Agradecimentos

Para alcançar os nossos objectivos temos que ser determinados. A convicção leva-nos a construir com confiança; a inconformidade detecta oportunidades; a decisão leva-nos à concretização dos nossos objectivos e dos nossos sonhos.

Cada aprendizagem é marcada por estágios e cada um traz as suas lições. Cada um de nós deve aproveitar os ensinamentos que nos são transmitidos pelos Professores e Orientadores. Este trabalho não teria sido possível, sem o apoio, os ensinamentos, a amizade e a dedicação de algumas pessoas. Quero deste modo, expressar a todos aqueles que de muitas formas o fizeram, o meu reconhecimento e sincero agradecimento.

Ao Professor Doutor Alberto António Caria Canelas Pais e ao Professor Doutor Jorge Luís Gabriel F.S. Costa Pereira pela orientação, disponibilidade e paciência, e pelos ensinamentos que me transmitiram. O trabalho e a pesquisa andam sempre a par, trazendo sempre algo de novo. Muito obrigado a ambos pelo entusiasmo contagiante; os bons hábitos podem levar a grandes êxitos. Podemos ser grandes em tudo o que fazemos e sermos especiais para quem nos rodeia.

Ao Professor Doutor Sebastião J. Formosinho Sanches Simões, pelo exemplo; a excelência não é um acto isolado, é uma arte que se alcança através do treino e do hábito, é nosso dever fazê-lo repetidamente, seguindo o exemplo.

A todos os colegas e amigas de curso, obrigado pela cumplicidade, paciência e ajuda em muitos momentos. Uma amizade como a nossa conhece o tempo exacto das palavras e dos silêncios. A amizade deixa-nos caminhar à vontade lado a lado; estimula e partilha a nossa maneira de olhar a vida.

À minha família, especialmente aos meus pais e avós que sempre me apoiaram, acreditaram em mim e respeitaram as minhas escolhas.

Se a primeira experiência é fundamental para formar uma opinião, se o presente é o reflexo do passado e o futuro o reflexo do presente, só posso dizer que este trabalho foi muito importante na minha formação, em grande parte devido aos Profissionais que encontrei.

Aos meus pais

"O conhecimento é o processo de acumular dados,

a sabedoria reside na sua simplificação."

Martin H. Fischer

Resumo

Este trabalho pretende mostrar a aplicabilidade dos métodos quimiométricos no tratamento e interpretação de vários tipos de problemas. Visa também, desenvolver e disponibilizar uma ferramenta de análise multivariada abrangente que possa ser executada de forma simples por vários utilizadores com ou sem formação especifica nesta área, assim como, evidenciar algumas características e limitações das ferramentas e procedimentos implementados.

Uma das principais contribuições das técnicas de tratamento da informação dirigida ao usuário é claramente auxiliar no diagnóstico médico. Estas técnicas são especialmente úteis para promover a detecção precoce e a utilização de métodos menos invasivos que reduzam os riscos e o sofrimento dos pacientes.

A análise multivariada é obrigatória quando vários parâmetros são obtidos em ensaios, e na maioria dos casos, a inspecção de todo o perfil de dados é claramente mais informativo que as avaliações parâmetro a parâmetro.

A nossa abordagem está relacionada com a utilização de algumas técnicas quimiométricas bem conhecidas, para avaliar o impacto das variáveis na descrição do sistema e procurar os melhores descritores que permitam encontrar os indicadores de desenvolvimento potencial e o grau de progressão da doença.

Especificamente, neste trabalho, abordamos uma variedade de dados relativos ao diagnóstico médico e a estudos epidemiológicos e mostramos que usando apenas estes métodos padrão, após uma selecção cuidadosa da abordagem para cada caso, é possível facilmente (i) separar classes ou categorias, recorrendo tanto a técnicas não supervisionadas como a técnicas supervisionadas, (ii) identificar redundâncias e relacionar variáveis, (iii) isolar factores e identificar padrões, incluindo marcas geográficas.

Os resultados mostram que os quatro métodos multivariados, a análise de agrupamento (HCA), a análise de componentes principais (PCA), a análise discriminante linear (LDA) e o método dos mínimos quadrados parciais (PLS), permitem tirar conclusões complementares e distinguem correctamente os grupos existentes. De facto, o procedimento descrito neste trabalho é totalmente automatizado e atinge um grau de confiabilidade comparável às abordagens mais sofisticadas.

Abstract

This study shows the applicability of chemometrics methods in the treatment and interpretation of several types of problems. It also aims to develop and provide a comprehensive multivariate analysis tool that can be simply performed by multiple users with or without specific training in this area, as well as highlight some features and drawbacks of the tools and the implemented procedures.

One of the key contributions of user-directed information treatment techniques is clearly directed to help in medical diagnosis. These are especially useful for promoting early detection and less invasive methods that reduce risks and patients suffering.

Multivariate analysis is mandatory when several parameters are obtained in assays, and in most cases the inspection of the whole profile of data is clearly more informative that parameter-by-parameter assessments.

Our approach is related to the use of some widespread chemometrics techniques to evaluate the impact of variables in the description of the system, and also to find out the best descriptors for achieving a ranking (scores) of indicators for potential disease development and degree of progression.

Specifically, in this work, we address a variety of data pertaining to medical diagnosis and epidemiological studies and show that using only these standard methods, after careful selection of the approach for each case, it is possible to easily (i) separate classes or ranks, resorting both to unsupervised an supervised techniques, (ii) pinpoint redundancies and link variables, (iii) isolate factors and identify patterns, including geographical marks.

Results show that the four multivariate methods, such as hierarchical cluster analysis (HCA), principal components analysis (PCA), discrimination analysis (DA) and partial least squares (PLS), can distinguish the existing groups correctly. Indeed, the procedure described here is completely automated and attains a degree of reliability comparable to the most sophisticated approaches.

Lista de abreviaturas

ANOVA – Analysis of Variance

CA – Cluster Analysis

CP – Componente Principal

ESI– Essential Sciences Indicators

LF – Latent Factor

HCA – Hierarchical Cluster Analysis

KNN – K-Nearest Neighbor

LDA – Linear Discriminant Analysis

OLS – Ordinary Least Squares

PCA – Principal Component Analysis

PC_i – Principal Component

PCR–Principal Component Regression

PLS – Partial Least Squares

PLS-DA – Partial Least Squares Discriminant Analysis

QDA – Quadratic Discriminant Analysis

NR – Noise Reduction

Rk – Ranking

UPDRS – Unified Parkinson´s Disease Rating Scale

Índice

Capítulo 1

Introdução

A maior dificuldade de se trabalhar com a informação não é a aquisição dos dados, mas sim encontrar informações úteis dentro da grande quantidade disponível nas diversas áreas.

Em qualquer decisão que tomamos nas nossas vidas, sempre levamos em conta um grande número de factores. Obviamente nem todos pesam da mesma maneira na hora de uma escolha. Quando tomamos uma decisão usando a intuição, não identificamos de maneira sistemática estes factores, nem quais as variáveis que afectaram a nossa decisão.

O grande volume desses dados em diversos meios e para diferentes domínios, tem desafiado a aptidão do ser humano em interpretar e compreender toda essa "mega" informação. Neste contexto, diversas ferramentas têm sido propostas e utilizadas para a extracção e interpretação da informação mais relevante de vários tipos de dados.

Uma grande quantidade de informação deve ser processada antes de ser transformada em conhecimento. Neste contexto, a Análise Multivariada corresponde a um grande número de métodos e técnicas que utilizam simultaneamente todas as variáveis na interpretação teórica do conjunto de dados. A necessidade de compreensão das relações entre as diversas variáveis faz com que a análise seja complexa ou até mesmo difícil.

Um dos objectivos deste trabalho é evidenciar a utilidade de alguns métodos de análise multivariada usando exemplos concretos.

Existem vários métodos de análise multivariada com finalidades distintas entre si. Quando o interesse é verificar como um conjunto de elementos se relacionam, o quanto estes são semelhantes segundo as variáveis utilizadas, destacam-se dois métodos: a análise de agrupamento hierárquico *(HCA[1])* e a análise de componentes principais *(PCA[2])* [1].

A extracção de informações dos dados envolve a análise de um grande número de variáveis. Muitas vezes, um pequeno número destas variáveis contém as informações mais relevantes, enquanto a maioria das variáveis adiciona pouco ou nada à interpretação dos resultados. A decisão sobre quais as variáveis mais importantes é feita, geralmente, com base na intuição ou na experiência, baseado em critérios que são mais subjectivos que objectivos.

[1] Do inglês *Hierarchical Cluster Analysis*
[2] Do inglês *Principal Component Analysis*

Assim, a redução de variáveis através de critérios objectivos, permitindo a construção de gráficos bidimensionais contendo maior informação estatística, pode ser conseguida através da análise de componentes principais. Também é possível construir agrupamentos entre os objectos de acordo com as suas similaridades, utilizando todas as variáveis disponíveis, e representá-los de forma bidimensional através de um dendrograma. A análise de componentes principais e de agrupamento hierárquico são, portanto, técnicas complementares [1, 2].

A crescente utilização dos métodos de análise multivariada é explicada quer pela diversidade de programas estatísticos que incluem estas metodologias, quer pela necessidade frequente de tratar, isto é resumir, grandes quantidades de dados. Actualmente, é mais fácil e económico possuir grandes bases de dados e, por isso, torna-se essencial encontrar ferramentas que retirem delas a informação relevante. Estes métodos desempenham um papel importante no tratamento da informação, sendo utilizados em áreas muito distintas que vão das Ciências Sociais à Química e Biologia, Medicina, ou das Ciências Económicas à Engenharia.

Actualmente, a análise multivariada de dados é uma realidade com a qual temos que lidar no sentido de processar e interpretar a informação mais relevante, relacionada com enormes conjuntos de dados. Muitos estudos têm sido desenvolvidos nesse sentido, começando numa fase inicial com a análise estatística descritiva, prosseguindo com diversas ferramentas sensíveis à inter-relação dos dados.

A Quimiometria desenvolveu-se inicialmente no sentido de dar resposta a questões relacionadas com o controlo da qualidade [3]. Grandes progressos foram rapidamente implementados, sobretudo na descrição de sistemas e respectiva modelação. Veja-se por exemplo os trabalhos pioneiros de *M. A. Sharaf* [4], *R. G. Brereton* [5] e *B.R. Kowalski* [6]. Desde então, a Quimiometria tem vindo a desenvolver-se cada vez mais, ultrapassando as fronteiras da química [3,7-9]. Contudo, surge uma dificuldade. Devido à enorme diversidade de algoritmos e suas variantes, à complexidade crescente e à multiplicidade de linguagens usadas, é difícil implementar soluções coerentes, eficazes e inovadoras, perdendo-se alguma informação. Além disso, acreditamos que existe um conjunto de ferramentas padrão que ainda não estão totalmente desenvolvidas, compreendidas e exploradas. Por esta razão, entendemos que a diversidade disponível não é um benefício claro e um maior esforço deveria ser investido no entendimento destas ferramentas. Por exemplo, a análise de componentes principais (PCA) surgiu em 1901 [3] e é uma ferramenta de compressão da informação multivariada, que permite extrair ainda mais informação, além da representação tridimensional da informação contida no

sistema.

Este trabalho procura demonstrar a aplicabilidade de alguns algoritmos clássicos no tratamento e interpretação da informação mais relevante de sistemas que ultrapassam a competência da química. Como exemplos, escolhemos algumas bases de dados relacionadas com o diagnóstico preliminar de doenças, como o cancro e a doença de Parkinson. Especificamente, procurámos combinar quatro ferramentas quimiométricas (HCA, PCA, LDA e PLS) para compreender qual a ferramenta e a abordagem mais promissora para extrair informações relevantes do sistema, permitindo a simplificação da previsão e a compressão da informação.

Como mencionado anteriormente, vamos apresentar e discutir alguns métodos padrão utilizados e os resultados disponíveis no campo do diagnóstico médico, nomeadamente no diagnóstico do cancro.

Em primeiro lugar, são apresentadas algumas estratégias para o tratamento e interpretação dos nossos casos de estudo, especialmente neste campo.

O cancro da mama é o tipo de cancro mais comum em mulheres, com cerca de uma em cada dez mulheres a desenvolver a doença. Este tipo de cancro, é a principal causa de morte por cancro em mulheres entre os 35 e os 54 anos de idade. O diagnóstico do cancro da mama baseia-se essencialmente no auto-exame, ou no exame clínico, na mamografia, ultra-som da mama (em conjunto com a mamografia) e na biopsia. Contudo, a ferramenta mais eficaz para combater o cancro da mama é a detecção precoce [8,9].

Actualmente, o cancro da mama é um tumor comum em mulheres em todo o mundo e representa a segunda causa de morte mais frequente por cancro, depois do cancro do pulmão. No entanto, quanto mais cedo o diagnóstico é feito, maiores são as probabilidades de cura, e a necessidade de remoção do peito pode ser evitada [10-12].

Os testes clínicos e os estudos epidemiológicos, produzem frequentemente grandes quantidades de dados de natureza multivariada. As técnicas clássicas de quimiometria fornecem a maior parte dos ingredientes para resolver este tipo de problemas. Em primeiro lugar, estes métodos permitem representar graficamente os dados multivariados, preservando as informações mais relevantes. Este é um ponto extremamente importante na avaliação exploratória de um determinado sistema. Em segundo lugar, é relativamente fácil realizar uma análise discriminante para classificação e estabelecer regras de decisão. Por último, estes métodos fornecem ferramentas para avaliar o impacto das variáveis na descrição do sistema e, assim, classificar e seleccionar as variáveis relevantes para a progressão e diagnóstico da

doença. O mesmo, pode ser transposto também para os estudos epidemiológicos, usando as mesmas ferramentas e alterando apenas o contexto.

A metodologia proposta foi demonstrada para (1) o problema da determinação do cancro da mama *(Winsconsin breast cancer)*, que apresenta nove variáveis associadas a características citológicas, (2) o problema de classificação de tecidos mamários, com dez variáveis associadas a parâmetros característicos inferidos de espectros de impedância, (3) a incidência de novos casos de cancro considerando apenas onze tipos de cancro seleccionados nos estados dos EUA, (4) avaliação do grau de progressão da doença de Parkinson, tendo em conta dezasseis medidas de voz, e por último (5) o estudo comparativo entre cinquenta e cinco universidades ibéricas, tendo em conta as suas áreas de *ranking*.

Este trabalho foi estruturado da seguinte forma: nos capítulos 2 e 3 encontram-se descritos os métodos quimiométricos clássicos (HCA, PCA, LDA e PLS) usados nesta abordagem e algumas das aplicações mais recentes no diagnóstico médico, assim como uma breve descrição da *toolbox* desenvolvida. O capítulo 4 apresenta de forma resumida a nossa abordagem, combinando estes métodos com o algoritmo para a limpeza de ruído/detecção de *outliers* [13]. No capítulo 5 são apresentados e discutidos os exemplos considerados e, por último, são apresentadas, no capítulo 6, as considerações finais.

Capítulo 2

Fundamentação teórica

Nesta secção são abordados alguns conceitos relativos à evolução da Quimiometria e à importância dos seus métodos no tratamento de vários tipos de problemas. Encontra-se também, uma descrição dos métodos quimiométricos utilizados no desenvolvimento deste trabalho assim como, algumas das suas principais aplicações de acordo com a literatura mais recente.

2.1 Quimiometria

O termo "quimiometria" foi proposto no final dos anos 70 para descrever as técnicas e operações associadas ao tratamento matemático e à interpretação de dados químicos. São considerados pais da Quimiometria o Professor Bruce Kowalski, da Universidade de Washington, já reformado, e o Professor Svante Wold, da Suécia, ainda em actividade [14].

A Quimiometria surgiu em resposta à necessidade de desenvolver novos métodos matemáticos e estatísticos para lidar com a vasta quantidade de dados produzida pelos instrumentos analíticos modernos.

Permite tratar dados mais complexos, que requerem a utilização de técnicas estatísticas multivariadas, álgebra matricial e análise numérica [15]. Envolve diferentes métodos, tais como a optimização e validação de metodologias analíticas, o planeamento experimental, a estimativa de parâmetros, o processamento de sinal, a análise de factores e também, a calibração multivariada.

Na última década, tem sido desenvolvida e utilizada em várias áreas da ciência. Tem uma vasta aplicação nas diferentes áreas da química (por exemplo, química analítica, química orgânica, química forense), na área alimentar, farmacêutica, área ambiental, agricultura e química industrial.

Na química medicinal, a Quimiometria é o grande esteio dos estudos QSAR nos quais se relaciona a estrutura molecular dos compostos com a sua actividade biológica. Na indústria, é usada para monitorização de processos e controlo de qualidade. Na área da química

ambiental, os métodos de ordem superior também se mostram úteis em estudos de poluição para a identificação de fontes [16].

Na área alimentar os métodos de análise exploratória e de reconhecimento de padrões são utilizados para discriminar diversos tipos de alimentos. Utilizando métodos de regressão é possível detectar e quantificar simultaneamente os teores de vários compostos usando dados de espectroscopia [17].

A disponibilidade generalizada de computadores mais potentes e menos dispendiosos contribuiu para a rápida evolução dos métodos quimiométricos, ao permitir uma análise mais flexível de grandes conjuntos de dados multivariados, o desenvolvimento de algoritmos computacionais mais eficientes e a difusão de software quimiométrico.

2.2 Métodos quimiométricos clássicos

Com a modernização das técnicas instrumentais de análise química, tornaram-se necessários métodos de tratamento de dados mais complexos do ponto de vista matemático e estatístico. Há muito tempo que a estatística univariada é aplicada a problemas químicos, mas a sua utilização tornou-se limitada. Nas últimas décadas, a análise multivariada foi introduzida no tratamento de dados químicos, aumentando a sua popularidade e dando origem à Quimiometria [18, 19].

A análise multivariada permite: a redução de dados ou simplificação estrutural, em que o fenómeno em estudo é representado da forma mais simples possível, sem sacrificar informações relevantes, tornando as interpretações mais simples; a ordenação e agrupamento de objectos ou variáveis similares; a investigação da dependência entre variáveis; a previsão, dado que, as relações entre variáveis devem ser determinadas para a previsão de uma ou mais variáveis com base na observação de outras variáveis; por último, a construção e teste de hipóteses.

O modelo estatístico dos métodos multivariados considera a correlação entre muitas variáveis analisadas simultaneamente, permitindo a extracção de uma quantidade muito maior de informação [19].

Os métodos quimiométricos possuem um enorme potencial no tratamento de diversos tipos de problemas. No entanto, até à segunda metade dos anos 80, a sua presença na literatura era escassa. A partir dos anos 90 nota-se um aumento significativo das suas aplicações, levando ao aumento da capacidade dos investigadores em extrair informações dos

dados. Está em curso um processo de substituição de muitos métodos tradicionais de análise univariada por métodos multivariados [18]. Este processo ainda está no início sendo necessário aumentar a sua divulgação, para que cada vez mais investigadores tomem contacto com esta ferramenta.

Nesta secção, são apresentadas algumas contribuições para o desenvolvimento destes métodos, com base na literatura mais recente.

Em quase todas as áreas de pesquisa várias variáveis são mensuradas e, em geral, devem ser analisadas conjuntamente. A análise multivariada trata desse tipo de estudo e existem vários métodos, cuja utilização depende do tipo de dados que se deseja analisar e dos objectivos da pesquisa.

Na literatura, são publicadas regularmente revisões detalhadas sobre a aplicação dos métodos quimiométricos em várias áreas do conhecimento. É difícil descrever em detalhe cada contribuição histórica para os primeiros anos da quimiometria. Contudo, essa informação pode ser consultada essencialmente em algumas referências clássicas [20-29]. Em geral, estes métodos têm sido aplicados com sucesso na visualização dos dados, na classificação, na resolução de curvas multivariadas e na predição em química analítica, química ambiental, engenharia, investigação médica e na indústria [30-36]. Recentemente, algumas abordagens mais complexas têm sido propostas para auxiliar em estudos de desenvolvimento como a genómica, proteómica, bioinformática e a metabonómica [37-50]. Embora muitas publicações apresentem métodos novos ou optimizados para o processamento de dados, é muito raro que esses métodos sejam comparados com os métodos mais estabelecidos [30]. Duas referências que fornecem uma visão geral dos métodos quimiométricos clássicos são as referências [51,52].

2.2.1 Métodos de classificação não supervisionados

A análise de componentes principais (PCA) é provavelmente a mais antiga e mais conhecida das técnicas de análise multivariada [3]. A ideia principal da análise de componentes principais é reduzir a dimensionalidade de um conjunto de dados com um grande número de variáveis inter-relacionadas, mantendo tanto quanto possível a variabilidade presente no conjunto de dados. Esta redução, é possível com a transformação num novo conjunto de variáveis, as componentes principais, não correlacionadas. Estas são ordenadas de forma que as primeiras retêm a maior variação presente nas variáveis originais [5, 53]. Normalmente, não é necessário considerar mais do que as primeiras 10 componentes. Por outras palavras, este método permite a representação do conjunto de dados original num novo sistema de referência caracterizado pelas novas variáveis. Cada componente (PC) tem a propriedade de explicar a maior variabilidade possível, contida no conjunto de dados original. As componentes principais (PCs), que são expressas como combinações lineares das variáveis originais, são ortogonais entre si e podem ser usadas para uma representação mais eficiente do sistema em estudo, com um número menor de variáveis do que na situação original. As coordenadas das amostras no novo sistema de referência são designadas por *scores*, enquanto o coeficiente da combinação linear que descreve cada PC, isto é, os pesos das variáveis originais em cada PC, são denominados por *loadings*.

A definição e o cálculo das componentes principais são simples, mas esta técnica, aparentemente simples, tem uma grande variedade de aplicações, assim como diferentes variantes [3].

Este método é recomendado como uma ferramenta exploratória para encontrar tendências em dados desconhecidos. É um método simples, não paramétrico, de extracção de informações relevantes a partir de conjuntos de dados multivariados [54,55].

A visão geral dos dados é uma das grandes categorias da análise de dados, especialmente relevante no desenvolvimento de ferramentas de diagnóstico de doenças. Neste contexto, a análise de componentes principais é primeiramente usada para fornecer uma visão geral dos dados, que irá revelar *outliers*, grupos e tendências nos dados [56,57].

Na investigação médica, particularmente no diagnóstico de cancro, a análise de componentes principais tem sido utilizada no processamento de dados, para extrair as componentes principais (variáveis significativas) e para estabelecer a classificação dos pacientes com cancro e de pessoas normais [30].

Relativamente aos métodos de agrupamento *(clustering),* o objectivo é dividir os dados em grupos de objectos semelhantes, utilizando um método de aprendizagem não supervisionado. Existem várias técnicas de agrupamento, como por exemplo, o agrupamento divisivo, hierárquico e o agrupamento baseado em densidade, assim como vários algoritmos de agrupamento para estas abordagens [58].

Os métodos hierárquicos baseiam-se em sucessivas divisões do conjunto de dados e o resultado final é uma sequência de agrupamentos, que pode ser representada num dendrograma [59,60]. Cada nível de associação do dendrograma representa uma divisão do conjunto de dados num número específico de grupos. Com base no dendrograma, é possível definir o número de grupos, sendo este passo muitas vezes baseado na intuição tendo em conta a estrutura de dados [13].

A análise de agrupamento hierárquico (HCA) é um método utilizado para encontrar a estrutura subjacente dos objectos através de um processo iterativo que associa (métodos aglomerativos) ou dissocia (métodos divisivos) objecto por objecto, até que todos os tenham sido processados. O procedimento aglomerativo inicia-se com cada objecto num grupo separado. Os grupos são combinados sequencialmente, reduzindo o número de grupos em cada etapa, até que todos os objectos pertencem a apenas um grupo. Isto significa que, para N objectos, o processo envolve N-1 etapas de agrupamento. No método HCA, existem duas opções importantes na definição do método: o tipo de medida de similaridade entre os objectos ou grupos, e o método de ligação [13,61].

Os métodos não supervisionados, mesmo se usados em situações clínicas, não são os mais comuns. É mais provável, nesses contextos, encontrar métodos supervisionados.

2.2.2 Métodos de classificação supervisionados

Existem três grandes diferenças entre os algoritmos de classificação supervisionados [31,62]. A primeira, baseia-se na distinção entre os métodos de discriminação, como a análise discriminante linear (LDA[3]), e aqueles que dão maior ênfase à similaridade dentro de cada classe, como por exemplo a modelação independente e flexível por analogia de classe (SIMCA[4]). A segunda diferença relaciona-se com os métodos lineares, como o LDA, e não lineares, tais como os métodos neurais (NNW[5]). A terceira distinção divide os cálculos paramétricos e não paramétricos. Nas técnicas paramétricas, como o LDA, os parâmetros estatísticos da distribuição normal das amostras são usados nas regras de decisão.

Os métodos clássicos para a classificação supervisionada são métodos baseados na correlação, métodos baseados na distância, LDA, SIMCA e análise discriminante por mínimos quadrados parciais (PLS-DA) [40,63].

A análise discriminante linear (LDA) é um método paramétrico, com características de discriminação [37,64]. Este método procura os limites ideais entre classes. Tal como a análise de componentes principais (PCA), é um método de redução de variáveis. No entanto, enquanto o PCA selecciona uma direcção que mantém a variabilidade máxima na menor dimensão entre os dados, o LDA selecciona as direcções que atingem um máximo de separação entre as diferentes classes [51,65]. O LDA clássico utiliza a distância euclidiana para classificar amostras desconhecidas.

A precisão dos métodos como o LDA e de outros métodos de classificação como a análise discriminante quadrática (QDA) e os métodos do k-vizinho mais próximo (KNN) tem sido avaliada para classificar as amostras provenientes de estudos clínicos [66]. No entanto, estes dois métodos não foram considerados no presente trabalho.

Na literatura, existem alguns estudos que comparam a exactidão dos diversos métodos de classificação, como por exemplo, o estudo desenvolvido por *Wu et al.* [67].

A selecção do método de classificação mais rigoroso depende do conjunto de dados e, portanto, vários métodos quimiométricos tem que ser testados.

[3] Do inglês *Linear Discriminant Analysis*

[4] Do inglês *Soft Independent Modeling of Class Analogy.*

[5] Do inglês *Neural Networks.*

Grande parte dos métodos discutidos na literatura foi comparada no contexto da utilização de dados de microarranjo *(microarray data),* para distinguir vários tipos de cancro [68]. Referências gerais sobre o tema da análise discriminante incluem os trabalhos de *Mardia et al.* [69], *McLachlan* [70] e *Ripley* [71].

A análise discriminante linear foi primeiramente descrita por Fisher [64] em 1936. Este método procura uma combinação linear $x_a$da intensidade da amostra $x = (x_1, \dots, x_p)$ que tem uma razão máxima de separação da classe para a variância dentro da classe, isto é, maximizando a relação $a^T.B_a/a^T.W_a$, onde *W* representa a matriz de covariância dentro da classe, ou seja, a matriz de covariância das variáveis centradas na média da classe, e *B* designa a matriz de covariância entre classes.

O critério usado na análise discriminante linear é bastante intuitivo. O LDA é um método paramétrico, que consiste numa forma especial da regra de discriminação de máxima verosimilhança para classes de densidade normal com a mesma matriz de covariância [72].

No método de mínimos quadrados parciais (PLS[6]) as regressões são calculados com algoritmos de mínimos quadrados [73]. O PLS é um método estatístico relacionado com a regressão de componentes principais, que em vez de encontrar hiperplanos de variância máxima entre a resposta e as variáveis independentes, encontra um modelo de regressão linear, projectando as variáveis previstas e as variáveis observáveis num novo espaço [74- 77].

A análise discriminante por mínimos quadrados parciais (PLS-DA) é uma variante do PLS utilizada quando a resposta *Y* é binária. O PLS é usado para encontrar relações fundamentais entre as duas matrizes *X* e *Y*, ou seja, trata-se de uma abordagem baseada na variável latente, para modelar as estruturas de covariância nestes dois espaços (*X* e *Y*).

Um modelo PLS procura a direcção multidimensional no espaço da matriz *X* que explica a direcção de máxima variação multidimensional no espaço da matriz *Y*. A regressão por mínimos quadrados parciais é particularmente indicada quando a matriz dos predictores (*X*) tem mais variáveis do que observações, e quando existe multicolinearidade entre os valores de *X* [78-80].

Tendo em conta a consistência destes métodos e a capacidade dos computadores pessoais da geração actual, é possível usar eficazmente estes métodos.

[6] Do inglês *Partial Least Squares*

2.2.3 Análise de agregados

A análise de agrupamento tem por finalidade reunir, por algum critério de classificação os objectos em grupos, de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos [81].

O processo de agrupamento envolve a estimativa de uma medida de dissimilaridade entre os indivíduos e a adopção de uma técnica de formação de grupos.

Diversas medidas de similaridade ou de dissimilaridade têm sido propostas e utilizadas, sendo a escolha entre elas baseada na preferência e/ou na conveniência do utilizador [82].

As técnicas de análise de agrupamento exigem a tomada de uma série de decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição e que podem representar diferentes agrupamentos. Além disso, o resultado dos agrupamentos pode ser influenciado pela escolha da medida de dissimilaridade, bem como pela definição do número de grupos [83-87].

2.2.3.1 Ferramentas e algoritmos de agrupamento hierárquico

A análise de agrupamento hierárquico (do inglês *Hierarchical Cluster Analysis, HCA*) interliga os objectos pelas suas associações, produzindo um dendrograma onde os objectos semelhantes, segundo as variáveis escolhidas, são agrupados entre si. Quanto menor a distância entre os objectos, maior a semelhança entre eles [82].

Existem muitas formas de procurar grupos no espaço n-dimensional. A forma mais simples consiste em agrupar os os pares de elementos que estão mais póximos, usando a distância euclidiana, e substituí-los por um novo ponto localizado na metade da distância entre eles. Este procedimento, quando repetido até todos os elementos serem agrupados num só grupo, leva à constução do dendrograma, onde, no eixo horizontal são colocados os elementos e no eixo vertical, as distâncias.

Diversos algoritmos e ferramentas para Análise de Agregados foram desenvolvidos e encontram-se disponíveis [83,84]. Muitos são de domínio público e podem ser encontrados na Internet. A seguir são apresentados os algoritmos clássicos utilizados na análise de agrupamento hierárquico, assim como outros algoritmos que foram desenvolvidos a partir destes.

Para uma descrição geral dos algoritmos de agrupamento existem na literatura vários exemplos [88-90].

Os vários algoritmos diferem no modo como estimam a distância entre grupos já formados e outros grupos ou elementos por agrupar. O processo de agrupamento de elementos já associados depende da similaridade e dissimilaridade entre os grupos. Portanto, diferentes definições destas distâncias poderão resultar em diferentes soluções finais [19].

A seguir, são apresentados, resumidamente alguns métodos de agrupamento. Não existe o que se possa chamar de melhor critério, mas alguns são mais indicados para determinadas situações do que outros [82]. É prática comum utilizar vários critérios e fazer a comparação dos resultados; se forem semelhantes, é possível concluir que possuem um elevado grau de estabilidade e, portanto, são confiáveis.

Os critérios de agrupamento hierárquico mais utilizados para determinar a distância entre grupos são: a ligação simples *(single linkage)*, a ligação completa *(complete linkage)*, o método do centróide *(centroid)*, a mediana *(median)*, a média das distâncias *(average linkage)* e a soma de erros quadráticos ou variância (método de *Ward*) [91-92].

Estes algoritmos (contidos no algoritmo HCA, de maior âmbito) encontram-se descritos no capítulo 3 deste trabalho e são referenciados como os algoritmos clássicos da literatura de agrupamento hierárquico.

O algoritmo "*single linkage*" descrito inicialmente por Sneath [93] (1957) e Johnson [86] (1967) é um dos mais simples. Neste método a distância entre dois grupos é determinada pela distância do par de objectos mais próximo, sendo cada objecto pertencente a um desses grupos.

Segundo Orlóci [94] (1978) e Mardia et al. [95] (1997) este método leva a grupos longos se comparados aos grupos formados por outros métodos de agrupamento. Os dendrogramas resultantes deste procedimento são, geralmente, pouco conclusivos, devido à informação dos objectos intermediários que não são evidentes. Alguns algoritmos que implementam esta estratégia são também descritos por Sibson [96] (1973) e Rohlf (1978) [97].

O algoritmo *"complete linkage"* ao contrário do anterior determina a distância entre dois grupos de acordo com a maior distância entre um par de objectos, sendo cada objecto pertencente a um grupo distinto. Segundo alguns autores, este método leva, geralmente a grupos compactos e discretos, com valores de dissimilaridade relativamente grandes [92].

Os métodos baseados na distância mínima e máxima (*single linkage* e *complete linkage*, respectivamente), representam dois extremos em termos de distância entre grupos. Os

procedimentos que envolvem esses extremos tendem a ser sensíveis à presença de *outliers*. Assim, o uso de uma abordagem intermédia é uma solução possível para o problema [97-99] .

No algoritmo *"average linkage"* a distância entre dois grupos é definida como a média das distância entre todos os pares de objectos em cada grupo.

O método do centróide foi proposto por Sokal e Michener [100](1958) e teve como origem, a caracterização da matriz de dados como pontos do espaço Mahalanobis. Neste método, a distância entre dois grupos é definida como a distância entre os seus centróides, pontos definidos pelas médias das variáveis caracterizadoras dos elementos de cada grupo, isto é, calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis. Um desvantagem deste método está relacionada com o tamanho dos grupos seleccionados para a junção, dado que, quando são muito diferentes, o centróide do novo grupo será semelhante ao centróide do grupo de maior dimensão, levando à perda das características do grupo mais pequeno.

Com base na metodologia proposta por Orlóci [94] (1978) o método da mediana é um caso particular do método do centróide. A determinação da distância entre dois agrupamentos através do cálculo do centro de massa não considera o número de elementos em cada um dos grupos. Assim, o vector médio que representa o novo grupo, pode eventualmente, ficar situado entre os elementos do grupo com maior número de elementos. Para contornar este problema, Gower [85] (1967) desenvolveu um procedimento de cálculo que pondera a medida de distância pelo número de elementos de cada agrupamento.

Ward [101] (1963) propõe um processo geral de classificação em que n elementos são progressivamente reunidos dentro de grupos através da minimização de uma função objectiva para cada (n-2) passos de fusão. Inicialmente, este algoritmo admite que cada um dos elementos se constituía num único grupo.

Considerando a primeira reunião de elementos num novo grupo, a soma dos desvios dos pontos representativos dos seus elementos, em relação à média do grupo, é calculada, e dá uma indicação de homogeneidade do grupo formado. Esta medida fornece a "perda de informação" que se produz ao reunir os elementos num grupo [101].

Como mencionado, os algoritmos descritos anteriormente são considerados os algoritmos clássicos da literatura de agrupamento hieráquico. Outros algoritmos foram elaborados a partir das ideias implementadas nos algoritmos clássicos: *AGNES –Agglomerative Nesting* [102] , *BIRCH – Balanced Iterative Reducing and Clustering Hierarchies* [103], *CURE – Clustering Using Representatives* [104], *CHAMELEON – Hierarchecal Clustering Algorithm Using Dynamic*

Modeling [105] , *DIANA- Divisive Analysis* [106], *ROCK- Robust Clustering Using Links* [107] e o *MONA – Monothetic Analysis* [108].

2.2.4 Análise de componentes principais

A Análise de Componentes Principais (do inglês *Principal Component Analysis, PCA*) é uma ferramenta quimiométrica que permite extrair, de um determinado conjunto de dados, informações relevantes para a sua interpretação. Pode ser utilizada para redução do número de variáveis e para fornecer uma visão estatisticamente privilegiada do conjunto de dados. Fornece as ferramentas adequadas para identificar as variáveis mais importantes no espaço das componentes principais [109].

Este método consiste em reescrever as variáveis originais em novas variáveis (componentes principais), através de uma transformação de coordenadas. A transformação de coordenadas é um processo trivial quando efectuado através de matrizes. A transformação matemática das coordenadas pode ser feita de diversas maneiras conforme o interesse do utilizador. Cada componente principal é uma combinação linear de todas as variáveis originais.

Aplicando este método, é possível efectuar uma simplificação e redução da dimensão original dos dados, modelação, detecção de *outliers*, selecção de variáveis importantes num determinado sistema, classificação e previsão [109].

Na literatura encontram-se inúmeras aplicações, que abrangem diversos ramos do conhecimento [110]. Estudos recentes, evidenciam o sucesso da aplicação deste método a dados de espectrofotometria, nas indústrias alimentar e farmacêutica e mais recentemente em análise de solos [111-113].

2.3 Métodos quimiométricos de segunda ordem

Os métodos multidimensionais de tratamento de dados foram desenvolvidos inicialmente por psicometristas (investigadores da área de psicologia que aplicam estatística multivariada aos seus dados) na década de 60, com destaque para L. Tucker, que propôs os "métodos de Tucker" [114-116]. Este desenvolvimento coincidiu com a tendência positivista, vigente na época, de valorizar excessivamente o papel da análise quantitativa nas ciências humanas. Desta forma, grandes quantidades de dados eram obtidos na forma de questionários ou testes aplicados a um grande número de indivíduos sob diferentes condições e analisadas

por psicometristas, que procuravam extrair componentes puros que deveriam representar influências ou padrões de comportamento.

No início dos anos 70, o investigador R. Harshman [117] desenvolveu, na área de linguística, um modelo que viria a encontrar aplicação na literatura química anos mais tarde, o *PARAFAC*. Carrol e Chang[118] propuseram na mesma época e de maneira independente um modelo idêntico, ao qual deram o nome de decomposição canónica ("*CANDECOMP, CANonical DECOMPosition")*.

A utilização de métodos multidimensionais em química foi relativamente tardia. O artigo de Ho, Christian e Davidson [119] (1978), propôs o método de análise de factores "*RAFA, Rank Annihilation Factor Analysis"* para tratamento de dados obtidos por espectrofluorimetria. Durante toda a década de 80, a aplicação destes métodos na literatura química permaneceu relativamente restrita. Em 1981, Appellof e Davidson [121], baseados no artigo de Carrol e Chang [120], apresentaram um modelo similar ao *PARAFAC*, ao qual não deram nome, para tratamento de dados cromatográficos com detecção espectrofluorimétrica. Numa série de três artigos publicados em 1988, Russell [122-124] e co-autores usaram o mesmo modelo, que denominaram por PCFA *("Principal Component Factor Analysis")*, para o tratamento de espectros de fluorescência molecular resolvidos no tempo.

Um destaque especial deve ser dado ao método generalizado *GRAM ("Generalized Rank Annihilation Method")* proposto em 1986 por Sanchez e Kowalski [125]. Este método, de solução algébrica, é baseado num modelo trilinear e numa decomposição dos dados em vectores próprios, mas possui a limitação de que uma das dimensões dos dados seja igual a dois. Em 1990, os mesmos autores propuseram uma extensão do método: a decomposição directa trilinear *("DTD, Direct Trilinear Decomposition")* [126], que superava a antiga limitação. No entanto, a inexistência de um critério de optimização bem definido (como o ajuste por mínimos quadrados) tem limitado a aplicação deste método a conjuntos de dados livres de ruído. O *DTD* produz resultados que não são robustos na presença de ruído instrumental.

Durante os anos 90, alguns métodos foram propostos, como a extensão da resolução multivariada de curvas para dados de segunda ordem *("MCR, Multivariate Curve Resolution")* [127,128] e o método de regressão dos mínimos quadrados bilineares ("BLLS, BiLinear Least Squares") [129].

Em 1989, Geladi [130] apresentou num artigo tutorial os métodos para análise de dados multidimensionais em química e chamou à atenção para a necessidade de sistematização e generalização. O autor identificou como factores limitantes para a difusão destes métodos a

ausência de um algoritmo na forma de um programa amigável e a ausência de generalização para as aplicações encontradas na literatura (" (…) aplicações que não vão além do próprio exemplo."). Em 1992, Smilde [131] também chamou a atenção para o potencial de alguns métodos de ordem superior, ainda pouco usados e conhecidos na literatura química. De 1995 a 1998, o autor iria orientar a tese de doutoramento de Bro [132], a qual veio contribuir para difundir a aplicação de alguns novos métodos quimiométricos, o *PARAFAC* [133] e o *N-PLS* [134]. Esta tese, que gerou mais de uma dezena de artigos científicos, descreve de modo bastante claro e didáctico os modelos e algoritmos para uma série de métodos multidimensionais e apresenta exemplos de aplicações em diferentes áreas, discutindo as suas generalizações. A contribuição talvez mais importante desse trabalho foi o facto de o autor disponibilizar gratuitamente os algoritmos usados, na forma de pacotes ("toolboxes") elaborados em ambiente Matlab (Mathworks), facilitando a difusão e utilização desses métodos.

Capítulo 3

Chemometrics ToolBox

A existência de um conjunto de programas de base permite, com facilidade, o tratamento prévio de um conjunto de dados, a utilização de novos métodos, e a combinação de vários métodos para o mesmo problema.

3.1 Funcionamento

Para o desenvolvimento desta ferramenta recorreram-se a algumas sub-rotinas do Octave Source Forge [135]. Algumas das interfaces e programas implementados foram elaborados localmente. Sempre que necessário e conveniente é possível incorporar novos métodos.

Esta ferramenta foi desenvolvida em linguagem Octave (Linux) compatível com MatLab (Windows). Tem incorporado um conjunto de ferramentas estatísticas para o tratamento de dados, sendo escolha do método dependente do objectivo do utilizador.

O correcto funcionamento desta ferramenta de trabalho envolve a instalação e integração de dois tipos de programas: o Octave e o GnuPlot.

