenas algumas variáveis associadas a elas.

Para alguns tipos de variáveis encontramos análises já conhecidas. Nos casos clássicos, sempre que se realiza a ACP das matrizes S_{LC} de variáveis quantitativas ou de variáveis número de ordem obtêm-se, respectivamente, os mesmos resultados a menos de uma translação que a ACP Normada e a Análise das Ordens.

Em todos os exemplos tratados, as matrizes de semelhanças P_L nunca são semidefinidas positivas. Por isso, a ACP da matriz P_L permite obter uma representação euclidiana aproximada. Pelo facto da transformação associada à função de distribuição da N(0,1) ser irregular, existe uma tendência para aproximar os valores que estão próximos dos extremos e para afastar os valores próximos dos "valores centrais"; contudo, a interpretação dos resultados obtidos com as primeiras componentes principais (as correspondentes aos valores próprios positivos) é boa e complementa bem a análise realizada com o coeficiente

s_{LC}. Quando se utilizam os coeficientes s_{LC} ou P_L, as representações gráficas respectivas são inevitavelmente diferentes, mas complementam-se bem.

Convém realçar o facto dos coeficientes s, s_{LC} e P_L permitirem relacionar variáveis de ordem parcial. Além de ser uma situação pouco tratada na literatura científica, também nos possibilita ultrapassar, em certos casos, o problema dos dados omissos, acrescentando a categoria "Não responde" às categorias das variáveis ordinais sem perder a informação da sua natureza.
Quando as Escalas são constituídas por itens com possibilidade de resposta "Não sabe"/"Não responde", há vantagem em recorrer à quantificação simbólica modal das escalas, que nos permite trabalhar com toda a informação fornecida pelos dados, aliada à utilização dos coeficientes s_{LC}, P_L ou da afinidade generalizada a dados simbólicos/complexos (e.g., Bacelar-Nicolau, 2002; Nicolau *et al*., 2007). Fica aqui a nossa sugestão para tratar este tipo de situações, muito frequentes na prática, particularmente em escalas utilizadas nas áreas da psicologia, da sociologia e na área da saúde como as que se encontram, por exemplo, no Questionário SERVQUAL Modificado.

Paralelamente, os algoritmos de ACHA baseados nos coeficientes s, s_{LC} e P_L, aqui utilizados, também nos permitiram visualizar as relações entre variáveis simbólicas e, de forma mais geral, entre variáveis heterogéneas. Além disso, complementaram, de forma satisfatória, as análises em componentes principais realizadas. É importante recorrer às várias análises referidas, pois as informações que elas dão, complementam-se. Não devemos esquecer que estamos a tratar dados multivariados, cuja estrutura pode ser certamente complexa – cada uma das análises pode fazer sobressair um aspecto diferente de uma realidade que é multifacetada tal como um diamante.

Bacelar-Nicolau (e.g., 1980; 1988) e Nicolau (1980) desenvolveram, no âmbito da ACHA, uma abordagem análoga para comparar, quer variáveis, quer indivíduos. Destacamos aqui os coeficientes de afinidade *a* (semelhança básica), a_W (semelhança padronizada) e o coeficiente probabilístico VAL_{AW}. Como se referiu, no caso dos atributos de descrição, o coeficiente P_L coincide com o coeficiente geral de semelhança do tipo VL para variáveis binárias p_{xy} proposto por Bacelar-Nicolau (1980). Em Análise Classificatória, tendo em conta os resultados obtidos por Bacelar-Nicolau (1980) sobre a equivalência distribucional de coeficientes para dados binários, aquele resultado significa que, utilizando o coeficiente P_L ou o coeficiente p_{xy} com qualquer método de agregação, os resultados serão os mesmos que os obtidos com algoritmos de ACHA que se baseiem noutros coeficientes probabilísticos do tipo VL para dados binários.

A generalização dos coeficientes s, s_{LC} e P_L à comparação das variáveis simbólicas (intervalares, modais e categóricas com valores múltiplos), aqui apresentada, permite-nos pois, realizar análises em componentes principais e análises classificatórias hierárquicas ascendentes de dados tridimensionais — o que constitui uma inovação para estes coeficientes.

Ainda no contexto da Análise Classificatória, a utilização da ACHA directamente sobre o coeficiente P_L permite-nos obter uma extensão da metodologia VL ao caso das variáveis heterogéneas apresentadas (nominais, do tipo ordinal, métricas e simbólicas). Esta extensão junta-se a outras que têm sido obtidas (e.g., Sousa, 2005; Nicolau *et al*., 2007), sabendo-se também que coeficientes probabilísticos baseados na função de distribuição de um coeficiente de semelhança básico têm propriedades importantes em análise classificatória de dados (e.g., Lerman, 1981; Bacelar-Nicolau, 1987, 2000; Nicolau e Bacelar-Nicolau, 1981, 1999; Sousa *et al*., 2005).

A maior parte dos novos métodos de Análise de Dados propostos para tratar dados simbólicos referem-se à representação gráfica das unidades estatísticas. Os coeficientes s, s_{LC} e P_L permitem a representação das variáveis. Por outro lado, o coeficiente de afinidade generalizado a dados complexos permite obter relações, quer entre unidades estatísticas, quer entre variáveis, o que possibilita a representação das unidades estatísticas ou das variáveis.

Os dados verídicos do Questionário SERVQUAL Modificado permitiram pôr à prova, e bem, os coeficientes s_{LC} e P_L. Não temos conhecimento de coeficientes que permitam comparar todos os tipos de variáveis analisados (nominais, ordinais com modalidades totalmente ordenadas e parcialmente ordenadas, simbólicas modais e simbólica categórica com valores múltiplos). Neste estudo, verificou-se que a ACP da matriz S_{LC} é pouco informativa e há vantagem em recorrer à ACHA aplicada directamente sobre a matriz S_{LC}. A ACP e a ACHA aplicadas directamente sobre a matriz de semelhanças P_L complementaram bem a análise destas variáveis heterogéneas. As análises multivariadas aqui realizadas permitiram fazer sobressair os traços predominantes do "pensar" e "sentir" dos cuidadores dos doentes, tendo em consideração as variáveis escolhidas pelo especialista para os descrever. Sobressaiu a dor e o sofrimento dos doentes e o sentimento que os prestadores de cuidados paliativos têm de que vale a pena cuidar deles, apesar de todos os sentimentos negativos que isso lhes traz; assim como a importância que os cuidadores dão aos vários aspectos relevantes nas escalas de percepções (em que algumas delas apresentam percentagem elevada de respostas "Não se aplica" e/ou "Não sabe/"Não responde"). Os

resultados obtidos com as análises multivariadas permitiram também fazer o estudo bivariado de forma mais racional.

A análise de parte do Questionário SERVQUAL Modificado com os coeficientes s_{LC} e P_L também permitiu atingir um objectivo que nos é caro e que está implícito em todo o nosso trabalho de investigação, assim como no de outros investigadores: - Que os métodos de análise de dados utilizados sejam instrumentos que permitam alertar para novas estratégias de acção, dando a possibilidade de podermos contribuir para o bem estar do "outro" que, frequentemente, sofre.

Os coeficientes de semelhança s, s_{LC}, P_L e VAL_{Aw} são mais apropriados do que o coeficiente linear r de Pearson, para relacionar ultramétricas. Os resultados obtidos com as análises em componentes principais e análises classificatórias hierárquicas ascendentes aplicadas directamente sobre as matrizes de semelhanças probabilísticas P_L e VAL_{Aw} com dados verídicos são muito semelhantes e revelaram uma estrutura classificatória hierárquica forte.

Um outro aspecto a realçar é a possibilidade de estudar a significância das relações encontradas com os coeficientes s e s_{LC} e P_L, sempre que as condições de aplicação da inferência estatística estejam satisfeitas. De forma análoga, é possível estudar a significância dos coeficientes de afinidade.

Por agora, a maior limitação da utilização destes coeficientes tem a ver com a dimensão das amostras ou das populações. Referimo-nos à capacidade de processamento do computador, que pode ser ultrapassada com amostras de grande dimensão. Já se analisaram amostras de dimensão 100 sem problemas de cálculo. Contudo, temos consciência que eles podem surgir com amostras de maior dimensão. Há possibilidade de ultrapassar este problema utilizando na programação as sugestões relativas aos cálculos práticos apresentadas por Le Calvé (1977) ou outras estratégias informáticas a estudar.

Quanto às perspectivas futuras, é aliciante pensar que a definição e aplicação dos coeficientes s, s_{LC} e P_L não termina aqui. Existem vários aspectos em aberto:

- Podemos tentar encontrar outras definições para os *scores* das variáveis (e.g., a proposta de Le Calvé (1977) para tratar respostas múltiplas ou recorrer a outros coeficientes de semelhança/dissemelhança entre unidades estatísticas) nos casos apresentados, assim como noutras situações que não foram aqui abordadas.
- Os bons resultados obtidos com os coeficientes s_{LC} e P_L na comparação de variáveis intervalares levam-nos a querer aprofundar mais este estudo comparando com o que

tem sido feito nesta área (e.g., Nicolau *et al*., 2007; Brito, 2008[201]). A este estudo junta-se com facilidade o estudo das variáveis modais e a aplicação dos coeficientes s_{LC} e P_L ao exemplo tratado por Ichino (1988).

- Habitualmente, os totais das escalas utilizadas em várias áreas e particularmente em saúde, que são constituídas por itens com categorias totalmente ordenadas, são a soma dos "valores" das categorias desses itens. Sugerimos que as unidades estatísticas sejam descritas pelos seus perfis de resposta aos itens das escalas e que os totais destas sejam variáveis modais, sendo as matrizes *score* respectivas obtidas com o coeficiente de afinidade ponderado generalizado a variáveis complexas. Finalmente, comparar os resultados obtidos desta maneira com os obtidos da forma tradicional através das análises multivariadas respectivas.
- Aprofundar o estudo das escalas ipsativas. Este assunto merece um estudo mais aprofundado, tendo em atenção a literatura que se debruça sobre ele. Pensamos resolver o problema da representação gráfica multivariada deste tipo de variáveis, utilizando os coeficientes s_{LC} e P_L, num futuro próximo.
- Fazer simulações sobre os coeficientes para perceber melhor o seu comportamento distribucional.
- Introduzir a possibilidade de ponderação das variáveis que pensamos poder melhorar o desempenho do coeficiente s_{LC} quando se comparam variáveis heterogéneas.
- Abordar assuntos tratados por Le Calvé (1977) e que não foram aqui desenvolvidos, tais como:
    - A questão da credibilidade dos dados introduzindo um factor de ponderação na expressão do coeficiente de semelhança s.
    - A possibilidade de generalização destes coeficientes a matrizes rectangulares, i.e., a variáveis que são relações de E em F (por exemplo, matrizes de comunicação, redes de transporte, tabelas de incidências).
- Estudar a possibilidade de utilizar os coeficientes com dados censurados (situação que surge com frequência em Medicina). Tendo em consideração a investigação que tem sido realizada sobre o coeficiente tau de Kendall com dados censurados (e.g., Beaudoin *et al*., 2007), pensamos que estes coeficientes poderão permitir realizar melhores análises em componentes principais ou análises classificatórias em análise de sobrevivência.

---

[201] Conferência convidada das JOCLAD 2008.

- Realizar mais estudos comparativos com os coeficientes de afinidade (e.g., Bacelar-Nicolau, 1980, 2002; Nicolau *et al*., 2007), sempre que possível.
- Estudar melhor o comportamento dos novos algoritmos de ACHA, aqui definidos, que se baseiam nestes coeficientes, fazendo estudos de validação.

Relativamente aos coeficientes de afinidade pretendemos:

- Obter representações euclidianas, quer dos indivíduos, quer das variáveis, recorrendo à ACP das matrizes de afinidade, assim como a técnicas de *multidimensional scaling*. No caso dos dados compositivos, esta poderá ser uma alternativa à ACP para este tipo de dados proposta por Aitchison (1983).
- Visualizar as relações entre indivíduos e entre variáveis através de *biplots*, utilizando a metodologia *multidimensional scaling* ponderada (Greenacre, 2005) sobre a matriz de distâncias obtidas a partir da afinidade.

Neste domínio não há soluções únicas, existindo diversos caminhos que levam a diferentes soluções. Nos "caminhos" da análise multivariada de variáveis heterogéneas ou não, estarão certamente, análises em componentes principais e análises classificatórias hierárquicas ascendentes aplicadas directamente sobre as matrizes de semelhanças S_{LC} e P_L, assim como, as que se baseiam na afinidade e nas suas generalizações.

> "A vida é curta, a arte é longa, a ocasião fugitiva, a experiência enganadora, o juízo difícil."
> (Hipócrates, *Aforismos*, Primeira Secção, séc. IV/V a. C.)

contudo,

> "Pelo Sonho é que vamos, comovidos e mudos.
> Chegamos? Não chegamos?
> .....
> Basta que a alma demos, com a mesma alegria,
> ao que desconhecemos e ao que é do dia-a-dia.
> Chegamos? Não chegamos?
> Partimos. Vamos. Somos."
> (in Sebastião da Gama, *Pelo sonho é que vamos*, 1953)

BIBLIOGRAFIA

ABDALLAH, H. e SAPORTA, G. (1998): Classification d'un ensemble de variables qualitatives. *Revue de Statistique Appliquée XLVI(4), 5-26*.

AITCHISON, J. (1983): Principal component analysis of compositional data. *Biometrika 70, 57-65*.

AL AYOUBI, B. (1991): *Analyse des Données en Distance de Type M_1. Théorie et Algorithmes d'Optimisation*. Thèse de doctorat, Université de Haute Bretagne, Rennes.

ANDERBERG, R. (1973): *Cluster Analysis for Applications*. Academic Press, New York and London.

BACELAR-NICOLAU, H. (1972): *Analyse d'un Algorithme de Classification*. Thèse de 3ème cycle, Université Pierre et Marie Curie, Paris.

BACELAR-NICOLAU, H. (1980): *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória*. Tese de doutoramento, Faculdade de Ciências da Universidade de Lisboa, Lisboa.

BACELAR-NICOLAU, H. (1982): L'Affinité: un coefficient de similarité. In: C. Perruchet (Ed.) *Edition Regroupée des Actes des Journées de Classification de Bruxelles 1982, Nancy 1981, Toulouse 1980.* Société Francophone de Classification, Paris, 81-87.

BACELAR-NICOLAU, H. (1985): The affinity coefficient in cluster analysis. *Methods of Operations Research 53, 507-512*.

BACELAR-NICOLAU, H. (1986): A probabilistic hierarchic cluster method based on the affinity coefficient. In: F. De Antoni, N. Lauro and A. Rizzi (Eds.): *COMPSTAT'86*. Physica-Verlag, Heidelberg, 19-20.

BACELAR-NICOLAU, H. (1987): On the distribution equivalence in cluster analysis. In: P. Devijver e J. Kittler (Eds.): *Pattern Recognition Theory and Applications*. NATO ASI Series, Springer-Verlag, New York, 73-79.