Toda a informação relativa à notação adoptada, fases de preparação, convenções e restrições encontram-se documentadas no respectivo manual (anexo A.1). Neste manual são resumidas as várias potencialidades da "*Chemometrics ToolBox*", facilitando a sua utilização. São descritas as ferramentas estatísticas disponíveis e as funções implementadas.

Nesta secção são apresentadas as ferramentas disponibilizadas: ANOVA, HCA PCA, LDA e PLS.

Para que este conjunto de ferramentas corra sem problemas foi necessário impor algumas convenções e restrições iniciais ao seu funcionamento (anexo A.1).

Esta ferramenta requer também que esteja instalado o pacote de análise estatística do Octave. A file de dados deve ser sempre uma matriz rectangular, sem omissões no seu interior.

3.2 Métodos implementados

Na ferramenta desenvolvida existem vários métodos de análise multivariada com finalidades diferentes entre si.

O utilizador deve ter uma noção prévia do que se pretende concluir e afirmar sobre os dados, isto é, saber que conhecimento pretende gerar.

Nesta secção são descritos alguns dos métodos implementados: ANOVA *(Analysis of Variance),* HCA *(Hierarchical Cluster Analysis),* PCA *(Principal Component Analysis)* LDA *(Linear Discriminant Analysis)* e PLS *(Partial Least Squares).*

3.2.1 Análise de variância

A Análise de Variância (ANOVA[7]) é uma ferramenta estatística importante para distinguir as diversas contribuições sobre a variância total observada [136]. A ANOVA permite distinguir dentro da variabilidade total de diversos conjuntos de valores experimentais as contribuições puramente aleatória e a contribuição sistemática entre amostras. Deste modo, permite verificar se as amostras (ou factores) exercem um efeito significativo fazendo com que estes se sobreponham à componente aleatória contribuindo para diferenças significativas entre si [136].

A ANOVA permite comparar em simultâneo várias médias (níveis diferentes do factor) e estimar as diversas contribuições de variabilidade: puramente aleatória (estimada dentro de cada amostra), variabilidade entre amostras, entre outras.

Como pressupostos assume-se que as distribuições em causa são normais e independentes e que existe homogeneidade de variância (variabilidade interna).

Na área de controlo de qualidade, existem vários exemplos de aplicação: amostragem, limite de repetibilidade e de reprodutibilidade, planeamento experimental e análise de factores, estudo de interferentes, robustez e coerência e validação do modelo de calibração.

A Análise de Variância foi implementada em três versões: de uma via ou de factor único, de duas vias ou de dois factores sem réplicas e com réplicas.

De um modo geral, a ANOVA de factor único estuda o efeito de um factor (aqui designado A) sobre a variabilidade do sistema em análise, enquanto a ANOVA de factor duplo sem réplicas destina-se a verificar o efeito simultâneo de dois factores (A e B).

[7] Do inglês *Analysis of Variance*

Na primeira abordagem, a soma de quadrados total (SS_T) pode ser decomposta nas componentes puramente aleatória (SS_{pe}) e na componente devida ao factor (SS_A),

$$SS_T = SS_{pe} + SS_A \quad (3.1)$$

A ANOVA de factor duplo sem réplicas permite a decomposição da variabilidade total (T) em três componentes: a puramente aleatória (pe), a devida ao factor linha (factor A) e a devida ao factor coluna (factor B) de acordo com a equação,

$$SS_T = SS_{pe} + SS_A + SS_B \quad (3.2)$$

Nestas duas abordagens a matriz de dados deve ser rectangular, sem omissões de valores, e estar organizada em N linhas e M colunas, sendo o número total de valores dado por N×M.

A ANOVA de factor duplo com réplicas permite a decomposição da variabilidade total (T) em quatro componentes: a puramente aleatória (pe), a devida ao factor linha (factor A), a devida ao factor coluna (factor B) e ao termo de interacção entre factores (AB),

$$SS_T = SS_{pe} + SS_A + SS_B + SS_{AB} \quad (3.3)$$

Neste caso a matriz de dados contém o factor A nas linhas e o factor B nas colunas sendo que cada conjunto de Q linhas reflecte o número de réplicas. A matriz de dados é igualmente uma matriz rectangular contendo N×Q linhas e M colunas, sem omissões de valores.

Os fundamentos inerentes as três versões encontram-se descritos no secção 1 do manual em anexo (anexo A.1).

Ao ser executada cada uma destas opções é criada a file correspondente que preserva os valores calculados através da tabela ANOVA respectiva.

3.2.2 Análise de agregados

Existem diversas versões de análises de agrupamentos sendo a versão por nós preferida a análise não supervisionada em modo hierárquico, já que o modo de associação é independente de critérios impostos.

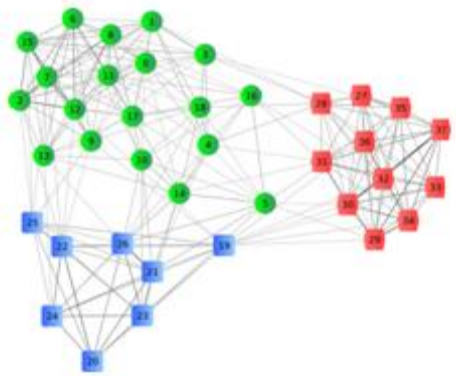


Figura 3.1 – Exemplificação de três agrupamentos [135].

O processo de agrupamento envolve basicamente duas etapas: a primeira refere-se à estimativa de uma medida de dissimilaridade entre os objectos e a segunda, refere-se à adopção de uma técnica de formação de grupos (algoritmo) [82, 88].

Existe um grande número de medidas de similaridade ou de dissimilaridade sendo a escolha entre elas baseada na preferência do utilizador. Com a definição da medida de dissimilaridade a ser utilizada, a etapa seguinte é a adopção do método de agrupamento para formação dos grupos. Para realização desta tarefa, existe um grande número de métodos disponíveis, sobre os quais o utilizador tem de decidir qual o mais adequado ao seu propósito (algoritmos diferentes podem levar a diferentes soluções) [82].

Para dados contínuos estão disponíveis e sistematizadas as métricas seguintes: *"Euclidean", "Seuclidean", "Mahalanobis", "Cityblock", "Minkowski", "Cosine", "Correlation", "Spearman", "Hamming", "Jaccard"* e *"Chebychev"* [138].

Para dados binários encontram-se disponíveis e sistematizadas 24 métricas distintas: *"Pattern difference", "Euclidean", "SEuclidean", "Variance", "Simple matching", "Manhattan", "Dice", "Antidice", "Lance and Williams", "Nei & Lei's", "Yule coefficient", "Cosine", "Sneath", "Forbes", "Hamman", "Jaccard", "Rogers", "Ochiai", "Anderberg", "Kulczynski", "Pearson", "Gower2", "Russel-Rao", e "Simpson"* [138-139].

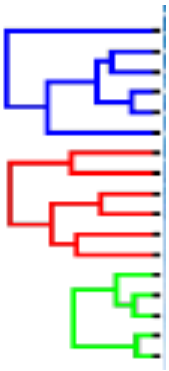
Esta opção exige a tomada de uma série de decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição e que podem representar diferentes agrupamentos.

Genericamente, a análise de agrupamento envolve as etapas seguintes:

1. Selecção dos objectos a agrupar;

2. Definição de um conjunto de variáveis a partir das quais serão obtidas informações necessárias ao agrupamento dos objectos;
3. Definição de uma medida de semelhança ou distância entre os objectos;
4. Escolha de um algoritmo de agrupamento;
5. Validação dos resultados obtidos.

3.2.2.1 Análise de agrupamento hierárquico



Nesta secção encontram-se descritas as etapas do processo de agrupamento, assim como as medidas de distância implementadas.

A HCA interliga os objectos pelas suas associações, produzindo um dendrograma onde os objectos semelhantes, segundo as variáveis escolhidas, são agrupados entre si.

Os dendrogramas são especialmente úteis na visualização de semelhanças entre objectos representados por pontos no espaço com dimensão maior do que três, onde a representação de gáficos convencionais não é possível.

Existem muitas maneiras de procurar agrupamentos no espaço n-dimensional. A forma mais simples consiste em agrupar os pares de pontos que estão mais póximos, usando a distância euclidiana, e substituí-los por um novo ponto localizado na metade da distância entre eles. Este procedimento, quando repetido até que todos os pontos sejam agrupados num só, leva à constução do dendrograma, onde, no eixo horizontal são colocados os objectos e no eixo vertical as distâncias.

A técnica de agrupamento hierárquico disponível nesta ferramenta, consiste numa série de agrupamentos sucessivos entre objectos. Parte-se de n grupos de apenas um objecto, que vão sendo agrupados, sucessivamente, até que se encontre apenas um grupo que incluirá a totalidade dos n objectos.

Este método conduz a uma estrutura que descreve uma hierarquia de agrupamentos sobre os dados (dendrograma). Para um número inicial de *n* objectos na base de dados, ao todo ocorrem *n-1* associações.

O ponto de partida é a construção da matriz de distâncias que é necessário calcular e armazenar, durante o processamento. Nesta matriz, cada elemento descreve o grau de diferença entre cada dois objectos com base nas variáveis escolhidas.

(i) Algoritmos de agrupamento

Nesta *toolbox*, para determinar a distância entre grupos encontram-se disponíveis os algoritmos seguintes: ligação simples *(single linkage)*, ligação completa *(complete linkage)*, média das distâncias *(average linkage),* centróide, mediana, e a soma de erros quadráticos ou variância *(Ward)*.

Ao ser executada a opção relativa ao tipo de dados a processar surge, com a ordem apresentada no respectivo menu, a possibilidade de escolher a métrica a usar e no menu seguinte surgem os critérios de ligação.

Método da ligação simples [*single linkage*]

O algoritmo da ligação simples, representado na Figura 3.2, é um dos mais simples e de rápida aplicação. A distância entre os grupos é definida como sendo aquela entre os objectos mais próximos (sendo cada objecto pertencente a cada um desses grupos). Conduz à formação de grupos longos se comparados aos grupos formados por outros critérios. Os dendrogramas resultantes são, geralmente, pouco elucidativos, dado que, a informação relativa aos objectos intermediários não é evidente.

Este método tende a formar longas cadeias[8], é sensível a *outliers*, pois tem tendência a incorporar os *outliers* num grupo já existente, e grupos muito próximos podem não ser identificados.

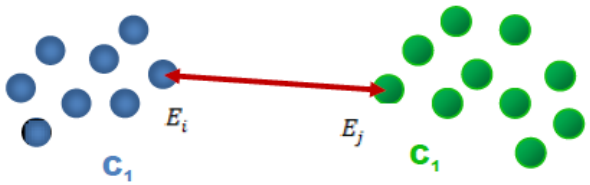


Figura 3.2 – Distância entre grupos obtida através da menor distância (*single linkage*)

Método da ligação completa [*complete linkage*]

Ao contrário do algoritmo anterior, o método da ligação completa, representado na Figura 3.3, determina a distância entre dois grupos de acordo com a maior distância entre um

[8] Situação em que há um primeiro grupo de um ou mais objectos que passa a incorporar um grupo de apenas um objecto, formando uma longa cadeia, onde se torna difícil definir um nível de corte para classificar os objectos em grupos.

par de objectos, sendo cada objecto pertencente a um grupo distinto. Geralmente, leva a grupos compactos e discretos, sendo os seus valores de dissimilaridade relativamente grandes.

Figura 3.3 – Distância entre grupos através da associação completa (*complete linkage*).

Método da ligação média [*average linkage*]

No algoritmo da ligação média, representado na Figura 3.4, a distância entre dois grupos é definida como a média das distâncias entre todos os pares de objectos em cada grupo.

Figura 3.4 – Distância obtida através da média das distâncias entre os objectos (*average linkage*).

Método do centróide [*centroid*]

No método do centróide, cada grupo é considerado um simples ponto representado pelo seu centro de massa (centróide). Este método utiliza uma função de agrupamento para medir a distância entre os centros de massa dos dados. Caracteriza-se pela redefinição, a cada passo, da matriz de dados, em que cada grupo é representado pelo vector médio das p variáveis envolvidas.

Uma nova matriz de distâncias é determinada a cada iteração. A distância entre dois grupos é definida como a distância entre os seus centróides, pontos definidos pelas médias das variáveis caracterizadoras dos objectos de cada grupo - calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis.

Uma desvantagem deste método é que se os dois grupos forem muito diferentes em termos de dimensão, o centróide do novo grupo estará mais próximo daquele que for maior e assim, as características do grupo menor tenderão a perder-se.

Uma característica importante deste algoritmo é o facto da distância entre grupos ser determinada pela distância entre os pontos representativos dos respectivos centros de massa. Outras características são a robustez à presença de *outliers* e o fenómeno da inversão[9].

Método da Mediana [*median*]

Este algoritmo é um caso particular do método do centróide. A determinação da distância entre dois grupos através do cálculo do centro de massa, não considera o número de elementos em cada um dos grupos, o vector médio que representa o novo grupo pode, eventualmente, ficar situado entre os elementos do grupo com maior número de objectos.

Este método pondera a medida de distância pelo número de elementos de cada grupo. Apresenta resultados satisfatórios quando os grupos têm tamanhos diferentes; pode apresentar resultados diferentes quando permutados os elementos na matriz de dissimilaridade; apresenta robustez à presença de *outliers* e também, o fenómeno da inversão.

Método de variância mínima [*Ward*]

O método da ligação *Ward*, esquematizado na Figura 3.5, baseia-se na análise de variância, associando os objectos aos grupos nos quais estes promovem a menor variância intra-grupo. Este algoritmo é altamente eficiente na formação de grupos.

Inicialmente, admite que cada um dos objectos constitui um único grupo. Considerando a primeira reunião de objectos num novo grupo, a soma dos desvios dos pontos representativos dos seus elementos, em relação à média do grupo, é calculada, e dá uma indicação de homogeneidade do grupo formado. Os grupos formados possuem uma elevada homogeneidade interna. No entanto, pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual; tem tendência a combinar grupos com poucos elementos e é sensível à presença de *outliers*.

[9] Ocorre quando a distância entre centróides é menor que a distância entre grupos já formados, isto fará com que os novos grupos sejam formados num nível inferior aos grupos já existentes, tornando o dendrograma confuso.

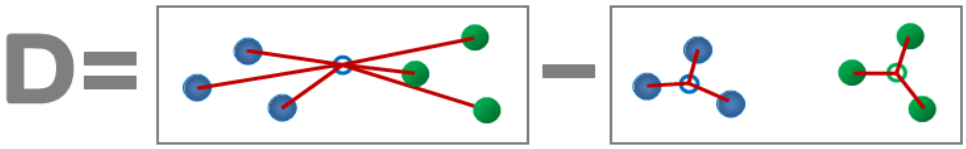
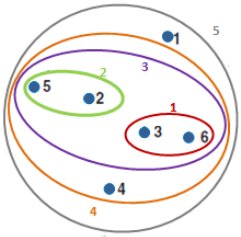
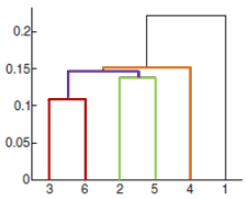
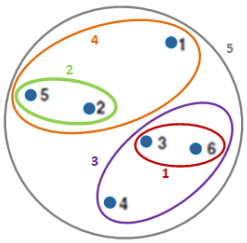
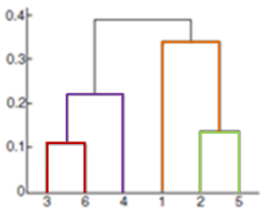
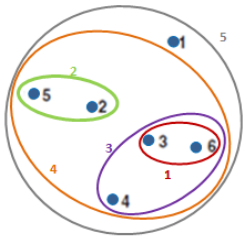
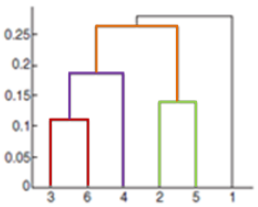
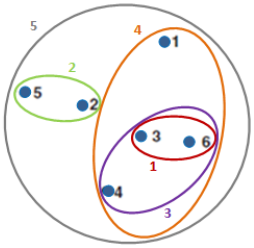
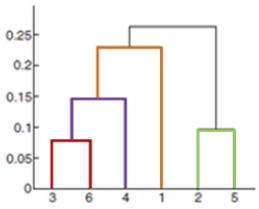


Figura 3.5 – Distância entre grupos obtida através do método da variância mínima (*Ward linkage*).

Na Tabela 3.1 encontra-se representado, num exemplo genérico, o efeito da aplicação destes algoritmos. São evidenciadas as diferenças nos agrupamentos obtidos, aplicando os diferentes critérios. Diferentes métodos de ligação conduzem a diferentes soluções.

Tabela 3.1 – Associações obtidas usando os diferentes critérios de ligação.

Critério de ligação	***Associação***	***Dendrograma***
Ligação simples (single linkage)		
Ligação completa (complete linkage)		
Ligação média (average linkage)		
Ward		

(ii) Etapas do processo

São várias as etapas que devem ser realizadas no processo de agrupamento: pré-processamento dos dados, selecção da medida de dissimilaridade, execução do algoritmo de agrupamento, avaliação dos resultados e interpretação dos grupos identificados.

As variáveis que caracterizam os objectos podem assumir tipos diferentes: contínuo, discreto ou binário.

As variáveis binárias assumem exactamente dois valores, 0 e 1, indicando a presença ou ausência de uma determinada característica; as variáveis discretas possuem um conjunto finito e pequeno de valores possíveis; as varáveis contínuas podem assumir qualquer valor real dentro de um intervalo pré-definido.

(iii) Medidas de distância

Para agrupar objectos, é necessário definir uma medida de similaridade (quanto maior o valor, maior a semelhança entre os objectos) ou dissimilaridade (quanto maior o valor, maior a diferença entre os objectos). Com base nessa medida, os objectos similares são agrupados e os outros são colocados em grupos separados.

As medidas de dissimilaridade têm um papel central nos algoritmos de agrupamento. Através destas medidas, são definidos critérios para avaliar se dois pontos estão próximos e podem fazer parte de um mesmo grupo, ou não.

De um modo geral, as medidas de similaridade e de dissimilaridade são facilmente, transformáveis entre si [92].

Existe um grande número de coeficientes de similaridade e/ou de dissimilaridade para dados binários disponíveis na literatura. Estes coeficientes podem ser, facilmente, convertidos para coeficientes de dissimilaridade: se a similaridade for denominada s, a medida de dissimilaridade será o seu complemento $(1 - s)$.

A maioria dos métodos de análise de agrupamento requer uma medida de similaridade ou dissimilaridade entre os elementos a serem agrupados, normalmente expressa como uma função distância ou métrica. Para que uma função *d* seja uma distância é necessário que as seguintes condicões sejam satisfeitas, para quaisquer objetos *i, j, k*:

1. $d\ (i,j)\ =\ d\ (j,i)$ (simétrica);
2. $d\ (i,j) > 0, se\ i \neq\ j$;
3. $d\ (i,j)\ =\ 0$, se e somente se, i = j;

4. $d\ (i,j)\ \leq\ d\ (i,k)\ +\ d\ (z,k)$ (desigualdade triangular).

A propriedade (1) implica que todos os elementos da matriz de dissimilaridade são positivos, a propriedade (2) implica que a diagonal da matriz de dissimilaridade é formada por zeros. A propriedade (3), por sua vez, implica que a matriz de dissimilaridade é simétrica em relação à diagonal.

Para que um índice de proximidade seja considerado uma métrica, este deve satisfazer, além das três propriedades anteriores, a propriedade (4) de desigualdade triangular.

Qualquer função que satisfaça estas quatro propriedades é chamada de distância.
Existem várias medidas que podem ser utilizadas como medidas de distâncias ou dissimilaridade entre elementos da matriz de dados: euclidiana ("Euclidean"), euclidiana quadrada ("Seuclidean"), "Mahalanobis", "Cityblock", "Minkowski", "Cosine", "Correlation", "Spearman", "Hamming", "Jaccard" e "Chebychev".
As medidas de distância mais importantes funções são:

- Distância Euclidiana:

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots + \left|x_{in} - x_{jn}\right|^2} \quad (3.4)$$

- Distância de Manhattan:

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{in} - x_{jn}\right| \quad (3.5)$$

- Distância de Minkowski:

$$d(i,j) = \sqrt[q]{\left|x_{i1} - x_{j1}\right|^q + \left|x_{i2} - x_{j2}\right|^q + \cdots + \left|x_{in} - x_{jn}\right|^q} \quad (3.6)$$

onde $q\ \geq 1$. Logo, a distância de Minkowski generaliza tanto a distância euclidiana (caso onde $q = 2$) quanto a distância de Manhattan (caso onde $q = 1$).

Antes de aplicar o algoritmo é preciso transformar a matriz de dados numa matriz de dissimilaridade. Os métodos de transformação dependem do tipo de valores que assumem os atributos dos objectos.

Se duas observações são caracterizadas por *p* variáveis que assumem o valor 1 se uma determinada característica está presente e 0, caso contrário, os resultados dessas observações podem ser analisados na Tabela 3.2, de contingência, seguinte:

$$X = \begin{bmatrix} 1 & 0 & 1 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 1 \end{bmatrix} \quad (3.7)$$

Tabela 3.2 – Resumo das informações.

		Objecto i		
		1	0	Soma
Objecto j	1	a	b	a+b
	0	c	d	c+d
	Soma	a+c	b+d	M

As variáveis binárias têm apenas dois valores: 0 ou 1. Para determinar a matriz de dissimilaridade nestes casos - determinar $d(i,j)$ para cada par de objectos i, j considera-se a tabela de contingência para i, j.

- a é o número de variáveis com valor 1 para i e j;
- b é o número de variáveis com valor 1 para i e 0 para j;
- c é o número de variáveis com valor 0 para i e 1 para j;
- d é o número de variáveis com valor 0 para i e 0 para j;
- M é o número total de variáveis (M = a+b+c+d)

Os valores a e d representam o número de variáveis em que as observações comparadas coincidem quanto à presença ou à ausência da característica de interesse, respectivamente.

Os valores b e c representam o número de variáveis em que os elementos analisados pertencem a categorias diferentes.

O coeficiente de concordância simples, dado por (a+d)/m , calcula a proporção de variáveis em que os objectos comparados têm o mesmo código. Neste caso, são atribuídos pesos iguais tanto à presença (par 1-1) quando à ausência (par 0-0) de um determinado atributo. Ambos os casos são considerados concordâncias.

Existe um grande número de medidas de dissimilaridade para dados binários.

Nesta *toolbox* estão disponíveis e sistematizadas 24 medidas diferentes, com a ordem apresentada no respectivo menu: “Pattern difference”, “Euclidean”, “SEuclidean"”, “Variance”, “Simple matching”, “Manhattan”, ”Dice”, “Antidice”, ”Lance and Williams”, “Nei & Lei´s”, “Yule coefficient”, “Cosine”, “Sneath”, “Forbes”, “Hamman”, “Jaccard”, “Rogers”, “Ochiai”, “Anderberg”, “Kulczynski”, “Pearson”, “Gower2”, "Russel-Rao”, e “Simpson”.

Quando a medida utiliza a ou d é uma medida de similaridade, se utiliza apenas b ou c, é uma medida de dissimilaridade. Na Tabela 3.3 descrevem-se as medidas de similaridade e de distância disponíveis:

Tabela 3.3 – Medidas de similaridade e de distância disponíveis para dados binários [138-141].

Nome	Equação	Variação
Pattern Difference	$bc/(a+b+c+d)\times 2$	[0,1]
Euclidean	$\sqrt{(b+c)}$	[0, ∞[
SEuclidean	$b+c$	[0, ∞[
Variance	$(b+c)/4(a+b+c+d)$	[0,1]
Simple Matching	$(a+d)/(a+b+c+d)$	[0,1]
Manhattan	$\frac{(b+c)}{a+b+c+d}$	[-1,-1]
Dice[10]	$2a/(2a+b+c)$	[0,1]
AntiDice[11]	$a/(a+2(b+c))$	[0,1]
Lance and Williams	$(b+c)/(2a+b+c)$	[0,1]
Neil & Leil´s	$2a/[(a+b)+(a+c)]$	[-]
Yule coefficient[12]	$(ad-bc)/(ad+bc)$	[-1, -1]
Cosine	$\frac{a}{\sqrt{(b+a)\times(c+a)}}$	[0,1]
Sneath	$2(a+d)/(2(a+d)+(b+c))$	[0,1]
Forbes	$\frac{a\times(a+b+c+d)}{(b+a)\times(c+a)}$	[0, ∞[
Hamman	$((a+d)-(b+c))/(a+b+c+d)$	[-1, -1]
Jaccard	$a/(a+b+c)$	[0,1]
Rogers	$(a+d)/((a+d)+2(b+c))$	[0,1]
Ochiai	$a/\sqrt{((a+b)(a+c))}$	[0,1]
Anderberg[13]	$(a/(a+b)+a/(a+c)+d/(c+d)+d/(b+d))/4$	[0,1]
Kulczynski[14]	$(a/(a+b)+a/(a+c))/2$	[0,1]
Pearson[15]	$\frac{(a\times d)-(b\times d)}{\sqrt{(b+a)\times(c+a)\times(b+d)\times(c+d)}}$	[-1, -1]
Gower2[16]	$ad/\sqrt{((a+b)(a+c)(d+b)(d+c))}$	[0,1]
Russel-Rao	$a/(a+b+c+d)$	[0,1}
Simpson	$a/min((a+b),(a.+c))$	[0,1]

[10] Também conhecido como medida de Czekanowski ou de Sorensen. Neste coeficiente são excluídas as ausências (correspondente aos valores d) e as coincidências (valores de a) têm peso duplo.

[11] Este coeficiente, ao contrário do anterior, atribui peso duplo às discordâncias (b e c).

[12] A formula para este coeficiente é indefinida quando um ou ambos os vectores ou são zeros ou tudo uns. O programa atribui o valor 1 quando b+c=0 - existe concordância completa. Quando a+d=0, o programa assume a medida como -1 - existe discordância completa. Por outro lado, se ad-bc=0, é atribuído valor 0. Estas regras, aplicadas antes de usar a fórmula, evitam os casos onde a mesma conduziria a um resultado indefinido.

[13] Neste caso, se ambos os vectores são tudo uns (ou tudo zeros), a medida de similaridade é 1. Se as somas a+b, a+c, c+d, b+d forem iguais a 0, então a medida é 0.

[14] A fórmula é indefinida se um ou ambos os vectores forem tudo zeros. Se ambos os vectores forem tudo zeros, programa atribui valor 1 à medida de similaridade. Se apenas um dos vectores é tudo zeros, a medida de similaridade tem valor 0.

[15] Este coeficiente segue as mesmas restrições que o coeficiente de Yule.

[16] Neste caso, se ambos os vectores são tudo uns (ou tudo zeros) a medida de similaridade é 1. Se ad=0, a medida é 0.

(iv) Validação dos resultados

Existem alguns procedimentos práticos para conferir, de maneira superficial, os resultados obtidos.

Nesta ferramenta encontra-se disponível um simulador de dados aleatórios correspondente à file "shuffle.m".

Apesar das tentativas de construção de vários testes para a confiabilidade estatística dos agrupamentos, nenhum procedimento totalmente comprovado está ainda disponível. A ausência de testes adequados provém da dificuldade de especificação de hipóteses nulas realistas.

3.2.3 Análise de componentes principais

A análise de componentes principais é um método que permite analisar grandes conjuntos de dados envolvendo um elevado número de variáveis, sem exigir quaisquer pressupostos complicados (relativos, por exemplo, ao tipo de distribuição subjacente aos dados).

O objectivo principal é a redução da dimensionalidade de grandes matrizes de dados - as m variáveis originais são substituídas por um outro subconjunto de p variáveis não correlacionadas, de menor dimensão, designadas por componentes principais, com perda de informação mínima.

De um modo geral, este método descobre novas variáveis (componentes principais) que consigam reunir a maioria da variabilidade dos dados. Estas novas variáveis são combinações lineares das variáveis originais. De entre todas as possíveis combinações lineares, escolhe-se, em cada caso, a de variância máxima, dado que as componentes principais devem reflectir, tanto quanto possível, as características dos dados, que eram expressas pela diferenciação que as variáveis originais permitiam estabelecer - devem explicar uma grande parte da variação associada às variáveis iniciais [109].

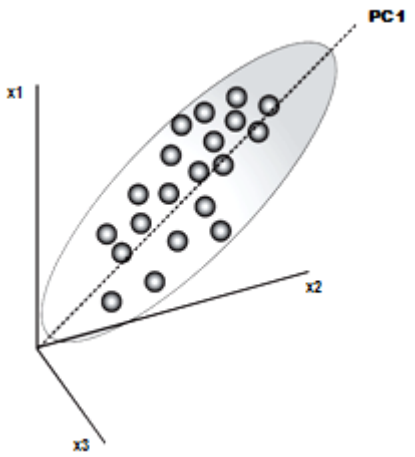


Figura 3.6 – Representação da primeira CP que justifica a variabilidade dos dados (adaptada de [137]).

O grau desta diferenciação entre elementos de uma população é medido pela variância: maior variância implica maior distinção. A variância de uma componente principal é uma medida da quantidade de informação explicada por essa componente principal.

A redução de dimensionalidade atinge-se considerando apenas algumas das componentes principais, as de maior variância. As que não se analisam são as que contribuem com pouca informação, dado que as suas variâncias são pequenas.

Como na maioria dos casos uma grande parte da variância é retida nas primeiras componentes principais, os dados após aplicação do PCA podem ser representados num gráfico de componentes principais a duas ou três dimensões, descrevendo grande parte da informação presente nos dados e facilitando a sua interpretação.

Para além da redução de dimensionalidade, outra vantagem deste método é o facto de as novas variáveis, as componentes principais, serem não correlacionadas: em vez de se analisar um elevado número de variáveis, as originais, com uma estrutura inter-relacional complexa (pois dizem respeito ao mesmo objecto), analisam-se apenas algumas variáveis não correlacionadas; poder-se-á prosseguir a análise aplicando outras técnicas estatísticas para variáveis não correlacionadas [109].

Para auxiliar a interpretação dos resultados da análise e permitir uma boa utilização destes, seja qual for a natureza dos dados, será vantajoso que se consiga atribuir um significado a cada uma das componentes principais (embora nem sempre isto seja possível).

Deste modo, a análise de componentes principais pode considerar-se uma técnica de análise de dados que pode ser útil para a melhor compreensão das relações existentes entre as variáveis em estudo.

Nesta secção, encontram-se descritos os fundamentos inerentes a este método; são definidos os objectivos, as fases de aplicação do método, assim como a interpretação das componentes principais.

3.2.3.1 Objectivos

De um modo geral, a análise de componentes principais visa essencialmente (i) a redução de dados, isto é, trabalhar com um menor número de variáveis e (ii) a identificação, interpretação e visualização de relações entre os dados, de forma a confirmar relações que suspeitávamos e outras que à partida não se identificam facilmente com outros métodos.

Por vezes, os resultados do PCA, novas variáveis ou componentes principais, são utilizados como dados de entrada na aplicação de outros métodos, como na regressão múltipla, na classificação, entre outros.

3.2.3.2 Passos na aplicação do PCA

A aplicação do PCA envolve uma série de etapas, descritas a seguir.

1. Escolha das variáveis a considerar, consoante o tipo e o objectivo do estudo;

   Ao executar esta opção é possível indicar o número de objectos e variáveis a considerar.

2. Escolha da matriz a processar - matriz de Covariância, se as variáveis iniciais tiverem a mesma unidade de medida e variâncias próximas ou a matriz de Correlação, se as variáveis iniciais não tiverem a mesma unidade de medida ou tiverem a mesma unidade mas variâncias muito distintas.

   Ao executar o programa surge a possibilidade de escolher, com a ordem apresentada no respectivo menu, uma das 3 opções disponíveis: Covariância, Pareto ou Correlação.

3. Execução do programa com o software escolhido (Octave neste caso).

   Ao executar o programa são criadas novas files que preservam os valores obtidos e que permitem a representação das variáveis (*loadings*) e dos objectos (*scores*) no sub-espaço das componentes principais

4. Escolha do número de componentes principais a analisar, através de três critérios distintos: Critério de Pearson, Critério de Kaiser e Scree Plot.

   Seleccionado o tipo de matriz a processar, o programa corre utilizando o critério correspondente, sendo o número de componentes principais dado pela instrução 'nfact'.

5. Análise e interpretação das componentes seleccionadas, de acordo com as representações gráficas e as contribuições para o espaço das variáveis e para o espaço dos objectos.

3.2.3.3 Componentes principais

Seja $X = [X_1 \dots X_p]'$ um vector aleatório com valor médio μ e matriz de covariância Σ,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix} \tag{3.8}$$

Onde,

$$\sigma_{ij} = Cov(X_i, X_j) \tag{3.9}$$

Seja $X_{(n\times p)}$ uma matriz de dados de dimensão n, de observações deste vector aleatório:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{3.10}$$

O resultado do PCA é uma rotação do sistema ortogonal de eixos associados às variáveis iniciais (matriz $X_{(n\times p)}$), de forma que, após a sua aplicação, se disponha de um novo sistema de dados $Y_{(n\times p)}$. As colunas da matriz Y são as novas variáveis Y_j , as componentes principais. O objectivo principal é encontrar um novo conjunto de p variáveis (componentes principais, $Y_1 \dots Y_p$) não correlacionadas, de variância máxima [81].

As componentes principais são combinações lineares das p variáveis da matriz X:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p \tag{3.11}$$

Onde, $j = 1, \dots, p$ e a_{ij} ($i = 1, \dots, p$; $j = 1, \dots, p$) são constantes.

Os coeficientes destas combinações lineares são determinados de modo a satisfazerem as condições seguintes:

1. $Var\,(Y_1) \geq Var\,(Y_2) \geq \cdots \geq Var\,(Y_p)$.
2. Quaisquer duas componentes principais são não correlacionadas: $Corr(Y_i, Y_j) = 0, \forall_{ij}$.
3. Em qualquer componente principal a soma dos quadrados dos coeficientes que engloba é 1 (para Y_i: $a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2 = 1$).

Das condições anteriores retiramos que, Y_1 é a componente principal com maior variância; Y_2 é a componente principal com a segunda maior variância, sujeita à condição de ser não correlacionada com Y_1; Y_3 é a componente principal com a terceira maior variância, sujeita à condição de ser não correlacionada com Y_1 e com Y_2 (e assim por diante). Assim, fica provado que as p componentes principais são as combinações lineares

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = a_j^{'}X\ ,\ \ j = 1, \ldots, p \tag{3.12}$$

Onde $a_1, a_2, \ldots, a_p$ são, respectivamente, os *p* vectores próprios associados aos *p* maiores valores próprios de $\Sigma(\lambda_1 > \lambda_2 > \cdots > \lambda_p)$ e $Var(Y_j) = \lambda_j$.

A covariância entre cada duas componentes principais $Y_j\ e\ Y_j$ é nula, pois todas as componentes principais foram determinadas de forma a serem não correlacionadas duas a duas. Tem-se então: $Cov(Y_j, Y_{j'}) = a_j^{'}\ \Sigma a_{j'} = a_j^{'}\lambda_j a_{j'} = \lambda_j a_j^{'} a_{j'} = 0$, que equivale a ter $a_j^{'}a_{j'} = 0$, o que indica que $a_j^{'}\ e\ a_{j'}\ (com\ j \neq j')$ são vectores ortogonais.

3.2.3.4 Obtenção das componentes principais a partir da matriz de correlação

Em muitas situações, as variáveis em estudo não são todas medidas na mesma unidade, na mesma escala, ou são até de natureza distinta. Surge assim, a necessidade de estabelecer uma certa uniformização, que se consegue através da divisão de cada valor pelo desvio padrão da variável centrada correspondente. Este procedimento conduz à obtenção de variáveis com valor médio nulo e variância unitária. As variáveis em estudo passam a ter todas a mesma variância e a influência das variáveis de variância pequena tende a ser inflacionada enquanto a influência das variáveis de variância elevada tende a ser reduzida.

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

$$Cov\left(\frac{X_i}{\sigma_i}, \frac{X_j}{\sigma_j}\right) = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j} = Corr(X_i, X_j) \tag{3.13}$$

Assim, a análise de componentes principais de um conjunto de dados deste tipo, é efectuada utilizando a matriz de correlação (*P*). As componentes principais serão determinadas tendo em conta os valores e vectores próprios da matriz *P*. Matematicamente, tudo se processa da mesma forma. No entanto, os vectores próprios de *P* não são iguais aos de Σ e as componentes principais também não serão as mesmas.

Matriz de correlação:

$$P = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ \rho_{p1} & \rho_{p2} & \ldots & 1 \end{bmatrix} \tag{3.14}$$

Onde,

$$\rho_{ij} = Corr(X_i, X_j) \tag{3.15}$$

3.2.3.5 Redução de dimensionalidade

Um dos principais objectivos desta análise, como já referido anteriormente, é a redução da dimensionalidade dos dados, que se atingirá substituindo as variáveis originais por algumas das componentes principais [81]. Falta agora ver quantas e quais se devem considerar.

Dado que, as componentes principais se podem ordenar por ordem decrescente da sua variância e que quanto maior esta for mais representativa dos dados originais será a correspondente componente principal, devemos considerar as primeiras componentes principais.

A variância de cada componente principal Y_j é dada por:

$$Var(Y_j) = \lambda_j \tag{3.16}$$

Assim, a soma das variâncias das componentes principais é dada por:

$$\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} \lambda_j \tag{3.17}$$

Além disso, como se sabe, numa matriz simétrica (que é o caso de Σ) a soma dos seus valores próprios é igual ao traço da matriz, pode-se desde logo dizer que:

$$tr(\Sigma) = \sum_{j=1}^{p} Var(X_j) \Rightarrow \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(X_j) \tag{3.18}$$

De onde vem:

$$\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} Var(X_j) \tag{3.19}$$

Isto significa que, a soma das variâncias das variáveis originais é igual á soma das variâncias das componentes principais (se considerarmos todas as componentes principais explicamos toda a variabilidade). Assim, a proporção da variância total que é explicada pela j-ésima componente principal Y_j e que indica a importância da mesma é dada por:

$$\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} = \frac{\lambda_j}{tr(\Sigma)} \tag{3.20}$$

Se estivermos a trabalhar com a matriz de correlação, a variância total será igual ao número de variáveis *p* (dado que a diagonal de *P* é toda formada por uns):

$$\sum_{j=1}^{p}\lambda_j = \sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} Var(X_j) = p \tag{3.21}$$

A proporção da variância total que é explicada pela j-ésima componente principal Y_j e que indica a importância da mesma é dada por:

$$\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} = \frac{\lambda_j}{p} \tag{3.22}$$

Existem vários critérios que podem ser usados para a escolha do número de componentes principais. Estes critérios encontram-se descritos a seguir.

(i) Critério de Pearson (ou regra dos 80%)[3]

Este critério é processado quando são seleccionadas as opções correspondentes à matriz de Covariância ou Pareto. O número de componentes principais é escolhido até termos mais de 80% da informação total ou variabilidade total. Por outras palavras, devem considerar-se tantas componentes principais quantas as necessárias para que a percentagem de variância por elas explicada seja superior a 80%; reter as primeiras r componentes principais de modo a que:

$$\sum_{j=1}^{r}\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} = \frac{\sum_{j=1}^{r}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \geq 0.80 \tag{3.23}$$

(ii) Critério de Kaiser (λ>1) [3]

O critério de Kaiser é processado quando é seleccionada a opção correspondente à matriz de Correlação. Neste caso, devem ser consideradas apenas as componentes com valor próprio superior à unidade (média do conjunto de valores próprios.

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

(iii) Scree Plot

Este terceiro critério, permite utilizar um gráfico onde se representam os pontos de abcissa j e ordenada igual ao j-ésimo valor próprio ou à percentagem de variância explicada pela j-ésima componente principal (pontos de coordenadas $(j,\ \lambda_j)$ ou $\left(j,\ \lambda_j/\sum_{j=1}^{p}\lambda_j\right)$ onde se distinguem as contribuições das diversas componentes principais. De acordo com este critério, devem-se considerar as r componentes principais que mais contribuem, destacando-se de forma acentuada das restantes [109].

3.2.3.6 Interpretação de uma componente principal

O significado de uma componente principal pode ser interpretado utilizando os coeficientes das combinações lineares (a_{ij}) e as correlações entre as variáveis iniciais e as componentes principais (*loading* da variável i para a CP j = $Corr(X_i, Y_j)$). A covariância entre a i-ésima variável (X_i) e a j-ésima CP (Y_j) é dada por:

$$Cov(X_i, Y_j) = a_{ij}\lambda_j \tag{3.25}$$

Dado que:

$$Cov(X_i, Y_j) = Cov\left(\sum_{k=1}^{p} a_{ik}\, Y_k, Y_j\right) = \sum_{k=1}^{p} a_{ik}\, Cov(Y_k, Y_j) = a_{ij} Cov(Y_j, Y_j) = a_{ij} Var(Y_j) = a_{ij}\lambda_j$$

O coeficiente de correlação entre a i-ésima variável (X_i) e a j-ésima CP (Y_j) é então definido por:

$$\rho_{ij} = \frac{a_{ij}\sqrt{\lambda_j}}{\sigma_i} \tag{3.26}$$

Dado que:

$$\rho_{ij} = Corr(X_i, Y_j) = \frac{Cov(X_i, Y_j)}{\sqrt{Var(X_i).Var(Y_j)}} = \frac{a_{ij}\lambda_j}{\sigma_i\sqrt{\lambda_j}} = \frac{a_{ij}\sqrt{\lambda_j}}{\sigma_i}$$

Se for usada a matriz de correlação, tem-se $\sigma_i = 1$, dando origem a,

$$\rho_{ij} = Corr(X_i, Y_j) = a_{ij}\sqrt{\lambda_j} \tag{3.27}$$

Assim, se o valor absoluto de um coeficiente de uma componente principal para uma dada variável for elevado pode-se concluir que, a correlação entre essa componente e a variável é elevada. As componentes principais são interpretadas através destas variáveis[17].