BACELAR-NICOLAU, H. (1988): Two probabilistic models for classification of variables in frequency tables. In: H.-H. Bock (Ed.): *Classification and Related Methods of Data Analysis.* Elsevier/North Holland, Amsterdam, 181-186.

BACELAR-NICOLAU, H. (1989): Sur l'équivalence distributionnelle entre coefficients d'association. *Bulletin of the International Statistical Institute, Proceedings of the 47th Session (1), 89-90.*

BACELAR-NICOLAU, H. (1991): Probabilistic models in cluster analysis based on L1 norm related to affinity coefficient. *Bulletin of the International Statistical Institute, Proceedings of the 48th Session (1), 49-50*.

BACELAR-NICOLAU, H. (2000): The affinity coefficient. In: H.-H. Bock e E. Diday (Eds.): *op. cit.*, 160-165.

BACELAR-NICOLAU, H. (2002): On the generalised affinity coefficient for complex data. *Biocybernetics and Biomedical Engineering 22(1), 31-42*.

BACELAR-NICOLAU, H. e DORIA, I. (1992): O coeficiente de afinidade em análise de dados multivariados. Uma aplicação à economia. In: M. Turkman e M. Carvalho (Eds.) Actas *da 1a Conferência em Estatística e Optimização*. INIC e DEIO, Lisboa, 384-404.

BACELAR-NICOLAU, H. e NICOLAU, F.C. (1978): *Convergence de la Loi Hypergéométrique Vers la Loi Normale. Conditions Optimales. Conséquences pour L'étude du Tableau de Contingence 2x2.* Série Notas e Comunicações, Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa.

BACELAR-NICOLAU, H. e NICOLAU, F.C. (1990): Affinity and correlation coefficients in cluster analysis. In: K. Momirovit, e V. Mildner (Eds.): *COMPSTAT'90*. Physica Verlag, Heidelberg, 15-16.

BACELAR-NICOLAU, H. e NICOLAU, F.C. (1993): Classifying integer scale data by the affinity coefficient: A probabilistic approach. In: J. Jansen e C.H. Skiadas (Eds.): *Proceedings of the Sixth International Symposium on Applied Stochastic Models and Data Analysis (ASMDA).* Vol. 1. World Scientific Publ. Co., New York, 63-74.

BACELAR-NICOLAU, H. e NICOLAU, F.C. (1994): Estatística e análise de dados multivariados: passado e futuro (Um exemplo da área da análise classificatória). In: D. Pestana, A. Turkman, J. Branco, L. Duarte e A. Pires (Eds.) *A Estatística e o Futuro e o Futuro da Estatística – Actas do I Congresso Anual da Sociedade Portuguesa de Estatística*. Edições Salamandra, 521-530.

BACELAR-NICOLAU, H., DIAS, O. e SOUSA FERREIRA, A. (2005): Humanização dos cuidados paliativos: Análise multivariada de um questionário de qualidade/satisfação ao apoio domiciliário. In: *Livro de Resumos das XII Jornadas de Classificação e Análise de Dados (JOCLAD 2005)*, 70-71.

BARTHÉLEMY, J.-P. e GUÉNOCHE, A. (1988): *Les Arbres et les Représentations des Proximités*. Masson, Paris.

BEAUDOIN, D., DUCHESNE, T. e GENEST, C. (2007): Improving the estimation of Kendall's tau when censoring affects only one of the variables. *Computational Statistics and Data Analysis 51(12), 5743-5764*.

BÉNASSÉNI, J. (1994): Partial additive constant. *Journal of Statististical Computing Simulation 49, 179-193*.

BÉNASSÉNI, J., BENNANI DOSSE, M. e JOLY, S. (2007): On a general transformation making a dissimilarity matrix euclidean. *Journal of Classification 24, 33-51.*

BENAYADE, M. e BENINEL, F. (2002): L_2-spherical dissimilarities: characterization and computing spherical image. *Rapport Cat-inist CNRS 171* (POI-02-171).

BENINEL, F. (1987): *Problèmes de Représentations Sphériques des Tableaux de Dissimilarité*. Thèse de 3ème cycle, Université de Haute Bretagne, Rennes.

BENINEL, F. (1999): Dissimilarités de type sphérique et positionnement multidimensionnel normé. *RAIRO Recherche Opérationnelle/Operations Research 33*, *569-581*.

BENINEL, F., QANNARI, A. e QANNARI, E.M. (1994): Distance à centre additive. *RAIRO Recherche Opérationnelle/Operations Research 28(4), 357-368.*

BENZÉCRI, J.P. (1967): Représentation Euclidienne d'un ensemble fini muni de masses et de distances. *ISUP, Septembre 1967* (edição concluída em 1969).

BENZÉCRI, J.-P. (Ed.) (1980a): *L'Analyse des Données.* Dunod, Paris.

BENZÉCRI, J.-P. (1980b): Les principes de l'analyse des données. In: J.-P. Benzécri (Ed.) (1980a), *op. cit.,* tomo 2, 3-17.

BENZÉCRI, J.-P. (1982): *Histoire et Préhistoire de l'Analyse des Données*. Dunod, Paris.

BHATTACHARYYA, A. (1943): On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society 35, 99-109.*

BHATTACHARYYA, A. (1946): On a measure of divergence between two multinomial populations. *Sankhyā 7*, *401-406*, *cit. in* Krzanowski (1983).

BISHOP, Y.M.M., FIENBERG, S.E. e HOLLAND, P.W. (1975): *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge (MA).

BLANC, F., CHARDY, P., LAUREC, A. e REYS, J.-P. (1976): Choix des métriques qualitatives en analyse d'inertie. Implications en écologie marine benthique. *Marine Biology (Berl.) 35(1), 49-67*.

BLUMENTHAL, L.M. (1936): New theorems and methods in determinant theory. *Duke Mathematical Journal 2*, *396-404*, *cit. in* Joly e Le Calvé (1986).

BLUMENTHAL, L.M. (1953): *Theory and Applications of Distance Geometry*. Oxford University Press, London, *cit. in* Caillez e Kuntz (1996).

BOCK, H.-H. (2000a): Symbolic data. In: H.-H. Bock e E. Diday (Eds.): *op. cit.*, 39-53.

BOCK, H.-H. (2000b): Dissimilarity measures for probability distributions. In: H.-H. Bock e E. Diday (Eds.): *op. cit.*, 153-160.

BOCK, H.-H. e Diday, E. (Eds.) (2000): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer-Verlag, Berlin/Heidelberg.

BORG, I. e GROENEN, P.J.F. (2005): *Modern Multidimensional Scaling*. Springer, Berlin.

BOUROCHE, J.-M. e SAPORTA, G. (1998): *L'Analyse des Données*. Presse Universitaires de France.

BRANDEN, K. (2005): *Robust Methods for High-Dimensional Data, and a Theoretical Study of Depth-Related Estimators*. Doctor in de wetenschappen, Katholieke Universiteit Leuven, Leuven.

BRITO, I., CELEUX, C. e SOUSA FERREIRA, A. (2006): Combining methods in supervised classification: a comparative study on discrete and continuous problems. *REVSTAT – Statistical Journal 4(3), 201-225*.

BRITO, P. (2008): Análise de dados simbólicos : Questões e perspectivas. O caso particular dos dados intervalares. In: *Livro de Resumos das XV Jornadas de Classificação e Análise de Dados (JOCLAD 2008),* 4-5.

BROSSIER, G. e LE CALVÉ, G. (1986): Analyse des dissimilarités sous l'éclairage $\sqrt{D}$. Application à la recherche d'arbres additifs optimaux. In: E. Diday, Y. Escoufier, L. Lebart, J.P. Pagès, Y. Schektman e R. Tomassone (Eds.): *Data Analysis and Informatics IV: Proceedings of the Fourth International Symposium on Data Analysis and Informatics*. North Holland, Amsterdam, 111-121.