Tendo em conta as relações anteriores, poder-se-á dizer que, as variáveis que devem ser usadas na interpretação da j-ésima componente principal são as que apresentarem coeficientes ρ_{ij} *(loadings)* que obedeçam a uma das regras (sendo a primeira a mais correcta):

1. $|\rho_{ij}| \geq \sqrt{\frac{\lambda_j}{p}}$ ou $\rho_{ij}{}^2 \geq \frac{\lambda_j}{p}$ $(quando\ se\ trabalha\ com\ a\ matriz\ de\ correlação)$
2. $|\rho_{ij}| \geq 0.5$ ou $\rho_{ij}{}^2 \geq 0.25$

   A importância relativa de uma variável X_i para a explicação de uma CP Y_j é dada por:

   $a_{ij}{}^2$, porque sendo o vector a_j normado, $\sum_{i=1}^{p} a_{ij}^2 = 1 \Rightarrow \frac{a_{ij}^2}{\sum_{i=1}^{p} a_{ij}^2} = a_{ij}^2$

O significado a dar a uma componente principal, útil para a interpretação, está intimamente associado com as variáveis às quais correspondam $a_{ij}{}^2$ elevados.

O quadrado do coeficiente de correlação entre X_i e Y_j, $(\rho_{ij}{}^2)$, pode ser interpretado como representando a proporção da variância da variável X_i que é explicada pela componente principal Y_j.

3.2.3.7 Scores

Nesta altura, sabemos que as componentes principais resultam de uma transformação sobre as variáveis em estudo (combinação linear) que, para a j-ésima componente principal se pode formalizar pela equação:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p \tag{3.28}$$

Podemos agora pensar em aplicar a mesma transformação aos dados, ou seja, aos vectores de observações $x_1, x_2, \ldots, x_p$ (colunas da matriz de dados X) das variáveis $X_1, X_2, \ldots, X_p$, respectivamente.

---

17 Uma regra possível será a de considerar elevada uma correlação *(loading)* cujo valor absoluto é igual ou superior a 0.5 [81].

Obtemos uma nova matriz de dados, a matriz *Y*, com dimensão $(n \times r)$ em que o ij-ésimo elemento será igual ao *score* do i-ésimo objecto para a j-ésima componente principal

$$y_{ij} = a_{1j}x_{i1} + a_{2j}x_{i2} + \cdots + a_{pj}x_{ip} \tag{3.29}$$

A matriz dos *scores* dos objectos é dada por:

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nr} \end{bmatrix} \tag{3.30}$$

3.2.3.8 Representações gráficas

As representações gráficas são um óptimo auxiliar na interpretação dos resultados do PCA. No desenvolvimento deste trabalho, recorreu-se ao Gnuplot para efectuar as representações.

(i) Representação das variáveis (*loadings*)

Na representação gráfica das *loadings,* cada ponto representa uma variável e o plano é definido por dois eixos correspondentes a um par de componentes principais. A cada variável é associado um ponto, cujas coordenadas são as correlações dessa variável com cada uma das componentes principais em causa.

(ii) Representação dos objectos no novo sistema de eixos (scores)

Na representação gráfica dos *scores,* cada ponto representa um objecto e tal como no caso anterior, o plano é definido por dois eixos correspondentes a um par de componentes principais. A cada objecto é associado um ponto, cujas coordenadas são os *scores* desse objecto para cada uma das componentes principais. Em geral escolhem-se preferencialmente as duas primeiras componentes, porque são as que mais contribuem para a explicação da variabilidade dos dados.

3.2.3.9 Utilização das componentes principais

A análise de componentes principais pode ser usada para avaliar a importância das variáveis originais com maior peso *(loadings)*. Portanto, a tarefa do químico que trabalha com estatística multivariada, consiste em interpretar a distribuição de pontos no gráfico de componentes principais e identificar as variáveis originais com maior peso na combinação linear das componentes principais mais importantes.

As componentes principais podem ser usadas para testar a normalidade das variáveis iniciais: se as componentes principais não forem normalmente distribuídas, as variáveis originais não o serão. Podem ser usadas, também, na detecção de *outliers*[18]. Por exemplo, num histograma de cada uma das componentes principais ou nas representações dos objectos efectuadas nos planos principais, podem ser identificados objectos a que correspondem valores demasiado elevados ou demasiado baixos, distinguindo-se da globalidade dos restantes. Por vezes, são utilizadas na análise de regressão: determinam-se as primeiras componentes principais, relativamente ao conjunto das variáveis independentes, aplicando-se depois a regressão às componentes seleccionadas. Esta técnica é útil sobretudo para ultrapassar o problema da multicolinearidade, dado que, as componentes principais não são correlacionadas. Outras importantes aplicações são a detecção de grupos e a classificação dos objectos. Nestes casos, se as duas primeiras componentes explicarem uma boa parte da variabilidade total, podemos representar os *scores* dos objectos no plano definido por estas duas componentes e tentar visualizar agrupamentos dos pontos obtidos. No caso de haver necessidade de utilizar mais do que duas componentes usam-se os *scores* dos objectos para as componentes mais importantes, em vez dos valores iniciais das variáveis (que eram em maior número), e constroem-se a partir deles os grupos, utilizando uma das técnicas de classificação.

[18] Valores discrepantes, que não pertencem a uma determinada distribuição.

3.2.4 Análise discriminante linear

A análise discriminante (DA) é uma técnica de classificação que permite encontrar as características que distinguem os membros de um grupo dos membros de outro grupo, de forma que conhecendo as características de um novo objecto se possa prever a que grupo esse objecto pertence. Especificamente, pode ser utilizada para avaliar diferenças entre grupos, determinar formas para distinguir grupos e classificar novos grupos. O objectivo principal consiste em encontrar uma ou mais dimensões que maximizem a distinção entre grupos, estimando uma ou mais funções discriminantes que vão permitir prever a que grupo pertencem os objectos não classificados. Neste processo, pode ser utilizada uma ou mais variáveis independentes e quantitativas (por exemplo, variáveis contínuas). As variáveis independentes devem ter um potencial predictor elevado.

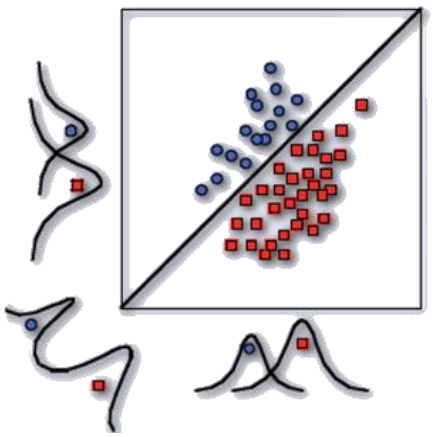


Figura 3.7 – Exemplificação da separação entre dois grupos de objectos por análise discriminante linear.

A análise discriminante exige alguns pressupostos que devem ser observados:

(i) O número de variáveis independentes deve ser muito inferior ao número de casos, dado que o poder discriminante aumenta com o número de casos se o número de variáveis se mantiver constante.

(ii) As variáveis independentes devem ter distribuição normal multivariada nas populações de onde provêm os diferentes grupos (existe uma regra empírica conhecida que prova que a análise discriminante é robusta a desvios da normalidade, desde que o tamanho do menor grupo seja maior que 20 e se o número de variáveis for menor do que 5).

(iii) A variabilidade dentro dos grupos deve ser a mesma, isto é, deve existir homogeneidade das matrizes de variância e covariância para todos os grupos.

(iv) Os *outliers* devem ser identificados e removidos, dado que a análise discriminante é muito sensível à inclusão dos mesmos.

A expressão "Análise Discriminante" tem sido utilizada para identificar diversas técnicas multivariadas que, no entanto, têm um objectivo comum. Parte-se do conhecimento de que os *n* objectos observados pertencem a diversos subgrupos e procura-se determinar

funções das *p* variáveis observadas que melhor permitam distinguir ou discriminar entre esses subgrupos ou classes.

As componentes principais relativas ao PCA, abordado na secção 3.2.3.3, não são necessariamente boas soluções para efeitos de discriminação, pois as direcções de maior variabilidade não têm que coincidir com as direcções de melhor discriminação [142]. Na análise discriminante coloca-se explicitamente o objectivo de separar subgrupos de indivíduos, subgrupos esses que são previamente conhecidos nos dados observados.

Nesta secção será abordada uma técnica discriminante, válida no contexto descritivo onde nos situamos, conhecida por análise discriminante linear, ou de *Fisher* [64]. Existem outras técnicas discriminantes, nomeadamente técnicas que se baseiam em modelos probabilísticos, que não serão abordadas neste trabalho.

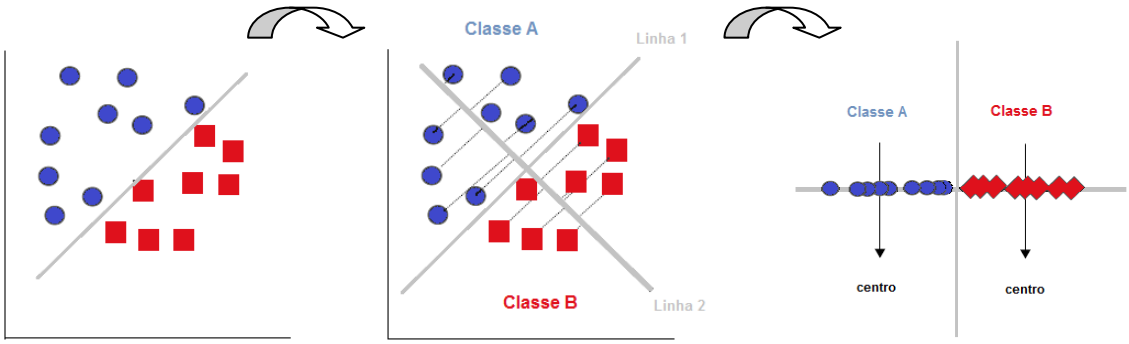
A análise discriminante linear (LDA) tem a virtude de ser facilmente visualizável em termos geométricos. Além disso, não exige hipóteses adicionais, ao contrário das técnicas baseadas em modelos probabilísticos. Tem também a vantagem de permitir discriminar mais do que dois subgrupos ou classes diferentes sem grande complexidade.

Na análise discriminante linear procuram-se as combinações lineares *Xa* das *p* variáveis observadas que melhor separam os subgrupos de indivíduos indicados, segundo um critério de separabilidade. As soluções *Xa* obtidas designam-se por funções ou eixos discriminantes. Estes eixos podem ser utilizados para obter uma representação gráfica que saliente a distinção entre as classes. Podem também ser úteis para classificar futuros indivíduos (observados nas mesmas variáveis), desde que seja desconhecido à partida o subgrupo a que pertencem.

De um modo geral, a análise discriminante linear tenta encontrar uma transformação linear através da maximização da distância entre classes e minimização da distância dentro de classes. Este método procura a melhor direcção de forma que quando os dados são projectados num plano, as classes possam ser separadas.

Considere-se como exemplo a realização de duas medições e a produção de um gráfico dos valores destas medições para os dois grupos, como ilustrado na Figura 3.8. Os objectos representados a azul são claramente distintos dos objectos representados a vermelho, mas nenhuma das duas medidas pode, por si só, discriminar entre estes dois grupos e, portanto, ambos são essenciais para a classificação. É possível no entanto, traçar uma linha entre os dois grupos. Neste caso concreto, se um objecto se encontrar acima da linha pertence à classe A, caso contrário pertence à classe B.

Graficamente podemos projectar os objectos sobre uma linha de projecção como representado na Figura 3.8b. A projecção pode se convertida numa posição ao longo da linha 2. A distância pode ser convertida num dado valor numérico, análogo a um *score*. Objectos com valores mais baixos pertencem à classe A, enquanto objectos com valores mais elevados pertencem à classe B. É possível determinar a associação de classe simplesmente de acordo com o facto de o valor estar acima ou abaixo do divisor. Por outro lado, é possível determinar o centro de cada classe ao longo da projecção e se a distância em relação ao centro da classe A é maior do que a distância ao centro da classe B, o objecto é colocado na classe B, dependendo da dispersão de cada classe.



(a) (b) (c)

Figura 3.8 – Discriminação entre duas classes e projecção sobre o eixo discriminante - linha 2 (adaptado de [52]).

3.2.4.1 Algoritmo LDA

Este tipo de análise supervisionada envolve um conjunto de aprendizagem a partir do qual são estimadas as propriedades que maximizam a resolução (discriminação) dos objectos. Com base nesta aprendizagem existem agora condições para se poder classificar os objectos.

O conjunto de aprendizagem, onde já se encontram definidos os grupos, é constituído por um conjunto com N objectos representados sobre M variáveis, previamente classificados em G grupos.

O algoritmo procura maximizar as distâncias entre grupos reduzindo em simultâneo as distâncias dentro dos grupos, aumentando deste modo o poder discriminante dos objectos.

Similarmente à ANOVA de uma via, a variabilidade total pode ser decomposta na variabilidade interna dos grupos (W) e na variabilidade externa (B) que define a separação relativa dos grupos.

A dispersão dos grupos (B) é dada pela matriz de distância de cada centróide dos G grupos ao centro global:

$$B_{(M\times N)} = n_i . b^T_{(M\times G)} . b_{G\times M} \tag{3.31}$$

Já a dispersão interna (W) reflecte a distância de cada objecto ao seu centróide:

$$W_{(M\times M)} = w^T_{(M\times N)} . w_{(N\times M)}) \tag{3.32}$$

A matriz discriminante (D) é dada pelo quociente entre B e W:

$$D_{(M\times M)} = B_{(M\times M)} / W_{(M\times M)} \tag{3.33}$$

Sendo esta matriz uma matriz quadrada e em tudo similar a uma matriz de variância (quociente de variâncias), faz sentido procurar decompor em valores e vectores próprios, sendo os mais relevantes o novo conjunto de eixos ortonormados que permitem maximizar a discriminação dos objectos, isto é, maximizar a variabilidade sobre o menor número de dimensões possível.

A maximização da discriminação dos objectos ($X_{(N\times M)}$) no sub-espaco de dimensao *d* é conseguida através da decomposição da matriz discriminante em valores e vectores próprios

$$X_{(N\times M)} = S_{(N\times d)} . \Lambda_{(d\times d)} . Q'_{(d\times M)} \tag{3.34}$$

onde *S* e *Q* são as matrizes dos *scores* discriminantes e das funções discriminantes desse sub-espaco.

Através do valor das contribuições (*loadings*) das funções discriminantes consegue-se avaliar a relevância de cada variável para a discriminação dos objectos[19].

Atendendo à propriedade de matriz ortonormada

$$Q'_{(d\times M)} . Q_{(M\times d)} = I_{(d\times d)} \tag{3.35}$$

a representação dos objectos no novo espaço discriminante (projecção sobre os vectores próprios) faz-se através do produto das suas componentes pelas componentes do espaço

$$S_{(N\times d)} = X_{(N\times M)} . Q_{(M\times d)} \tag{3.36}$$

[19] No caso de a matriz discriminante ($D_{(M\times M)}$) não ser simétrica, a nova base vectorial não é ortonormada.

3.2.5 Análise por mínimos quadrados parciais

O método dos mínimos quadrados parciais (PLS) é o método estatístico mais usado em calibração multivariada e difere da PCR[20] por usar a informação de *Y* no cálculo das chamadas variáveis latentes (equivalentes às componentes principais na análise PCA). As matrizes *X* e *Y* são decompostas simultaneamente numa soma de variáveis latentes.

O PLS é usado para encontrar relações fundamentais entre as duas matrizes *X* e *Y*, ou seja, trata-se de uma abordagem baseada na variável latente, para modelar as estruturas de covariância nestes dois espaços (*X* e *Y*) [74-77]. Por outras palavras, procura a direcção multidimensional no espaço da matriz *X* que explica a direcção de máxima variação multidimensional no espaço da matriz *Y*. A regressão por mínimos quadrados parciais é particularmente indicada quando a matriz dos predictores (*X*) tem mais variáveis do que observações, e quando existe multicolinearidade entre os valores de *X* [78-80]. É portanto, indicado para prever a resposta do sistema quando o número de variáveis (predictores) é muito grande.

Este algoritmo actua através da maximização da correlação entre os predictores (sub-espaço de *X*) e a resposta (sub-espaço das respostas de *Y*). Inicialmente são analisados em cada sub-espaço, predictores e respostas, no sentido de encontrar combinações lineares de variáveis, designadas de factores latentes, que se encontram correlacionadas e permitem descrever ao máximo a variabilidade dos resultados de cada sub-grupo de valores.

A Figura 3.9 ilustra um esquema do método. A ideia principal é usar os factores para prever as respostas na população. Isto é obtido indirectamente através da extracção de variáveis latentes (*T* e *U*) seleccionadas a partir dos factores e das repostas, respectivamente. Os factores *T* extraídos (também designados por *X-scores*), são usados para prever as respostas U (também designados por *Y-scores*), que, por sua vez, serão usadas para prever a resposta.

[20] Do inglês *Principal Component Regression*

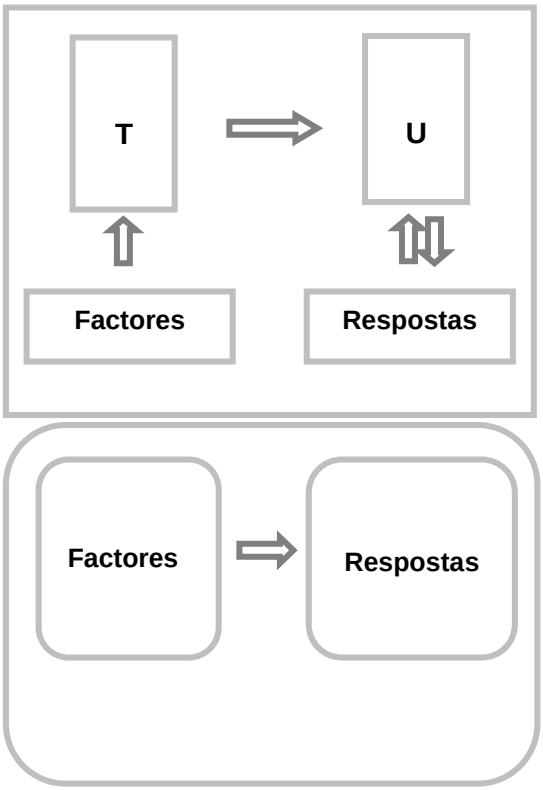


Figura 3.9 – Esquema correspondente ao método PLS (modelação implícita)

No conjunto de dados inicial cada objecto é avaliado sob um conjunto de variáveis (m+p). Este conjunto de valores constitui dois sub-espaços: de predictores, $X_{(n\times m)}$

$$X_{(n\times m)} = \begin{Bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{Bmatrix} \tag{3.37}$$

e de respostas, $Y_{(n\times p)}$

$$Y_{(n\times p)} = \begin{Bmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nj} & \cdots & y_{np} \end{Bmatrix} \tag{3.38}$$

Sendo n o número de objectos, m o número de variáveis do sub-espaço predictor e p o número de variáveis do sub-espaço resposta.

O sub-espaço dos predictores apresenta os objectos representados sob o espaço constituído pelas variáveis independentes que originam as respostas do sistema. Estes são os valores de entrada no sistema. O sub-espaço das respostas apresenta os mesmos objectos representados sobre as respectivas variáveis dependentes que resultam das condições impostas ao sistema em estudo. Estas variáveis dependem, através de uma função previamente desconhecida, dos estímulos fornecidos ao sistema.

Designa-se por factor latente a combinação linear de variáveis que consegue descrever uma fracção da variabilidade de cada sub-espaço de valores.

O algoritmo PLS é uma versão mais complexa do parente mais simples OLS[21]. O ajuste polinomial simples através de OLS baseia-se num modelo polinomial do tipo

$$\eta_i = a_0 + a_1.x_{1i} + a_2.x_{2i} + a_{12}.x_{1i}.x_{2i} + \ldots \tag{3.39}$$

onde a resposta prevista (η_i) depende linearmente de um termo constante (a_0) e das variáveis do sistema (x_{ij}).

O algoritmo PLS assume que o conjunto de respostas (η_{ij}) é descrito através de uma combinação linear de variáveis independentes (x_{ij}),

$$q_f.y_{ij} = \sum_f a_f.w_f.x_{ij} \tag{3.40}$$

onde w_f representa o factor latente do sub-espaço dos predictores ($X_{(n\times m)}$) que está relacionado com o factor latente do sub-espaço das respostas ($Y_{(n\times p)}$).

O algoritmo procura maximizar a correlação entre factores latentes dos sub-espaços predictor e resposta no sentido de procurar justificar o máximo da variabilidade da resposta obtida. Após encontrada a melhor relação linear, esse efeito é subtraído aos respectivos sub-espaços. O modelo, obtido por via implícita, vai sendo construído iterativamente permitindo obter uma descrição tanto mais fiável quanto se pretenda.

Após a remoção da resposta característica, cada sub-espaço deve conter ruído aleatório, valores independentes de qualquer interacção.

[21] Do inglês *Ordinary Least Squares*

3.2.5.1 Algoritmo PLS

Sendo necessário maximizar a correlação entre os sub-espaços X e Y, calculam-se as matrizes de covariância de $X_{(n\times m)}$ em $Y_{(n\times p)}$,

$$(n-1).Cx_{(mxm)} = X^T_{(mxn)} Y_{(nxp)} Y^T_{(pxn)} X_{(nxm)} = \left| Y^T_{(pxn)} X_{(nxm)} \right|^2 \tag{3.41}$$

e de covariância de $Y_{(n\times p)}$ em $X_{(n\times m)}$,

$$(m-1).Cy_{(pxp)} = Y^T_{(pxn)} X_{(nxm)} X^T_{(mxn)} Y_{(nxp)} = \left| X^T_{(mxn)} Y_{(nxp)} \right|^2 \tag{3.42}$$

Chama-se traço de uma matriz ao produto dos termos da diagonal principal. É necessário que o valor obtido em ambas as matrizes de covariância, *Cx* e *Cy*, seja igual. Este valor corresponde à variância global do sistema.

Em qualquer instante do processo iterativo, a soma de quadrados de cada sub-espaço pode ser estimada com base nas equações

$$SSx = \sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}^2 \qquad SSy = \sum_{i=1}^{p}\sum_{j=1}^{p} y_{ij}^2 \tag{3.43}$$

À medida que os factores são calculados e subtraídos à matriz de covariância estas tendem para matrizes nulas. Sendo as variáveis previamente centradas e escaladas, a soma de quadrados de *X* e de *Y* é dada pela soma dos termos de *X* e *Y* ao quadrado já que estas matrizes se encontram normalizadas.

O primeiro valor próprio de *Cx* ($\lambda x1$) é o valor próprio mais significativo, com maior valor

$$\left| Cx_{(mxm)} - \lambda x_1 . I_m \right| = 0 \tag{3.44}$$

assim como o primeiro valor próprio de *Cy* (λy_1)

$$\left| Cy_{(pxp)} - \lambda y_1 . I_p \right| = 0 \tag{3.45}$$

também será o maior valor em causa.

Atendendo às operações efectuadas em 3.41 e 3.42, estes valores próprios estão intimamente relacionados – a sua correlação foi maximizada através da covariância cruzada entre os sub-espaços predictor-resposta.

Esta decomposição matricial através dos seus valores próprios faz com que surja agora um novo sistema de vectores directores ortonormalizado definido através dos respectivos vectores próprios

$$\left\{Cx_{(mxm)} - \lambda x_1 . I_m\right\} . w_{1(mx1)} = 0 \quad \wedge \quad \left|w_1\right| = 1 \tag{3.46}$$

$$\left\{Cy_{(pxp)} - \lambda y_1 . I_p\right\} . q_{1(px1)} = 0 \quad \wedge \quad \left|q_1\right| = 1 \tag{3.47}$$

Para um determinado factor latente *f*, extraído por via decrescente de importância, vai existir uma relação estabelecida entre os sub-espaços *X* e *Y* que está explicitamente definida através dos respectivos pesos *(loadings)*. Assim, o vector próprio w_f do sub-espaço predictor e q_f do sub-espaço resposta

$$w_f\,(m \times 1) = \begin{Bmatrix} w_{f1} \\ \vdots \\ w_{fi} \\ \vdots \\ w_{fm} \end{Bmatrix} \qquad q_f\,(p \times 1) = \begin{Bmatrix} q_{f1} \\ \vdots \\ q_{fi} \\ \vdots \\ q_{fp} \end{Bmatrix} \tag{3.48}$$

são constituídos por componentes, w_{fj} e q_{fk}, que traduzem o impacto das variáveis originais sobre esse factor latente e, deste modo indicam quais as inter-relações entre variáveis de ambos os sub-espaços.

Se existirem *p* respostas, é previsível que o número de factores a considerar para descrever esse sistema seja próximo deste valor.

Este algoritmo visa essencialmente simplificar o problema multidimensional inicial. Assim, importa estabelecer critérios para recuperar uma fracção significativa da variabilidade inicial do sub-espaço resposta.

É portanto necessário estabelecer critérios de aceitação e de rejeição de factores latentes no sentido de estabelecer qual o número mínimo de factores latentes que devem ser considerados para descrever o sistema em causa.

Um dos critérios mais utilizados consiste em tentar reproduzir cerca de 80% da resposta original. Sendo este critério bastante significativo em termos de reprodução do sistema, é geralmente difícil obter índices tão elevados de desempenho com sistemas estocásticos onde as inter-relações das variáveis são obscurecidas através de causas desconhecidas.

Outro critério também utilizado consiste em representar o valor paramétrico obtido na modelação ou a variância incremental descrita por esse factor através de um gráfico de desempenho (*scree plot*). Os factores latentes significativos são reconhecidos por estarem anormalmente acima da contribuição basal, relacionada com a contribuição aleatória. Esta alternativa é, regra geral, bastante robusta para evidenciar os factores latentes mais relevantes.

Estes factores latentes eleitos podem agora ser utilizados para evidenciar informação preciosa sobre a interacção entre variáveis do sistema em causa, através das respectivas *loadings* mais significativas.

Para se reconhecer as contribuições mais relevantes estabelecidas por determinado factor latente que cruza os sub-espaços causa-efeito, é necessário efectuar uma inspecção minuciosa às respectivas contribuições (*loadings)*. É necessário catalogar as contribuições de cada sub-espaço por ordem decrescente de importância e garantir que uma fracção significativa do módulo do vector latente é recuperada.

3.2.6 Algoritmo *convex hull*

A determinação da "fronteira convexa" *(convex hull)* de um conjunto finito de pontos no plano é um dos mais antigos problemas considerados na definição da geometria computacional. Uma das suas principais aplicações é o reconhecimento de padrões [157].

O *convex hull* de um conjunto S de pontos é intuitivamente fácil de descrever. Um conjunto S de pontos do plano é convexo, se para quaisquer dois pontos x e y de S, o segmento entre x e y está totalmente contido em S. Na Figura 3.10 encontram-se representados dois exemplos de conjuntos convexos e não convexos.

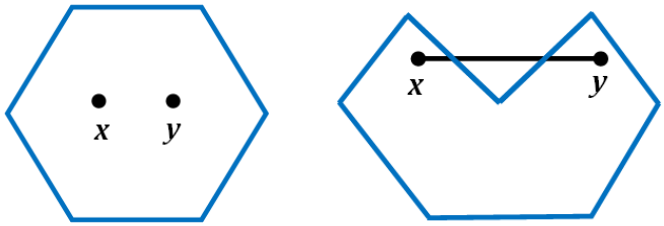


Figura 3.10 – Exemplo de um conjunto convexo (a) e de um conjunto não convexo (b).

Existem diversos algoritmos para a obtenção dos conjuntos convexos, tais como o algoritmo de Graham, o algoritmo de Jarvis e o algoritmo *QuickHull*. O algoritmo desenvolvido neste trabalho, o *Convexhull2D*, baseia-se nos princípios do algoritmo *QuickHull* e será descrito a seguir [158]. A Figura 3.11 ilustra o conceito inerente ao cálculo geral deste algoritmo.

Dado um conjunto de pontos qualquer, na Figura 3.11a, o algoritmo *Convexhull2D* inicia-se pela procura dos pontos externos numa das direcções, por exemplo, ao longo do eixo das abcissas. Traça-se um segmento de recta ligando esses pontos, conforme ilustrado na Figura 3.11b, e considera-se o ponto de maior distância em relação ao segmento de recta obtido anteriormente, num dos sentidos da direcção perpendicular a este segmento. A partir do ponto extremo encontrado é gerado um triângulo com os pontos extremos do segmento de recta. Este passo é demonstrado na Figura 3.11c. Posteriormente, encontra-se, no sentido contrário, o ponto perpendicularmente mais distante à recta, conforme representado na Figura 3.11d. A partir do ponto obtido no passo anterior, é traçado um segundo triângulo ligando-o aos pontos extremos da recta. A união dos dois triângulos, representada na Figura 3.11e forma um quadrilátero. Adicionalmente, os pontos internos deste quadrilátero não precisam de ser novamente analisados, restando para apenas os pontos externos ao polígono. Para cada lado do quadrilátero, representado por um segmento de recta, considera-se o ponto de maior distância em relação à perpendicular do segmento de recta, ligando-o aos pontos extremos do mesmo. Desta forma, o quadrilátero é ampliado em cada um de seus lados que têm pontos exteriores, formando um polígono. O passo anterior é repetido para cada lado do polígono, até que não existam mais pontos externos. Desta forma, é então possível formar um conjunto convexo com os pontos dados, como ilustrado na Figura 3.11f.

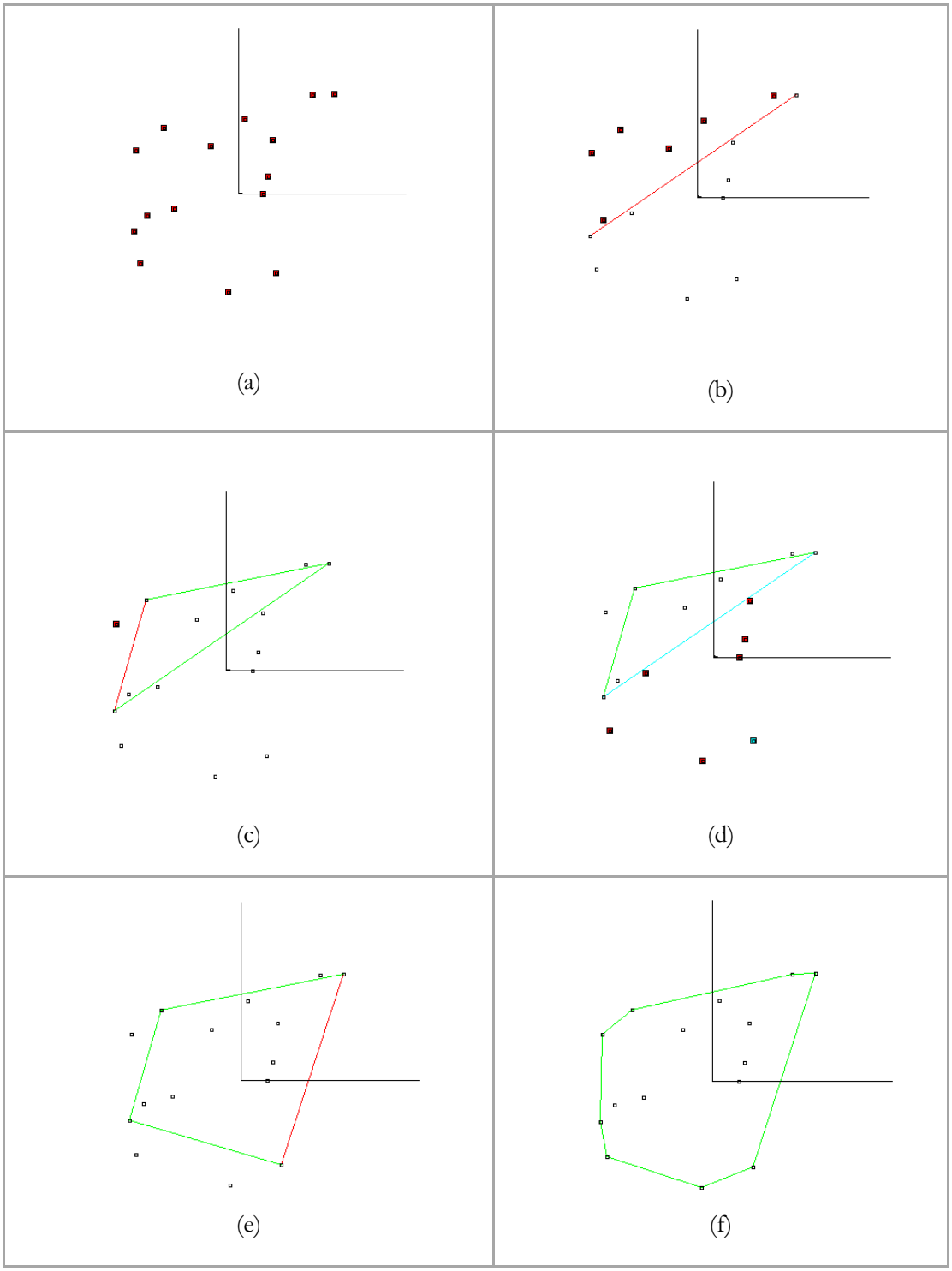


Figura 3.11 – Representação das etapas de formação do conjunto convexo obtido pelo algoritmo convexhull2D: (a) pontos aleatórios no plano bidimensinal; (b) procura de pontos externos no eixo das abcissas; (c) representação da recta que une os pontos; (d) construção do triângulo com os pontos externos; (e) formação do quadrilátero; (f) formação do conjunto convexo.

Capítulo 4

Metodologia proposta e bases de dados

A metodologia proposta neste trabalho baseia-se na utilização de quatro métodos clássicos, HCA, PCA, LDA e PLS, no tratamento de dados provenientes de diferentes domínios. São apresentadas e discutidas três aplicações fundamentais, bastante diferentes entre si. As duas primeiras, estão associadas a dados de diagnóstico médico, e a terceira, insere-se num outro campo de aplicação, a cientometria[22]. Especificamente, na área do diagnóstico médico são tratados três exemplos distintos, relativos ao cancro e à doença de Parkinson. No último exemplo, na área da cientometria, é feito um estudo comparativo entre várias universidades ibéricas.

Na secção anterior, foram descritos detalhadamente cada um dos métodos utilizados. Pretendemos também, mostrar as vantagens da utilização prévia da limpeza de ruído/detecção de *outliers* (NR) na análise de dados [13].

Situações em que os dados contêm *outliers* ou valores menos característicos tornam a estrutura dos dados menos definida e o número de grupos que se formam pode ser muito elevado ou muito reduzido, dependendo da situação real e do algoritmo. A identificação e remoção de *outliers*, numa etapa preliminar, são sugeridas através de uma técnica em que as propriedades do sistema são usadas para efectuar a especificação automática dos parâmetros necessários. A presença de *outliers*, discutida no próximo capítulo, pode ter consequências diferentes em diferentes algoritmos. Podem ser considerados como objectos ou pequenos grupos de objectos localizados em zonas de baixa densidade, contrastando com a densa estrutura intragrupo. Numa perspectiva ligeiramente diferente, os *outliers* podem ser considerados como objectos com baixa conectividade na região intragrupo [13].

[22] Definida como o estudo da mensuração do progresso científico e tecnológico e que consiste na avaliação quantitativa e na análise das inter-comparações da actividade, produtividade e progresso científico. Por outras palavras, a cientometria consiste em aplicar técnicas numéricas analíticas para estudar a ciência.

Uma forma eficiente de visualizar os grupos formados em 2D é alcançada através da representação na forma de *convex hull*. A complexidade dos algoritmos correspondentes é, geralmente, estimada em termos do número de pontos de entrada *(input)*, e o número de pontos no *convex hull* [143-145].

Todos os algoritmos utilizados neste trabalho foram optimizados pelos autores e executados com o software GNU Octave (versão 3.2.4). As representações gráficas foram realizadas recorrendo ao Gnuplot (versão 3.2.4).

Neste trabalho, propomos uma abordagem automatizada que consiste em quatro etapas principais: (i) definição da estrutura dos dados baseada no HCA, (ii) visão geral dos dados e selecção de variáveis baseada no PCA, (iii) LDA para a resolução de diferentes classes e (iv) PLS para encontrar relações entre as variáveis previstas e observáveis. Uma etapa adicional, que pode auxiliar na caracterização do conjunto de dados consiste na remoção dos pontos menos característicos (NR[23]) do conjunto de dados a tratar. A sua implementação depende do problema em estudo.

Na apresentação dos resultados, no Capítulo 5, será dada ênfase ao tipo de abordagem seleccionada em cada caso, nomeadamente na análise PCA. O procedimento usado na aplicação deste método consiste na selecção prévia do tipo de matriz a processar (covariância ou correlação), seguida da escolha do número de componentes principais com base nos três critérios disponíveis (critério de Pearson, critério de Kaiser e o *scree plot*). Após a representação dos objectos no novo sistema de eixos *(scores)*, que permite detectar possíveis grupos e as suas posições relativas, faz-se a análise das contribuições (*loadings)* para seleccionar as variáveis responsáveis pela distribuição desses grupos. Esta selecção é feita com base em dois métodos descritos e discutidos no próximo capítulo.

No diagnóstico médico, mais especificamente no diagnóstico do cancro, esta metodologia atinge um grau de confiabilidade comparável às abordagens mais sofisticadas. No entanto, nem sempre é necessário considerar a sequência completa destes métodos, fazendo-se uma selecção cuidadosa da abordagem para cada caso.

[23] Do inglês *Noise Reduction*

4.1 Bases de dados

Os dados relativos ao diagnóstico do cancro da mama e da doença de Parkinson foram obtidos por consulta da base de dados da Universidade da Califórnia, designada por *UCI Machine Learning Repository* [146].

Os dados referentes ao estudo epidemiológico foram obtidos a partir da ACS (*American Cancer Society, Surveillance and Health Policy Research),* que disponibiliza os dados gratuitamente na Internet [147].

Os dados relativos às Universidades ibéricas podem ser consultados na *Essencial Science Indicators*, base de dados também disponível na Internet [148].

O primeiro exemplo tratado com a metodologia proposta, corresponde ao conhecido problema da determinação de cancro da mama no estado de *Winsconsin*, tendo em conta algumas características citológicas, que determinam se o tumor é benigno ou maligno.

O segundo exemplo diz respeito à classificação de tecidos mamários através de parâmetros inferidos de espectros de impedância.

No exemplo seguinte, o conjunto de dados resulta de uma estimativa da incidência de novos casos de cancro nos EUA, no ano de 2010, considerando uma selecção de alguns tipos de cancro. No último exemplo, é feito um estudo comparativo entre várias universidades ibéricas tendo em conta as suas áreas de *ranking*. Em concordância com a base de dados original, mantiveram-se as designações correspondentes a cada conjunto de dados.

4.1.1 *Wisconsin breast cancer database*

Um dos aspectos importantes no diagnóstico de cancro é distinguir os tumores malignos dos tumores benignos [149]. A primeira aplicação corresponde a um estudo realizado em 699 pacientes, com nove atributos, num conjunto organizado em duas classes. Este conjunto de dados, refere-se ao diagnóstico de cancro da mama baseado em observações microscópicas associada à fisiologia de células, incluindo a espessura, a uniformidade de tamanho e da forma das células, a adesão marginal, o tamanho das células epiteliais, o número de núcleos vazios, a estrutura da cromatina, a quantidade de nucléolos normais e o grau de mitoses (para uma descrição mais detalhes consultar as referências [150-152]).

Por simplicidade de leitura e compreensão deste problema foi atribuída uma notação específica a todas as variáveis descritas na Tabela 4.1.

Tabela 4.1 – Descrição das variáveis em estudo tendo em conta as designações originais da base de dados [151].

Attributes	***Notation***
Clump Thickness	x1
Uniformity of Cell Size	x2
Uniformity of Cell Shape	x3
Marginal Adhesion	x4
Single Epithelial Cell Size	x5
Bare Nuclei	x6
Bland Chromatin	x7
Normal Nucleoli	x8
Mitosis	x9
Class	*1 for benignant, 10 for malignant*

Esta base de dados reflecte o agrupamento cronológico de alguns casos clínicos. Cada variável foi convertida num atributo numérico com valores variando de 1 a 10. A inspecção prévia do conjunto de dados revelou que 16 pacientes apresentavam *"missing values"*. Estes casos foram removidos, levando a um total de 683 pacientes, sendo 239 correspondentes a casos malignos e 444 a casos benignos. Esta base de dados será designada por "*breast cancer 1*" na análise que se segue.

4.1.2 *Breast cancer classification data*

Este segundo exemplo corresponde a um problema de classificação do tecido mamário, com base em nove características associadas a medições de impedância eléctrica em amostras de tecido fresco retirado do peito. Os detalhes sobre o procedimento de recolha de dados, assim como a classificação dos casos e as frequências utilizadas estão descritos em detalhe nas referências [153,154]. Este conjunto de dados contém informação sobre seis tecidos que se encontram agrupados em duas classes principais de tecidos normais e patológicos. Do grupo de tecidos normais fazem parte, o tecido glandular (com 16 amostras), o tecido conjuntivo (com 14 amostras) e o tecido adiposo (com 22 amostras), representados na Figura 4.1. O grupo de tecidos patológicos é constituído pelo tecido característico do carcinoma (21 amostras), o fibro-adenoma (15 amostras) e o tecido característico de mastopatia[24] (18 amostras).