BRUCE, N., POPE, D. e STANISTREET, D. (2008): *Quantitative Methods for Health Research*. John Wiley & Sons, Ltd.

CAILLEZ[202], F. e KUNTZ, P. (1996): A contribution to the study of the metric and euclidean structures of dissimilarities. *Psychometrika 61(2)*, *241-253*.

CAILLIEZ, F. (1983): The analytical solution of the additive constant problem. *Psychometrika 48*, *305-308*.

CAILLIEZ, F. e PAGÈS, J.P. (1976): *Introduction à L'Analyse des Données*. SMASH, Paris.

CAMIZ, S. e LE CALVÉ, G. (2001): *Recent experimentation on euclidean approximations of biased euclidean distances*. In: S. Borra, R. Rocci, M. Vichi, and M. Schader (Eds.): *Advances in Classification and Data Analysis*. Springer, Berlin, 77-84.

CARROLL, J.D. e ARABIE, P. (1980): Multidimensional scaling. *Annual Review of Psychology 31, 607-649*, *cit. in* Everitt e Rabe-Hesketh (1997).

CHAH, S. (1984): *Agrégation des Préordonnances*. Étude F–063, IBM, Paris.

CHAH, S. (1985): Critères de classification sur données hétérogènes. *Revue de Statistique Appliquée 33(2)*, 19-36.

CHANDON, J.L. e PINSON, S. (1981): *Analyse Typologique: Théories et Applications*. Masson, Paris.

[202] Nesta referência, aparece escrito no original, por lapso, Caillez em vez de Cailliez.

CHARDY, P., GLEMAREC, M. e LAUREC, A. (1976): Application of inertia methods to benthic marine ecology: practical implication of the basic options. *Estuarine and Coastal Marine Science 4, 179-205.*

CHOUAKRIA, A., CAZES, P. e DIDAY, E. (2000): Symbolic principal component analysis. In: H.-H. Bock e E. Diday (Eds.): *op. cit.*, 200-212.

COHEN, J. (1960): A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, *37-46*.

COHEN, J. (1968): Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin 70*, *213-220*.

CONOVER, W.J. (1999): *Practical Nonparametric Statistics*. 3rd Ed., John Wiley and Sons, Inc., New York.

COX, D.R. (1972): The analysis of multivariate binary data. *Applied Statistics 21, 113-120*.

COX, T. e COX, M. (2000): A general weighted two-way dissimilarity coefficient. *Journal of Classification 17*, *101-121*.

CRITCHLEY, F. (1986): Dimensionality theorems in multidimensional scaling and hierarchical cluster analysis. In: E. Diday, Y. Escoufier, L. Lebart, J.P. Pagès, Y. Schektman e R. Tomassone (Eds.): *Data Analysis and Informatics IV: Proceedings of the Fourth International Symposium on Data Analysis and Informatics*. North Holland, Amsterdam, 45-70.

CRITCHLEY, F. (1988): The Euclidean structure of a dendogram, the variance of a node and the question: “How many clusters really are there?”. In: H.H. Bock (Ed.): *Classification and Related Methods of Data Analysis*. Elsevier/North-Holland, Amsterdam, 75-84.

CUADRAS, C.M. e ARENAS, C. (1990): A distance based regression model for prediction with mixed data. *Communications in Statistics - Theory and Methods 19, 2261-2279*.

DANIELS, H.E. (1944): The relation between measures of correlation in the universe of sample permutations. *Biometrika 33*, 129-135.

DAVIDSON, G.S., HENDRICKSON, B., JOHNSON, D.K., MEYERS, C.E. e WYLIE, B.N. (1998): Knowledge mining with VxInsight: Discovery through interaction. *Journal ot Intelligent Information Systems 11, 259-285*, c*it. in* Van Eck e Waltman (2006).

DE LEEUW, J. e VAN RIJCKEVORSEL, J. (1980): HOMALS and PRINCALS: Some generalizations of principal component analysis. In: E. Diday, L. Lebart, J.P. Pagès e R. Tomassone (Eds.): *Data Analysis and Informatics: Proceedings of the Second International Symposium on Data Analysis and Informatics.* North-Holland, Amsterdam, 231-242.

DESDEVISES, Y., LEGENDRE, P., AZOUZI, L. e MORAND, S. (2003): Quantifying phylogenetically structured environmental variation*. Evolution 57(11), 2647-2652.*

DEZA, M. e MAEHARA, H. (1990): Metric transforms and euclidean embeddings. *Transactions of the American Mathematical Society 317 (2)*, *661-671*.

DIAS, O. (1994): Apresentação do programa LEAS. LEAS-F e LEAS-P: Comparação na análise classificatória. In: D. Pestana, A. Turkman, J. Branco, L. Duarte e A. Pires (Eds.) *A Estatística e o Futuro e o Futuro da Estatística – Actas do I Congresso Anual da Sociedade Portuguesa de Estatística*. Edições Salamandra, 435-445.

DIAS, O., SOUSA FERREIRA, A. e BACELAR-NICOLAU, H. (2006): Avaliação da qualidade dos cuidados paliativos em contexto domiciliário. In: *Livro de Resumos das XIII Jornadas de Classificação e Análise de Dados (JOCLAD 2006),* 38-39.

DOMENGES, D. e VOLLE, M. (1979): Analyse factorielle sphérique: une exploration. *Annales de l'INSEE 35, 3-83*, *cit. in* Bacelar-Nicolau (1980).

DORIA, I. (1989): *Contribuições ao Estudo de Coeficientes de Semelhança Baseados na Noção de Afinidade em Classificação Hierárquica.* Tese de mestrado, Faculdade de Ciências da Universidade de Lisboa, Lisboa.

DORIA, I., DIAS, O., SOUSA FERREIRA, A., LE CALVÉ, G. e BACELAR-NICOLAU, H. (2007): Comparison of methodologies of multivariate analysis in a palliative care context. In: *Livro de Resumos da XI Conferencia Española de Biometria e Primer Encuentro Iberoamericano de Biometria (CEIB2007)*, 101-102.

DORIA, I., LE CALVÉ, G. e BACELAR-NICOLAU, H. (1999): Comparing several similarity indices on dichotomics based on the associated ultrametrics obtained with real data. In: H. Bacelar-Nicolau, F. Costa Nicolau, J. Janssen (Eds.): *Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society*. Instituto Nacional de Estatística, Lisboa, 142-147.

DORIA, I., LE CALVÉ, G. e BACELAR-NICOLAU, H. (2000): Comparision of ultrametrics obtained with real data, using the P_L and VAL_{AW} coefficients. In: H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen e M. Schader (Eds.): *Data Analysis, Classification, and Related Methods*. Springer, Berlin, 107-112.

DORIA, I., LE CALVÉ, G. e BACELAR-NICOLAU, H. (2006a): Uma abordagem ao estudo de variáveis com estruturas diversas baseada nos coeficientes s, s_{LC} e P_L. In: L. Canto e Castro, E.G. Martins, C. Rocha, M.F. Oliveira, M.M. Leal e F. Rosado (Eds.): *Ciência Estatística – Actas do XIII Congresso Anual da Sociedade Portuguesa de Estatística*. Edições SPE, 309-320.

DORIA, I., LE CALVÉ, G., BACELAR-NICOLAU, H., SOUSA FERREIRA, A. e DIAS, O. (2006b): Os coeficientes s, s_{Lc} e P_L na análise de um questionário de qualidade e satisfação na saúde. In: *Livro de Resumos das XIII Jornadas de Classificação e Análise de Dados (JOCLAD 2006),* 40-41.

DUTILLEUL, P., STOCKWELL, J.D., FRIGON, D. e LEGENDRE, P. (2000): The Mantel test versus Pearson's correlation analysis: Assessment of the differences for biological and environmental studies. *Journal of Agricultural, Biological and Environmental Statistics 5*, 131-150.

ECKART, C. e YOUNG, G. (1936): The approximation of one matrix by another of lower rank. *Psychometrika 1, 211-218*, *cit. in* Le Clavé (1976b).

ESCOFIER, B. (1979): Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. *Les Cahiers de l'Analyse des Données 4(2), 137-146.*

ESCOFIER, B. e PAGÈS, J. (1998): *Analyses Factorielles Simples et Multiples*. Dunod, Paris.

ESCOUFIER, Y. (1973): Le traitement des variables vectorielles. *Biometrics 29, 751-760.*

ESCOUFIER, Y., CAILLIEZ, F. e PAGÈS, J.P. (1978): Géometrie et techniques particulières en analyse factorielle. *European Meeting on Psychometrics and Mathematical Psychology*, *Sweden: Uppsala: cit in Pagès, J.-P.; Cailliez, F.; Escoufier, Y. (1979) Analyse factorielle: un peu d'histoire et de géométrie. Revue de Statistique Appliquée 27(1), 5-28.*

ESPOSITO, F., MALERBA, D. e TAMMA, V. (2000): Dissimilarity measures for symbolic objects. In: H.-H. Bock e E. Diday (Eds.): *op. cit.*, 165-185.

ESPOSITO, F., MALERBA, D., TAMMA, V. e BOCK, H.H. (2000): Classical resemblance measures. In: In: H.-H. Bock, e E. Diday (Eds.): *op. cit.*, 139-152.

EVERITT, B.S. e DUNN, G. (1997): *Applied Multivariate Data Analysis*. 5th ed., Arnold, London.

EVERITT, B.S. e RABE-HESKETH, S. (1997): *The Analysis of Proximity Data*. Arnold, London.

FAITH, D.P. (1983): Asymmetric binary similarity measures. *Oecologia (Berl.) 57, 287-290*, *cit. in* Legendre e Legendre (2000).

FAVRE, C. (1999): *Analyse en Norme L_1 et L_0 des Distances et des Préférences. Planification en Analyse Sensorielle. Application au Confort d'Accueil de Sièges Automobiles*. Thèse de 3ème cycle, Université de Rennes 2, Rennes.

FICHET, B. (1983): *Analyse Factorielle sur Tableaux de Dissimilarités*. Thèse d'État de Biologie Humaine, Université d'Aix-Marseille II, Marseille.

FICHET, B. (1986): Approximation d'une dissimilarité quelconque par une distance à centre au sens des moindres carrés. In: F. De Antoni, N. Lauro and A. Rizzi (Eds.): *COMPSTAT1986: Proceedings in Computational Statistics*. Springer-Verlag, Heidelberg.

FICHET, B. e LE CALVÉ, G. (1984): Structures geométriques des principaux indices de dissimilarité sur signes de presence-absence. *Statistique et Analyse des Données 9 (3), 11-44.*

FRÉCHET, M. (1935): Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert. *Annals of Mathematics 36, 705-718*, *cit. in* Joly e Le Calvé (1986).

FUKUNAGA, K. (1990): *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego.

GIFI, A. (1983): *PRINCALS User's Guide*. University of Leiden, Leiden.

GIFI, A. (1990): *Nonlinear Multivariate Analysis*. Wiley, New York.

GORBAN, A.N., KÉGL, B., WUNSCH, D.C. e ZINOVYEV, A. (Eds.) (2008): *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer-Verlag, Berlin Heidelberg.

GORDON, A.D. (1990): Constructing dissimilarity measures. *Journal of Classification 7*, *257-269*.

GORDON, A.D. (1999): *Classification.* 2nd Ed., Chapman and Hall, Boca Raton Florida.

GOWER, J.C. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*, *325-338*.

GOWER, J.C. (1967): Multivariate analysis and multidimensional geometry. *Statistician 17*, *13-28*.

GOWER, J.C. (1971): A general coefficient of similarity and some of its properties. *Biometrics 23, 623-637*.

GOWER, J.C. e BLASIUS, J. (2005): Multivariate prediction with nonlinear principal component analysis: theory. *Quality and Quantity 39, 359-372*.

GOWER, J.C. e LEGENDRE, P. (1986): Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification 3, 5-48*.

GREENACRE, M.J. (1978): *Quelques Méthodes Objectives de Représentation Graphique d'un Tableau de Données*. Thèse de 3ème cycle, Université de Paris, Paris.

GREENACRE, M.J. (1984): *Theory and Applications of Correspondence Analysis*. Academic Press, London.

GREENACRE, M.J. (2005): Weighted metric multidimensional scaling. In: M. Vichi, P. Monari, S. Mignani, e A. Montanari (Eds.) *New Developments in Classification and Data Analysis.* Springer, Heidelberg, 141-150.

GURU, D.S., KIRANAGI, B.B. e NAGABHUSHAN, P. (2004): Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. *Pattern Recognition Letters 25(10), 1203-1213.*

HAJDU, L.J. (1981): Graphical comparison of resemblance measures in phytosociology. *Vegetatio 48, 47-59.*

HARTIGAN, J.H. (1975): *Clustering Algorithms*. John Wiley and Sons, New York, c*it. in* Beninel (1987).

HEISER, W.J. e MEULMAN, J.J. (1994): Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In: M.J. Greenacre e J. Blasius (Eds.): *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*. Academic Press, London, 179-209.

HELM, C.E. (1964): Multidimensional ratio scaling analysis of perceived color relations. *Journal of the Optical Society of America 54, 256-262*.

HOLMAN, E.W. (1972): The relation between hierarchical and euclidean models for psychological distances. *Psychometrika 37*, *417-423*, *cit. in* Joly e Le Calvé (1986).

HOTELLING, H. (1933): Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24, 417-441 e 498-520.*

HUBÁLEK, Z. (1982): Coefficients of association and similarity based on binary (presence-absence) data: An evaluation. *Biological Revue 57*, *669-689*.

ICHINO, M. (1988): General metrics for mixed features – The Cartesian space theory for pattern recognition. In: *Proceedings of the 1988 IEEE International Conference on Systems, Man and Cybernetics.* Pergamon, Oxford, 494-497.

ICHINO, M. e YAGUCHI, H. (1994): Generalized Minkowski metrics for mixed feature type data analysis. *IEEE Transactions on Systems, Man and Cybernetics 24(4), 698-708.*

JACOBY, W.G. (1991): *Data Theory and Dimensional Analysis*. Sage, Newbury Park (CA), *cit. in* Everitt e Rabe-Hesketh (1997).

JOHNSON, R.M. (1963): On a theorem stated by Eckart and Young. *Psychometrika 28, 259-263, cit. in* Le Calvé (1976b).

JOLLIFFE, I.T. (1986): *Principal Component Analysis*. Springer-Verlag, New York.

JOLY, S. e LE CALVÉ, G. (1985): *Rapport Technique nº1*. Laboratoire d'Analyse des Données, Université de Haute Bretagne, Rennes.

JOLY, S. e LE CALVÉ, G. (1986): Étude des puissances d'une distance. *Statistique et Analyse des Données 11(3), 30-50.*

JOLY, S. e LE CALVÉ, G. (1992): *Rapport de Recherche 92-1: Realisable 0-1 Matrices and City Block Distance*. Laboratoire d'Analyse des Données, Université de Haute Bretagne, Rennes, France.