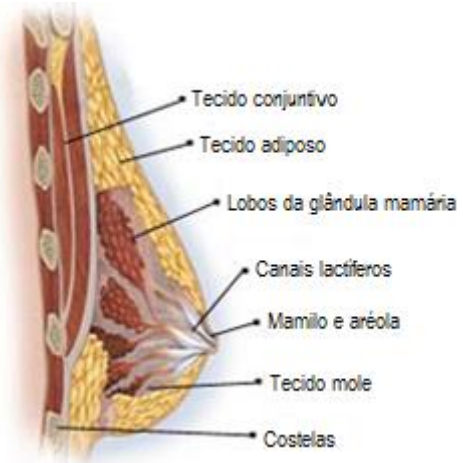


Figura 4.1 – Esquema representativo dos principais tecidos constituintes da mama: o tecido glandular (lobos), o tecido conjuntivo e o tecido adiposo (adaptado de [155]).

Na Tabela 4.2 encontram-se descritas as classes de tecidos em estudo e o correspondente número de amostras, respeitando as designações originais da base de dados.

Tabela 4.2 – Descrição das classes em estudo e o respectivo número de amostras tendo em conta as designações originais da base de dados [154].

Pathological tissue classes	***Number of cases***
Carcinoma	21
Fibro-adenoma	15
Mastopathy	18
Normal tissue classes	
Glandular tissue	16
Connective tissue	14
Adipose tissue	22

[24] Nome que se dá ao conjunto de modificações patológicas da glândula mamária.

As seis classes de tecidos, descritas na Tabela 4.2, foram estudadas através de medidas de impedância eléctrica, usando nove variáveis como predictores: a medida de impedância (ohm) na frequência zero (I0), o ângulo de fase em 500 KHz (PA500), o declive do ângulo de fase de alta frequência (HFS), a distância entre as extremidades do espectro de impedância (DA), a área sob o espectro (AREA), a área normalizada pela distância DA (A/DA), o máximo do espectro (IP MAX), a distância entre a impedância e a parte real do ponto de frequência máxima (DR) e o comprimento da curva espectral (P). Estes parâmetros foram obtidos a partir de espectros de impedância eléctrica, sobre o plano de Argand[25] como ilustrado no gráfico da Figura 4.2.

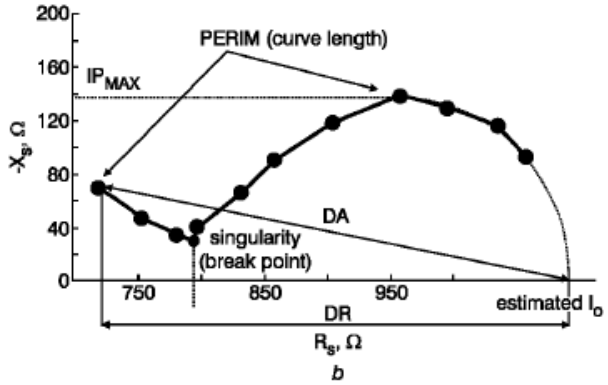


Figura 4.2 – Medidas de impedância eléctrica representadas no plano de Argand, estudadas para as seis classes de tecidos (retirado de [154]).

À semelhança do caso anterior, todas as variáveis serão representadas por x1 a x9, respectivamente. A Tabela 4.3 apresenta as atribuições efectuadas às variáveis em estudo, respeitando as designações originais da base de dados. Esta base de dados será designada por "*breast cancer 2*" no estudo que segue.

[25] Plano cartesiano usado para representar números complexos geometricamente. Neste diagrama, a parte imaginária de um número complexo é representada pela ordenada e a parte real pela abcissa.

Tabela 4.3 – Atribuição imposta às variáveis de acordo com as designações originais da base de dados [154].

Attributes	***Notation***
Impedivity (ohm) at zero frequency – I0	x1
Phase angle at 500 KHz – PA500	x2
High-frequency slope of phase angle – HFS	x3
Impedance distance between spectral ends – DA	x4
Area under spectrum – AREA	x5
Area normalized by DA – A/DA	x6
Maximum of the spectrum – MAX IP	x7
Distance between impedivity (I0) and real part of the maximum frequency point – DR	x8
Length of the spectral curve – P	x9

4.1.3 *New cancer cases estimated data*

Os dados analisados neste caso correspondem aos números de incidência de novos casos de cancro estimados em 2010, para 11 tipos de cancro seleccionados e apresentados pelos residentes dos 51 estados dos EUA. Os tipos de cancro seleccionados são: o cancro da mama, do colo uterino, do cólon e recto, do *corpus* do útero, leucemia, cancro do pulmão e brônquios, melanoma da pele, linfoma não-Hodgkin[26] e os cancros da prostata e bexiga. A décima primeira variável contém "outros casos"obtidos a partir da diferença entre o total de novos casos e o número de casos registados em cada estado. Os tipos de cancro seleccionados estão descritos na Tabela 4.4 de acordo com as designações originais da base de dados [147].

[26] Neoplasia maligna de células linfóides, correspondente a um dos vários tipos de cancro do sistema linfático.

Tabela 4.4 – Tipos de cancro seleccionados mantendo as designações originais da base de dados [147].

Cancer types	***Notation***
Female breast cancer	*x1*
Uterine cervix	*x2*
Colon & Rectum	*x3*
Uterine Corpus	*x4*
Leukemia	*x5*
LeukemiaLung & Bronchus	*x6*
Melanoma of the skin	*x7*
Non-Hodgkin Lymphoma	*x8*
Prostate	*x9*
Urinary Bladder	*x10*
Other cancer types	*x11*

4.1.4 *Parkinsons Telemonitoring dataset*

O conjunto de dados correspondente a este exemplo é constituído por uma gama de medidas de voz obtidas a partir de 42 indivíduos com a doença de Parkinson em estado inicial. Os pacientes foram recrutados para um estudo de seis meses, usando um dispositivo para a monitorização à distância da progressão dos sintomas da doença. As gravações foram automaticamente recolhidas em casa de cada paciente. Os detalhes inerentes ao procedimento de recolha de dados podem ser consultados na referência [156]. As variáveis em estudo são: a idade e o género, o intervalo de tempo a partir da data de recrutamento, UPDRS[27] motor, UPDRS total e 16 medidas de voz, definidas na Tabela 4.5, de acordo com as designações originais da base de dados. Cada linha no conjunto de dados corresponde a uma das 5875 gravações de voz destes indivíduos. Existem cerca de 200 gravações por paciente.

[27] Do inglês *Unified Parkinson´s Disease Rating Scale*, reflecte a presença e o grau de progressão da doença, mas não mede as causas subjacentes.

Tabela 4.5 – Definição das variáveis de acordo com as designações originais da base de dados [156].

Attributes	*Definition*	*Notation*
Age	*Subject age*	*x1*
Sex	*Subject gender : '0' - male, '1' - female*	*x2*
Test time	*Time since recruitment into the trial*	*x3*
Jitter(%),Jitter(Abs), Jitter:RAP,Jitter:PPQ5,Jitter:DDP	*Several measures of variation in fundamental frequency*	*x4-x8*
Shimmer,Shimmer(dB),Shimmer:APQ3, Shimmer: APQ5, Shimmer: APQ11, Shimmer: DDA	*Several measures of variation in amplitude*	*x9-x14*
NHR, HNR	*Two measures of ratio of noise to tonal components in the voice*	*x15,x16*
RPDE	*A nonlinear dynamical complexity measure*	*x17*
DFA	*Signal fractal scaling exponent*	*x18*
PPE	*A nonlinear measure of fundamental frequency variation*	*x19*

4.1.5 As universidades ibéricas e as suas áreas de *ranking*

Neste exemplo é feito um estudo comparativo entre 55 universidades ibéricas tendo em conta 22 áreas de *ranking,* dadas pela *Essential Science Indicators* (ESI), base de dados disponível na Internet (consulta em Janeiro e Julho de 2010). São consideradas na base ESI as citações científicas de países, como os países ibéricos com, pelo menos, cem mil citações para o conjunto das 22 áreas científicas e contempladas por esta base de dados num período de 10 anos. As actualizações da ESI têm sempre um certo diferencial em relação à base de dados originais (*Thomson Scientific–indexed journal articles*), e a partir da qual são elaboradas. Os dados da ESI de Julho de 2010 cobrem o período de 10 anos+4-meses, 1 de Janeiro de 2000 a 30 de abril de 2010. A ESI apresenta as áreas científicas de cada instituição (*Field Rankings*) desde que se localizem em citações no *top* 1% a nível mundial da respectiva área. As áreas de *ranking* são apresentadas na Tabela 4.6 em concordância com esta base de dados.

Tabela 4.6 – Áreas científicas contempladas pela ESI, respeitando as designações desta base de dados[148].

Scientific disciplines	
Agricultural Science	*Mathematics*
Biology & Biochemistry	*Microbiology*
Chemistry	*Molecular Biology & Genetics*
Clinical Medicine	*Multidisciplinary **
Computer Science	*Neuroscience & Behavior*
Economics & Business	*Pharmacology*
Engineering	*Physics*
Environment/ Ecology	*Plant & Animal Science*
Geosciences	*Psychiatry/Psychology*
Immunology	*Social Sciences – general*
Materials Science	*Space Science*

Capítulo 5

Resultados e discussão

O procedimento descrito no capítulo anterior foi aplicado a cinco sistemas diferentes. Nesta secção apresentamos os resultados mais relevantes obtidos pela metodologia proposta. Para maior clareza e simplicidade, cada caso será analisado separadamente, no entanto, as principais conclusões serão dadas em conjunto.

5.1 *Breast cancer 1*

No primeiro caso, o conjunto de dados inicial é constituído por 683 casos diagnosticados através de 9 predictores. De acordo com diagnóstico médico, 444 indivíduos correspondem a casos benignos (atribuído o valor 1) e 239 a casos malignos (atribuído o valor 10), numa proporção aproximada de 2:1.

A Figura 5.1 apresenta a frequência dos valores para cada variável. Com excepção das variáveis correspondentes à estrutura da cromatina (x7) e à espessura do conjunto de células (x1), as restantes sete variáveis exibem uma incidência elevada nos valores de 1 e 2, este último, apenas na variável correspondente ao tamanho das células epiteliais (x5).

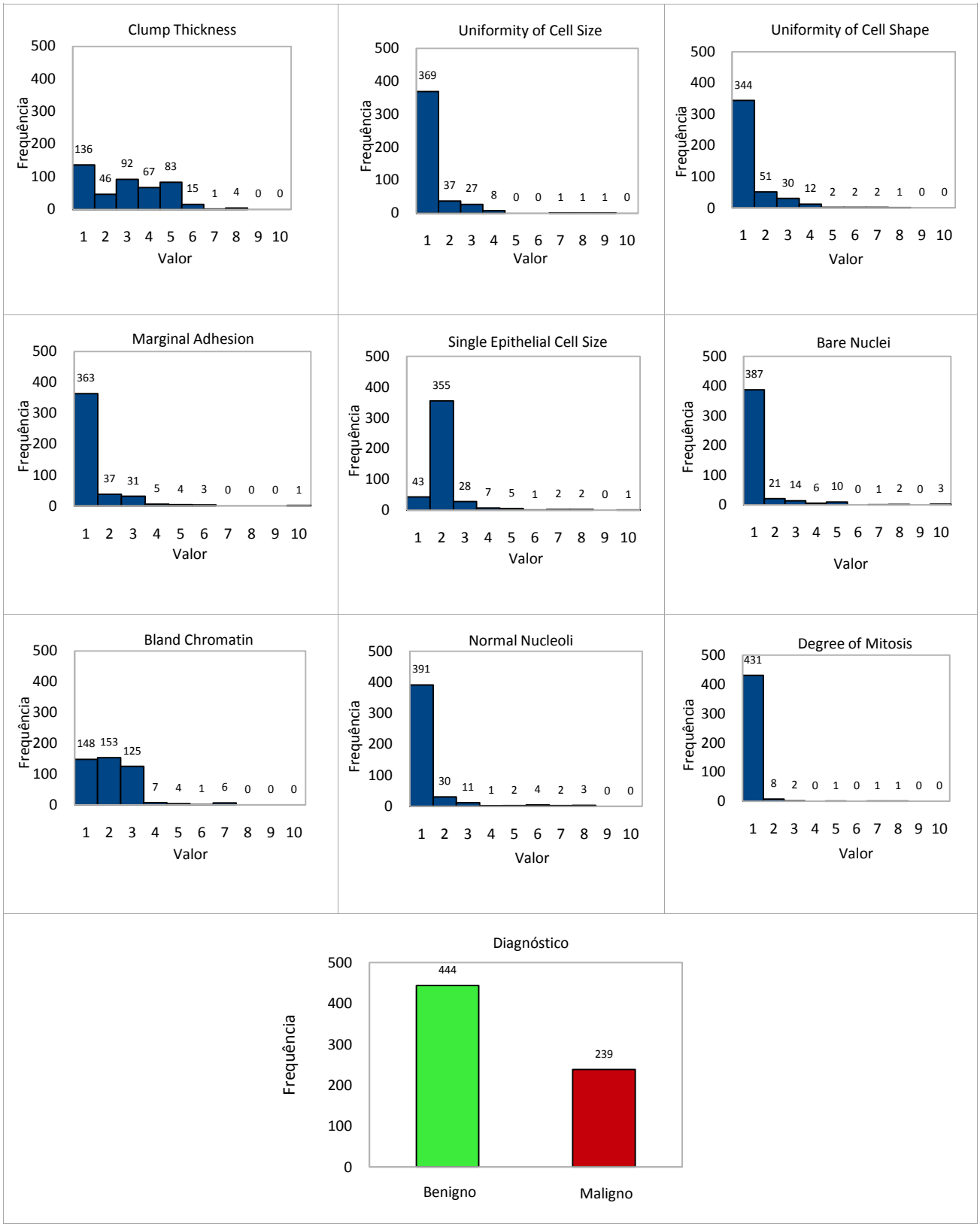


Figura 5.1 – Frequência de valores para cada variável na forma de histograma, incluindo o diagnóstico final.

Neste caso particular, todas as variáveis são discretas e não apresentam um comportamento normal. A Tabela 5.1 apresenta os resultados mais relevantes relativos à aplicação directa do PCA sobre as variáveis do conjunto de dados, usando a matriz de correlação. Nesta Tabela, é possível observar que as primeiras duas e três componentes principais conseguem recuperar cerca de 74.2% e 80.2% da variabilidade dos dados, respectivamente. Isto significa, que uma representação gráfica baseada nestas duas ou três componentes é claramente significativa. Outro critério para a selecção das componentes principais com base na matriz de correlação é o critério do $\lambda>1$, discutido na secção 3.2.3.5 do capítulo 3. Tendo em conta este critério, podemos concluir que apenas uma componente seria suficiente para descrever correctamente os dados.

Tabela 5.1 – Valores próprios e evolução da percentagem de recuperação da informação inicial em relação ao número de componentes principais para o conjunto de dados original (N = 683).

#	Valor próprio (λ_i)	Variância explicada (%)	Variância explicada cumulativa (%)
PC_1	**5.899**	65.5	**65.5**
PC_2	0.776	8.6	**74.2**
PC_3	0.539	6.0	80.2
PC_4	0.460	5.1	85.3
PC_5	0.380	4.2	89.5
PC_6	0.302	3.5	92.8
PC_7	0.294	3.3	96.1
PC_8	0.261	2.9	99.0
PC_9	0.088	1.0	100

A Figura 5.2 apresenta uma visão global dos objectos representados no novo sistema de eixos (as componentes principais). Ambas as representações, a duas e a três dimensões, permitem a discriminação visual entre os casos benignos e malignos. Também é possível observar, que a terceira componente não fornece informações adicionais significativas.

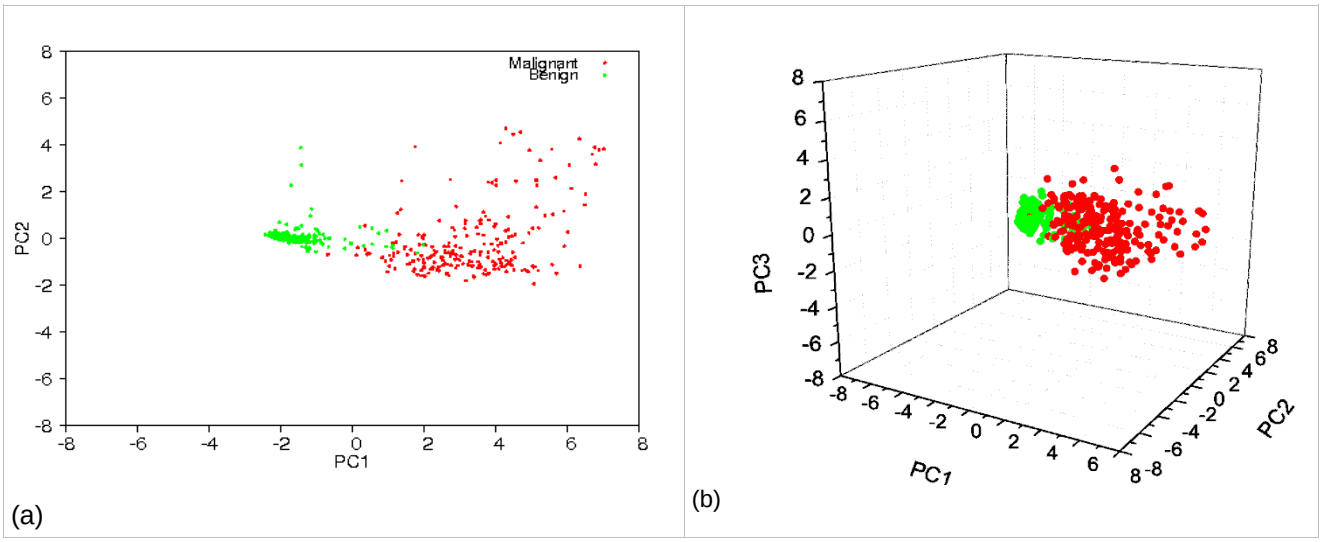


Figura 5.2 – Representação dos objectos no novo sistema de eixos (as componentes principais) após a aplicação do PCA sobre os casos de cancro da mama: (a) e (b) correspondem à representação a 2D e a 3D da PC_1 e PC_2 (com uma recuperação de 74.2% da informação inicial) e da PC_1, PC_2 e PC_3 (com uma recuperação de 80.2% da informação inicial), respectivamente, antes da limpeza de ruído. Os pontos de cor verde referem-se aos casos benignos e os pontos de cor vermelha correspondem aos casos malignos.

Na Figura 5.2, a representação dos objectos no novo sistema de eixos indica que os casos malignos correspondem a um grupo de dados de maior dimensão e com dispersão elevada, enquanto os casos benignos correspondem um grupo menor e muito denso. Curiosamente, o número de casos pertencentes ao grupo benigno é bastante superior ao número de casos presentes no grupo maligno (2:1), mas encontram-se muito condensados.

Uma explicação para a distribuição destes dados está provavelmente relacionada com a evolução do grau de malignidade ao longo da primeira componente.

A Figura 5.3 explora ainda mais os resultados do PCA, exibindo a representação, com base em histogramas de frequência, da projecção dos objectos sobre as duas primeiras componentes principais. Os resultados indicam que o maior poder de discriminação reside na primeira componente. Esta primeira componente, é de facto muito significativa, como pode ser observado na Tabela 5.1. O primeiro valor próprio está acima da média (1) e é mais de 7 vezes superior ao segundo. Além disso, os histogramas sobrepostos na Figura 5.3 também mostram que a separação dos dois grupos é possível ao longo da primeira componente (PC_1). O mesmo não se verifica para a segunda (PC_2) dado que, existe uma sobreposição acentuada destes grupos nesta componente. Esta observação leva-nos a concluir que a PC_1 contém a informação mais relevante para discriminar entre casos benignos e casos malignos.

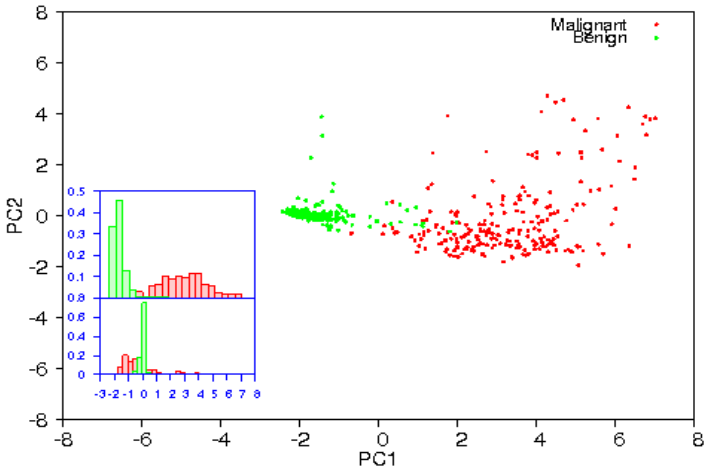


Figura 5.3 – Representação dos objectos no novo sistema de eixos, as componentes principais, após a aplicação do PCA sobre os casos de cancro da mama, mostrando, por correspondência de cores os histogramas de frequência relativos às projecções dos objectos sobre as duas primeiras componentes principais. As barras de cor verde referem-se aos casos benignos e as barras de cor vermelha correspondem aos casos malignos.

A Tabela 5.2 traduz o impacto e a dependência de cada variável original nas duas primeiras componentes principais. Esta tabela traduz o impacto de cada variável original na descrição da variabilidade dos dados. O critério para a selecção das contribuições (*loadings)* mais significativas é baseado na comparação com o valor médio esperado. Sabendo que, uma componente principal constitui uma base vectorial ortonormada, para um caso m dimensional esperamos um valor médio de $1/\sqrt{m}$. A primeira componente retém essencialmente informações sobre as variáveis x2, x3, x5, x6, x7 e x8.

Os coeficientes de correlação indicam que as variáveis x2, x3, x5, x6, x7 e x8 estão relacionados, sendo a maior correlação entre as variáveis x2-x3 (0.907) e a menor entre x3-x6 (0.714). A segunda componente está mais relacionada com a variável x9. Outras conclusões podem ser tiradas a partir do facto de que os valores da primeira componente são todos positivos, sugerindo que esta componente representa uma medida do grau de malignidade. Este é um ponto crucial.

Tabela 5.2 – Impacto das variáveis originais sobre as duas primeiras componentes principais.

	PC_1	PC_2
x1	0.302	-0.141
x2	**0.381**	-0.047
x3	**0.378**	-0.082
x4	**0.333**	-0.052
x5	**0.336**	0.164
x6	**0.335**	-0.261
x7	**0.346**	-0.228
x8	0.336	0.034
x9	0.230	**0.906**

Na maioria dos conjuntos de dados, obtidos experimentalmente, existem *outliers* ou objectos menos representativos. Vamos agora avaliar o impacto da remoção desses pontos nos resultados da análise PCA.

Para ilustrar o impacto da aplicação do algoritmo de limpeza de ruído (NR) nos limites dos grupos, consideramos três situações diferentes, representadas na Figura 5.4 sob a forma de *convex hull*. A Figura 5.4a corresponde à representação dos dados originais sem qualquer tratamento e as restantes duas mostram o impacto do NR nos limites dos grupos. Este resultado é obtido de duas formas diferentes: aplicando o NR sobre os dados originais, Figura 5.4b, ou sobre os *scores* do PCA, Figura 5.4c. Em todos os casos, os grupos são tratados separadamente. Observando as Figura 5.4 é evidente que o grupo de casos benignos foi drasticamente reduzido, deixando uma zona mais densa e isolada, enquanto os objectos mais dispersos e menos densamente distribuídos foram eliminados (cerca de metade). O grupo de casos malignos permanece isolado, pelo menos no que diz respeito os limites respectivos.

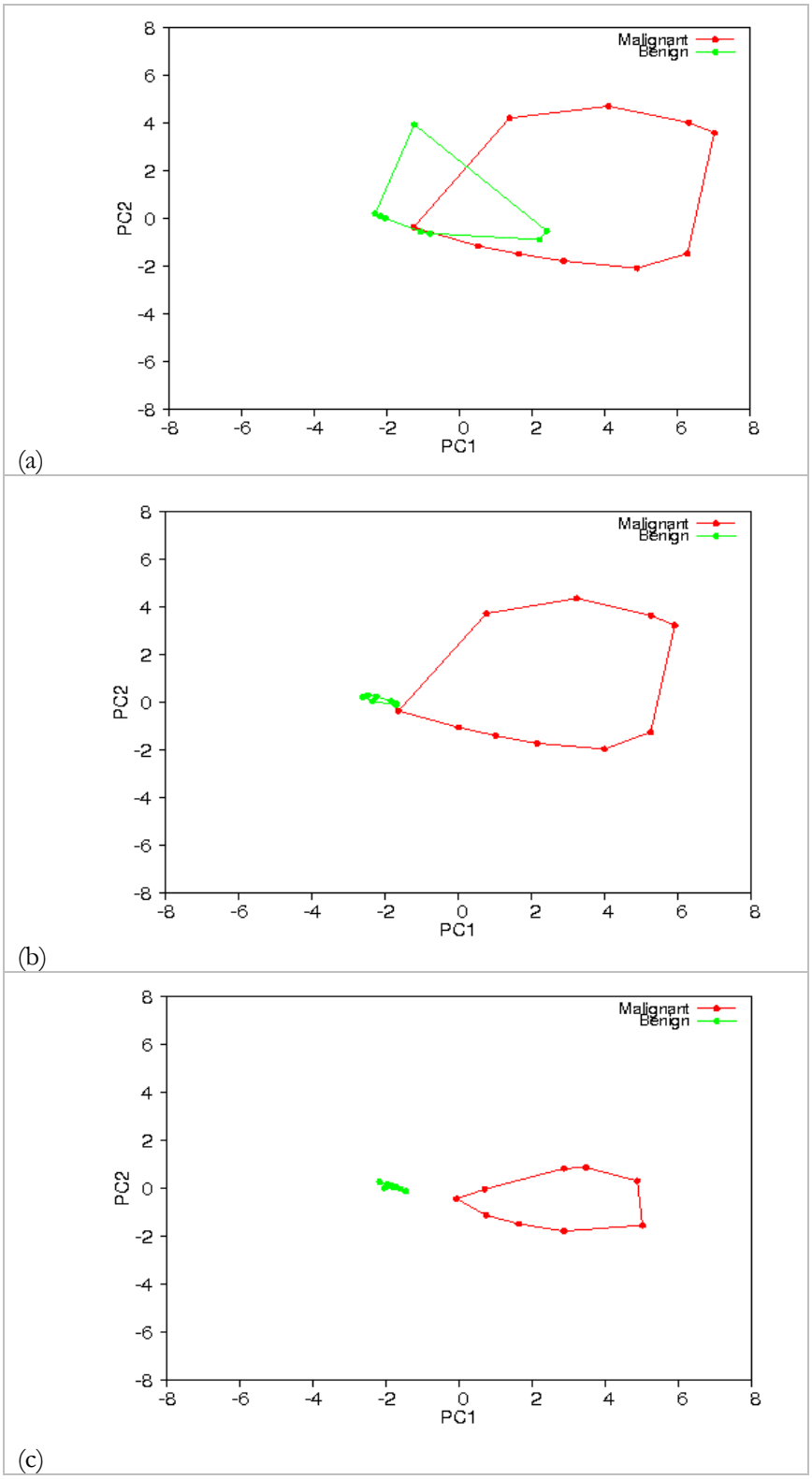


Figura 5.4 – Representação dos dois grupos em estudo (casos benignos e malignos) a 2D sob a forma de *convex hull*, mostrando o impacto do algoritmo NR nos limites grupos: antes da filtragem (a), após a aplicação do NR sobre os dados originais, para cada grupo separadamente (b), e depois da aplicação do NR sobre os *scores* do PCA, para cada grupo separadamente (c). O grupo de cor verde refere-se ao grupo de casos benignos e o grupo de cor vermelha ao grupo de casos malignos.

Com base nestas observações, e lembrando que cada variável só pode ter alguns valores, um esforço é agora feito sobre a possibilidade de impor regras de decisão directamente sobre cada variável, ou seja, estabelecer um valor limite para distinguir os casos malignos dos casos benignos. Tabela 5.3 traduz a percentagem de erro na classificação tendo em conta o diagnóstico médico.

Tabela 5.3 – Percentagem de erro na classificação dos pacientes de acordo com o diagnóstico médico, através da imposição directa de regras de decisão sobre os valores de cada variável, com base no valor limite T (Xn ≤ T , benigno, Xn > T, maligno) .

Valor limite (T)	x1	x2	x3	x4	x5	x6	x7	x8	x9
<1	65.0	65.0	65.0	65.0	65.0	65.0	65.0	6i5.0	65.0
1	45.8	11.9	15.2	16.5	58.9	10.8	43.6	13.8	21.2
2	39.7	7.6	8.8	14.2	**10.2**	**9.1**	22.5	**10.2**	24.0
3	28.0	**7.3**	**7.8**	**13.3**	12.4	**9.1**	**9.5**	13.8	**28.3**
4	20.2	10.5	10.5	16.7	16.8	10.1	13.2	16.1	30.0
5	**14.6**	14.6	14.6	18.9	21.1	11.9	17.0	18.3	30.6
6	15.1	18.3	18.3	21.1	26.9	12.4	17.7	20.4	31.0
7	17.9	20.8	21.8	23.0	28.0	13.3	26.6	22.1	32.1
8	23.1	24.6	25.5	26.6	30.5	15.8	30.5	24.6	32.9
9	25.2	25.2	26.5	27.2	30.7	17.1	32.1	26.2	32.9
10	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0

Como era esperado, observando os resultados da Tabela 5.3, verifica-se que as variáveis apresentam diferentes habilidades para serem usadas isoladamente no diagnóstico do cancro da mama. Esta habilidade é avaliada a partir dos níveis de erro mínimo de classificação obtidos em cada caso. Da Tabela 5.3, podemos verificar que, para este critério de selecção mais adequado, as variáveis x2 (uniformidade do tamanho da célula) e x3 (uniformidade da forma da célula) são capazes de prever a resposta com erros de classificação na ordem dos 7.3 e 7.8%, respectivamente. Curiosamente, a variável x9 associada ao grau de mitoses, apresenta o erro de classificação mais elevado (28.3%), porque esta variável traduz informação diferente e complementar à das outras variáveis (correlação baixa) e não contribui para a separação dos grupos. Tomamos nota de que se os pontos identificados como *outliers* forem removidos, os valores de erro na classificação são ainda menores (dados não representados).

Em conclusão, relativamente à discriminação de classes, a informação obtida a partir do PCA é suficiente para identificar as variáveis mais importantes e estabelecer uma regra de decisão baseada numa única variável (ou numa dimensão muito baixa). Até mesmo uma técnica adicional de baixa complexidade, como o LDA é algo supérflua nesta situação.

Vamos agora considerar uma abordagem mais sofisticada baseada no método PLS. Neste caso, recorre-se ao PLS para seleccionar apenas alguns factores latentes, responsáveis pela maior variação na resposta, ou seja, no diagnóstico. A ideia geral é tentar extrair esses factores latentes, representando o máximo de variação possível, enquanto a resposta é também modelada. Resultados semelhantes para a previsão da resposta usando estes 9 predictores podem, também, ser obtidos através do PLS. A Tabela 5.4 resume a descrição da resposta, correspondente ao diagnóstico do cancro da mama.

Tabela 5.4 – Resultados do PLS relativos aos predictores e à descrição da resposta.

# Factores latentes (n_f)	V_X%	UV_X%	CV_X%	V_Y%	UV_Y%	CV_Y%	β
1	34.48	**65.52**	65.52	18.60	**81.40**	**81.40**	23.56
2	26.89	**7.58**	**73.11**	16.03	**2.58**	**83.97**	4.19
3	21.57	5.32	78.43	15.70	0.32	84.30	1.48
4	16.86	4.72	83.14	15.67	0.03	84.33	0.45
5	13.38	3.48	86.62	15.67	0.00	84.33	0.18
6	10.43	2.95	89.57	15.67	0.00	84.33	0.13
7	7.31	3.12	92.69	15.67	0.00	84.33	0.02
8	3.34	3.97	96.66	15.67	0.00	84.33	0.00
9	0.00	3.34	100.00	15.67	0.00	84.33	0.00

V_X,V_Y – informação residual contida no sub-espaço de X e de Y (percentagem da soma de quadrados); UV_X, UV_Y – informação do sub-espaço de X e de Y usada na descrição; CV_X, CV_Y – informação cumulativa correspondente a n_f factores latentes; β – parâmetro do modelo PLS.

Pela análise da Tabela 5.4 é possível verificar que o primeiro factor latente (FL_1) usa 65.5% da informação do sub-espaço de predictores para descrever mais de 81% da resposta, com uma eficiência de 124.3%. O segundo factor latente requer mais 7.6% de informação adicional para descrever apenas 2.6% da resposta, com uma eficiência de 34.0%, muito inferior ao caso anterior. Mais de 83.0% da informação é descrita no sub-espaço da resposta

com estes dois factores latentes (n_f = 2). Esforços adicionais para descrever a resposta são irrelevantes, dado que, apenas 84.3% da resposta pode ser descrita.

Resultados semelhantes podem ser obtidos pela avaliação directa do parâmetro β. Os três primeiros valores, 23.56, 4.19, e 1.48 revelam a maior relevância do FL_1 sobre os outros dois (FL_1>> FL_2> FL_3). Os parâmetros subsequentes rapidamente convergem para zero.

Apesar do facto de o PLS falhar em cerca de 16% ($n_f \geq 2$) na recuperação da resposta, pode ser útil na identificação das variáveis mais relevantes para fins de diagnóstico, consulte a Tabela 5.5.

Tabela 5.5 – Relevância do sub-espaço de predictores nos dois primeiros factores latentes, FL_1 e FL_2.

Predictor	FL_1	FL_2
x1	0.305	**0.372**
x2	**0.381**	-0.065
x3	**0.378**	0.006
x4	0.332	-0.079
x5	**0.334**	-0.300
x6	**0.339**	**0.465**
x7	**0.347**	0.085
x8	**0.335**	-0.135
x9	0.225	-0.720

Considerando-se um critério semelhante ao critério utilizado no PCA no que diz respeito à identificação das contribuições (*loadings)* mais significativas, o primeiro factor latente (FL_1) contém informações das variáveis x2, x3, x7, x6, x8 e x5. Este resultado está de acordo com resultados obtidos anteriormente com a aplicação do PCA.

5.2 *Breast cancer 2*

No segundo caso, 106 amostras representando 6 diferentes tecidos extraídos da mama (52 relativas a tecidos normais e 54 relativas a tecidos patológicos), são avaliadas através de 9 variáveis.

O exemplo anterior mostrou, que o PCA substitui grande parte das tarefas da EDA[28], e fornece uma boa visualização dos dados. Como tal, procedemos directamente à apresentação dos resultados relativos a esta técnica.

A Tabela 5.6 resume os resultados do PCA, agora usando as duas abordagens: a covariância e a correlação, após a centragem dos dados.

Tabela 5.6 – Resultados do PCA para as três primeiras componentes principais.

	Covariância			Correlação		
#	Valor próprio (λ_i)	Variância explicada (%)	Variância explicada cumulativa (%)	Valor próprio (λ_i)	Variância explicada (%)	Variância explicada cumulativa (%)
PC_1	3.46×10^8	99.76	99.76	5.46	40.7	60.7
PC_2	7.88×10^5	0.23	99.99	1.81	20.1	80.8
PC_3	7.55×10^4	0.01	100	0.78	8.61	89.5

No caso da covariância, a primeira componente principal permite a descrição de 99.8% da informação contida no sub-espaço de predictores. No caso de ser usada a matriz de correlação, o que implica a normalização das variáveis, pelo menos duas componentes ($p = 2$) são necessárias para contabilizar aproximadamente a mesma informação. A primeira componente explica apenas 60.7% da variabilidade dos dados.

A Figura 5.5 apresenta uma visão geral dos dados referentes a cada tipo de tecido, considerando o no novo sistema de eixos. A abordagem considerada baseia-se na matriz de correlação e corresponde a uma sobreposição muito acentuada dos grupos e uma região muito densa na qual, pelo menos quatro grupos são indistinguíveis. Este facto, indica que teremos algumas dificuldades em encontrar um diagnóstico correcto.

[28] Do inglês *Exploratory Data Analysis*

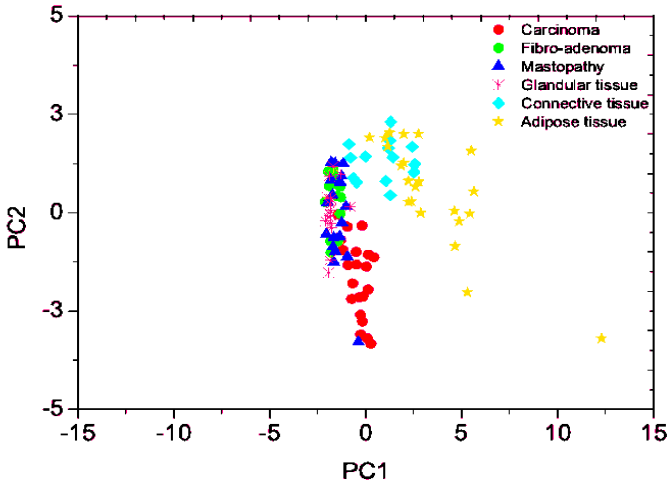


Figura 5.5 – Representação das amostras de tecidos no novo sistema de eixos, após efectuar a normalização das variáveis. As cores e os símbolos estão relacionados com cada tipo de tecido.

Considerando os valores próprios apresentados na Tabela 5.6, para a matriz de correlação, apenas os dois primeiros valores (5.46 e 1.81) representam aproximadamente 80.8% da variabilidade dos dados descrita com as duas primeiras componentes principais. Outros valores subsequentes, tais como 0.78 são menos significativos, representando apenas 8.6% de recuperação da variabilidade. Isto significa que, uma representação gráfica baseada nas duas primeiras componentes reflecte, claramente, a estrutura dos dados.

A fim de recuperar algumas informações relevantes para as variáveis mais discriminantes, a Tabela 5.7 apresenta o impacto (*loadings*) de cada variável, nas duas primeiras componentes principais, para a matriz de correlação.

À semelhança do caso anterior, o critério para a selecção das contribuições *(loadings)* significativas é baseado na comparação com o valor médio esperado.

A primeira componente principal retém essencialmente informação sobre x1, x4 x5, x6, x7, x8 e x9, variáveis que são, em alguns casos auto-relacionadas, tais como as variáveis x1-x9 (0.989), x4-x8 (0.974), x7-x9 (0.862), x5-x6 (0.830), x1-x7 (0.824), x1-x4 (0.820) e x4-x7 (0.753). Para este grupo de variáveis, a menor inter-relação ocorre entre as variáveis x1-x3 (0.028). A segunda componente está relacionada principalmente com x2 e x3, que também podem ser inter-relacionadas (0.509).

Tabela 5.7 – Impacto das variáveis originais nas duas primeiras componentes principais (para a correlação).

	PC_1	PC_2
x1	**0.387**	0.240
x2	-0.047	**-0.665**
x3	0.094	**-0.586**
x4	**0.395**	-0.059
x5	**0.352**	-0.167
x6	**0.355**	-0.276
x7	**0.392**	-0.098
x8	**0.358**	0.092
x9	**0.389**	0.179

Este caso de correlação foi ainda estudado, para evidenciar o grupo de variáveis mais significativo na descrição dos dados. Foram testados dois métodos de selecção [3] baseados nas componentes principais, para determinar o conjunto de variáveis que preservam a maior variação no conjunto de dados, e que melhor separam as seis classes de tecidos mamários (carcinoma, fibro-adenoma , mastopatia, tecido glandular, tecido conjuntivo e tecido adiposo).

O primeiro método iterativo inicia-se removendo a variáveis de maior peso (maior *loading*) em valor absoluto, associado à componente menos significativa. Um novo PCA é realizado sobre as restantes variáveis. Em cada processo, os dados são representados a 2D e a 3D, de forma a avaliar o impacto da remoção das variáveis na distribuição dos grupos. Este procedimento é repetido até que ocorra uma perda significativa de informação, levando à sobreposição de grupos, que não requer mais eliminações.

No segundo método as variáveis com maior valor absoluto, associadas às primeiras componentes principais são preservadas, enquanto as variáveis com coeficientes pouco significativos, no conjunto das componentes principais são eliminadas. Tal como no método anterior, este procedimento é repetido até que não sejam necessárias mais eliminações.

Comparando os resultados obtidos através destes dois métodos, conclui-se que um conjunto constituído por apenas três das nove variáveis estudadas é suficiente para separar os dois grupos principais (variáveis x1, x5 e x9). Cada uma dessas variáveis, isolada, preserva a maior quantidade de informação dada pelo grupo correspondente.

Como no caso anterior, vamos agora avaliar o impacto da aplicação do algoritmo NR nos resultados do PCA. Para comparação directa, consideramos a abordagem de correlação em duas situações diferentes, retratadas na Figura 5.6 sob a forma de *convex hull*. A Figura 5.6a representa os dados originais e a Figura 5.6b ilustra o impacto da aplicação do NR nos limites de cada grupo. Cada grupo é tratado separadamente. Verifica-se que existe claramente uma sobreposição significativa de objectos entre os grupos relativos a cada tipo de tecido, indicando que a discriminação neste caso é mais difícil que no caso anterior. Mesmo depois de efectuar a limpeza de ruído sobre os *scores* do PCA, na Figura 5.6b, a sobreposição dos grupos permanece excessiva. Em ambas as situações (a) e (b) parece haver uma separação relativa entre o grupo normal, incluindo os tecidos adiposo e conjuntivo e o grupo patológico, que inclui o carcinoma, fibro-adenoma e o tecido característico da mastopatia, ao longo da primeira componente. A classe normal dos tecidos mamários encontra-se no lado positivo do eixo PC_1 e a classe dos tecidos patológicos encontra-se no lado negativo do eixo PC_1. Embora seja possível separar estes dois grupos principais, os tecidos correspondentes encontram-se fortemente ligados.