JOLY, S. e LE CALVÉ, G. (1994): Similarity functions. In: B. Van Cutsen, (Ed.): *Classification and Dissimilarity Analysis.* Springer-Verlag, New York, 67-86.

KENDALL, M.G. (1970): *Rank Correlation Methods.* 4th Ed., Charles Griffin, London.

KETTENRING, J.R. (2006): The practice of cluster analysis. *Journal of Classification 23*, *3-30*.

KRAEMER H. C. (2006): Correlation coefficients in medical research: from product moment correlation to the odds ratio. *Statistical Methods in Medical Research 15, 525-545*.

KRISTOFF, W. (1970): A theorem on the trace of certain matrix products and some applications. *Journal of Mathematical Psychology 7, 515-530*, *cit. in* Le Calvé (1976b).

KRUSKAL, J. B. e WISH, M. (1978): *Multidimensional Scaling*. Sage, Beverly Hills, *cit. in* Everitt e Dunn (1997).

KRZANOWSKI, W.J. (1983): Distance between populations using mixed continuous and categorical variables. *Biometrika 70(1)*, *235-243*.

KUSIAK, A. e CHO, M. (1992): Similarity coefficient algorithms for solving the group technology problems. *International Journal of Production Research 30 (11), 2633-2646*, *cit. in* Sarker e Saiful Islam (1999).

LAPOINTE, F.-J. e LEGENDRE, P. (1992): Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees. *Systematic Biology 41(3)*, *378-384*.

LE CALVÉ, G. (1976a): *Problèmes d'Analyse des Données*. Thése d'État (2ème partie), Université de Rennes I, Rennes.

LE CALVÉ, G. (1976b): Quelques remarques sur certains aspects de l'analyse factorielle. In: *Cahier 2 du Laboratoire d'Analyse et de Traitement des Données en Sciences Humaines.* Université de Haute-Bretagne, Rennes II.

LE CALVÉ, G. (1977): Un indice de similarité pour des variables de type quelconque. *Revue Statistique et Analyse des Données 01/02*, *39-47.*

LE CALVÉ, G. (1985): Distances à centre. *Statistique et Analyse des Données 10(2), 29-44.*

LE CALVÉ, G. (1988): Similarities functions. In: D. Edwards e N.E. Raun (Eds.): *COMPSTAT'88*, Physica-Verlag, Heidelberg, 341-347.

LE CALVÉ, G. (1993): *Cours de DEA de Biostatistiques*. Université Montpellier 2, Montpellier.

LE CALVÉ, G. (2000): About some problems arising in sensory analysis and some propositions to solve them. *Food Quality and Preference 11, 341-347.*

LEBART, L., MORINEAU, A. e PIRON, M. (1995): *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

LEE, S.-C. e HUH, M.Y. (2003): A measure of association for complex data. *Computational Statistics and Data Analysis 44, 211-222*.

LEFORT-BUSON, M. e VIENNE, D. (Eds.) (1985): *Les Distances Génétiques. Estimations et Applications*. INRA, Paris.

LEGENDRE, P. (2000): Comparison of permutation methods for the partial correlation and partial Mantel tests. *Journal of Statistical Computation and Simulation 67, 37-73*.

LEGENDRE, P. e CHODOROWSKI, A. (1977): A generalization of Jaccard's association coefficient for Q-analysis of multi-state ecological data matrices. *Ekologia Poshha 25, 297-308*.

LEGENDRE, P. e LEGENDRE, L. (2000): *Numerical Ecology.* Elsevier, New York.

LEGENDRE, P. e TROUSSELLIER, M. (1988): Aquatic heterotrophic bacteria: modelling in the presence of spatial autocorrelation, *Limnology and Oceanography 33,1055-1067*.

LEGENDRE, P. LAPOINTE, F.-J. e CASGRAIN, P. (1994): Modeling brain evolution from behaviour: a permutational regression approach. *Evolution 48, 1487-1499.*

LERMAN, I.C. (1970): *Sur l'Analyse des Données Préalable à une Classification Automatique. Proposition d'une Nouvelle Mesure de Similarité*. MSH, rapport 32 (8e. année), Paris.

LERMAN, I.C. (1973): *Étude Distributionnelle de Statistiques de Proximité entre Structures Algébriques Finies de Même Type. Application à la Classification Automatique*. Cahiers du Bureau Universitaire de Recherche Opérationnelle, n. 19. Institut de Statistique de l'Université de Paris, Paris.

LERMAN, I.C. (1977): Formal analysis of a general notion of proximity between variables. In: J.R. Barra, F. Brodeau, G. Romier e B. Van Cutsem, (Eds.): *Recent Developments in Statistics: Proceedings of European Meeting of Statisticians.* Elsevier/North-Holland, Amsterdam, 787-795.

LERMAN, I.C. (1981): *Classification et Analyse Ordinale des Données*. Dunod, Paris.

LERMAN, I.C. (1987): Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification. *Revue de Statistique Appliquée XXXV (2), 39-60.*

LERMAN, I.C. (1992a): Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles I. *Mathématiques, Informatique et Sciences Humaines 118, 33-52.*

LERMAN, I.C. (1992b): Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles II. *Mathématiques, Informatique et Sciences Humaines 119, 75-100.*

LERMAN, I.C. (1999): Comparing classification tree structures: A special case of comparing q-Ary relations/comparaison d'arbres de classification: Un cas spécifique de la comparaison de relations q-aires. *RAIRO-Operations Research 33, 339-365.*

LERMAN, I.C. e PETER, P. (1985): Élaboration et logiciel d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application à la recherche d'un consensus en classification. *IRISA 262*.

LERMAN, I.C. e PETER, P. (2003): Indice probabiliste de vraisemblance du lien entre objets quelconques: Analyse comparative entre deux approches. *Revue de Statistique Appliquée I(1), 5-35.*

LEV, J. (1949): The point biserial coefficient of correlation. *Annals of Mathematical Statistics 20(1)*, *125-126*.

LINGOES, J. (1971): Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika 36, 195-203*.

LOVE, R.F. e MORRIS, J.G. (1972): Modelling intercity road distances by mathematical functions. *Operational Research Quaterly 23(1), 61-71.*

MANLY, B. (1992): *The Design and Analysis of Research Studies*. Cambridge University, Cambridge.

MANLY, B.F.J. (1994): *Multivariate Statistical Methods: a Primer*. 2nd Ed. Chapman & Hall, London.

MANN, H.B. e WHITNEY, D.R. (1947): On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *50-60*.

MANTEL, N. (1967): The detection of disease clustering and a generalised regression approach. *Cancer Research 27, 209-220*.

MARCOTORCHINO, J.-F. (1984): *Utilisation des Comparaisons par Paires en Statistiques des Contingences* (Partie 1 et 2). Étude du centre scientifique IBM France, Paris, *cit. in* Abdallah e Saporta (1998).

MARCOTORCHINO, J.-F. e MICHAUD, P. (1979): *Optimisation en Analyse Ordinale des Données*. Masson, Paris.

MARCOTORCHINO, J.-F. e MICHAUD, P. (1981): *Agrégation de Similarité en Classification Automatique*. Étude F–012, IBM, Paris, *cit. in* Beninel (1987).

MARDIA, K.V. (1978): Some properties of classical multi-dimensional scaling. *Communications in Statistics – Theory and Methods 47(13), 1233-1241.*

MARDIA, K.V., KENT, J.T. e BIBBY, J.M. (1979): *Multivariate Analysis*. Academic Press, London.

MATUSITA, K. (1951): Decision rules based on distance for problems of fit, two samples and estimation. *Annals of Mathematical Statistics 3*, *1-30*.

MATUSITA, K. (1955): On the theory of statistical decision functions. *Annals of Mathematical Statistics 26*, *631-640*.