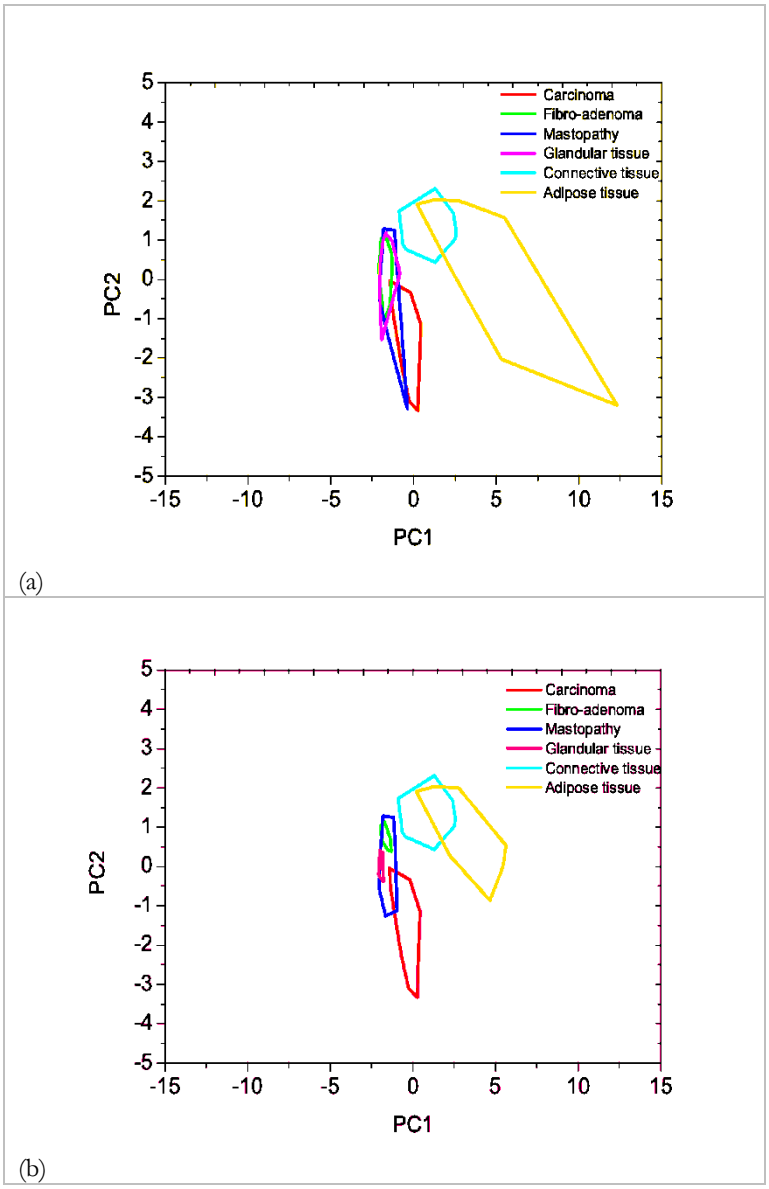


Figura 5.6 – Representação das classes de tecidos a 2D sob a forma de *convex hull*, ilustrando o impacto da aplicação do NR nos limites dos grupos: (a) antes da aplicação do NR sobre o sistema PCA e (b) após a aplicação do NR sobre os *scores* do PCA (b). Para efeitos de comparação directa, são mantidas as cores correspondentes a cada tipo de tecido.

Observando a Figura 5.6b verifica-se que o algoritmo da limpeza de ruído (NR) torna a estrutura dos dados mais definida, ou seja, remove os pontos correspondentes à sobreposição entre os grupos, que correspondem aos valores menos discriminativos, permitindo a separação entre o tecido glandular e fibro-adenoma. Nestes casos, a redução de ruído não pode ser encarada como uma ferramenta para eliminar outliers, mas sim como uma forma de reter apenas os pontos mais relevantes e característicos de cada grupo.

Mais uma vez, vemos que só é possível separar o grupo patológico do normal com excepção do tecido glandular. Este tecido cai na região do grupo patológico e necessita de mais investigação. O tecido glandular tem características morfológicas semelhantes aos tecidos pertencentes ao grupo patológico. Assim, esta característica pode justificar a sua posição no grupo patológico.

Outra forma de visualizar esta sobreposição consiste em representar as distribuições dos dados de acordo com as frequências de cada amostra de tecido ao longo da PC_1 e da PC_2, na Figura 5.7. Esta Figura representa as distribuições dos tecidos no sistema PCA original, na Figura5.7a, e depois da limpeza de ruído (NR), na Figura 5.7b. Podemos observar que, a aparência geral das distribuições surge alterada, principalmente nos tecidos correspondentes ao fibro-adenoma, à mastopatia e ao tecido adiposo.

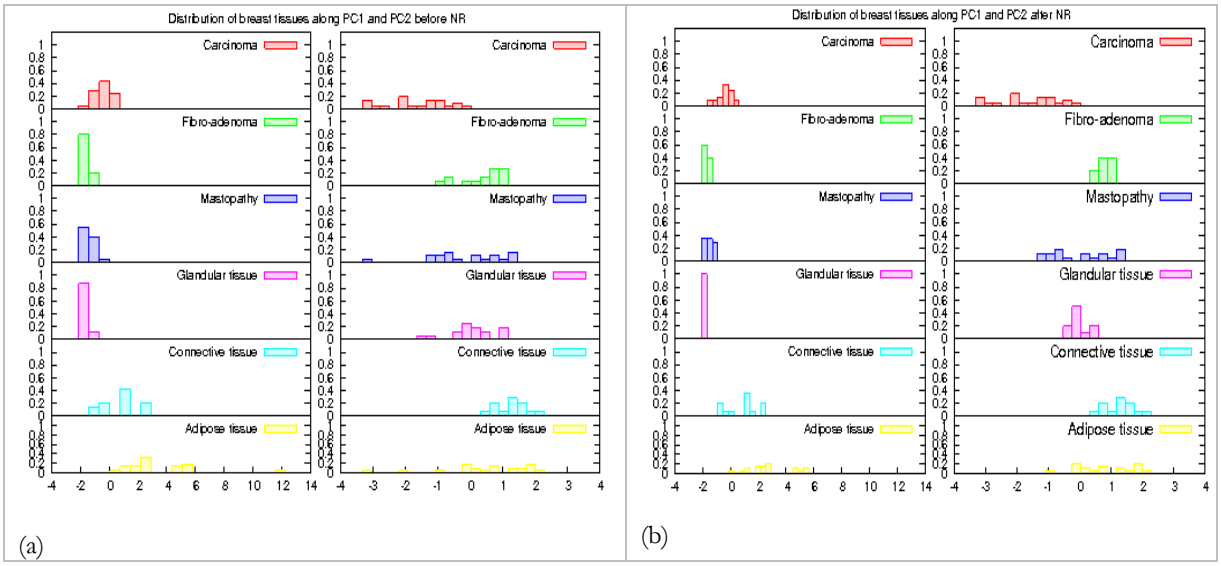


Figura 5.7 – Histogramas relativos aos tecidos mamários antes (a) e depois da aplicação do algoritmo NR (b). As distribuições apresentam a cor correspondente a cada tipo de tecido. As barras nos histogramas representam, por correspondência de cores, as frequências de *scores* referentes a cada grupo ao longo das duas componentes PC_1 e PC_2.

Vamos de seguida avaliar os resultados obtidos recorrendo à matriz de covariância, descritos na Tabela 5.8. Neste caso, existe apenas uma variável dominante, com uma contribuição (*loading)* muito significativa na primeira componente principal (x5). Este resultado decorre directamente do maior valor absoluto desta variável, e reduz a informação transmitida por esta análise. A segunda componente é caracterizada por apenas duas outras variáveis (x1 e x9).

Tabela 5.8 – Impacto das variáveis originais nas duas primeiras componentes principais (para a matriz de covariância).

	PC_1	PC_2
x1	0.023	**-0.701**
x2	0.000	0.000
x3	0.000	0.000
x4	0.008	-0.102
x5	**0.999**	0.034
x6	0.001	-0.006
x7	0.003	-0.047
x8	0.007	-0.081
x9	0.024	**-0.699**

Comparando os resultados da Figura 5.6a, com os resultados obtidos usando a matriz de covariância, na Figura 5.8 é evidente que, neste último há uma separação mais eficiente entre o tecido conjuntivo e o tecido adiposo, pertencentes ao grupo normal, e entre o carcinoma e os outros tecidos, no grupo patológico. Esta separação é feita principalmente ao longo da segunda componente principal. Não podemos separar, no entanto, os restantes três grupos de tecidos: o tecido relativo à mastopatia, fibro-adenoma e o tecido glandular.

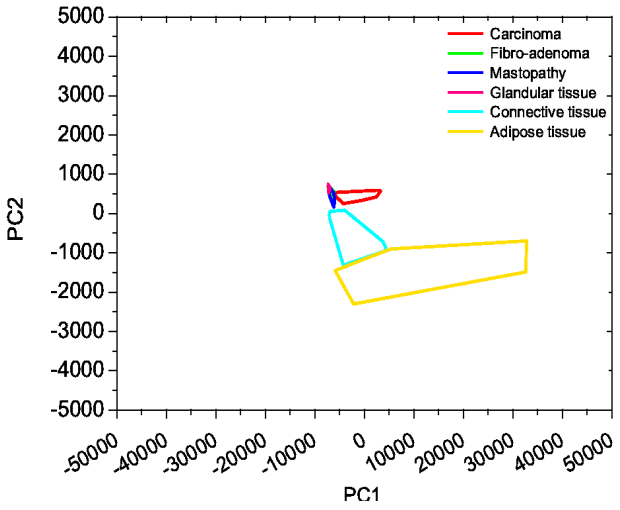


Figura 5.8 – Representação das classes de tecidos a 2D sob a forma de *convex hull*, a partir do sistema PCA original, tendo em conta a centragem das variáveis.

A classificação precisa de tumores num problema específico é uma tarefa extremamente importante para o diagnóstico correcto, tratamento e acompanhamento clínico de pacientes com cancro. Neste contexto, a análise discriminante linear foi realizada de duas maneiras: (1) aplicando o LDA sobre os dados originais e (2) aplicando o LDA após o tratamento de cada grupo com o NR.

A Tabela 5.9 descreve a percentagem de erro na classificação dos tecidos, fazendo combinações de pares de tecidos normais e patológicos (nove combinações no total), cada um pertencente a um grupo diferente. Observando os resultados da Tabela 5.9, conclui-se que, em geral, há um aumento da percentagem de objectos correctamente classificados, após a realização de NR. Como esperado, o tecido adiposo tem uma precisão de 100% em ambas as situações. Isto significa, que o tecido adiposo tem propriedades diferentes dos outros tecidos. Além disso, após o NR é possível separar completamente o carcinoma do tecido glandular. Os maiores erros de classificação foram obtidos para a combinação do tecido glandular com o tecido característico da mastopatia (16.91%) e com o fibro-adenoma (11.67%).

Estes resultados sugerem que existem duas hipóteses para a escolha do tecido mais característico da classe normal: o tecido conjuntivo ou o tecido adiposo. No entanto, dada a proximidade do tecido conjuntivo aos tecidos patológicos, podemos afirmar que a melhor escolha será utilizar o tecido adiposo como referência para a classe de tecidos normais. Adicionalmente, descobrimos que o tecido glandular não pode ser usado como referência, devido à sua posição na região dos tecidos patológicos.

Tabela 5.9 – Percentagens de erro na classificação dos tecidos, obtidas por aplicação do LDA sobre a combinação de pares de tecidos, antes e depois da limpeza de ruído (NR).

		Tecidos Normais					
Erro na classificação (%)		Tecido Glandular		Tecido Conjuntivo		Tecido Adiposo	
		Sem NR	Com NR	Sem NR	Com NR	Sem NR	Com NR
Tecidos Patológicos	Carcinoma	4.00	0.00	0.80	0.31	0.00	0.00
	Fibro-adenoma	11.67	9.38	0.54	0.60	0.00	0.00
	Mastopatia	16.91	5.93	2.56	2.58	0.00	0.00

Finalmente, devemos notar que, o aumento da precisão na classificação dos casos estudados (1 e 2), entre os casos benignos e malignos (caso 1), entre os tecidos normais e

patológicos (caso2), era esperado, uma vez que, o pressuposto da redução de ruído (NR) é preservar os dados mais característicos de cada grupo.

5.3 *New cancer cases estimated data*

Vamos agora abordar um problema diferente, relacionado com a distribuição de novos casos de cancro nos EUA.

Seguindo o esquema proposto, vamos em primeiro lugar, aplicar a análise de agrupamento hierárquico (HCA), que fornece um meio visual para estimar as relações entre os dados. Foi usada a distância euclidiana para representar a dissimilaridade entre os estados.

Uma questão fundamental, diz respeito à normalização dos dados. Neste caso, a nossa opção é auto-normalizar cada objecto, que corresponde a um estado específico. O número de casos para cada tipo de cancro é, portanto, dividido pelo número total de casos previsto para aquele estado. Como tal, cada estado é descrito por um conjunto de variáveis que são a fracção de ocorrências previstas para cada tipo de cancro. Os estados mais semelhantes são aqueles que têm o mesmo perfil de casos, independentemente da magnitude de incidência. Neste estudo, foi seleccionado a ligação *Ward* como método padrão.

O dendrograma construído com base na informação total existente para os 51 estados é representado na Figura 5.9.

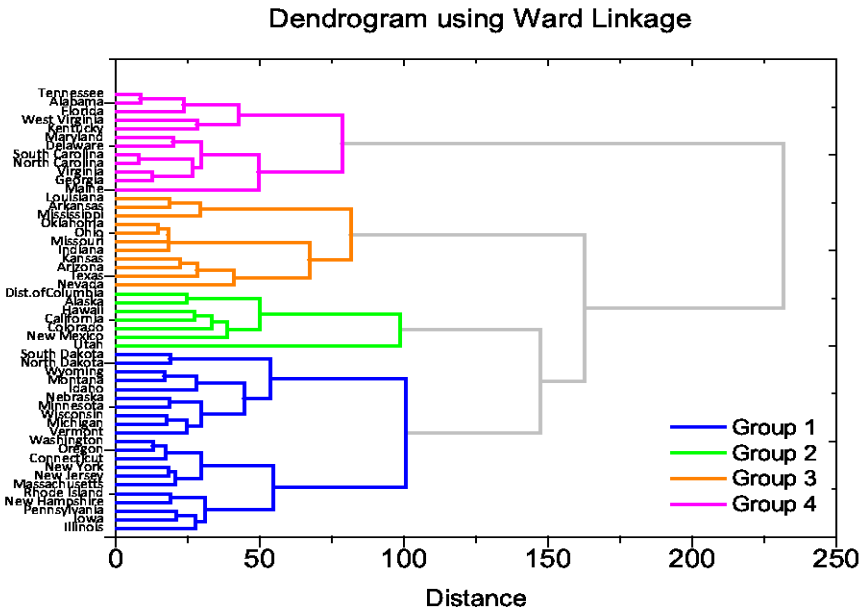


Figura 5.9 – Dendrograma (método da ligação *Ward*) construído por agrupamento hierárquico sobre o conjunto de dados correspondente a 51 estados dos EUA. São usadas cores diferentes para representar os grupos de estados diferentes, de acordo com as suas semelhanças.

Pela análise desta Figura, é evidente que os dados possuem uma super-estrutura na qual, são visíveis quatro grupos de estados, com uma estrutura semelhante. Estes quatro grupos estão sobrepostos na Figura 5.10, correspondente ao mapa dos EUA. Os grupos estão (i) localizados na região norte, (ii) na costa leste, (iii) na região central e (iv) na parte sudoeste. Os dois últimos grupos têm um certo grau de sobreposição.

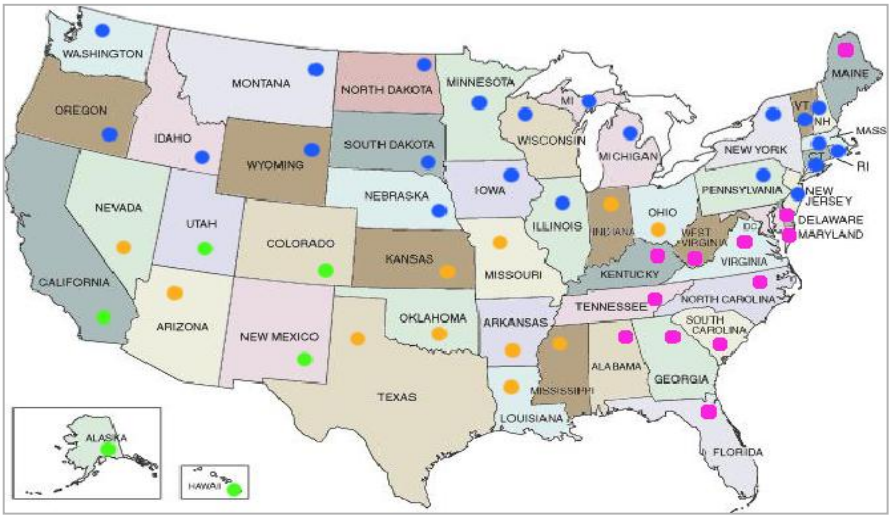


Figura 5.10 – Representação geográfica dos grupos formados a partir do HCA e do PCA.

Depois de estabelecer o número de grupos, foi então aplicada directamente a análise de componentes principais, sem qualquer tratamento prévio com o NR a fim de revelar a relação (variáveis centradas na média) entre os estados e os tipos de cancro dentro e entre estes quatro grupos.

Como resultado preliminar do PCA, a representação dos *scores*, na Figura 5.11, é directamente concordante com os resultados obtidos através do HCA. Em ambos os casos, os grupos são identificados pelas cores correspondentes.

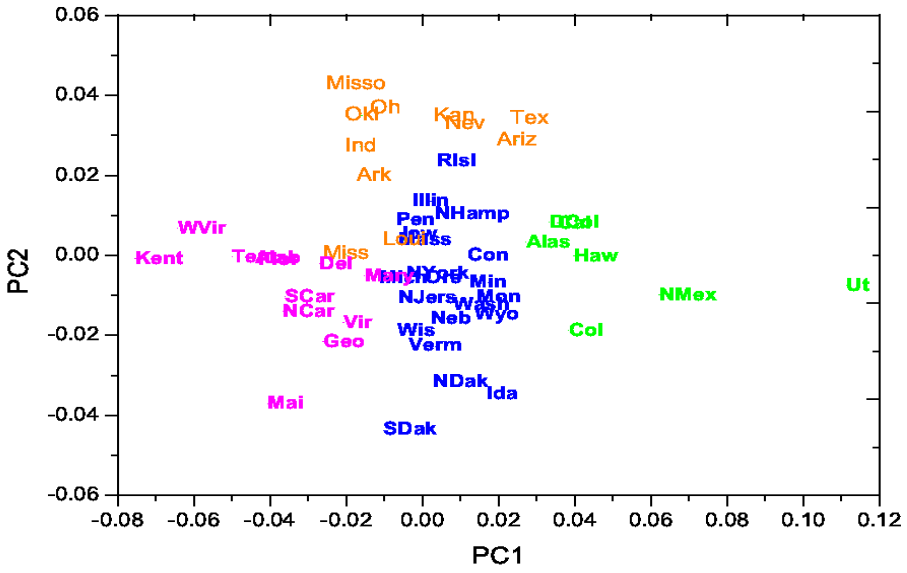


Figura 5.11 – Representação dos estados americanos no novo sistema de eixos PC_2 vs PC_2, com 81.0% de informação recuperada. O conjunto de dados contém 51 estados e 11 tipos de cancro diferentes. As cores estão relacionadas com os grupos correspondentes. Por conveniência, utilizaram-se abreviaturas para os nomes dos estados.

As duas primeiras componentes principais seleccionadas descrevem aproximadamente 81.0% da variabilidade total. Observando os *scores* na Figura 5.11 e as contribuições (*loadings)* de cada variável, na Tabela 5.10, para as duas componentes principais mais relevantes, concluímos que, o cancro do pulmão e dos brônquios (x6) é o mais importante dos restantes tipos de cancro seleccionados (*loadings* de 0.724 e 0.403 para a primeira e segunda componente principal, respectivamente).

Tabela 5.10 – Impacto das variáveis originais nas duas primeiras componentes principais (para a matriz de covariância).

	PC_1	PC_2
x1	0.000	0.000
x2	0.000	0.006
x3	0.000	0.000
x4	0.019	-0.032
x5	0.023	-0.017
x6	**-0.724**	**0.403**
x7	0.007	-0.131
x8	0.005	-0.018
x9	0.189	-0.633
x10	-0.020	-0.086
x11	0.652	0.625

Nas duas primeiras componentes principais, as contribuições (*loadings)* reflectem, em valor absoluto, a influência dos tipos de cancro sobre o agrupamento dos estados. Como já foi mencionado, o cancro do pulmão e dos brônquios (variável x6) é responsável pelo posicionamento (distribuição) dos estados americanos ao longo da primeira componente (PC_1). Observando o gráfico de dispersão na Figura 5.11, podemos afirmar que a separação dos grupos é feita principalmente ao longo da primeira componente principal.

Estados de grande população como a Califórnia estão posicionados no lado positivo do eixo PC_1. No lado oposto, estão localizados estados como a Flórida. Estados como Nova York e Texas estão incluídos nos grupos de maior dispersão ao longo da segunda componente.

Outro resultado interessante é que a distribuição dos estados no PCA é semelhante ao mapa geográfico real dos EUA. Como podemos ver na Figura 5.10, grupos semelhantes são grupos vizinhos.

5.4 *Parkinsons Telemonitoring dataset*

O exemplo que se segue está também relacionado com a aplicação dos métodos quimiométricos no diagnóstico médico. No entanto, passamos agora a lidar com um problema relacionado com a evolução dos sintomas da doença de Parkinson em doentes ainda no estado inicial, tendo em conta um conjunto de 16 medidas de voz. Existem outras três variáveis, duas delas associadas às características individuais de cada paciente como a idade e o género, e a terceira associada ao intervalo de tempo de recrutamento.

Seguindo o esquema proposto nos exemplos anteriores, efectuou-se, em primeiro lugar, a análise de agrupamento hierárquico (HCA), que permite visualizar as relações entre os vários pacientes de acordo com as suas semelhanças. Foi usado o método da ligação *Ward* e a distância euclidiana para representar a diferença entre os pacientes.

O dendrograma construído com base na informação total existente para os 42 pacientes é representado na Figura 5.12.

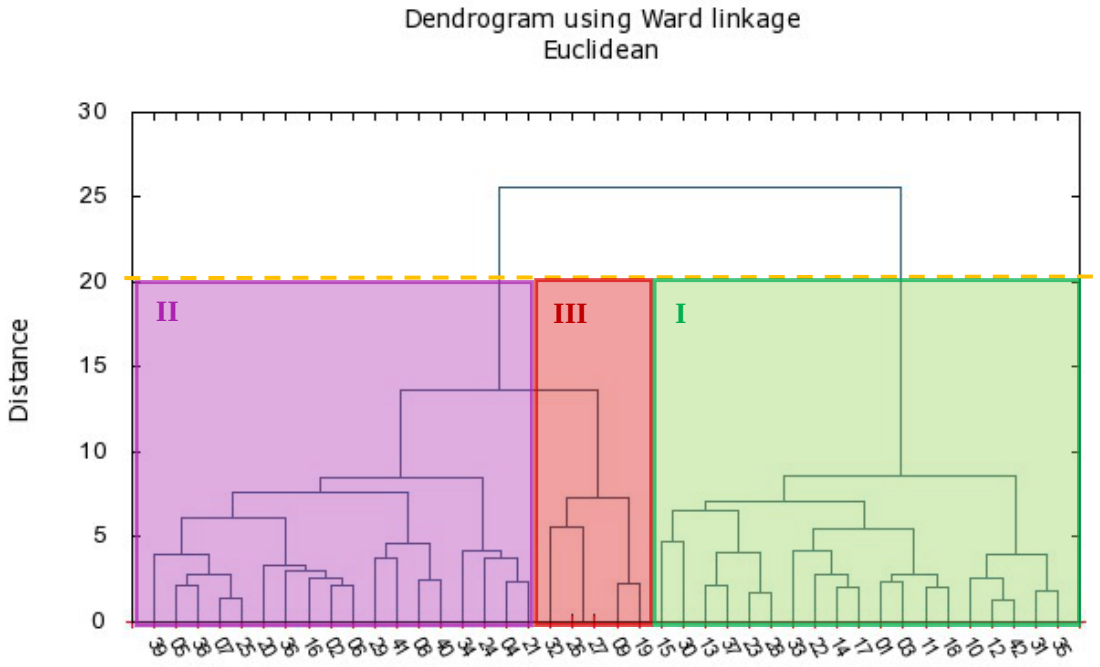


Figura 5.12 – Dendrograma construído por agrupamento hierárquico (método da ligação *Ward*) sobre o conjunto de dados de 42 pacientes com sintomas da doença de Parkinson. São usadas cores diferentes para representar os grupos dos vários pacientes, de acordo com as suas semelhanças.

Pela análise do dendrograma da Figura 5.12, são visíveis os três grupos de pacientes, com estrutura semelhante. Estes três grupos estão também representados na Figura 5.13, correspondente ao novo sistema PCA, considerando a matriz de correlação.

A análise de componentes principais foi aplicada directamente sem qualquer tratamento prévio com o NR a fim de revelar a relação entre os pacientes e as medidas de voz nestes três grupos.

Os grupos estão distribuídos ao longo da primeira componente, sem qualquer sobreposição. Esta representação revela uma evolução dos sintomas da doença ao longo da primeira componente. O grupo de pacientes representado a verde (grupo I), localizado no eixo negativo do PC_1 constitui os casos de menor gravidade dos sintomas. O grupo intermédio (grupo II) caracteriza-se por pacientes com sintomas num grau intermédio. O último grupo (III) localizado no extremo positivo do eixo PC_1 constitui os casos de maior gravidade, pois dele fazem parte os doentes com os sintomas mais graves, característicos da doença.

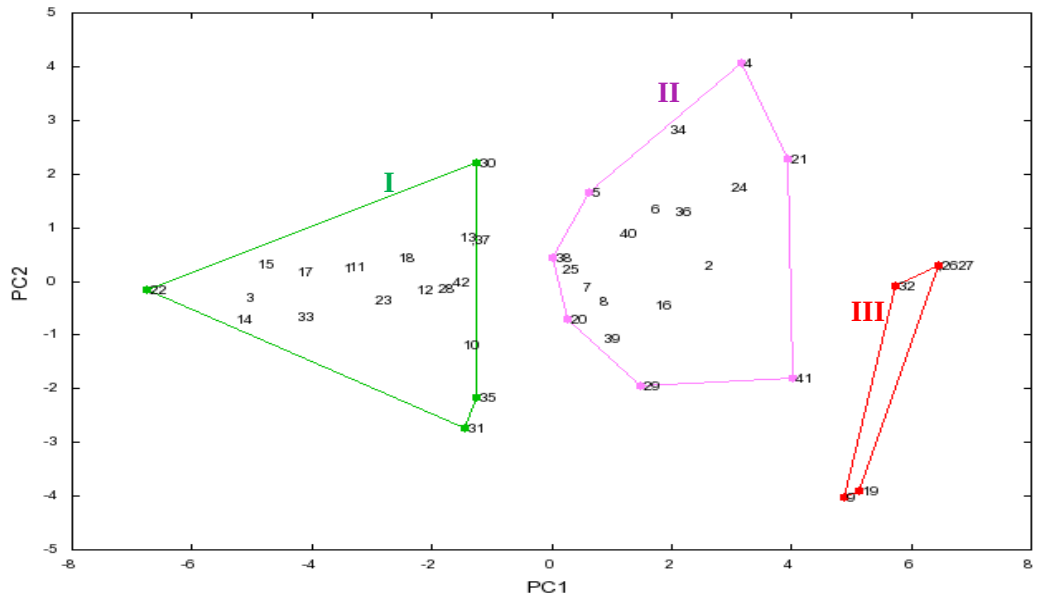


Figura 5.13 – Representação dos 42 pacientes com sintomas da doença de Parkinson no novo sistema de eixos PC_1 vs PC_2, com uma recuperação de 73.0% da variabilidade inicial. O conjunto de dados contém 16 medidas de voz. As cores atribuídas estão relacionadas com os grupos correspondentes na análise HCA.

A representação dos *scores*, na Figura 5.13, resultante a análise PCA é concordante com os resultados obtidos através do HCA. Em ambos os casos, os grupos são identificados pelas cores correspondentes.

A Tabela 5.11 resume os resultados do PCA para as primeiras seis componentes principais.

Tabela 5.11 – Resultados do PCA para as seis primeiras componentes principais.

	Correlação		
#	Valor próprio (λ_i)	Variância explicada (%)	Variância explicada cumulativa (%)
PC_1	11.43	60.1	60.1
PC_2	2.49	13.1	73.3
PC_3	1.29	6.8	80.1
PC_4	1.20	6.3	86.4
PC_5	0.964	5.1	91.4
PC_6	0.569	3.0	94.4

A primeira componente principal descreve apenas 60.1% da informação inicial. Neste caso, em que é usada a matriz de correlação (normalização das variáveis) pelo menos três componentes (p = 3) são necessárias para contabilizar aproximadamente 80% da informação inicial. No entanto, a terceira componente não fornece informações adicionais significativas.

Usando o mesmo critério para a selecção das contribuições (*loadings)* significativas, baseado na comparação com o valor médio esperado, verifica-se, pela análise da Tabela 5.12 que a primeira componente principal retém essencialmente informação sobre dois conjuntos de variáveis: x4 a x8, associadas a várias medidas de frequência e x9 a x14, associadas a várias medidas de amplitude. Neste caso, as variáveis destes dois conjuntos apresentam uma elevada interdependência estando altamente auto-relacionadas. Para grupo de variáveis, associadas às medidas de frequência a maior inter-relação ocorre entre as variáveis x4-x7 (0.991). Para o grupo das variáveis associadas às medidas de amplitude a maior inter-relação ocorre entre as variáveis x9-x10 (0.997). A segunda componente está relacionada também com estes dois conjuntos de variáveis.

Tabela 5.12 – Impacto das variáveis originais nas duas primeiras componentes principais (para a matriz de correlação).

	PC_1	PC_2
x1	0.057	0.206
x2	-0.089	-0.013
x3	0.055	0.010
x4	**0.256**	**-0.296**
x5	**0.249**	**-0.254**
x6	**0.245**	**-0.307**
x7	**0.259**	**-0.284**
x8	**0.245**	**-0.307**
x9	**0.267**	**0.264**
x10	**0.267**	**0.261**
x11	**0.254**	**0.295**
x12	**0.260**	**0.286**
x13	**0.263**	**0.227**
x14	**0.254**	**0.295**
x15	0.221	-0.020
x16	-0.262	-0.088
x17	0.220	-0.060
x18	0.166	**-0.233**
x19	0.264	-0.156

Na Figura 5.14 encontram-se representadas as variáveis em estudo no sistema formado pelas duas primeiras componentes principais, após a rotação varimax[29]. Esta representação suporta as conclusões anteriores, dado que são visíveis os dois conjuntos de variáveis responsáveis pela distribuição dos grupos ao longo da primeira e da segunda componente. Sugere também, que seria necessário apenas uma medida representativa de cada grupo (amplitude e frequência) para separar os três grupos de pacientes.

[29] Método de maximização que consiste numa rotação ortogonal de eixos, pretendendo que, para cada componente principal, sejam maximizadas as contribuições dos valores mais significativos em detrimento dos menos expressivos, preservando a ortogonalidade da base vectorial inicial.

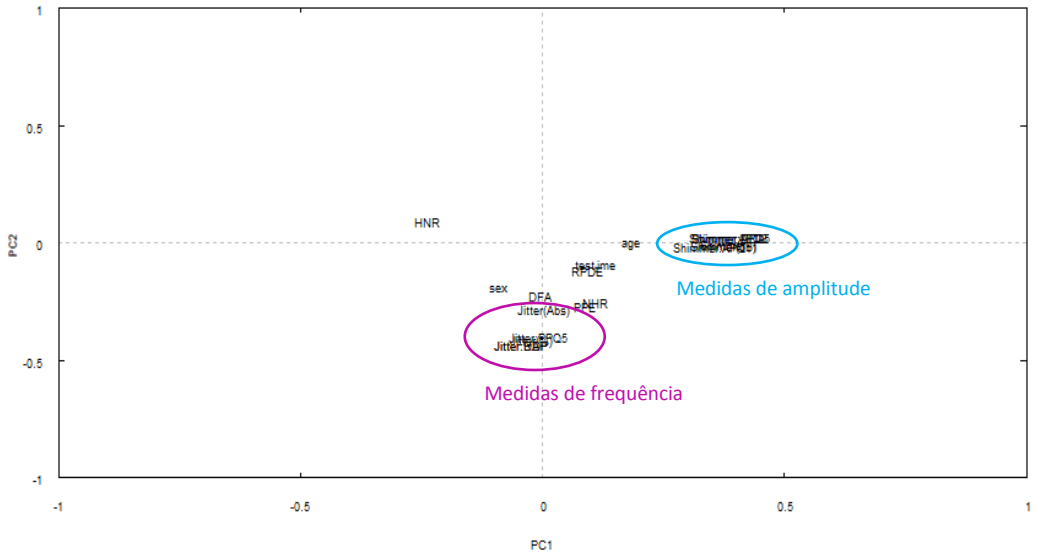


Figura 5.14 – Representação das variáveis originais associadas às medidas de voz e às características individuais dos pacientes no novo sistema de eixos (PC_1 *vs* PC_2), após a rotação varimax.

Para reforçar as conclusões tiradas anteriormente, encontram-se representadas, na Figura 5.15, os dados originais tendo em conta apenas duas das variáveis de maior relevância para as duas componentes principais: uma medida de amplitude (x4) e uma medida de frequência (x9). Podemos concluir, pela análise desta Figura, que usando apenas estas duas variáveis (correlacionadas) conseguimos separar na totalidade os três grupos.

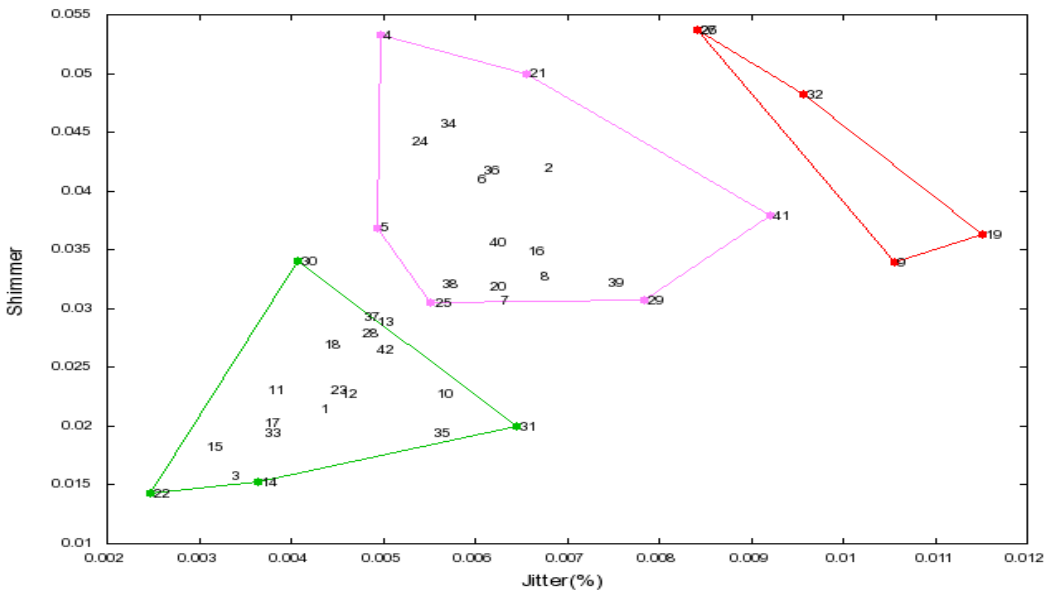


Figura 5.15 – Representação dos 42 pacientes com sintomas da doença de Parkinson tendo em conta apenas duas das variáveis originais responsáveis pela distribuição dos três grupos formados.

5.5 As universidades ibéricas e as suas áreas de ranking

Vamos agora tratar de um outro exemplo, diferente dos anteriores, e que ilustra claramente a versatilidade dos métodos quimiométricos, apresentados nos capítulos 2 e 3, na resolução de outro tipo de problemas, nomeadamente no tratamento e interpretação dos dados da ciência cultivada em Portugal e Espanha e constantes do *Essential Science Indicators* (ESI). Especificamente, vamos sempre considerar na base ESI as citações científicas de países com, pelo menos, cem mil citações para o conjunto das 22 áreas científicas e contempladas por esta base de dados num período de 10 anos. Este é o caso dos dois países ibéricos.

Uma questão importante é a escolha das citações como o melhor indicador para a ciência. As citações correspondem a um exame realizado a nível internacional, que se repete ao longo de vários anos. Por outro lado, os artigos, também apresentados por esta base de dados, correspondem a um exame efectuado a nível internacional, apenas numa determinada ocasião.

As citações são igualmente um indicador muito apropriado para avaliar o impacto científico das instituições de investigação e nomeadamente das Universidades, que constituem o verdadeiro motor da produção científica em quase todos os países do mundo. Reconhecendo este facto, o ESI apresenta as áreas científicas de cada instituição (*Field Rankings*) desde que se localizem em citações no *top* 1% a nível mundial da respectiva área: as denominadas "*áreas de ranking*". Com este requisito de competitividade internacional acima do limiar de *top* 1%, num período temporal de 10 anos, as áreas de *ranking* são um "indicador de qualidade" dinâmico, e consequentemente, não haverá o risco de todas as universidades ibéricas alcançarem uma saturação nos seus *Rk*. Seguindo a perspectiva já referida para as citações, podemos afirmar que os *Rk* correspondem a um exame a nível internacional, realizado de dois em dois meses pelas universidades, academias, institutos de investigação, hospitais e laboratórios do Estado nas suas diferentes áreas científicas, sendo seleccionadas apenas os melhores 1% [148].

5.5.1 Posições relativas das universidades ibéricas

Verificado que no final do ano de 2009 as universidades portuguesas haviam alcançado uma certa estabilidade no número, *Rk*, das suas áreas de *ranking*, para a

actualização de ESI de Janeiro de 2010 propusemo-nos fazer um estudo das universidades ibéricas em termos das áreas de *ranking*, atendendo ao seu número e natureza.

O conjunto de dados analisados corresponde a um estudo comparativo entre 55 universidades ibéricas tendo em conta 21 áreas de *ranking* (base de dados da *ESI* - consultada em Janeiro de 2010). A informação relativa às 55 universidades independentes é descrita por 21 variáveis associadas às áreas de *ranking*, descritas na Tabela 4.6 do Capítulo 4. Neste estudo foi excluída a área "Multidisciplinar" que cobre artigos publicados na *Nature*, *Science* e nos *Proceedings of the National Academy of Sciences*, mas dos quais 98% são distribuídos pelas restantes 21 áreas científicas com base nas citações[148].

Como referido anteriormente, o ESI apresenta as áreas científicas de cada instituição desde que estas se localizem no *top* 1% em citações, a nível europeu da respectiva área. Nos casos em que as áreas científicas se encontram no *top* 1% é atribuído o valor 1 a essas variáveis. Caso contrário, é atribuído o valor 0. A Figura 5.16 apresenta o extenso dendrograma das 55 instituições consideradas. O dendrograma é apresentado colorido para mais fácil separação dos diferentes grupos, que se estabeleceu com base numa distância de corte próxima de d=4. O método que melhor discrimina as instituições é o método da ligação *Ward* (variância mínima) para distâncias euclidianas.

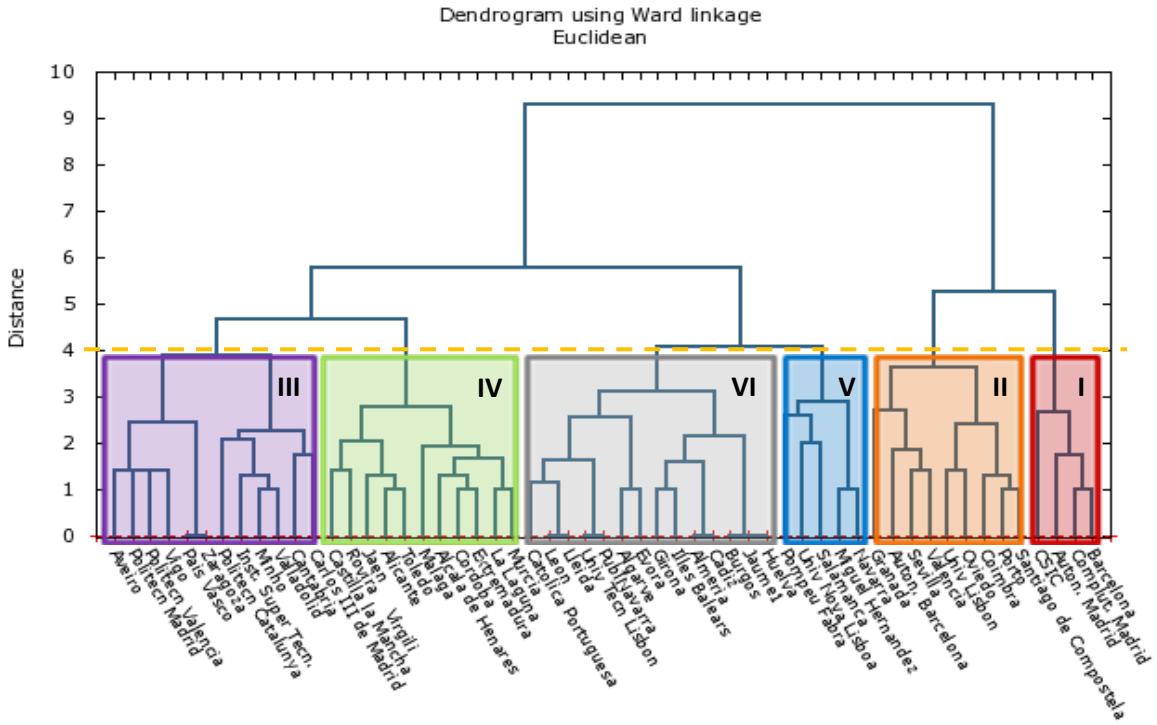


Figura 5.16 Dendrograma para as 55 universidades ibéricas; ESI Janeiro de 2010.

O dendrograma da Figura 5.16 separa no extremo da direita as "Universidades *Top*" (grupo I) no impacto da investigação, cujo grupo inclui o "CSIC" (Consejo Superior de Investigaciones Científicas), e as boas "Universidades Clássicas" (grupo II). No extremo do lado esquerdo figuram as "Universidades Politécnicas" (grupo III), grupo assim denominado por incorporar todas as universidades politécnicas espanholas.

A Espanha dispõe do CSIC (*Rk*=18), que é a maior instituição pública dedicada à investigação no país e a terceira a nível europeu, e três "Universidades *Top*" como a universidade de "Barcelona" (*Rk*=19), "Complutense de Madrid" (*Rk*=16) e "Autonoma de Madrid" (*Rk*=17).

No grupo das "Universidades Clássicas", em Espanha encontram-se seis universidades das quais a "Universidade Autonoma de Barcelona" é a que apresenta um *Rk* mais elevado, com *Rk*=15, seguida da universidade de "Valencia", com *Rk*=14. No entanto, o agrupamento não depende unicamente do valor de *Rk*, ou seja, procede também da natureza das áreas de *ranking*. Por exemplo, "Granada" e "Sevilla", ambas com *Rk*=12 e "Oviedo" com *Rk*=7, a de *ranking* mais baixo, encontram-se no mesmo grupo. A

"Universidade de Santiago de Compostela" com um *Rk*=10, tem o mesmo nível de *ranking* que as melhores universidades portuguesas, "Porto" e "Coimbra".

Na Figura 5.17 encontram-se representados, por correspondência de cores, os grupos formados a duas dimensões (2D). Esta representação revela a posição relativa dos grupos e das suas instituições.

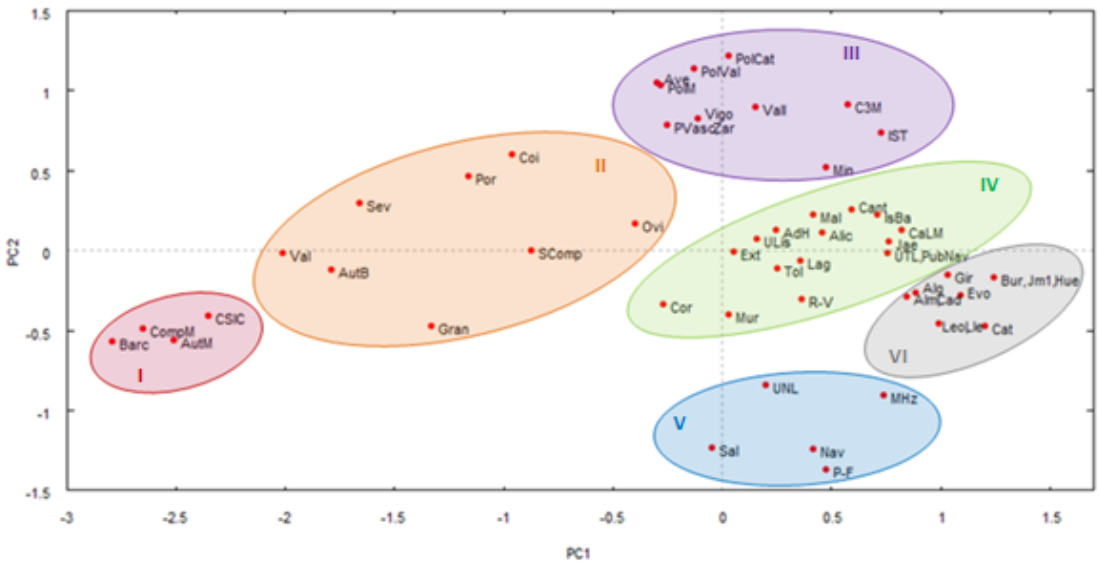


Figura 5.17 – Distribuição relativa das instituições ao longo dos eixos correspondentes às duas primeiras componentes principais (com uma recuperação de 46% da informação inicial).

Na actualização de Julho de 2010, algumas universidades perderam ou ganharam áreas de *ranking* e o dendrograma da Figura 5.18 sofre alteração, mesmo na disposição relativa dos grupos. As cores dos grandes grupos originais são mantidas para facilitar a identificação das alterações. Verificam-se, também, transições de universidades de um grupo para outro grupo vizinho: por exemplo, o grupo das "Universidades *Top*" (grupo I) vê-se enriquecido com mais uma instituição, correspondente à "Univ. Autonoma Barcelona". Do grupo das "Universidades Clássicas" (grupo II) saiu por subida de nível a "Univ. Autonoma Barcelona" e saíram "Granada", "Oviedo" e "Lisboa" que evoluíram para grupos de menor impacto. No entanto, tendo em conta a Análise de Componentes Principais (PCA), "Granada" mantém-se na órbita deste grupo, e por isso deve continuar a ser incluída no mesmo, dado ter um *Rk* elevado e não ter grupo vizinho onde se incluir.

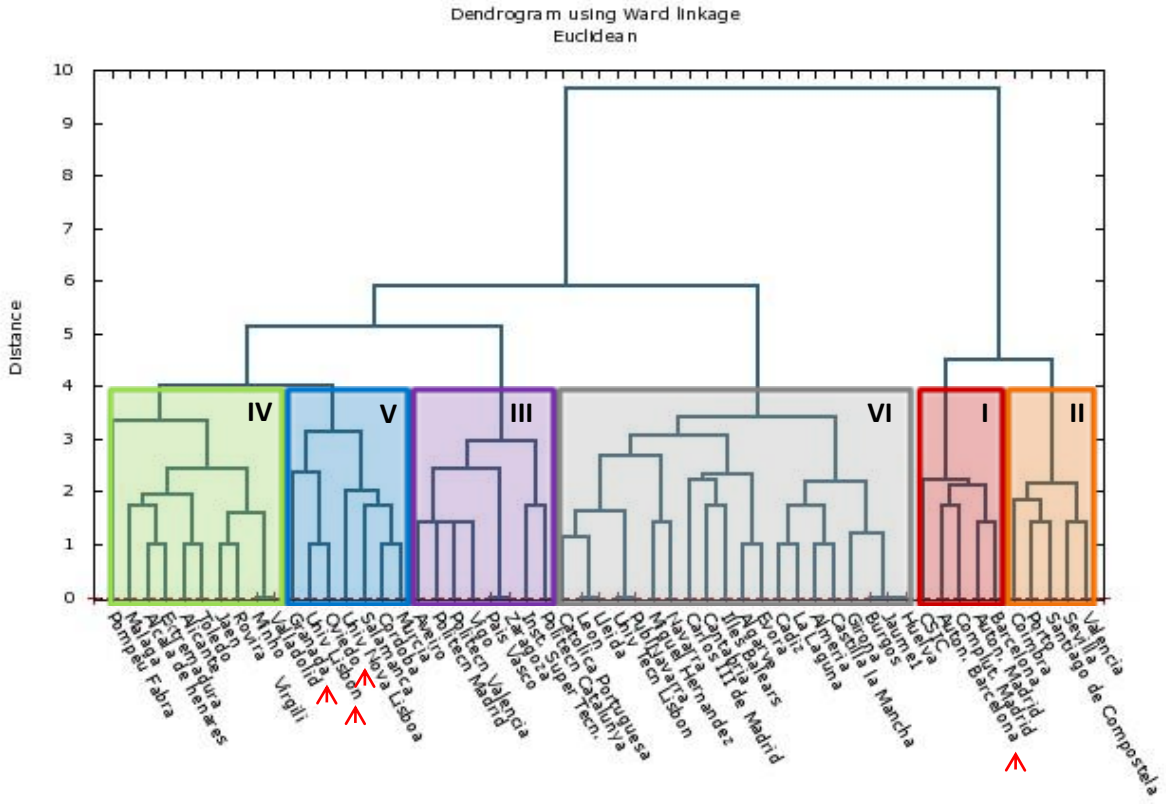


Figura 5.18 - Dendrograma para as 55 universidades ibéricas; ESI Julho de 2010. Encontram-se assinaladas as Universidades que sofreram alterações relativamente a Janeiro do mesmo ano.

5.5.2 Distinção entre as universidades

Para complementar a informação obtida com o método HCA e os respectivos dendrogramas, foi aplicada a análise de componentes principais (PCA) com base na matriz de covariância, usando toda a informação existente. Pelo critério de Pearson foram seleccionadas 8 componentes principais, dado que, só com a oitava componente recuperamos mais de 80% da informação total ou variabilidade total. Com duas componentes principais, recupera-se apenas 46% da informação inicial, e com três componentes principais apenas 55%.

Comecemos por analisar o PCA referente à actualização de Julho de 2010. Tendo em conta a variabilidade explicada pelas 8 componentes principais, podemos afirmar que as áreas de *ranking* mais discriminantes são: "*Agricultural Science*", "*Environment/Ecology*", "*Clinical Medicine*" e "*Physics*" (mantemos as designações originais, por coerência com a base de dados).

Na Figura 5.19 estão representados os *scores* relativos às universidades ibéricas a 2D (para mais informação apresenta-se na Figura 5.17 o tratamento correspondente, mas estabelecido em Janeiro de 2010). Sobre ele, a terceira dimensão (PC_3) está representada por barras cuja amplitude pode ser lida sobre o eixo PC_2.

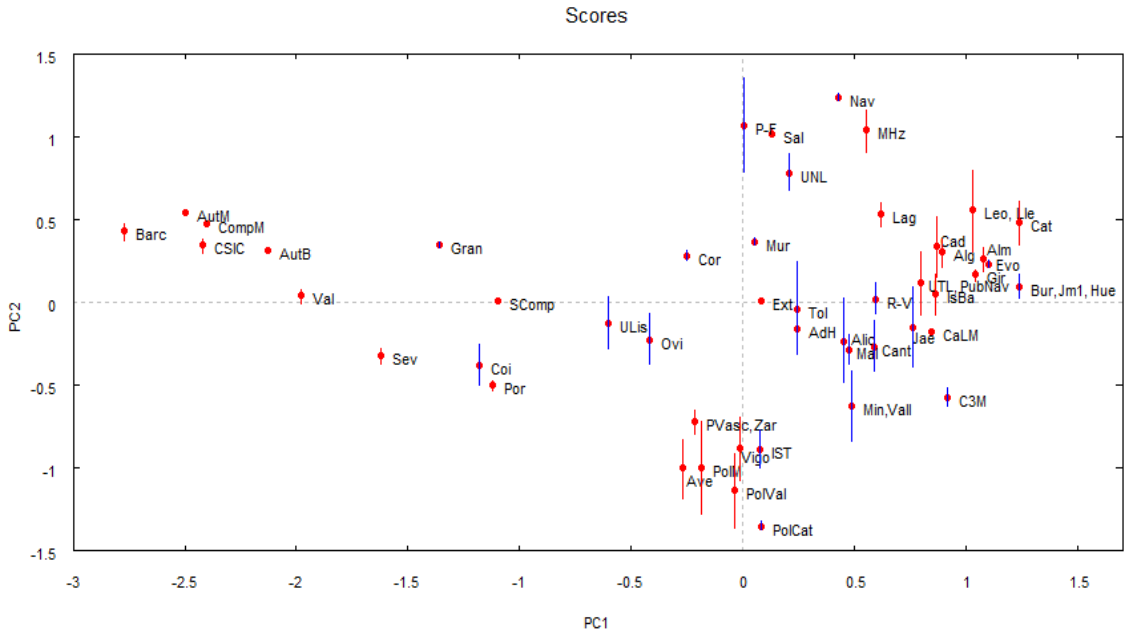


Figura 5.19 - PCA para as universidades ibéricas com representação dos *scores* relativos às universidades em estudo a 2D. Os *scores* positivos e negativos encontram-se representados na terceira dimensão por barras de cor vermelha e azul, respectivamente (ESI Julho de 2010). Por conveniência, utilizaram-se abreviaturas para os nomes das instituições.

Antes de prosseguirmos, há que atender há possível existência de correlações entre áreas científicas que por vezes surgem associadas no *ranking* de uma mesma universidade, porque uma qualidade elevada num dado domínio pode propulsionar qualidade em ramos afins. Na hierarquia das ciências, a Química é uma *ciência central* a partir da qual há uma *tri-furcação* de ramos de diferentes especialidades, como se ilustra no esquema da Figura 5.20, que inclui as Xenociências, a Biologia e as Nanociências. A Química também é considerada uma *ciência de transferências*, um nó de comunicação e uma ligação entre o simples e o complexo, entre as leis da física e as regras da vida, entre o fundamental e o aplicado.

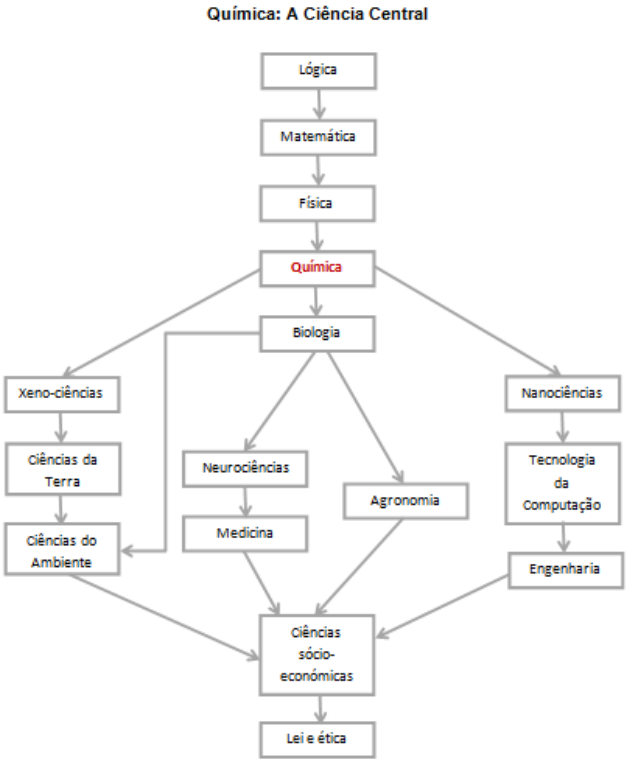


Figura 5.20 – Fluxograma representativo das áreas científicas, considerando a química como a ciência central.

Observando em detalhe a Figura 5.19 verificamos que os grupos das "Universidades *Top*" (grupo I) e das "Universidades Clássicas" (grupo II) apresentam uma variação significativa na primeira componente (PC_1), enquanto que na segunda componente (PC_2) a gama de variação é inferior. O grupo III das "Universidades Politécnicas "é muito pouco sensível na PC_1, com uma inclinação praticamente nula; as suas universidades têm impacto elevado em áreas do ramo das Nanociências.

O grupo IV/VI formado por universidades de *Rk*=1 ou *Rk*=2 e outras de *Rk* superior (Rk=6 ou Rk=7) apresenta uma dispersão significativa nos dois eixos. A "Univ Extremadura" com *Rk*=6 nas áreas de *"Agricultural Science"*, *"Chemistry"*, *"Clinical Medicine"*, "*Engineering*", "*Environment / Ecology*" e "*Plant & Animal Science*" encontra-se próximo da origem do sistema de eixos PC_1, PC_2.

A primeira componente principal (PC_1) discrimina as universidades de um modo quantitativo, em função do valor de *Rk*.

No ramo negativo do eixo PC_1 encontram-se as *"Universidades Top"* com os valores de *Rk* mais elevados (entre 15 e 19). No ramo oposto, encontram-se distribuídas as universidades com valores de *Rk* muito pequenos (por exemplo, as "Universidade de Evora" e "Católica

Portuguesa"). Um valor de *Rk* elevado (ramo negativo do eixo PC_1 com as "Universidades *Top*"), que cubra quase todas as áreas perde valor de discriminação. Do mesmo modo, valores de *Rk* muito pequenos, discriminam pouco as universidades.

Na Figura 3.21 é notória a distribuição das instituições ao longo do eixo PC_1 tendo em conta o número de áreas de *ranking* (valores de *Rk*).

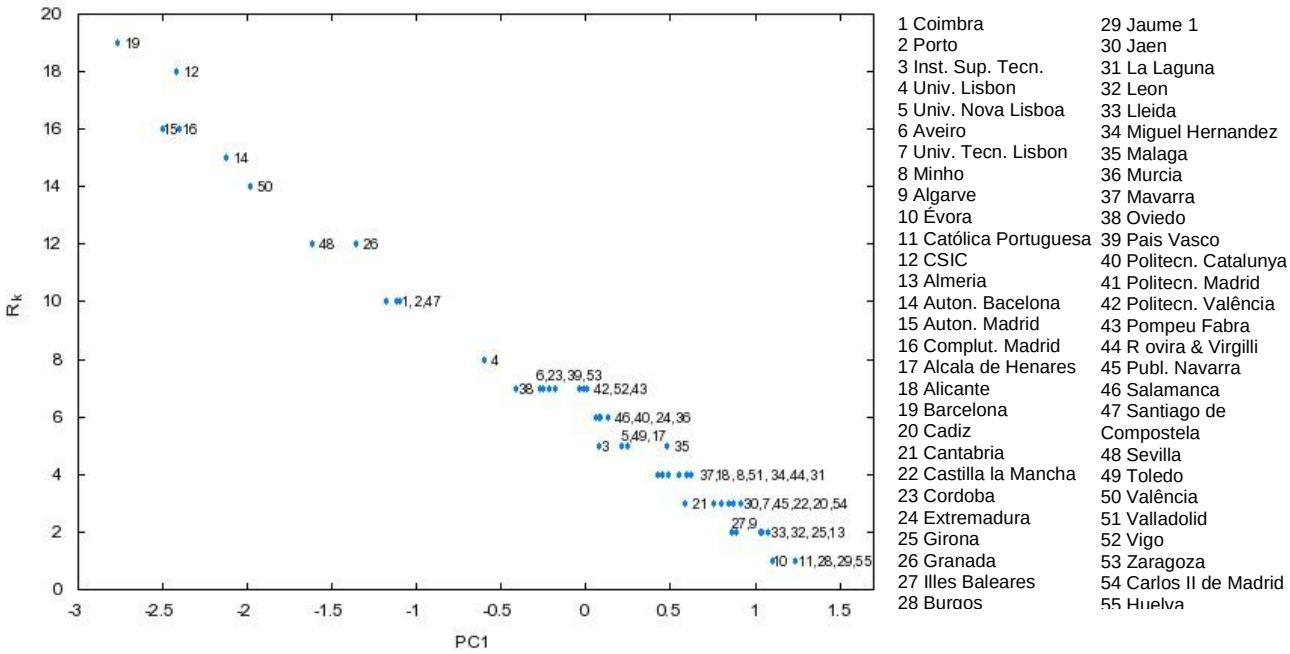


Figura 5.21 – Distribuição das universidades em estudo ao longo do eixo PC_1 em função dos valores de *Rk*. No extremo negativo do eixo PC_1 localizam-se as "Universidades *Top*" com valores de *Rk* muito elevados (*Rk*=15 a 19). As universidades com valores de Rk muito pequenos estão localizadas no extremo oposto.

Para caracterizar a segunda componente (PC_2) foram analisados os três grupos que apresentam maior dispersão ao longo do eixo PC_2: o grupo III das "Universidades Politécnicas" com predomínio do ramo das Nanociências, o grupo IV/VI (geral) de impacto intermédio a baixo, e o grupo V de impacto intermédio com predomínio do ramo da biologia.

A segunda componente (PC_2) caracteriza-se pela especificidade das áreas científicas, isto é, reflecte a influência da natureza específica das áreas de *ranking*. Tendo em conta a média das áreas de maior contribuição para a PC_2 (*"Material Science"*, *"Engineering"*, *" Physics"*, *" Neuroscience & Behavior"*, *"Biology & Biochemistry", "Chemistry" e "Molecular Biology & Biochemistry"*) representadas no gráfico da Figura 5.22, procurámos avaliar a importância destas áreas para a distribuição das instituições nos três grupos considerados.

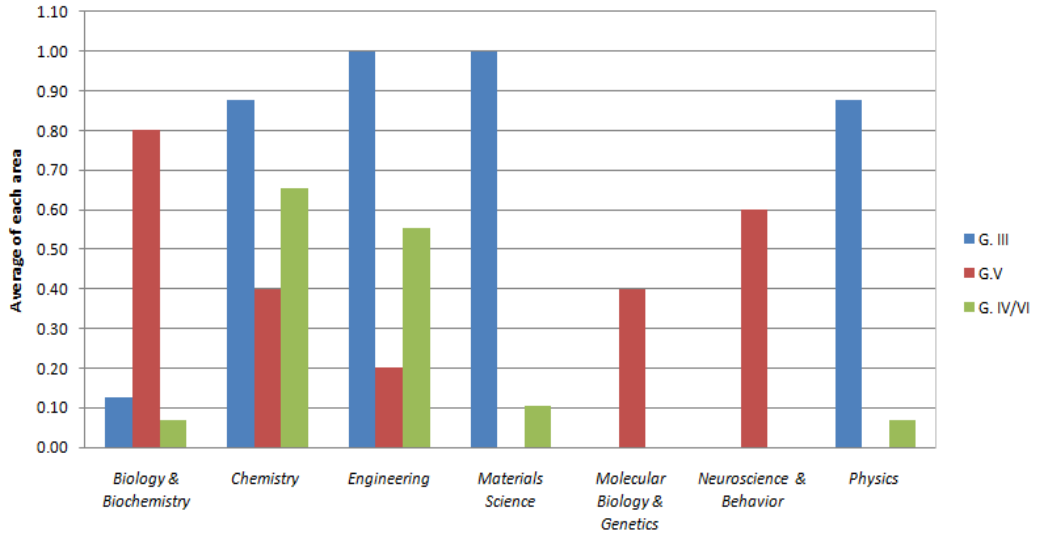


Figura 5.22 - Importância relativa das áreas científicas de maior contribuição em PC_2, baseada no valor médio dos *rankings* de cada área para os três grupos.

Pela análise do gráfico da Figura 5.22, podemos concluir que no grupo III, das "Universidades Politécnicas" (ramo negativo do PC_2), predominam áreas do ramo das Nanociências, como *"Materials Science"* e *"Engineering",* para além de *"Chemistry"* e *"Physics"*. Estas áreas são pouco dominantes ou mesmo ausentes (caso de *"Materials Science"* e *"Physics")* no grupo V (ramo positivo do PC_2). Neste último grupo, destacam-se as áreas do ramo da biologia (*"Biology Biochemistry"*, *"Neuroscience & Behavior"* e *"Molecular Biology & Genetics"*). O grupo IV/VI caracteriza-se por uma dominância intermédia das áreas consideradas.

Considerando a *loading* de maior peso para a segunda componente (PC_2), *"Material Science",* podemos verificar que a média correspondente influencia, significativamente, a posição dos três grupos no eixo PC_2.

Considere-se como exemplo, as universidades de *"Navarra"* (grupo V) e *"Politecn. Catalunya"* (grupo III) localizadas em ramos opostos do eixo PC_2. *"Navarra"* apresenta áreas como *"Molecular Biology & Genetics"* e *"Neuroscience & Behavior",* estando as mesmas ausentes nos grupos III e IV/VI. Por outro lado, na universidade "Politecn. Catalunya" dominam as áreas de *"Materials Science"*, *"Engineering"*, "*Physics"* e *"Chemistry"*, ausentes na instituição anterior. Com base neste exemplo, podemos concluir que a PC_2 traduz a influência da natureza específica das áreas na distribuição dos grupos.

Na Figura 5.19, e como referimos anteriormente, a terceira dimensão do PCA está representada por barras de amplitude proporcional ao valor respectivo (lida sobre o eixo PC_2), e a cores diferentes consoante a componente é positiva ou negativa. A hipótese de trabalho

que vamos considerar é que a terceira componente é caracterizada pela combinação de determinadas áreas científicas.

As áreas correspondentes a *"Agricultural Science"* e *"Plant & Animal Science"* quando combinadas com áreas de outros ramos das Nanociências, como *"Computer Science", "Engineering", "Materials Science"* e *"Physics"* conduzem a *scores* positivos na PC_3, apresentando a barra correspondente uma amplitude muito elevada. Como exemplos deste tipo de combinação destacam-se as universidades *"Politecn. Madrid"*, *"Politecn. Valência"* e *"Aveiro"*.

Os *scores* negativos surgem como resultado de combinações de áreas da biologia e outras como *"Economics & Business"*, *"Social Sciences-General"* e *"Engineering"*. Como exemplo deste tipo de combinação destaca-se *"Pompeu Fabra"*. Esta instituição apresenta um PC_3 negativo e de elevada amplitude, devido à combinação das áreas de *"Economics & Business"* e *" Social Sciences"*. A universidade de *"Pompeu Fabra*" é a única universidade ibérica que dispõe da combinação das áreas de *"Economics & Science"* e *"Social Science, General"*.

Universidades como *"Navarra*", *"Murcia", "Salamanca", "Santiago de Compostela"* ou *"Granada"* têm *scores* praticamente nulos na PC_3, dado que, apresentam uma distribuição de áreas muito equilibrada no ramo da biologia.

A discrepância encontrada para as universidades de *"Salamanca"* e *"Pompeu Fabra"* entre o dendrograma e a análise de *scores* a 2D é também explicada pela terceira componente.

Considerando a média das áreas de maior contribuição para a PC_3 (*"Agricultural Science", "Plant & Animal Science", "Clinical Medicine", "Chemistry" e "Engineering"*) procurámos avaliar a importância destas áreas para a discriminação das instituições na terceira dimensão (*scores* positivos e negativos).

Na Figura 5.23 encontra-se representado o impacto das áreas de maior contribuição para a PC_3 na discriminação dos *scores* positivos e negativos relativos às universidades ibéricas.

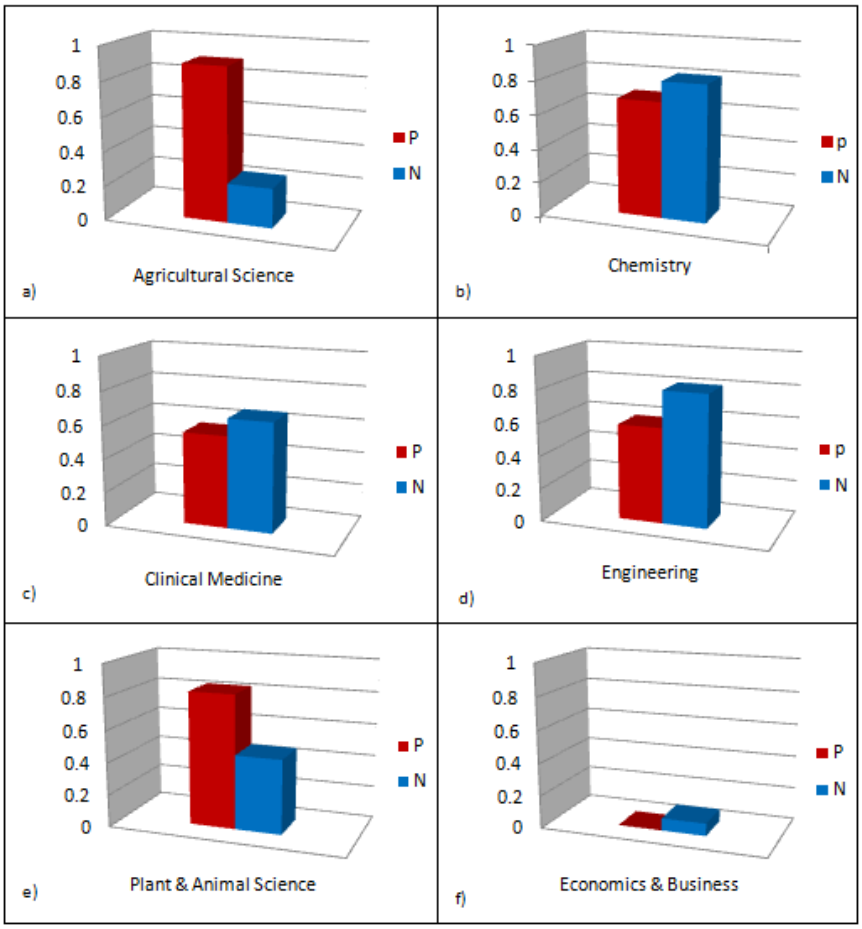


Figura 5.23 - Impacto das áreas de maior peso na PC_3 na discriminação das universidades em estudo. A média das variáveis para as universidades com *scores* positivos está representada a vermelho. As barras azuis representam a média das variáveis para as universidades com *scores* negativos.

Os gráficos da Figura 5.23a e 5.23e sustentam as conclusões retiradas anteriormente: as áreas científicas de *"Agricultural Science"* e *" Plant & Animal Science"* têm grande influência na discriminação dos *scores* na terceira componente.

Procuremos, nesta fase, explicar o comportamento das *loadings* nas três primeiras componentes principais. Na Figura 5.24 encontram-se representadas as contribuições das áreas científicas para estas componentes.

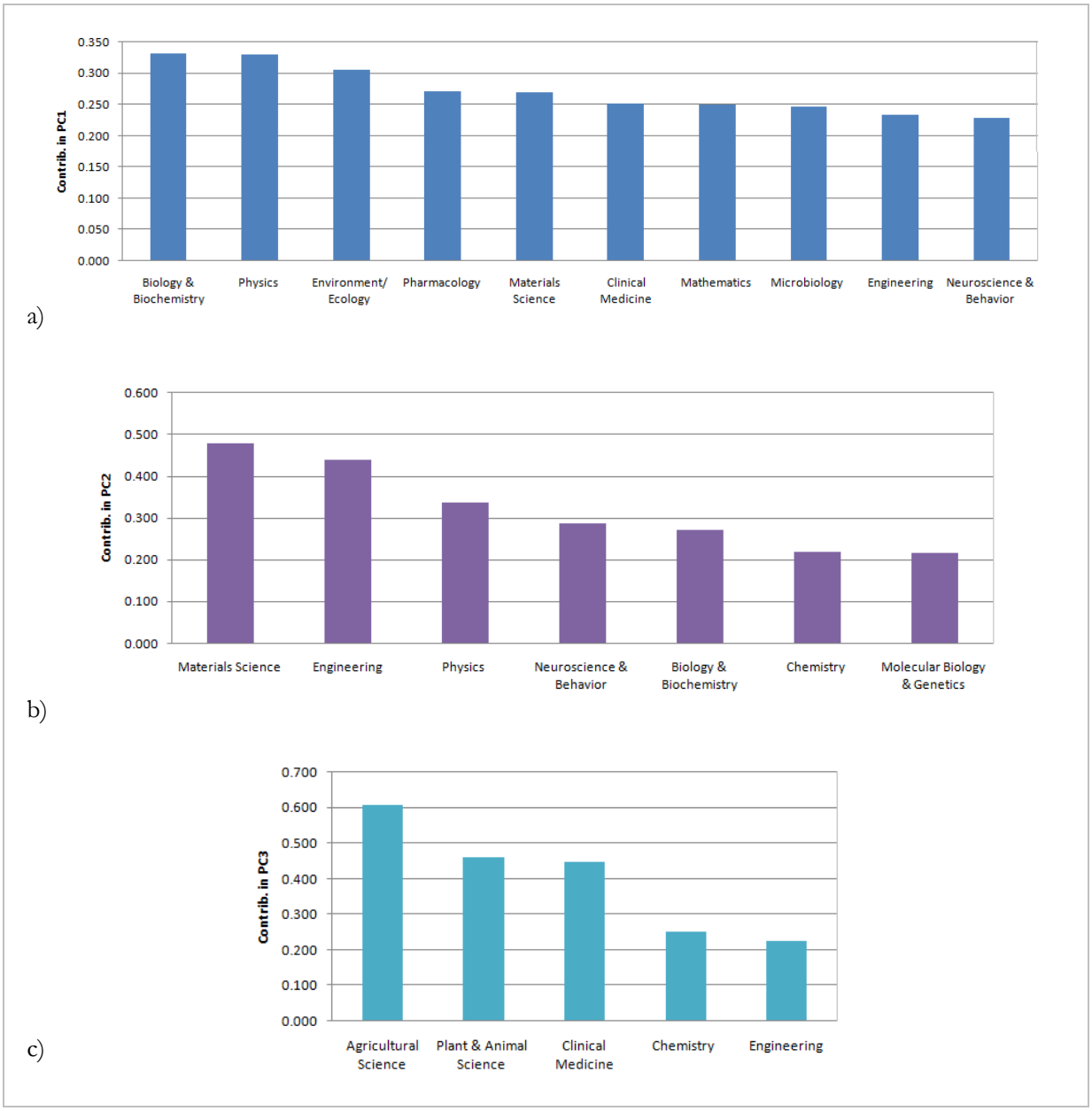


Figura 5.24 - Representação da contribuição das áreas científicas em estudo nas três primeiras componentes principais. A primeira componente (a) apresenta uma recuperação de apenas 36% da informação inicial e a segunda componente (b) apenas 46%. Adicionando a terceira componente (c), temos uma recuperação de 55% da informação.

Considerando as duas primeiras componentes principais (PC_1 e PC_2), podemos afirmar que as variáveis de maior peso (de maior importância para o valor da variância) na primeira componente (PC_1) são as áreas de *"Biology & Biochemistry"*, "*Physics*" e de *"Environment/Ecology"*; as áreas de *"Materials Science"*, *"Engineering" e* de "*Physics*" correspondem às variáveis de maior contribuição para a segunda componente (PC_2). Outras áreas com alguma relevância nestas componentes encontram-se também representadas.

As áreas de maior contribuição na terceira componente (PC_3) são: *"Agricultural Science", "Plant & Animal Science" e "Clinical Medicine".*

A visualização das *loadings* a 2D (PC_1 vs PC_2) não é esclarecedora, dado que, *"Agricultural Science",* de variabilidade máxima, apresenta maior peso na terceira componente. Tudo indica, que a terceira componente principal (PC_3) pode trazer informações estatísticas relevantes para o entendimento do sistema em estudo.

Analisando a Figura 5.25, que representa a variância em função da média das áreas científicas, podemos considerar que a variabilidade está associada ao afastamento em relação ao valor médio 0.5 (entre variáveis binárias, 0 e 1). Quanto maior a proximidade ao valor médio, maior a variabilidade associada às variáveis em estudo.

As *loadings* mais significativas (de variabilidade elevada) têm valor médio próximo de 0.5. Por exemplo, *"Agricultural Science"* e "*Environment /Ecology"* são as áreas científicas que apresentam maior variância (0.248), encontrando-se ambas à mesma distância do valor médio (0.58 e 0.42, respectivamente). Seguem-se *"Clinical Medicine"* e *"Physics"* com variância mais elevada (0.244). Por outro lado, as *loadings* menos significativas, como *"Space Science"* e *"Economics & Business"* apresentam valores de variância muito próximos de zero (0,018 e 0,036, respectivamente), encontrando-se muito afastados do valor médio. Deste modo, podemos concluir que o que distingue as áreas científicas é o afastamento em relação à média 0.5.

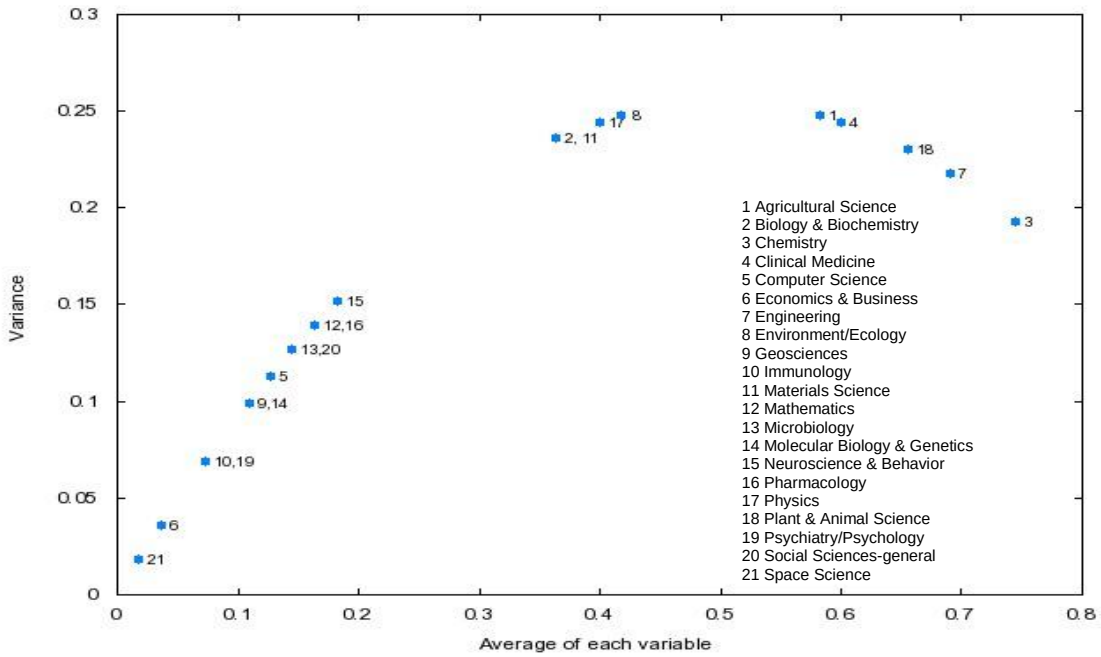


Figura 5.25 – Representação da variância em função da média de cada variável, para o conjunto de instituições em estudo. (ESI Julho 2010).

Podemos agora apresentar uma palavra final sobre o dendrograma correspondente à actualização de Julho de 2010 e o Grupo das "Universidades Clássicas". Como referido, apesar de " Granada" ficar um pouco marginal a este grupo, foi incluída no mesmo pelo seu *Rk* elevado. Já "Lisbon" e "Oviedo" estão situadas algures entre as "Universidades Clássicas" e um grupo geral de menor impacto. Observando os resultados do PCA na Figura 5.19, verificamos que a "Univ Lisbon" se aproximou do grupo das "Universidades Clássicas", carecendo ainda de mais uma área de *ranking* para entrar neste grupo. Parece ser relativamente indiferente a natureza da nova área de *ranking*. O mesmo é válido relativamente a "Oviedo", que carece de mais duas áreas de *ranking*. Este resultado deve-se, a uma definição mais apertada do grupo das "Universidades Clássicas" que esta metodologia conseguiu operar com os dados da actualização de Julho de 2010. Quanto à "Univ Granada" a questão não se prende com o valor de *Rk*, mas carece de um PC_2 um pouco mais negativo, ou seja, de um maior reforço no ramo das Nanociências.

Relativamente a Janeiro de 2010, em Julho de 2010 o grupo das "Universidades Politécnicas" ficou mais definido, dado que perdeu as universidades "Carlos III Madrid", "Valladolid", "Minho", assim como "Cantabria".

Capítulo 6

Considerações finais

A metodologia proposta neste trabalho encara a quimiometria como uma alternativa eficaz às ferramentas mais complexas para o tratamento de dados em diferentes domínios, tais como no diagnóstico médico e implementação clínica, em situações reais multivariáveis e em outras áreas do conhecimento.

Especificamente, foram tratados e discutidos detalhadamente uma variedade de dados relativos ao diagnóstico de doenças, como o cancro e a doença de Parkinson, a estudos epidemiológicos e a estudos relativos à cientometria.

Os resultados obtidos nos diferentes casos mostram a versatilidade nos métodos quimiométricos clássicos na resolução de problemas específicos e de diferentes naturezas. Revelam também, que usando apenas estes métodos padrão, após uma selecção cuidadosa da abordagem para cada caso, é possível facilmente (i) separar classes, recorrendo tanto técnicas supervisionadas como a técnicas não supervisionadas, (ii) identificar variáveis redundantes, (iii) isolar e identificar padrões e factores, incluindo marcas geográficas.

Deste modo, uma sequência de métodos clássicos como o HCA, o PCA, o LDA e o PLS pode auxiliar no diagnóstico de doenças, nomeadamente o cancro, de acordo com as propriedades disponíveis, e a abordagem mais conveniente para cada caso específico. Esta sequência é, também, extremamente eficaz no tratamento de outros tipos de problemas, mas nem sempre é necessário aplicar todos os métodos na interpretação do mesmo conjunto de dados. Em certas situações podem ocorrer excepções. No entanto, a aplicação é estatisticamente correcta. A excepção não distorce o significado destes métodos no auxílio do diagnóstico médico.

O PCA, além de permitir uma boa visualização dos dados, constitui uma ferramenta essencial no tratamento dos problemas considerados, fornecendo praticamente toda a informação necessária à interpretação dos mesmos.

Uma abordagem quimiométrica, em combinação com uma ferramenta que inclui vários métodos, pode ser aplicada de forma eficiente a casos específicos, desde o inicio da análise até todo o processo de interpretação.