MATUSITA, K. (1956): Decision rule, based on the distance, for the classification problem. *Annals of the Institute of Statistical Mathematics 8, 67-77, cit. in* Krzanowski (1983).

MATUSITA, K. (1977): Cluster analysis and affinity of distributions. In: J.R. Barra, F. Brodeau, G. Romier e B. Van Cutsem, (Eds.): *Recent Developments in Statistics: Proceedings of European Meeting of Statisticians.* Elsevier/North-Holland, Amesterdam, 537-544.

MEULMAN, J.J. e HEISER, W.J. (1999): *SPSS Categories 10.0*. SPSS Inc., Chicago.

MEULMAN, J.J., VAN DER KOOIJ, A.J. e HEISER, W.J. (2004): Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In:

D. Kaplan (Ed.): *Handbook of Quantitative Methodology for the Social Sciences*. Sage, London, 49-70.

MIRKIN, B. (2008): The iterative extraction approach. In: A.N. Gorban *et al.* (Eds.): *op. cit.*, 151-177.

NEI, M. (1978): Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics 89, 583-590, cit. in* Beninel (1987).

NICOLAU, F.C. (1980): *Critérios de Análise Classificatória Hierárquica Baseados na Função de Distribuição*. Tese de doutoramento, Faculdade de Ciências da Universidade de Lisboa, Lisboa.

NICOLAU, F.C. (1981): Análise classificatória e função de distribuição. In: Departamento de Matemática da Universidade de Coimbra (Org.): *Actas das VIII Jornadas Luso-Espanholas de Matemática.* Vol. II. Universidade de Coimbra, Coimbra, 369-377.

NICOLAU, F.C. (1994): *Estudo Distribucional de Coeficientes de Semelhança por Simulação*. Relatório Interno, LADERF / Unidade de Biometria. Universidade de Aveiro, Aveiro.

NICOLAU, F.C. e BACELAR-NICOLAU, H. (1981): Nouvelles méthodes d'agrégation basées sur la fonction de répartition. In: *Collection Séminaires INRIA, Classification Automatique et Perception par Ordinateur.* Versailles, 45-60.

NICOLAU, F.C. e BACELAR-NICOLAU, H. (1998): Some trends in the classification of variables. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock e Y. Baba (Eds.): *Data Science, Classification and Related Methods*. Springer, Heidelberg, 89-98.

NICOLAU, F.C. e BACELAR-NICOLAU, H. (1999): Clustering symbolic objects associated to frequency or probability laws by the weighted affinity coefficient. In: H. Bacelar-Nicolau, F.C. Nicolau e J. Janssen (Eds.): *Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society*. Instituto Nacional de Estatística, Lisboa, 155-158.

NICOLAU, F.C. e BACELAR-NICOLAU, H. (2003): Teaching and learning hierarchical clustering probabilistic models for categorical data. In: *Bulletin of the International Statistical Institute, Proceedings of the 54th Session*, 346-349.

NICOLAU, F.C. e SOROMENHO, G. (1988): *Manual de Utilização do Programa para a EXEC Cluster de Análise Classificatória Hierárquica*. Notas e Comunicações do CEAUL, 12. Universidade de Lisboa, Lisboa.

NICOLAU, F.C., BACELAR-NICOLAU, H., SOUSA, A., BACELAR-NICOLAU, L., SILVA, O. e MAGALHÃES, B (2007): Probabilistic models in three way cluster analysis. In: *Bulletin of the International Statistical Institute, Proceedings of the 56th Session* (CD Proceedings of ISI 2007).

NICOLAU, L. (2002): *Caracterização dos Sistemas de Informação das Organizações com Base no Modelo de Nolan – Aplicação de Modelos de Classificação Hierárquica aos Organismos da Administração Pública*. Tese de mestrado, ISEGI, Universidade Nova de Lisboa, Lisboa.

NICOLAU, L. e BACELAR-NICOLAU, H. (2005): Métodos de classificação hierárquica e análise de perfis. Um estudo de caso em educação médica. In: C. Braumann, P. Infante, M.M. Oliveira, R. Alpízar-Jara, e F. Rosado (Eds.): *Estatística Jubilar – Actas do XII Congresso Anual da Sociedade Portuguesa de Estatística*. Edições SPE, 489-496.

NIITSUMA, H. e OKADA, T. (2005): Covariance and PCA for categorical variables. In: T.B. Ho, D. Cheung e H. Liu (Eds.): *Advances in Knowledge Discovery and Data Mining* (*PAKDD 2005)*. Springer-Verlag, Berlin Heidelberg*,* 523-528.

ORLÓCI (1978): *Multivariate Analysis in Vegetation Research*. W. Junk, The Hague. *cit. in* Legendre e Legendre (2000).

OUALLI ALLAH, M. (1991a): *Analyse en Préordonnances des Données Qualitatives – Applications aux Données Numériques et Symboliques*. Thése de doctorat, Université de Rennes 1, Rennes.

OUALLI ALLAH, M. (1991b): AVARE: Un programme de calcul des associations entre variables relationnelles. *Publication Interne IRISA 591*.

PAGÈS, J. (2002): Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de Statistique Appliquée 50(4), 5-37*.

PAGÈS, J. (2004): Analyse factorielle de données mixtes. *Revue de Statistique Appliquée 52(4), 93-111*.

PEARSON, K. (1901): On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 6(2), 559-572*.

PEARSON, K. (1905): Skew variation, a Rejoinder. *Biometrika 4*, *169-212*, *cit. in* Yule e Kendall (1950).

PEARSON, K. e HERON, D. (1913): On theories of association. *Biometrika, 9, 159-315.*

PERRIER, X. (1998): *Analyse de la Diversité Génétique: Mesures de Dissimilarités et Représentations Arborées.* Thèse de 3ème cycle, Université de Montpellier II, Montpellier.

RAJASEKARAN, S., THAPAR, V., DAVE H. e HUANG, C.-H. (2004). A randomized algorithm for distance matrix. Calculations in multiple sequence alignment. In: J.A. López, E. Benfenati e W. Dubitzky (Eds.): *Knowledge Exploration in Life Science Informatics (KELSI 2004)*. Springer-Verlag, Berlin Heidelberg, 33-45.

RAMAKRISHNAN, S. e SELVAN, S. (2006): A new statistical model based on wavelet domain singular value decomposition for image texture classification. *GVIP Journal 6(3), 15-22*.

RAO, C.R. (1964): The use and interpretation of principal component analysis in applied research. *Sankhyā A 26, 329-358.*

RODRIGUES, P.C. e BRANCO, J.A. (2007): Análise de componentes principais sobre dados dependentes. In: M.E. Ferrão, C. Nunes e C. Braumann (Eds.): *Estatística: Ciência Interdisciplinar – Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística*. Edições SPE, 653-662.

ROHLF, F.J. (1982): Consensus indices for comparing classifications. *Mathematical Biosciences 59, 131-144.*

ROSENBERG, S. (1982): The method of sorting in multivariate research with applications selected from cognitive psychology and person perception. In: N. Hirchberg e L.G. Humphreys (Eds.): *Multivariate Applications in the Social Sciences*. L. Erlbaum Assoc., University of Illinois, Urbana-Champaign, 117-142, *cit. in* Mirkin (2008).

ROUX, G. e ROUX, M. (1980): A propos de quelques méthodes de classification en phytosociologie. In: J.-P. Benzécri (Ed.) (1980a), *op. cit.,* tomo 1, 360-374.

SAMMON, J.W. (1969): A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers C-18 (5), 401-409*, *cit. in* Van Eck e Waltman (2006).

SAPORTA, G. (1990): *Probabilités, Analyse des Données et Statistiques*. Éditions Technip, Paris.

SARKER, B.R. e SAIFUL ISLAM K.M. (1999): Relative performances of similarity and dissimilarity measures. *Computers and Industrial Engineering 37, 769-807*.

SATO-ILIC, M. e OSHIMA, J. (2006): On weighted principal component analysis for interval-valued data and its dynamic feature. *International Journal of Innovative Computing, Information and Control 2(1)*, *69-82.*

SCHOENBERG, I.G. (1937): On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in Hilbert space. *Annals of Mathematics 38(4), 787-793.*

SCHWARZ, N. (1999): How the questions shape the answers. *American Psychologist*, *93-105.*

SHANNON, W.D., WATSON, M.A., PERRY, A. e RICH, K. (2002): Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genetic Epidemiology 23, 87-96.*

SHEPARD, R.N. (1962a): The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika 27, 125-140*.

SHEPARD, R.N. (1962b): The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika 27, 219-246*.

SIEGEL, S. e CASTELLAN JR., N.J. (1989): *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Singapore.

SILVA, A.L. (2005): *Tratamento de Dados Omissos e Métodos de Imputação em Classificação*. Tese de doutoramento, ISEG-UTL, Lisboa e CNAM, Paris.

SOARES, A. S. (1999): *Medidas e Modelos de Concordância em Dados Categorizados*. Tese de Mestrado, Faculdade de Ciências, Universidade de Lisboa, Lisboa.

SOHAIL KHALID, M., UMAR ILYAS, M., SAQUIB SARFARAZ, M. e ASIM AJAZ, M. (2006): Bhattacharyya coefficient in correlation of gray-scale objects. *Journal of Multimedia 1(1)*, *56-61*.

SOROMENHO, G. e BACELAR-NICOLAU, H. (1999): The weighted affinity coefficient in Gaussian mixtures models. In: H. Bacelar-Nicolau, F. Costa Nicolau, J. Janssen (Eds.): *Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society*. Instituto Nacional de Estatística, Lisboa, 148-154.

SOUSA FERREIRA, A. (2000): *Combinação de Modelos em Análise Discriminante sobre Variáveis Qualitativas*. Tese de doutoramento, Universidade Nova de Lisboa, Lisboa.

SOUSA FERREIRA, A., CELEUX G. e BACELAR-NICOLAU, H. (2001): New developments on combining models in discrete discriminant analysis by a hierarchical coupling approach. In: G. Govaert, J. Janssen e N. Limnios (Eds.): *Applied Stochastic Models and Data Analysis (ASMDA 2001)*. Université de Technologie de Compiègne, Compiègne, 430-435.

SOUSA FERREIRA, A., DORIA, I., DIAS, O. e BACELAR-NICOLAU, H. (2003): Estudo de um questionário de cuidados paliativos em contexto domiciliário através de métodos de análise de dados multivariados. In: *Livro de Resumos das X Jornadas de Classificação e Análise de Dados (JOCLAD 2003), 112-113.*

SOUSA, A. (2003): *Programa Cluster – Manual do Utilizador*. Departamento de Matemática, Universidade dos Açores*,* Ponta Delgada.

SOUSA, A. (2005): *Contribuições à Metodologia VL e Índices de Validação para Dados de Natureza Complexa*. Tese de doutoramento, Universidade dos Açores, Ponta Delgada.

SOUSA, A., BACELAR-NICOLAU, H. e NICOLAU, F.C. (2007): A metodologia VL e a validação em análise classificatória de dados de natureza complexa: uma aplicação a dados reais. In: M.E. Ferrão, C. Nunes e C. Braumann (Eds.): *Estatística: Ciência Interdisciplinar – Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística*. Edições SPE, 775-788.

SOUSA, A., SILVA, O., BACELAR-NICOLAU, H. e NICOLAU, F. (2003): Estudo distribucional empírico do coeficiente de afinidade com base em simulações. In: *Livro de Resumos das X Jornadas de Classificação e Análise de Dados (JOCLAD 2003),* 128-129.

SOUSA, A., SILVA, O., BACELAR-NICOLAU, H. e NICOLAU, F. (2005): Estudo distribucional empírico do coeficiente de afinidade com base em simulações. In: C.

Braumann, P. Infante, M.M. Oliveira, R. Alpízar-Jara, e F. Rosado (Eds.): *Estatística Jubilar – Actas do XII Congresso Anual da Sociedade Portuguesa de Estatística*. Edições SPE, 767-776.

SOUSA, F. (2000): *Novas Metodologias e Validação em Classificação Hierárquica Ascendente*. Tese de doutoramento, Universidade Nova de Lisboa, Lisboa.

SOUSA, F. e NICOLAU, F.C. (2002): Validação em classificação hierárquica ascendente – alguns resultados. In: L. Carvalho, F. Brilhante, e F. Rosado (Eds.): *Novos Rumos em Estatística – Actas do IX Congresso Anual da Sociedade Portuguesa de Estatística.* Edições SPE, 403-413.

STREINER, D.L. e NORMAN, G.R. (2003): *Health Measurement Scales. A Practical Guide to Their Development and Use*. Oxford University Press, Oxford.

TÉROUANNE, E. (1998) Corrélation entre variables nominales, ordinals, métriques ou numériques. *Mathématiques, Informatique et Sciences Humaines 142, 5-16*.

TIAGO DE OLIVEIRA, J. (1982): The δ-method for obtention of asymptotic distributions: applications. *Publications de l'Institut de Statistique de l'Université de Paris, XXVII, 49-70*, *cit. in* Bacelar-Nicolau (1980).

TORGERSON, W.S. (1952): Multidimensional scaling: I. Theory and method. *Psychometrika 17, 401-419.*

TORGERSON, W.S. (1958): *Theory and Methods of Scaling*. Wiley, New York. *cit. in* Joly e Le Calvé (1986).

VAN ECK, N.J. e WALTMAN, L. (2006): *VOS - A New Method for Visualizing Similarities Between Objects*. Technical Report ERS-2006-020-LIS. Erasmus Research Institute of Management, Erasmus University, Rotterdam.

VIDAL, R. (2003): *Generalized Principal Component Analysis (GPCA)*. PhD thesis, University of California, Berkeley.

WALD, A. e WOLFOWITZ, J. (1944): Statistical tests based on permutations of the observations. *Annals of Mathematical Statistics 15, 358-372*.

WANG, W. e WELLS, M.T. (2000): Estimation of Kendall's tau under censoring. *Statistica Sinica 10, 1199-1215*.

WIWANITKIT, V. (2004): Dengue virus nonstructural-1 protein and its phylogenetic correlation to human fibrinogen and thrombocytes: A study to explain hemorrhagic complications. *The Internet Journal of Genomics and Proteomics 1(2)* (*disponível em* http://www.ispub.com/ostia/index.php?xmlFilePath=journals/ijgp/vol1n2/dengue.xml)

WOOD, J. M. (2007): Understanding and computing Cohen's Kappa: A tutorial. *WebPsychEmpiricist.* Web Journal (*disponível em* http://wpe.info/)

YAGUCHI, H. e ICHINO, M. (1992): A generalized principal component analysis for mixed measurement level data. *Trans. IEICE* Japan *J75-A (10), 1580-1589* (em japonês no original) *cit. in* Ichino e Yaguchi (1994).

YOUNESS, G. e SAPORTA, G. (2004): Une méthodologie pour la comparaison de partitions. *Revue de Statistique Appliquée 52(1)*, *97-120*.

YULE, G.U. (1912): On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society 75,* 579-642.

YULE, G.U. e KENDALL, M.G. (1950): *An Introduction to the Theory of Statistics*. Charles Griffin, London.