No cenário actual, é importante dispor de uma metodologia de análise de dados que possa, de um modo mais criterioso e científico, extrair o maior número de informações de um conjunto de dados.

Referências bibliográficas

[1] Brown D. S., Stephen T., Despagne F. *Anal. Chem.* 1996; **68**: 21-61.

[2] Cazar R. A. *J. Chem.,* Ed.*,* Madison, WI, 2003; **80(9)**: 1026-1029.

[3] Jolliffe I. T. *Principal Component Analysis*, 2nd ed., Springer: New York, 2002.

[4] Sharaf M.A., Illman D. L. and Kowalski B. R. *Chemometrics*. John Wiley & Sons: New York, 1986.

[5] Brereton R. G. *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*. Ellis Horwood: Chichester, 1990.

[6] Kowalski B. R. (editor). *Chemometrics: Mathematics and Statistics in Chemistry*. Reidel: Dordrecht, 1984.

[7] Otto M. *Chemometrics: Statistics and Computer Applications in Analytical Chemistry*. Wiley VCH: Weinheim, 1998.

[8] Jossinet J., Lavandier B. *Bioelectroch. and Bioener.* 1998; **45**: 161–167.

[9] Sawarkar S. D. , Ghatol A. A., A. P. Pande. In *Proceedings of the 7th WSEAS International Conference on Neural Networks*, Cavtat, Croatia, June 12-14, 2006: 158-163.

[10] Thurfjell E. L., Lernevall K. A., Taube A. A. S. Benefit of independent double reading in a population-based mammography screening-program, *Radiology* 1994; **191**: 241–244.

[11] Mautner B. D., Schmidt K. V., Brennan M. B. New diagnostic techniques and treatments for early breast cancer. *Semin. Oncol. Nurs.* 2000; **16**: 185-196.

[12] Piacenti M. et al. *Spectrochimica Acta B* 2009; **64**: 587-592.

[13] Almeida J. A. S., Barbosa L. M. S., Pais A. A. C. C., Formosinho S. J. *Chemom. Intell. Lab. Syst.* 2007; **87**: 208-217.

[14] Wold S., Esbensen K., Geladi P. *Chemom. Intell. Lab.* 1987; **2**: 37.

[15] Martens H., Martens M. *Multivariate Analysis of Quality: An Introduction*. Wiley: New York, 2001: 68-73.

[16] Wu L., Jiang H., Lin Q. *Curr. Comput.* 2006; **2(3)**: 255-266.

[17] Munck L., Norgaard L., Engelsen S. B. *Chemom. Intell. Lab. Syst.* 1998; **44**: 31-60.

[18] Duin J. R., Mao J. *IEEE Trans. Pattern Anal.* 2000; **22(1)**: 63-68.

[19] Anderson T. W. *An introduction to multivariate statistical analysis*. John Wiley & Sons: New York, 1984: 675.

[20] Cochran W., Cox G. *Experimental Designs*. Wiley: New York, 1950.

[21] Scheffe H. *The Analysis of Variance*. Wiley: New York, NY, 1959.

[22] Scarle S. *Linear Models*. Wiley: New York, NY, 1971.

[23] Box G., Hunter W., Hunter S. *Statistics for experimenters: An introduction to Design, Data Analysis and Model Building*. Wiley: New York, NY, 1978.

[24] Box G., Draper N. *Empirical Model-Building and Response Surfaces*. Wiley: New York, 1987.

[25] Morgan E. *Chemometrics: Experimental Design*. Wiley: Chichester, 1991.

[26] Deming S., Morgan S. *Experimental Design: A Chemometric Approach*. 2nd ed., Elsevier: Amsterdam, 1993.

[27] Meyers R., Montgomery D. *Response surface methodology, Process and Product Optimization Using Designed Experiments*, Wiley: New York, 1995.

[28] Carlson R. *Design and Optimization in Organic Synthesis*. Elsevier: Amsterdam, Netherlands, 1992.

[29] Lewis G., Mathieu D., Phan-Tan-Luu R. *Pharmaceutical Experimental Design*. Marcel Dekker: New York, 1998.

[30] Liu S., Kokot S., Will G. *J. Photochem. Photobio. C: Photochemistry Reviews* 2009; **10**: 159-172.

[31] Roggo Y., Chalus P., Maurer L., Martinez C. L., Edmond A., Jent N. *J. Pharmaceut. Biomed. Anal.* 2007; **44**: 683–700 .

[32] Kim S. B. et al. *Expert Syst. Appl.* 2010; **37**: 3863-3869.

[33] Ni Y., Kokot S. *Anal. Chim. Acta* 2008; **26**: 130-146.

[34] Silva M. P., Zucchi O. L. A. D., Silva A. R., Poletti M. E. *Spectrochim. Acta B* 2009; **64**: 587-592.

[35] Lewis P. D., Manshian B., Routledge M. N., Scott G. B., Burns P. A. *Carcinogenesis,* 2008; **29(4)**: 772–778.

[36] Benninghoff L., Czarnowski D., Denkhaus E., Lemkem K.. *Spectrochim. Acta B* 1997; **52**: 1039-1046 .

[37] Madsen R., Lundstedt T., Trygg J. *Anal. Chim. Acta* 2010; **659**: 23-33.

[38] J. Trygg, E. Holmes, T. Lundstedt. *J. Proteome. Res.* 2007; **6**: 469.

[39] Madsen R., Lundstedt T., Trygg J. *Anal. Chim. Acta* 2010; **659**: 23-33.

[40] Woo H. M., Kim K.M., Choi M. H. et al. *Clin. Chim. Acta* 2009; **400**: 63-69.

[41] Lanfranchi G. et al. *Hum. Mol. Gen.* 2003; **12(8)**: 823-836.

[42] Li W. et al. *Clin. Chim. Acta* 2009; **401**: 8-13.

[43] Wang C. et al. *Comput. Biol. Chem.* 2004; **28**: 235-244.

[44] Altman R. B. et al. *J. Biomed Inform.* 2010; **43**: 932-944.

[45] Cano C., García F., López F. J., Blanco A. *Expert Syst. Appl.* 2009; **36**: 4654-4663.

[46] J. Gui et al. *Artif. Intell. Med.* 2010; **50**: 181–191.

[47] Xiong M., Jin L., Li W., Boerwinkle E., *Biotechniques* 2000; **29**: 1264–1268.

[48] Furey T. S., Cristianini N., Duffy N., Bednarski D.W., Shummer M., Haussler D. *Bioinformatics* 2000; **16**: 906-914.

[49] Tibshirani R., T. Hastie, Narasimhan B., Chu G. *Proc. Natl. Acad. Sci.* USA 2002; **99**: 6567-6572.

[50] Nguyen D.V., Rocke D.M. *Bioinformatics* 2002; **18**: 39–50.

[51] Massart D. L., Vandeginste B., Buydens L., de Jong S., Lewi P. J., Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics*: *Part B*. Elsevier: Amsterdam, 1998.

[52] Brereton R. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant.* Wiley: Chichester, 2003.

[53] Massart D. L., Vandeginste B. G. M., Deming S. N., Michotte Y., Kaufman L. *Chemometrics: a Textbook*. Elsevier: Amsterdam, 1988.

[54] Campanella L., de Angelis G., Visco G. *Anal. Bioanal. Chem.* 2003; **376**: 467.

[55] Kokot S., Grigg M., Panayiotou H., Phuong T. D. *Electroanalysis* 1998; **10**: 108.

[56] Wold S., Esbensen K., Geladi P. *Chemom. Intell. Lab. Syst.* 1987; **2**: 37.

[57] Jackson E. *A User's Guide to Principal Components*. John Wiley & Sons, Inc.: New York, 1991.

[58] Acara E., Brob R., Schmidt B. *J. Chemometrics* 2008: **22**: 91–100.

[59] Downs G. M., Barnard J. M. *Clustering methods and their uses in computational chemistry*. In: Lipkowitz K. B. Boyd D.B. (Eds.), *Reviews in Computational Chemistry*. Wiley: United Kingdom, 2002; **18**: 1-40.

[60] Daszykowski M., Walczak B., Massart D. L. *Anal. Bioanal. Chem*. 2004; **380**: 370-372.

[61] Bratchell N. Cluster analysis. *Chemom. Intell. Lab. Syst.* 1989: 105–125.

[62] Danzer K., Hobert H., Fischbacher C., Jagemann K. U. *Chemometrics*. Springer: Berlin, 2001.

[63] Lilanda K. H., Indahl U. G. Powered partial least squares discriminant analysis, *J. Chemometrics* 2009; **23**: 7–18.

[64] Fisher R. A. The use of multiple measurements in taxonomic problems. *Annals Eugen.* 1936; **7**: 179-188.

[65] Sharaf M., Illman D., Kowalski B. *Chemometrics*. Wiley/Interscience: New York, 1986.

[66] Candolfi A., Wu W., Massart D. L., Heuerding S. *J. Pharm. Biomed. Anal.* 1998; **16**: 1329-1347.

[67] Wu W., Mallet Y., Walczak B., Penninck W., Massart D. L., Heuerding S., Erni F. *Anal. Chim. Acta* 1996; **329**: 257-265.

[68] Dudoit S., Fridlyand J., Speed T. P. *Comparison of discrimination methods for the classification of tumors using gene expression data*. *J. Am. Stat. Assoc.* 457; **97**: 77-87.

[69] Mardia K. V., Kent J. T., Bibby J. M. *Multivariate Analysis*. Academic Press, Inc.: San Diego, 1979.

[70] McLachlan G. J. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley: New York, 1992.

[71] Ripley B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press: Cambridge, New York, 1996.

[72] Wu B., Abbott T., Fishman D., McMurray W., Mor G., Stone K., Ward D., Williams K., Zhao H. *Bioinformatics* 2003; **19(13)**: 1636–1643.

[73] Adams M. J. *Chemometrics in Analytical Spectroscopy*. Royal Society of Chemistry: Cambridge, UK, 1995.

[74] Wold S., Trygg J., Berglund A., Antti H. *Chemom. Intell. Lab. Syst*. 2001; **58**: 131-150.

[75] Fearn T. *Chemom. Intell. Lab. Syst.* 2000; **50**: 47–52.

[76] Markopoulou C. K., Malliou E. T., Koundourellis J. E. *J. Pharm. Biomed. Anal.* 2005; **37**: 249-258.

[77] Jiang J. H., Berry R. J., Siesler H. W., Ozaki Y. *Anal. Chem.* 2002; **74**: 3555–3565.

[78] Tenenhaus M., Vinzia V. E., Chatelinc Y. M., Laurob C. PLS path modeling. *Comput. Stat. Data An.* 2005; **48**: 159-205.

[79] Vinzi V., Chin W., Henseler J., Wang H. (eds), *Handbook of Partial Least Squares*, 2010.

[80] Wold S., Ruhe A., Wold H., Dunn W. J. The collinearity problem in linear regression: the partial least squares (PLS) approach to generalized inverses. *SIAM, J. Sci. Stat. Comp*. 1984; **5**: 735-743.

[81] Sneath P. H. A., Sokal R. R. *Numeric taxonomy: the principles and practice of numerical classification*. San Francisco: W. H. Freeman, 1973: 573.

[82] Kaufman L., Rousseeuw P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley: New York, 1990.

[83] Gower J. C. *Biometrics* 1967; **23**: 623-637.

[84] Edwards A. W. F, Cavalli-Sforza L. L. *Biometrics* 1965; **21(2)**:362–375.

[85] Gower J. C., Legendre P. *J. Classif.* 1986; **3**: 5-48.

[86] Johnson R. A., Wichern D. W. *Applied multivariate statistical analysis*, 3rd ed., Prantice Hall: New Jersey, 1992: 642.

[87] Duarte M. C., Santos J. B., Melo L. C. *Genet. Mol. Biol.* 1999; **22(3)**: 427-432.

[88] Wu J., Mo C., Gan G. *Data Clustering: Theory, Algorithms and Applications*. *SIAM* 2007: 3-17, 19-24, 67-71, 74-96.

[89] Berry M. J. A., Linoff G. *Data mining techniques*. John Wiley & Sons, 1997.

[90] Khattree R., Naik D. N. *Multivariate data reduction and discrimination with SAS Software*. Wiley InterScience, 2000.

[91] Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2001.

[92] Anderberg M. R. *Cluster analysis for applications*. Academic Press: New York, 1973: 359.

[93] Sneath P. H. A. *Gen. Microbiol.* 1957; **1(17)**: 201-226.

[94] Orlóci L. *Multivariate analysis in vegetational research*. 2nd ed., The Hague, Dr. W. Junk B. V. Publishers, 1978: 451.

[95] Mardia A. K. V., Kent J. T., Bibby J. M. *Multivariate analysis*. Academic Press: London, 1997: 518.

[96] Sibson R. *Comput. J.* 1973; **1(16)**: 30-34.

[97] Rohlf F. J. *Comput. J.* 1978; **1(16)**: 93-95.

[98] Duda R. O., Hart P. E., Stork D. G. *Pattern Classifrication*. Wiley Inter Science: New York, 2nd ed., 2000.

[99] Everitt B. *Cluster analysis*. Heinemann Educational Books: London, 1974: 136.

[100] Sokal R. R., Michener C. D. A statistical method for evaluating systematic relationships. *Bull. Soc. U. Kans.* 1958; **38**: 109-143.

[101] Ward J. H. *J. Am, Stat. Assoc.* 1963; **58**: 236-244.

[102] Kaufman and Rousseeuw P. J. L. *Agglomerative nesting (program AGNES)*. 1999; **1(5)**: 199-252.

[103] Zhang T., Ramakrishman, M. R. Livny, BIRCH: an efficient data clustering method for very large databases. In *Second International Conference on Knowledge Discovery and Data Mining*, *Am. Assoc. Artif. Intell.* 1996: 103-114.

[104] Guha S., Rastogi and Shim K. R. CURE: an efficient clustering algorithm for large databases. In *ACM Sigmod International Conference on Management of Data*. 1998: 73-84.

[105] Karypis G., Kumar V. E. H. S. CHAMELEON: Hierarchical clustering using dynamic modeling. *Computer* 1999; **32(8)**: 68-75.

[106] Kaufman and Rousseeuw P. J. L. *Divisive analysis (program DIANA)*. 1990; **1(6)**: 253-279.

[107] Guha S., Rastogi and Shim K. R. ROCK: A robust clustering algorithm for categorical attributes. *Inf. Syst.* 2000; **25(5)**: 345-366.

[108] Kaufman and Rousseeuw P. J. L. *Monothetic analysis (program MONA)*. 1990; **1(7)**: 280-311.

[109] Davies A. M. C., Fearn T. Back to basics: the principles of Principal Component Analysis. *Spectrosc. Eur.* 2004; **16**: 20-23.

[110] DelValls T. A., Forja J. González-Mazo M., E., Gómez-Parra A., Blasco J. *Trends Anal. Chem.* 1998; **17**: 181.

[111] Blanco M., Coello J., Eustaquio A., Iturriaga H., Maspoch S. *Anal. Chim. Acta* 1999; **392**: 237-246.

[112] Haiyan C., Young H. *Trends Food Sci. Tech., Review* 2007; **18 (2)**: 72-83.

[113] Rossel R. A.V., Walvoort D. J. J., McBartney A. B., Janik L. J. *Geoderma,* 2006; **136**: 59-75.

[114] Tucker L. R. *Psychometrika* 1966; **31**: 279.

[115] Kroonenberg P. M. *Three Mode Principal Component Analysis: Theory and Applications*, DSWO: Leiden, 1983.

[116] Kiers H. A. L. *Psychometrika* 1991; **56**: 449.

[117] Harshman R. A. *UCLA Work. Pap. Phonet.* 1970; **1**: 16.

[118] Carroll J. D., Chang J. J. *Psychometrika* 1970; **35**: 283.

[119] Ho C. N., Christian G. D., Davidson E. R. *Anal. Chem.* 1978; **50**: 1108.

[120] Carroll J. D., Chang J. J. *Psychometrika* 1978; **35**: 283.

[121] Appellof C. J., Davidson E. R. *Anal. Chem.* 1981; **53**: 2053.

[122] Russell M. D., Gouterman M. *Spectrochim. Acta A* 1988; **44**: 857.

[123] Russell M. D., Gouterman M. *Spectrochim. Acta A* 1988; **44**: 863.

[124] Russell M. D., Gouterman M., Vanzee J. A. *Spectrochim. Acta A* 1988; **44**: 873.

[125] Sanchez E., Kowalski B. R. *Anal. Chem.* 1986; **58**: 496.

[126] Sanchez E., Kowalski B. R. *J. Chemom.* 1990; **4**: 29.

[127] Anderson C. A., Bro R. *Chemom. Intell. Lab. Syst.* 2000; **1**: 52.

[128] Tauler R. *Chemom. Intell. Lab. Syst.* 1995; **30**: 133.

[129] Linder M., Sundberg R. *Chemom. Intell. Lab. Syst.* 1998; **42**:159.

[130] Geladi P. *Chemom. Intell. Lab. Syst*. 1989; **7**: 11.

[131] Smilde A. K. *Chemom. Intell. Lab. Syst*. 1992; **15**:143.

[132] Bro R. *Tese de Doutoramento*, Universidade de Amsterdam, Holanda, 1998.

[133] Bro R. *J. Chemom*. 1996; **10**: 47.

[134] Bro R. *Chemom. Intell. Lab. Syst*. 1997; **38**: 149.

[135] http://octave.sourceforge.net

[136] Pereira J. L. G. F. S. C. *Caracterização e Validação de Métodos Analíticos*. Coimbra, 2008: 63-67.

[137] Adams M.J. *Chemometrics and Statistics: Multivariate Classification Thecniques*. Elsevier: Australia, 2005: 21-26.

[138] Goodman L. A., Kruskal W. H. *J. Am. Stat. Assoc.* 1963; **58**: 310-364.

[139] Choi S., Cha S., Tappert C. C. *J. Pattern Recogn.* 2000; **35**: 515-525.

[140] Hubalek Z. *Biol. Reviews* 1982; **57(4)**: 669-689.

[141] Willett P. *Biochem. Soc. Trans.* 2003; **31**: 603–606.

[142] Belhumeur P. N., Hespanha J. P., Kriegman D. J. *IEEE Trans. Pattern An. Machine Intell.*, 1997; **19 (7)**: 711-719.

[143] Devroye L., Toussaint G. A note on linear expected time algorithms for finding convex hulls, Computing 1981; **26**: 361-366.

[144] Avis D., Bremner D., Seidel R. How good are convex hull algorithms? *Comput. Geom. Theo. Appl.* 1997; **7 (5-6)**: 265–301

[145] http://mathworld.wolfram.com/ConvexHull.html (consultado em Fevereiro de 2011)

[146] *UCI Machine Learning Repository*, http://www.ics.uci.edu/~mlearn/MLRepository.html - (consultado em Novembro de2010).

[147] *American Cancer Society (ACS)*, http://www.cancer.org/Research/CancerFactsFigures/CancerFactsFigures/cancer-facts-and-figures-2010 (consultado em Novembro de 2010).

[148] *Essencial Science Indicators* (ESI), http://esi.isiknowledge.com/allmenus.cgi?option=I (consultado em Julho de 2010).

[149] Chan E. C., Koh P.K., Mal M., Cheah P.Y., Eu K.W., Backshall A., Cavill R., Nicholson J.K., Keun H.C. *J. Proteome. Res.* 2009; **8**: 352.

[150] Wolberg W. H., Mangasarian O.L. MultiSurface method of pattern separation for medical diagnosis Applied to breast cytology. In *Proc. Natl. Acad. Sci.* 1990; **87**: 9193-9196.

[151] Mangasarian O. L., Wolberg W. H. Cancer diagnosis via linear programming. *SIAM News* 1990; **23 (5)**: 1-18.

[152] Zhang J. Selecting typical instances in instance-based learning. In *Proc. Int. Machine Learn. Conf.* Morgan Kaufmann: Aberdeen, Scotland, 1992: 470-479.

[153] Jossinet J. Variability of impedivity in normal and pathological breast tissue. *Med. Biol. Eng. Comput.* 199; **34**: 346-350.

[154] Silva J. E., Sá J. P., Jossinet J. Classification of Breast Tissue by Electrical Impedance Spectroscopy. *Med. Biol. Eng. Comput.* 2000; **38**: 26-30.

[155] http://www.mentorwwllc.com/global-pt/breast-reconstruction/mastectomy-anatomy.htm (consultado em Julho de 2011).

[156] Tsanas A., Little M. A., McSharry P. E., Ramig L. O. Accurate telemonitoring of Parkinson.s disease progression by non-invasive speech tests. *IEEE Trans. Biomed- Eng.* 2009; **56 (4)**: 1015-1022.

[157] Barber C. B., Dobkin D. P. The convekhulls Algothim for convexhulls. *ACM Trans. Math. Soft.* 1996; **22(4)**: 469-483.

[158] Hussain Z. A fast approximation to convex hull. *ACM Trans. Math. Soft.* 1988; **8(5)**: 289-294.

Anexo A.1

CHEMTOOL
Chemometrics ToolBox

Ver.02

Índice

Introdução

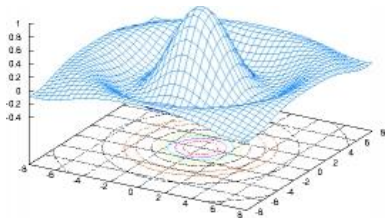
Este manual resume as diferentes potencialidades da *Chemometrics ToolBoox*, procurando, deste modo, documentar e facilitar a sua utilização já que serão aqui descritas as ferramentas disponíveis e as opções implementadas.

Nos capítulos seguintes são apresentadas as diferentes funcionalidades desta ToolBox bem como os respectivos parâmetros de entrada e de saída.

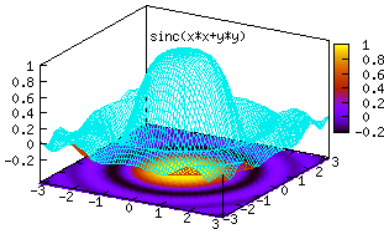
Devido ao seu grau de complexidade, para que o conjunto de ferramentas corra sem problemas foi necessário impor algumas convenções e restrições iniciais ao seu funcionamento para evitar perdas de tempo futuras. Mais informações encontram-se na secção 1.5.

O correcto funcionamento desta ferramenta de trabalho envolve a instalação e integração de dois tipos de programas – o Octave[30] e o GnuPlot[31].

Octave



Gnuplot



Notação

Para facilitar o entendimento é necessário impor algumas convenções de representação no texto.

- No caso de existir diferenças no modo de funcionamento entre as plataformas Windows e Linux os comandos anteriores surgirão com a respectiva referência (Windows / Linux)
- Os nomes de directórios e de ficheiros representam-se com a cor azul-claro e sob a forma "directório" ou "nome.ext", respectivamente.
- Esta ToolBox é composta por files de comando MatLab (m-files) que se reconhecem pela extensão "✳.m" e meta-files de comandos do Gnuplot com extensão "✳.gnu" (Windows) ou "✳.gp" (Linux).
- As files de leitura para importação de valores têm a extensão "✳.dat" enquanto que as files de saída de resultados a extensão "✳.csv"[32].
- Os comandos dados dentro do Octave surgem sob a forma [o > comando]
- Comandos dentro do Gnuplot surgem [g > comando].

[30] Programa livre que funciona de interpretador de comandos do tipo MatLab.

[31] Programa livre de representação gráfica de dados.

[32] Formato compatível para a importação de dados ASCII em folhas de cálculo.

Preparação

Esta ToolBox foi desenvolvida em linguagem MatLab (Windows), compatível com Octave (Linux)[33].

Para colocar esta ferramenta em funcionamento é necessário proceder à *a)* instalação do interpretador MatLab, *b)* criação de directórios, *c)* cópia da informação, *d)* activação local da informação e *e)* execução da ferramenta e *f)* representação dos resultados.

i) Instalação do Octave

O primeiro passo consiste na instalação do interpretador de files de comando MatLab.

➢ Para os utilizadores da plataforma **Ubuntu** (Linux)

Através do gestor de pacotes de software (Synaptic) deve-se solicitar a instalação da mais recente versão do Octave disponível na sua distribuição tendo o cuidado de solicitar também a instalação do respectivo pacote estatístico (octave-statistics).

No caso de a distribuição Linux não apresentar este pacote deve procurar localizar a versão compatível em http://www.gnu.org/software/octave/download.html

Deve ainda instalar o Gnuplot para conseguir gerar os respectivos gráficos criados durante a execução.

➢ Para os utilizadores do **Windows** (Windows)

O primeiro passo consiste em fazer o download do respectivo binário em

http://www.gnu.org/software/octave/download.html

De seguida deve-se executar este binário tendo o especial cuidado de seleccionar a opção "Octave Forge" para garantir que os pacotes disponíveis são instalados e disponibilizados.

O pacote instalado já contém, por defeito, o Gnuplot associado não sendo necessário instalar separadamente.

[33] Sendo o MatLab um programa comercial e o Octave um programa de distribuição livre, optou-se por desenvolver a opção Octave.

ii) Criação de directórios

Após instalação do Octave deve-se criar um directório "bin" na unidade principal de trabalho[34]. Dentro deste directório deve-se criar um sub-directório "chemtool" sendo a localização final da ToolBox o percurso "c:\bin\chemtool" (Windows) ou "~/bin/chemtool" (Linux).

Deste modo a informação da ToolBox ficará preservada de eventuais perdas devidas a apagamento ou sobre-escrita involuntária de ficheiros.

iii) Cópia da informação

De seguida deve-se colocar todas as files desta ToolBox neste directório de modo a manter as files preservadas de eventuais perdas involuntárias de informação.

NOTA:

Este directório apenas deve conter as files originais e não deve ser utilizado no processamento de dados.

iv) Activação local da informação

Esta ferramenta trabalha com o Octave e Gnuplot em simultâneo e por isso deve-se activar duas consolas destinadas a estes dois programas.

A título de exemplo vamos assumir que as files que se pretende tratar se encontram em "c:\tmp\work" (Windows) / "~/tmp/work" (Linux) na área de trabalho do "/home/username".

- Na janela destinada ao Octave:

a) inicializar o programa;

b) redireccionar a sua posição para o directório de trabalho onde constam as files de dados, por exemplo através do comando [o > cd c:\tmp\work] (Windows) / [o > cd ~/tmp/work] (Linux);

c) activação da localização da toolbox através do comando [o > addpath 'c:\bin\chemtool'] (Windows) / [o > addpath '~/bin/chemtool'] (Linux).

- Na janela destinada ao Gnuplot :

a) mudança para a directoria local de trabalho;

b) activação do Gnuplot;

c) activação da localização da toolbox através do comando [g > set loadpath 'c:\bin\chemtool'] (Windows) / [g > set loadpath '/home/username/bin/chemtool'] (Linux).

[34] No Linux este directório corresponderá a algo do tipo "//home/username" enquanto que no Windows à unidade "c:\".

v) Execução da ChemTool

Ainda dentro da mesma sessão de Octave, iniciada em d), o comando seguinte deverá ser [o > chemom] para invocar a ferramenta desenvolvida.

Este menu principal permite ao utilizador identificar as ferramentas disponíveis nesta toolbox.

Durante a sua execução surgirão menus secundários com as opções específicas de cada ferramenta seleccionada.

vi) Representação gráfica

A representação gráfica bem como a obtenção de imagens JPEG é realizada de modo autónomo pelo Gnuplot.

De um modo genérico, a execução desta ToolBox gera files de resultados (✳.csv) que podem ser representadas pelo Gnuplot, de forma semi-automática, através das instruções contidas na respectiva file de comandos "✳.gnu" (Windows) / "✳.gp" (Linux) invocada[35].

Para tal, basta manter activa a janela Gnuplot em que foi previamente activada a directoria da chemtool.

Convenções

Na concepção desta ferramenta procurou-se que a entrada dos dados fosse o mais simples possível - os dados são lidos "em bloco" do ficheiro chamado "x.dat" que deve conter a informação organizada em "objectos" (linhas) e "variáveis" (colunas).

No caso de ser pertinente a identificação de objectos o programa procura localizar uma file com o nome "labels.dat" lendo o mesmo número de rótulos (labels) que o número de objectos carregados. Se não localizar esta file, cria os rótulos de uma forma numérica simples.

Se for necessário especificar as variáveis em uso, o programa procura ler os respectivos rótulos da file "labels_var.dat". Se não conseguir localizar este ficheiro cria os próprios rótulos de uma forma numérica simples, sendo o número de rótulos igual ao número de variáveis.

[35] Por exemplo para construir o dendrograma deve invocar o comando Gnuplot [g > load 'dendro.gnu'].

Restrições

As files de dados carregadas pelo Octave através da instrução “load” para “leitura em bloco” não suportam omissão de valores nem dimensões inadequadas de matrizes por exemplo, entre dados e labels.

Para tal, a file de dados “x.dat” deve ser sempre uma matriz rectangular, sem omissões no seu interior, e as dimensões de linhas e colunas devem ser compatíveis com “labels.dat” e “labels_var.dat” respectivamente.

Ferramentas

Na toolbox desenvolvida existem vários métodos de análise multivariada com finalidades diferentes entre si. Assim, é importante conhecer previamente as suas potencialidades bem como ter uma noção prévia do que se pretende averiguar e concluir a respeito dos dados a avaliar.

Neste capítulo são apresentadas as ferramentas disponibilizadas na ChemTool. Estas estão sistematizadas de acordo com a forma de surgimento no respectivo menu:

ANOVA
PCA
PLS
Clustering
LDA
Regression Analysis
Time series

1. ANOVA

A Análise de Variância (ANOVA[36]) foi implementada em três versões: de uma via ou de factor único, de duas vias ou de dois factores sem réplicas e com réplicas[37].

1.1 Factor único

A ANOVA de factor único estuda o efeito de um factor (aqui designado A) sobre a variabilidade do sistema em análise.

[36] Do Inglês, *Analysis Of Variance*.

[37] Provavelmente no futuro existirá a possibilidade MANOVA (Multi-ANOVA).

File de dados

A matriz de dados deve ser rectangular, sem omissões de valores, e estar organizada com as M réplicas em linhas e N níveis de factor por colunas, sendo o número total de valores dado por NxM. Esta matriz deve constar na file "x.dat".

Fundamentação

Segundo esta abordagem estatística, a soma de quadrados total (SS_T) pode ser decomposta nas componentes puramente aleatória (SS_{pe}) e na componente devida ao factor (SS_A),

$$SS_T = SS_{pe} + SS_A \tag{1.1}$$

Dividindo os termos desta equação pelos respectivos graus de liberdade, (N.M-1), (N(M-1)) e (N-1), obtêm-se as variâncias total, puramente aleatória e devida ao factor, eq.(2.2).

$$\sigma_T^2 = \sigma_{pe}^2 + \sigma_A^2 \tag{1.2}$$

O teste ANOVA pretende verificar se o factor em causa (A) é responsável por introdução de variabilidade nos dados além da contribuição puramente aleatória.

No caso de o factor não possuir efeito, a contribuição do factor em causa aproxima-se da componente puramente aleatória; caso contrário o factor manifesta-se através da sua contribuição específica ($\square_A^2$), sendo esta superior à contribuição aleatória.

As hipóteses de trabalho são:

$$\begin{aligned} H0: &\quad \sigma_A^2 \leq \sigma_{pe}^2 \\ H1: &\quad \sigma_A^2 > \sigma_{pe}^2 \end{aligned} \tag{1.3}$$

A hipótese nula (H0) vai no sentido de que o factor não apresenta qualquer efeito sobre a variabilidade observada enquanto que a hipótese alternativa (H1) sugere que o factor apresenta efeito sobre a variabilidade observada nos dados.

O **valor de prova** (α = p[H0]) traduz a probabilidade de aceitação da hipótese nula e fornece uma importante indicação sobre o efeito do factor[38].

NOTA: Ao ser executada esta opção é criada a file "anova1w_res.csv" que preserva os valores calculados através de uma tabela ANOVA de uma via.

[38] Valores de prova superiores a 0.05 indicam franca aceitação de H0 enquanto que valores inferiores a 0.01 sugerem a sua rejeição. Valores intermédios revelam que a aceitação de H0 é dúbia podendo ser utilizado o valor de referência de 0.03 como termo de desempate – superior a 0.03 aceitação dúbia, inferior a 0.03 rejeição dúbia.

1.2 Factor duplo sem réplicas

Esta abordagem destina-se a verificar o efeito simultâneo de dois factores.
A matriz de dados é ainda uma matriz rectangular contendo N linhas e M colunas, sem omissões, que é carregada através da file de dados "x.dat".

Neste caso em concreto, a ANOVA permite a decomposição da variabilidade total (T) em três componentes: a puramente aleatória (pe), a devida ao factor linha (factor A) e a devida ao factor coluna (factor B) de acordo com a equação,

$$SS_{\mathrm{T}} = SS_{\mathrm{pe}} + SS_{\mathrm{A}} + SS_{\mathrm{B}} \tag{1.4}$$

Dividindo os termos desta equação pelos respectivos graus de liberdade (N.M-1), N(M-1), (N-1) e (M-1) obtêm-se a equação,

$$\sigma^2_{\mathrm{T}} = \sigma^2_{\mathrm{pe}} + \sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}} \tag{1.5}$$

permitindo a individualização de cada uma das contribuições da variabilidade. Deste modo, podem ser testados individualmente os efeitos dos factores A e B, designados de "F", através das hipóteses:

$$\begin{aligned} \mathrm{H0}: &\quad \sigma^2_{\mathrm{F}} \le \sigma^2_{\mathrm{pe}} \\ \mathrm{H1}: &\quad \sigma^2_{\mathrm{F}} > \sigma^2_{\mathrm{pe}} \end{aligned} \tag{1.6}$$

A hipótese nula (H0) vai no sentido de que o factor F não apresenta qualquer efeito sobre a variabilidade observada enquanto que a hipótese alternativa (H1) sugere que o factor apresenta efeito sobre a variabilidade observada nos dados.

O **valor de prova** (α= p[H0]) traduz a probabilidade de aceitação da hipótese nula e fornece uma importante indicação sobre o efeito do factor.

NOTA: Ao ser executada esta opção é criada a file "anova2w_res.csv" que preserva os valores calculados através de uma tabela ANOVA de duas vias.

1.3 Factor duplo com réplicas

Neste caso a matriz de dados contém o factor A nas linhas e o factor B nas colunas sendo que cada conjunto de Q linhas reflecte o número de réplicas.

A matriz de dados é igualmente uma matriz rectangular contendo NxQ linhas e M colunas, sem omissões, que é carregada pelo programa através da file de dados "x.dat". Das NxQ linhas, apenas N referem-se aos níveis distintos do factor A (linhas).

Neste caso a ANOVA permite a decomposição da variabilidade total (T) em quatro componentes: a puramente aleatória (pe), a devida ao factor linha (factor A), a devida ao factor coluna (factor B) e ao termo de interacção entre factores (AB),

$$SS_{\mathrm{T}} \quad = \quad SS_{\mathrm{pe}} + SS_{\mathrm{A}} + SS_{\mathrm{B}} + SS_{\mathrm{AB}} \tag{1.7}$$

Dividindo os termos desta equação pelos respectivos graus de liberdade (N.M.Q-1), N.M.(Q-1), (N-1), (M-1) e (N-1)(M-1) obtêm-se a equação,

$$\sigma^2_{\mathrm{T}} \quad = \quad \sigma^2_{\mathrm{pe}} + \sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}} + \sigma^2_{\mathrm{AB}} \tag{1.8}$$

permitindo a individualização de cada uma das contribuições da variabilidade.

De igual modo podem ser testados individualmente os efeitos dos factores A e B e respectiva interacção AB através das hipóteses:

$$\begin{aligned} \mathrm{H0}: &\quad \sigma^2_{\mathrm{F}} \leq \sigma^2_{\mathrm{pe}} \\ \mathrm{H1}: &\quad \sigma^2_{\mathrm{F}} > \sigma^2_{\mathrm{pe}} \end{aligned} \tag{1.9}$$

A hipótese nula (H0) vai no sentido de que o factor F não apresenta qualquer efeito sobre a variabilidade observada enquanto que a hipótese alternativa (H1) sugere que o factor apresenta efeito sobre a variabilidade observada nos dados.

O **valor de prova** (α = p[H0]) traduz a probabilidade de aceitação da hipótese nula e fornece uma importante indicação sobre o efeito do factor.

NOTA: Ao ser executada esta opção é criada a file "anova2wRep_res.csv" que preserva os valores calculados através de uma tabela ANOVA de duas vias com réplicas.

2. Análise de Componentes Principais

A Análise de Componentes Principais [PCA[39]] aplica-se essencialmente a exploração de sistemas multivariados no sentido de estudar a relevância das variáveis utilizadas e auxiliar na representação dos objectos. A redução da dimensionalidade do sistema multivariado conseguida através da análise PCA permite:

1. Obter informação da interdependência das variáveis

2. Verificar a importância relativa das variáveis na explicação do sistema em análise

3. Detectar variáveis redundantes (supérfluas)

4. Obter uma representação dos objectos no sub-espaço dos factores.

[39] Do Inglês, Principal Component Analysis.

2.1 Fundamentação

A análise de componentes principais é recomendada na análise multivariada já que permite tratar grandes conjuntos de dados[40] sem exigir quaisquer pressupostos complicados.

O conjunto de variáveis que definem um determinado sistema multivariado pode ser classificado em variáveis "relevantes", "redundantes" e "mudas".

As variáveis "relevantes" constituem o sub-conjunto que preserva a informação pertinente do sistema permitindo a discriminação dos objectos.

Já as variáveis "redundantes", pouca ou nenhuma informação acrescentam no sentido de descrever o sistema e distinguir objectos – regra geral estas variáveis estão correlacionadas com as variáveis relevantes.

As variáveis "mudas" ou "pseudo-variáveis" são de facto "falsas" variáveis que não apresentam variabilidade significativa, podendo inclusivamente, em certos casos, apresentarem valor constante.

Com base nesta constatação, o objectivo principal da análise PCA consiste na redução da dimensionalidade do problema – as M variáveis originais são substituídas por um outro sub-conjunto de p ($p < M$) variáveis não correlacionadas, chamadas de componentes principais, de forma a preservar o máximo da informação original.

2.2 O Algoritmo

A análise PCA convencional estuda a relevância das M variáveis na definição/descrição de um sistema multivariado contendo N objectos representados nesse espaço de dimensão M, X(NxM).

A matriz de dados X(NxM) é composta por N objectos (linhas) observados segundo M variáveis (colunas).

As componentes principais[41] constituem um novo sistema de eixos ortonormados[42] em que cada vector (u_i) está representado em função das M variáveis originais,

$$u_i = q_{i1}.X_1 + q_{i2}.X_2 + ... + q_{iM}.X_M \qquad (2.1)$$

Os coeficientes q_{ij} designam-se de LOADS[43] e reflectem o impacto/contribuição da variável X_j sobre a componente principal i.

Por sua vez, os objectos podem ser representados no sub-espaço das componentes principais através do respectivo gráfico de SCORES[44].

[40] A análise multivariada define-se como sendo a análise de matrizes de grande dimensão – possuem elevado número de objectos (linhas) e de variáveis (colunas).

[41] Vectores próprios da matriz de covariância/correlação das variáveis.

[42] Sistema de vectores ortogonais (linearmente independentes) e com norma unitária.

[43] Do Inglês, LOADS significam cargas ou impacto.

[44] Do Inglês, SCORES significam as coordenadas dos objectos no sub-espaço das Componentes Principais.

Este processamento pode ser visto como a sucessão de 6 etapas.

a) Pré-acondicionamento dos dados

A matriz de dados não pode conter valores omissos. Variáveis "falsas" devem ser evitadas já que estas, por assumirem sempre valor constante, não contribuem para discriminar objectos e apresentam dispersão nula.

Cada variável, representada pela coluna j ($j \in \{1,...,N\}$), possui como estimativas paramétricas a média ($\bar{x}_j$) e o respectivo desvio padrão (s_j).

Dado que se procura maximizar a variância[45] das variáveis originais sobre as componentes principais os dados têm que ser no mínimo centrados. A centragem corresponde a subtrair a cada valor (x_{ij}) a respectiva média de coluna ($\bar{x}_j$),

$$e_{ij} = \left(x_{ij} - \bar{x}_j\right) \tag{2.2}$$

No caso de as variáveis possuírem ordens de grandeza muito distintas[46], para que a análise PCA não seja afectada por este fenómeno, é preferível realizar a normalização, dividindo a variável centrada pela respectiva dispersão,

$$z_{ij} = \frac{e_{ij}}{s_j} = \frac{\left(x_{ij} - \bar{x}_j\right)}{s_j} \tag{2.3}$$

Saliente-se ainda que a variável normalizada é adimensional o que apresenta ainda a vantagem de poder ser encarada como genérica.

NOTA: As "pseudo-variáveis" (valores constantes) não podem ser normalizadas já que a sua dispersão é nula.

Pode ainda optar-se por não ser tão drástico entre estas duas situações – pretende-se corrigir de algum modo as disparidades de escala mas salvaguardar ainda alguma desta informação. Nestes casos utiliza-se a transformação de Paretto,

$$r_{ij} = \frac{e_{ij}}{\sqrt{s_j}} = \frac{\left(x_{ij} - \bar{x}_j\right)}{\sqrt{s_j}} \tag{2.4}$$

Estas são as transformações mais correntes para o pré-tratamento das variáveis.

[45] A variância corresponde a uma estimativa paramétrica do momento de segunda ordem de uma distribuição. Este estimador paramétrico depende por sua vez da posição (estimativa central) da distribuição de modo que, para que esta estimativa seja independente da posição, os valores originais têm que ser "centrados" - subtraídos do seu valor central.

[46] Para se ter uma noção deste fenómeno, pode-se admitir um triângulo rectângulo com lados de comprimento 1 e 10; a respectiva hipotenusa terá a dimensão de 10.05, condicionada essencialmente pelo maior cateto.

b) Obtenção da matriz de covariância

A variabilidade do sistema em estudo pode ser estimada através da matriz de covariância ($A_{(M\times M)}$) que pode ser facilmente obtida por produto matricial, utilizando as variáveis pré-tratadas (Xc),

$$A_{(\mathrm{M}\times\mathrm{M})} = \left(Xc^{T}_{(\mathrm{M}\times\mathrm{N})}.Xc_{(\mathrm{N}\times\mathrm{M})}\right)/\left(N-1\right) \quad (2.5)$$

De acordo com o pré-tratamento utilizado, esta matriz calculada assumirá significados diferentes.

Se for realizada a centragem, esta corresponde a uma matriz de covariância paramétrica onde os valores da diagonal principal correspondem à variância e os restantes valores às covariâncias; se for realizada a normalização esta matriz corresponde a uma matriz de correlação – os elementos da diagonal principal são unitários e os restantes valores limitados ao domínio [-1,1].

c) Maximização da variabilidade

O algoritmo PCA está intimamente relacionado com a decomposição da matriz de covariância em valores próprios[47]. A matriz de covariância ($A_{(\mathrm{M}\times\mathrm{M})}$) é um exemplo de matriz quadrada e simétrica com dimensão M e a sua decomposição de valores singulares conduz à determinação de M valores próprios (λ_i) que anulam o seguinte determinante

$$\left|A_{(\mathrm{M}\times\mathrm{M})} - I_{(\mathrm{M}\times\mathrm{M})}.\lambda_i\right| = 0 \quad (2.6)$$

Cada valor próprio (λ_i) tem associado um respectivo vector próprio ($u_{i(\mathrm{M}\times 1)}$), designado de Componente Principal (CP). Esse valor próprio expressa a variabilidade descrita pela componente na representação do sistema em causa.

Os vectores próprios caracterizam-se por constituírem uma nova base ortonormada de eixos de forma que,

$$\left(A_{(\mathrm{M}\times\mathrm{M})} - I_{(\mathrm{M}\times\mathrm{M})}.\lambda_i\right).u_{i(\mathrm{M}\times 1)} = 0 \quad (2.7)$$

Esta decomposição em valores singulares permite ainda a representação dos objectos sobre as componentes principais,

$$Xc_{(\mathrm{N}\times\mathrm{M})} = S_{(\mathrm{N}\times p)}.\Lambda_{(p\times p)}.Q^{T}_{(p\times\mathrm{M})} \quad (2.11)$$

[47] Do Inglês, *Eigenvalue decomposition*.

onde $Xc_{(\mathrm{N}\times\mathrm{M})}$ é a matriz de dados (pré-tratada), $S_{(\mathrm{N}\times p)}$ a matriz dos SCORES[48], $\Lambda_{(p\times p)}$ a matriz diagonal contendo os valores próprios e $Q_{(\mathrm{M}\times p)}$ a matriz dos respectivos LOADS[49].

d) Escolha do número de componentes principais

Sendo A uma matriz de covariância, os seus valores próprios expressam a variabilidade descrita por cada componente principal na representação do sistema em causa e podem, deste modo servir de critério para decisão na estimativa do número de componentes principais necessárias para descrever o sistema.

As componentes principais são por isso ordenadas por ordem decrescente de representatividade.

Trata-se agora de decidir qual o número mínimo (mais conveniente) que permite reter o máximo de informação do sistema de modo a permitir a sua racionalização mais fácil.

Existem essencialmente três critérios para auxiliar esta escolha: Pearson (regra dos 80%), Kaiser (significância) e representação dos valores próprios (Scree plot).

Critério de Pearson (ou regra dos 80%) – o número de CP (***p***) é escolhido até se atingir a descrição de pelo menos 80% da informação total (variabilidade total)

$$\frac{\sum \lambda_i}{\sum \lambda_i} \geq 0.80 \qquad (2.8)$$

Critério de Kaiser – o número de componentes é definido pela quantidade de valores próprios que são significativos (acima do valor médio esperado)[50].

Scree Plot – este gráfico representa, de uma forma ordenada, os valores próprios em função do número de componentes, da mais significativa para a menos significativa. As componentes relevantes são aquelas que se destacam acima da linha de tendência basal que converge para zero.

NOTA: O número de componentes principais (CP) necessárias para descrever a informação contida nos dados dá uma preciosa indicação sobre variáveis redundantes[51].

[48] Esta matriz corresponde à representação dos N objectos no sub-espaço das componentes principais.

[49] A matriz dos LOADS traduz a relação entre as M variáveis originais e as p componentes principais.

[50] No caso de ter sido feita a normalização das variáveis, a matriz de covariância é de facto uma matriz de correlação cujo traço corresponde à dimensão da matriz; neste caso os valores próprios significantes são superiores ou iguais à unidade.

[51] Num espaço de dimensão M e p componentes principais existem (M-p) variáveis redundantes.

e) Estudo dos *LOADS*

A matriz dos *LOADS* ($Q_{(\mathrm{M}\times p)}$) contém os *p* vectores próprios representados em colunas e expressos em função das M variáveis do sistema em análise. Assim, cada componente principal pode ser vista como a combinação linear das variáveis do sistema,

$$u_j = q_{j1}.v_1 + q_{j2}.v_2 + ... + q_{jM}.v_M \tag{2.9}$$

As componentes de cada vector, LOADS (qij), podem servir para verificar o impacto de cada variável (vi) sobre cada uma das componentes principais.

Uma vez que os vectores próprios possuem norma unitária, a contribuição de uma determinada variável sobre cada uma das componentes principais é significativa se o módulo dessa componente exceder o valor médio,

$$q_{ij} \geq \bar{q} \qquad \frac{1}{\sqrt{\mathrm{M}}} \tag{2.10}$$

f) Representação dos *SCORES*

A matriz dos *SCORES* ($S_{(\mathrm{N}\times p)}$) permite a visualização dos N objectos no sub-espaço das p componentes principais facultando a) a observação de arranjos, b) a identificação de eventuais agregados e c) a observação de eventuais valores discrepantes (valores isolados).

2.3 Modo de funcionamento

A versão implementada permite ao utilizador escolher dos dados do ficheiro "x.dat" o número de objectos e de variáveis que pretende analisar.

Ao executar a análise PCA são criadas novas files relacionadas com as variáveis ("loandings.dat" e "loandings_varimax.dat") e com a representação dos objectos no sub espaço das componentes principais ("scores.dat" e "scores_varimax.dat") de forma a preservar os valores obtidos na execução do programa.

As files "loandings.gnu" e "scores.gnu" contêm toda a informação necessária para a representação das loadings e dos scores através do Gnuplot.

Para que ambos os programas funcionem, tem que existir no mesmo directório a file de dados "x.dat" e, caso existam as files de "labels" estas devem possuir dimensões compatíveis.

Ao executar cada um dos programas são realizadas várias operações tais como:

a) Leitura dos dados (x.dat);

b) Leitura dos rótulos dos objectos (labels.dat) e das variáveis (labels_var.dat), caso existam as files correspondentes;

c) Cálculo da matriz de covariância ('imat==1'), pareto ('imat==2') ou correlação ('imat==3');

d) Escolha do número de CP's aplicando dois critérios distintos: o critério de Kaiser[52], no caso de ser usada a matriz de correlação; e o critério de Pearson[53], no caso de serem usadas a matriz de covariância ou pareto.

e) Aplicação da rotação varimax[54];

f) Relacionar as variáveis após a rotação;

g) Criação das files para a representação (com o Gnuplot) das loadings (antes e após a rotação) e dos scores a 2D.

Algumas notas sobre a execução do algoritmo PCA:

- **Escolha da matriz a processar –** a análise da matriz de Covariância é válida para variáveis com a mesma ordem de grandeza; contrariamente, a matriz de Correlação é mais adequada no estudo de sistemas com escalas muito distintas. A abordagem Paretto corresponde a uma situação intermédia destas. Esta escolha afecta o critério utilizado na escolha do número CP's.
- **Representação dos *LOADS* –** gráfico bidimensional em que cada ponto representa uma variável; o plano é definido pelos dois eixos correspondentes ao primeiro par de componentes principais.

O gráfico das *loadings* é obtido no Gnuplot através da instrução [g > load 'loadings.gnu'] (Windows) / [g > load 'loadings.gp'] (Linux). Este gráfico é preservado em formato JPEG com o nome "pca.jpg".

- **Representação dos *SCORES* –** gráfico bidimensional em que cada ponto representa um objecto sobre o primeiro par de componentes principais. O gráfico dos *scores* é obtido no Gnuplot através da instrução [g > load 'scores.gnu'] (Windows) / [g > load 'scores.gp'] (Linux). O gráfico é preservado em formato JPEG com o nome "scores.jpg".

[52] LOAD significativo se for superior ao valor médio.

[53] Recuperação de 80% da variabilidade total.

54 A Varimax é um método de maximização que consiste numa rotação ortogonal de eixos, pretendendo que, para cada componente principal, sejam maximizadas as contribuições dos valores mais significativos em detrimento dos menos expressivos, preservando a ortogonalidade da base vectorial inicial.

3. Análise de Agregados

A análise de agregados permite ao utilizador, de uma forma automática, procurar reconhecer padrões e definir grupos de objectos atendendo à proximidade relativa dos valores que os caracterizam.

Existem diversas versões de análises de agrupamentos sendo a por nós preferida a análise não-supervisionada em modo Hierárquico já que o modo de associação é independente de critérios impostos.

3.1 Fundamentação

A análise de agrupamentos tem por finalidade reunir, por algum critério de classificação os objectos em grupos, de tal forma que, exista alguma homogeneidade dentro do grupo e heterogeneidade entre grupos.

O processo de agrupamento envolve basicamente duas etapas: a primeira refere-se à estimativa de uma medida de similaridade[55] (ou dissimilaridade[56]) entre os objectos e a segunda, refere-se à adopção de um procedimento de formação de grupos (algoritmo).

De um modo geral, as medidas de similaridade e de dissimilaridade são facilmente, transformáveis entre si. Contudo, existe um grande número de algoritmos disponíveis, dos quais o utilizador tem de decidir qual o mais adequado ao seu propósito já que procedimentos diferentes pode levar a diferentes soluções.

Esta opção exige a tomada de uma série de decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição e que podem representar diferentes agrupamentos.

[55] A similaridade exprime-se por um valor numérico que é tanto maior quanto maior for a semelhança entre os objectos comparados.

[56] No caso de uma medida de dissimilaridade, quanto menor for esse valor, tanto menor será a diferença entre os objectos revelando a sua proximidade.

Genericamente, a análise de agrupamento envolve as etapas seguintes:

1. Selecção dos objectos a agrupar;
2. Definição de um conjunto de variáveis a partir das quais serão obtidas informações necessárias ao agrupamento dos objectos;
3. Definição de uma medida de semelhança ou distância entre os objectos;
4. Escolha de um algoritmo de agrupamento;
5. Validação dos resultados obtidos.

Chama-se dendrograma à representação bidimensional do esquema de associação sucessiva dos objectos, atendendo à sua similaridade, até culminar na fusão de todos os grupos num único grupo final.

Estes gráficos são especialmente úteis na visualização de semelhanças entre objectos representados por pontos no espaço com dimensão maior do que três, onde a representação de gráficos convencionais não é possível.

3.2 Etapas do processo

São várias as etapas que devem ser realizadas no processo de agrupamento: a) selecção da medida de dissimilaridade, b) escolha do algoritmo de agrupamento, c) avaliação dos resultados e d) interpretação dos grupos identificados.

As variáveis que caracterizam os objectos podem assumir tipos diferentes: contínuo, discreto ou binário. Variáveis binárias assumem exactamente dois valores, 0 e 1, indicando a presença ou ausência de uma determinada característica; variáveis discretas possuem um conjunto finito e pequeno de valores possíveis; variáveis contínuas podem assumir qualquer valor real.

3.2.1 Medidas de dissimilaridade

Para agrupar objectos, é necessário definir uma medida de dissimilaridade. Com base nessa medida, os objectos mais similares são agrupados e os outros são colocados em grupos separados. Este procedimento depende do tipo de variável em causa.

Dados Binários

As variáveis binárias caracterizam-se por apenas assumirem dois valores 0 ou 1 conforme uma determinada característica está ausente (0) ou presente (1)
Neste caso é possível definir a seguinte tabela de contingência, Tabela 3.1.

Tabela 3.1 – Resumo das informações.

		Objecto i		
		1	0	Soma
Objecto j	1	a	b	a+b
	0	c	d	c+d
	Soma	a+c	b+d	M

Toda a informação comparativa de quaisquer 2 objectos (i, j) é dada através de uma matriz quadrada (2x2):

- a – concordância “11” (presença simultânea da característica) - número de variáveis com valor “1” em i e j;
- b – discordância “10” - número de variáveis com valor “1” para i e “0” para j;
- c – discordância “01” - número de variáveis com valor “0” para i e “1” para j;
- d – concordância “00” (ausência simultânea da característica) - número de variáveis com valor “0” em i e j;

O número total de variáveis M corresponde ao somatório destes elementos (M = a+b+c+d).
Com esta aproximação podem ser tiradas algumas conclusões, por exemplo, o coeficiente de concordância simples, dado por (a+d)/M, calcula a proporção de variáveis em que os objectos comparados têm o mesmo valor[57].
Existe um grande número de medidas de dissimilaridade para dados binários.
Nesta Toolbox estão disponíveis e sistematizadas 24 medidas diferentes, com a ordem apresentada no respectivo menu: 'Pattern difference', 'Euclidean', 'SEuclidean', 'Variance', 'Simple matching', 'Manhattan', 'Dice', 'Antidice', 'Lance and Williams', 'Nei & Lei´s', 'Yule coefficient', 'Cosine', 'Sneath', 'Forbes', 'Hamman', 'Jaccard', 'Rogers', 'Ochiai', 'Anderberg', 'Kulczynski', 'Pearson', 'Gower2', 'Russel-Rao', e 'Simpson'.
Quando a medida utiliza a ou d é uma medida de similaridade, se utiliza apenas b ou c, é uma medida de dissimilaridade.
Na tabela seguinte descrevem-se as medidas de similaridade e de distância disponíveis:
No programa estão implementadas 24 medidas de distância (dissimilaridade) entre elementos da matriz de dados, ver tabela (3.2), das quais as mais importantes são a distância Euclidiana, e a distância de Manhattan.
A distância de Minkowski é apenas utilizada no caso de variáveis contínuas.

[57] Neste caso, são atribuídos pesos iguais tanto à presença (par 1-1) quanto à ausência (par 0-0) de um determinado atributo - ambos os casos são considerados concordâncias.

Tabela 3.2 – Medidas de similaridade e de distância disponíveis para dados binários.

Nome	Equação	Variação
Pattern Difference	$2bc/(a+b+c+d)$	[0, 1]
Euclidean	$\sqrt{b+c}$	[0, ∞[
SEuclidean	$b+c$	[0, ∞[
Variance	$(b+c)/4(a+b+c+d)$	[0, 1]
Simple Matching	$(a+d)/(a+b+c+d)$	[0, 1]
Manhattan	$(b+c)/(a+b+c+d)$	[-1, 1]
Dice[58]	$2a/(2a+b+c)$	[0, 1]
AntiDice[59]	$a/(a+2(b+c))$	[0, 1]
Lance and Williams	$(b+c)/(2a+b+c)$	[0, 1]
Neil & Leil´s	$2a/((a+b)+(a+c))$	[0 -]
Yule coefficient[60]	$(ad-bc)/(ad+bc)$	[-1, 1]
Cosine	$a/\sqrt{(b+a).(c+a)}$	[0, 1]
Sneath	$2(a+d)/(2(a+d)+(b+c))$	[0, 1]
Forbes	$a.(a+b+c+d)/((b+a).(c+a))$	[0, ∞]
Hamman	$((a+d)-(b+c))/(a+b+c+d)$	[-1, 1]
Jaccard	$a/(a+b+c)$	[0, 1]

[58] Também conhecido como medida de Czekanowski ou de Sorensen. Neste coeficiente são excluídas as ausências (correspondente aos valores d) e as coincidências (valores de a) têm peso duplo.
[59] Este coeficiente, ao contrário do anterior, atribui peso duplo às discordâncias (b e c).
[60] A formula para este coeficiente é indefinida quando um ou ambos os vectores ou são zeros ou tudo uns. O programa atribui o valor 1 quando b+c=0 - existe concordância completa. Quando a+d=0, o programa assume a medida como -1 - existe discordância completa. Por outro lado, se ad-bc=0, é atribuído valor 0. Estas regras, aplicadas antes de usar a fórmula, evitam os casos onde a mesma conduziria a um resultado indefinido.

Rogers	$(a+d)/((a+d)+2(b+c))$	[0, 1]
Ochiai	$a/\sqrt{(a+b)(a+c)}$	[0, 1]
Anderberg[61]	$(a/(a+b)+a/(a+c)+d(c+d)+d/(b+d))/4$	[0, 1]
Kulczynski[62]	$(a/(a+b)+a/(a+c))/2$	[0, 1]
Pearson[63]	$d(a-b)/\sqrt{(a+b)(a+c)(b+d)(c+d)}$	[-1, 1]
Gower2[64]	$ad/(a+b)(a+c)(b+d)(c+d)$	[0, 1]
Russel-Rao	$a/(a+b+c+d)$	[0, 1]
Simpson	$a/min((a+b),(a+c))$	[0, 1]

Dados Discretos

A ferramenta desenvolvida ainda não possui implementado o tratamento de variáveis discretas não binárias mas está previsto para versões posteriores.

Dados Contínuos

No caso de as variáveis serem contínuas, estão previstas as métricas de distância Euclidiana [Euclidean], Euclidiana quadrática [Seuclidean], Mahalanobis, Cityblock, Minkowski, Cosine, Correlation, Spearman, Hamming, Jaccard e Chebychev.

[61] Neste caso, se ambos os vectores são tudo uns (ou tudo zeros), a medida de similaridade é 1. Se as somas a+b, a+c, c+d, b+d forem iguais a 0, então a medida é 0.

[62] A fórmula é indefinida se um ou ambos os vectores forem tudo zeros. Se ambos os vectores forem tudo zeros, programa atribui valor 1 à medida de similaridade. Se apenas um dos vectores é tudo zeros, a medida de similaridade tem valor 0.

[63] Este coeficiente segue as mesmas restrições que o coeficiente de Yule.

[64] Neste caso, se ambos os vectores são tudo uns (ou tudo zeros) a medida de similaridade é 1. Se ad=0, a medida é 0.

3.2.2 Algoritmos

Os algoritmos de agrupamento disponíveis para determinar a distância entre grupos são: ligação simples *(Single Linkage)*, ligação completa *(Complete Linkage)*, médias das distâncias *(Average Linkage),* centróide, mediana, e a soma de erros quadráticos ou variância *(Ward)*.

Método de ligação simples (Single Linkage)

A distância entre os grupos é definida como sendo aquela entre os objectos mais próximos (sendo cada objecto pertencente a cada um desses grupos).

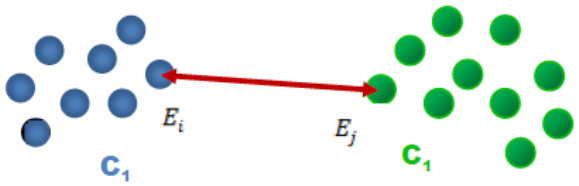


Figura 3.1 – Distância entre grupos obtida através de "Single Linkage" (menor distância).

Este método de associação pode levar a algumas anomalias já que os dendrogramas resultantes são, geralmente, pouco informativos - a informação relativa aos objectos intermediários não é evidente.

Apresenta ainda as seguintes características:

- grupos muito próximos podem não ser identificados;

- pouca tolerância a *outliers*,

- tendência a formar longas cadeias[65].

Método da ligação completa **[Complete Linkage]**

Determina a distância entre dois grupos de acordo com a maior distância entre um par de objectos, sendo cada objecto pertencente a um grupo distinto. Geralmente, leva a grupos compactos e discretos, tendo os seus valores de dissimilaridade relativamente grande.

[65] Situação em que há um primeiro grupo de um ou mais objectos que passa a incorporar um grupo de apenas um objecto, formando uma longa cadeia, onde se torna difícil definir um nível de corte para classificar os objectos em grupos.

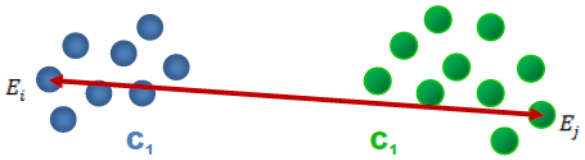


Figura 3.2 – Distância entre grupos através da associação completa (Complete Linkage).

Método da ligação média [Average Linkage]

No algoritmo *Average Linkage*, a distância entre dois grupos é definida como a média das distâncias entre todos os pares de objectos em cada grupo.

Figura 3.3 – Distância obtida através de Average Linkage (média das distância entre objectos).

Método do Centróide

No método do centróide cada grupo é considerado um simples ponto, representado pelo seu centro de massa (centróide). Este método, utiliza uma função de agrupamento para medir a distância entre os centros de massa dos dados. Caracteriza-se pela redefinição, a cada passo, da matriz de dados, em que cada grupo é representado pelo vector médio das p variáveis envolvidas.

Uma desvantagem desse método é que se os dois grupos forem muito diferentes, em termos de dimensão, o centróide do novo grupo estará, mais próximo daquele que for maior e, assim as características do grupo menor, tenderão a diluir-se.

Uma característica importante deste algoritmo é o facto da distância entre grupos ser determinada pela distância entre os pontos representativos dos respectivos centros de massa. Outras características são a sua robustez à presença de *outliers* e probabilidade de ocorrência de fenómenos de inversão.[66]

[66] Ocorre quando a distância entre centróides é menor que a distância entre grupos já formados, isto fará com que os novos grupos sejam formados num nível inferior aos grupos já existentes, tornando o dendrograma confuso.

Método da Mediana

Este algoritmo é um caso particular do método do centróide. A determinação da distância entre dois grupos através do cálculo do centro de massa, não considera o número de elementos em cada um dos grupos - o vector médio que representa o novo grupo pode, eventualmente, ficar situado entre os elementos do grupo com maior número de objectos. Este método pondera a medida de distância pelo número de elementos de cada grupo.

Características:

- resultado satisfatório quando os grupos possuem tamanhos diferentes;

- pode apresentar resultado diferente quando permutados os elementos na matriz de dissimilaridade;

- apresenta robustez à presença de *outliers*

- fenómeno da inversão.

Método de variância mínima [Ward]

Baseia-se na análise de variância tendendo a associar os objectos aos grupos de forma a minimizar a variância intra-grupo. Este algoritmo é altamente eficiente na formação de grupos. Inicialmente, admite que cada um dos objectos constitui um único grupo. Considerando a primeira reunião de objectos num novo grupo, a soma dos desvios dos pontos representativos dos seus elementos, em relação à média do grupo, é calculada, e dá uma indicação de homogeneidade do grupo formado. Os grupos formados possuem uma elevada homogeneidade interna. No entanto, pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual; tem tendência a combinar grupos com poucos elementos; é sensível à presença de *outliers*.

3.2.3 Validação dos resultados

Existem alguns procedimentos práticos para conferir, de maneira superficial, os resultados obtidos.

Nesta ToolBox encontra-se disponível um simulador de dados aleatórios correspondente à file "shuffle.m".
Apesar das tentativas de construção de vários testes para a confiabilidade estatística dos agrupamentos, nenhum procedimento totalmente comprovado está ainda disponível. A ausência de testes adequados provém da dificuldade de especificação de hipóteses nulas realistas.

3.3 Procedimento

A análise de agregados Hierárquico[67] foi implementada em três versões: '*cluan'*, 'cluan2' e 'cluan3'. As três versões do programa foram desenvolvidas para processarem dados contínuos (cluan), binários (cluan2) e discretos (cluan3).

Como já foi previamente esclarecido, para que a ferramenta funcione tem que ser activada no Octave o conjunto de m-files desta ferramenta posicionadas no sub-directório "chemtool".

Leitura de dados

Os dados da matriz X(NxM) (N objectos expressos sobre M variáveis), são lidos do ficheiro "x.dat".

Na identificação de objectos, o programa procura localizar a file "labels.dat" lendo o mesmo número de rótulos que o número de objectos carregados (N). Na ausência desta, o programa cria os rótulos de uma forma numérica simples.

Funcionamento

Para os dados contínuos estão disponíveis e sistematizadas as métricas seguintes: 'Euclidean', 'Seuclidean', 'Mahalanobis', 'Cityblock', 'Minkowski', 'Cosine', 'Correlation', 'Spearman', 'Hamming', 'Jaccard' e 'Chebychev'.[68]
Para os dados binários encontram-se disponíveis e sistematizadas 24 métricas distintas: 'Pattern difference', 'Euclidean', 'SEuclidean', 'Variance', 'Simple matching', 'Manhattan', 'Dice', 'Antidice', 'Lance and Williams', 'Nei & Leí´s', 'Yule coefficient', 'Cosine', 'Sneath', 'Forbes', 'Hamman', 'Jaccard', 'Rogers', 'Ochiai', 'Anderberg', 'Kulczynski', 'Pearson', 'Gower2', 'Russel-Rao', e 'Simpson'.[69]
Ao executar o programa são realizadas várias operações tais como:

a) Cálculo da matriz de distâncias;
b) Cálculo das ligações entre os objectos (connections);
c) Formação dos grupos;
d) Cálculo da posição dos objectos no dendrograma;
e) Reorganização dos objectos;
f) Construção do dendrograma com os objectos reorganizados[70].

[67] Do Inglês, *Hierarchical Clustering*.

[68] Por defeito o programa calcula a distância euclidiana, correspondente à instrução *Idist="euclidean*" e utiliza o algoritmo *Single Linkage,* correspondente à instrução *link="Single"*.

[69] Provavelmente no futuro estarão disponíveis nesta ToolBox novas opções relativas a este tipo de coeficientes.

[70] Na versão cluan2 o dendrograma é construído num único passo, dado que os objectos já se encontram reorganizados.

Resultados

Ao executar o programa são criadas as files "plotfile1.csv" e "plotfile2.csv". Na primeira encontram-se preservados os valores coordenadas xy usados na construção do dendrograma e na segunda a informação relativa a organização dos objectos no dendrograma, assim como as respectivas labels (rótulos).

O dendrograma é obtido no Gnuplot através da instrução [g > load 'dendro.gnu'] (Windows) / [g > load 'dendro.gp'] (Linux) e é automaticamente preservado em formato JPEG com o nome "dendro.jpg".

NOTA: Conforme o tipo de método utilizado deve-se alterar o título dentro da meta-file Gnuplot, alterando a linha da instrução "set title", ex:

set title "Dendrogram using Ward linkage\n\Euclidean"

4. Métodos dos Mínimos Quadrados Parciais

Designa-se de PLS[71] a análise multivariada de sistemas através do método de mínimos quadrados parciais. Este algoritmo explora um sistema através da maximização da correlação entre os predictores (sub-espaço das causas) e o sub-espaço das respostas. Inicialmente são analisados em cada sub-espaço, predictores e respostas, no sentido de encontrar combinações lineares de variáveis, designadas de factores latentes, que se encontram correlacionadas e permitem descrever ao máximo a variabilidade dos resultados de cada sub-grupo de valores.

4.1 Fundamentação

O conjunto de dados inicial consiste numa tabela de valores onde cada cada objecto é avaliado sob um conjunto de variáveis (*m*+*p*). Este conjunto de valores constitui dois sub-espaços: de predictores, $X_{(n\times m)}$

$$X_{(n\times m)} = \begin{Bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{Bmatrix} \tag{4.1}$$

e de respostas, $Y_{(n\times p)}$

[71] Do inglês, PLS significa Partial Least Squares.

$$Y_{(n\times p)} = \begin{Bmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nj} & \cdots & y_{np} \end{Bmatrix} \tag{4.2}$$

sendo *n* o número de objectos, *m* o número de variáveis do sub-espaço predictor e *p* o número de variáveis do sub-espaço resposta.

O sub-espaço dos predictores apresenta os objectos representados sob o espaço constituído pelas variáveis independentes que originam as respostas do sistema. Estes são os valores de entrada no sistema. O sub-espaço das respostas apresenta os mesmos objectos representados sobre as respectivas variáveis dependentes que resultam das condições impostas ao sistema em estudo. Estas variáveis dependem, através de uma função previamente desconhecida, dos estímulos fornecidos ao sistema.

Designa-se por factor latente a combinação linear de variáveis que consegue descrever uma fracção da variabilidade de cada sub-espaço de valores.

O algoritmo PLS é uma versão mais complexa do parente mais simples OLS[72]. O ajuste polinomial simples através de OLS baseia-se num modelo polinomial do tipo

$$\eta_i = a_0 + a_1.x_{1i} + a_2.x_{2i} + a_{12}.x_{1i}.x_{2i} + \ldots \tag{4.3}$$

onde a resposta prevista (η_i) depende linearmente de um termo constante (a_0) e das variáveis do sistema (x_{ij}).

O algoritmo PLS assume que o conjunto de respostas (η_{ij}) é descrito através de uma combinação linear de variáveis independentes (x_{ij}),

$$q_f.y_{ij} = \sum_f a_f.w_f.x_{ij} \tag{4.4}$$

onde w_f representa o factor latente do sub-espaço dos predictores ($X_{(n\times m)}$) que está relacionado com o factor latente do sub-espaço das respostas ($Y_{(n\times p)}$).

O algoritmo procura maximizar a correlação entre factores latentes dos sub-espaços predictor e resposta no sentido de procurar justificar o máximo da variabilidade da resposta obtida. Após encontrada a melhor relação linear, esse efeito é subtraído aos respectivos sub-espaços. O modelo, obtido por via implícita, vai sendo construído iterativamente permitindo obter uma descrição tanto mais fiável quanto se pretenda.

Após a remoção da resposta característica, cada sub-espaço deve conter ruído aleatório, valores independentes de qualquer interacção.

[72] Do inglês *Ordinary Least Squares*

4.1.1 Verificação inicial

Sendo este tipo de análise de factores efectuada através de operações matriciais, é necessário garantir previamente que não vão surgir singularidades de cálculo com valores mal acondicionados nem vão existir incompatibilidades nessas operações matriciais. Assim é necessário verificar, objecto a objecto, se todos os valores foram correctamente inseridos sobre as respectivas variáveis e se as linhas estão completas.

4.1.2 Pré-acondicionamento

Na modelação PLS assume-se que os modelos são lineares em relação aos factores latentes – apenas é assumido o primeiro grau. Assim, o termo constante (a_0) previsto para um modelo do tipo polinomial, eq.5.3, deve ser suprimido. Esta simplificação do modelo é conseguida através da centragem de todas as variáveis.

Por outro lado, uma segunda dificuldade experimental pode inviabilizar este tipo de tratamento. No caso de as variáveis apresentarem escalas muito diferentes, a sua influência sobre os parâmetros estimados do modelo seria nítida obscurecendo o efeito das restantes variáveis. Assim, é usual efectuar uma normalização prévia das variáveis

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \tag{4.5}$$

Este escalamento prévio torna possível verificar o efeito de todas as variáveis, independentemente da sua variabilidade relativa. Para evitar singularidades no auto-escalamento das variáveis é necessário garantir que cada variável não é constante[73].

[73] Uma variável não constante possui variância não nula.

4.1.3 O Algoritmo

Sendo necessário maximizar a correlação entre os sub-espaços X e Y, calculam-se as matrizes de covariância de $X_{(n\times m)}$ em $Y_{(n\times p)}$

$$(n-1).Cx_{(mxm)} = X^T_{(mxn)} Y_{(nxp)} Y^T_{(pxn)} X_{(nxm)} = \left| Y^T_{(pxn)} X_{(nxm)} \right|^2 \tag{4.6}$$

e de covariância de $Y_{(n\times p)}$ em $X_{(n\times m)}$

$$(m-1).Cy_{(pxp)} = Y^T_{(pxn)} X_{(nxm)} X^T_{(mxn)} Y_{(nxp)} = \left| X^T_{(mxn)} Y_{(nxp)} \right|^2 \tag{4.7}$$

Chama-se traço de uma matriz ao produto dos termos da diagonal principal. É necessário que o valor obtido em ambas as matrizes de covariância, *Cx* e *Cy*, seja igual. Este valor corresponde à variância global do sistema.

Em qualquer instante do processo iterativo, a soma de quadrados de cada sub-espaço pode ser estimada com base nas equações

$$SSx = \sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}^2 \qquad SSy = \sum_{i=1}^{p}\sum_{j=1}^{p} y_{ij}^2 \tag{4.8}$$

À medida que os factores são calculados e subtraídos à matriz de covariância estas tendem para matrizes nulas. Sendo as variáveis previamente centradas e escaladas, a soma de quadrados de *X* e de *Y* é dada pela soma dos termos de *X* e *Y* ao quadrado já que estas matrizes se encontram normalizadas.

O primeiro valor próprio de *Cx* ($\lambda x1$) é o valor próprio mais significativo, com maior valor

$$\left| Cx_{(mxm)} - \lambda x_1 . I_m \right| = 0 \tag{5.9}$$

assim como o primeiro valor próprio de *Cy* (λy_1)

$$\left| Cy_{(pxp)} - \lambda y_1 . I_p \right| = 0 \tag{4.10}$$

também será o maior valor em causa.

Atendendo às operações efectuadas em 3.41 e 3.42, estes valores próprios estão intimamente relacionados – a sua correlação foi maximizada através da covariância cruzada entre os sub-espaços predictor-resposta.

Esta decomposição matricial através dos seus valores próprios faz com que surja agora um novo sistema de vectores directores ortonormalizado definido através dos respectivos vectores próprios

$$\left\{ Cx_{(mxm)} - \lambda x_1 . I_m \right\} . w_{1(mx1)} = 0 \quad \wedge \quad \left| w_1 \right| = 1 \tag{4.11}$$

$$\left\{Cy_{(pxp)} - \lambda y_1 . I_p\right\} . q_{1(px1)} = 0 \quad \wedge \quad \left|q_1\right| = 1 \tag{4.12}$$

Para um determinado factor latente *f*, extraído por via decrescente de importância, vai existir uma relação estabelecida entre os sub-espaços *X* e *Y* que está explicitamente definida através dos respectivos pesos *(loadings)*. Assim, o vector próprio w_f do sub-espaço predictor e q_f do sub-espaço resposta

$$w_f\,(m \times 1) = \begin{Bmatrix} w_{f1} \\ \vdots \\ w_{fi} \\ \vdots \\ w_{fm} \end{Bmatrix} \qquad q_f\,(p \times 1) = \begin{Bmatrix} q_{f1} \\ \vdots \\ q_{fi} \\ \vdots \\ q_{fp} \end{Bmatrix} \tag{4.13}$$

são constituídos por componentes, w_{fj} e q_{fk}, que traduzem o impacto das variáveis originais sobre esse factor latente e, deste modo indicam quais as inter-relações entre variáveis de ambos os sub-espaços.

Se existirem *p* respostas, é previsível que o número de factores a considerar para descrever esse sistema seja próximo deste valor.

Este algoritmo visa essencialmente simplificar o problema multidimensional inicial. Assim, importa estabelecer critérios para recuperar uma fracção significativa da variabilidade inicial do sub-espaço resposta.

É portanto necessário estabelecer critérios de aceitação e de rejeição de factores latentes no sentido de estabelecer qual o número mínimo de factores latentes que devem ser considerados para descrever o sistema em causa.

Um dos critérios mais utilizados consiste em tentar reproduzir cerca de 80% da resposta original. Sendo este critério bastante significativo em termos de reprodução do sistema, é geralmente difícil obter índices tão elevados de desempenho com sistemas estocásticos onde as inter-relações das variáveis são obscurecidas através de causas desconhecidas.

Outro critério também utilizado consiste em representar o valor paramétrico obtido na modelação ou a variância incremental descrita por esse factor através de um gráfico de desempenho (*scree plot*). Os factores latentes significativos são reconhecidos por estarem anormalmente acima da contribuição basal, relacionada com a contribuição aleatória. Esta alternativa é, regra geral, bastante robusta para evidenciar os factores latentes mais relevantes.

Estes factores latentes eleitos podem agora ser utilizados para evidenciar informação preciosa sobre a interacção entre variáveis do sistema em causa, através das respectivas *loadings* mais significativas.

Para se reconhecer as contribuições mais relevantes estabelecidas por determinado factor latente que cruza os sub-espaços causa-efeito, é necessário efectuar uma inspecção minuciosa às respectivas *loadings*. É necessário catalogar as *loadings* de cada sub-espaço por

ordem decrescente de importância e garantir que uma fracção significativa do módulo do vector latente é recuperada.

5. Análise Discriminante

A análise discriminante [DA[74]] permite ao utilizador encontrar as características que distinguem os membros de um grupo dos membros de outro grupo, de forma que conhecendo as características de um novo objecto se possa prever a que grupo esse objecto pertence. Neste processo, pode ser utilizada uma ou mais variáveis independentes e quantitativas (por exemplo, variáveis contínuas). As variáveis independentes devem ter um potencial predictor elevado.

5.1 Pressupostos

A análise discriminante exige alguns pressupostos que devem ser observados:

(v) O número de variáveis independentes deve ser muito inferior ao número de casos, dado que, o poder discriminante aumenta com o número de casos se o número de variáveis se mantiver constante.

(vi) As variáveis independentes devem ter distribuição normal multivariada nas populações de onde provêm os diferentes grupos (existe uma regra empírica conhecida que prova que a análise discriminante é robusta a desvios da normalidade, desde que o tamanho do menor grupo seja maior que 20 e se o número de variáveis for menor do que 5).

(vii) Homogeneidade das matrizes de variância e covariância para todos os grupos, ou seja, a variabilidade dentro dos grupos deve ser a mesma.

(viii) Como a análise discriminante é muito sensível à inclusão de *outliers,* os mesmos devem ser identificados e removidos.

5.2 Fundamentação

Nesta toolbox encontra-se implementada a análise discriminante linear. Este programa procura as combinações lineares *Xa* das *p* variáveis observadas que melhor separam os subgrupos de indivíduos indicados, segundo um critério de separabilidade. As soluções *Xa* obtidas designam-se por funções ou eixos discriminantes. Estes eixos podem ser utilizados para obter uma representação gráfica que saliente a distinção entre as classes. Podem também ser úteis para classificar futuros indivíduos (observados nas mesmas variáveis), desde que seja desconhecido à partida o subgrupo a que pertencem.

[74] Do inglês *Discriminant Analysis*

De um modo geral, o programa tenta encontrar uma transformação linear através da maximização da distância entre classes e minimização da distância dentro de classes. Este método procura a melhor direcção de forma que quando os dados são projectados num plano, as classes possam ser separadas.

5.3 O Algoritmo

Este tipo de análise supervisionada envolve um conjunto de aprendizagem a partir do qual são estimadas as propriedades que maximizam a resolução (discriminação) dos objectos. Com base nesta aprendizagem existem agora condições para se poder classificar os objectos.

O conjunto de aprendizagem, onde já se encontram definidos os grupos, é constituído por um conjunto com *N* objectos representados sobre *M* variáveis, previamente classificados em G grupos.

O algoritmo procura maximizar as distâncias entre grupos reduzindo em simultâneo as distâncias dentro dos grupos, aumentando deste modo o poder discriminante dos objectos.

Similarmente à ANOVA de uma via, a variabilidade total pode ser decomposta na variabilidade interna dos grupos (W) e na variabilidade externa (B) que define a separação relativa dos grupos.

A dispersão dos grupos (B) é dada pela matriz de distância de cada centróide dos G grupos ao centro global:

$$B_{(M\times N)} = n_i . b^T_{(M\times G)} . b_{G\times M} \tag{5.1}$$

Já a dispersão interna (W) reflecte a distância de cada objecto ao seu centróide:

$$W_{(M\times M)} = w^T_{(M\times N)} . w_{(N\times M)}) \tag{5.2}$$

A matriz discriminante (D) é dada pelo quociente entre B e W:

$$D_{(M\times M)} = B_{(M\times M)} / W_{(M\times M)} \tag{5.3}$$

Sendo esta matriz uma matriz quadrada e em tudo similar a uma matriz de variância (quociente de variâncias), faz sentido procurar decompor em valores e vectores próprios, sendo os mais relevantes o novo conjunto de eixos ortonormados que permitem maximizar a discriminação dos objectos, isto é, maximizar a variabilidade sobre o menor número de dimensões possível.

A maximização da discriminação dos objectos ($X_{(N\times M)}$) no sub-espaco de dimensao *d* é conseguida através da decomposição da matriz discriminante em valores e vectores próprios

$$X_{(N\times M)} = S_{(N\times d)} . \Lambda_{(d\times d)} . Q'_{(d\times M)} \tag{5.4}$$

onde *S* e *Q* são as matrizes dos *scores* discriminantes e das funções discriminantes desse sub-espaco.

Através do valor das contribuições (*loadings*) das funções discriminantes consegue-se avaliar a relevância de cada variável para a discriminação dos objectos[75].

Atendendo à propriedade de matriz ortonormada

$$Q'_{(d\times M)}.Q_{(M\times d)} = I_{(d\times d)} \tag{5.5}$$

a representação dos objectos no novo espaço discriminante (projecção sobre os vectores próprios) faz-se através do produto das suas componentes pelas componentes do espaço

$$S_{(N\times d)} = X_{(N\times M)}.Q_{(M\times d)} \tag{5.6}$$

6. Análise de Regressão

7. Análise de Séries Temporais

[75] No caso de a matriz discriminante ($D_{(M\times M)}$) não ser simétrica, a nova base vectorial não é ortonormada.