1. ABORDAGEM INTUITIVA DA ANÁLISE DE DADOS

QUADROS DE PARTIDA

Em diferentes domínios do conhecimento, tanto nas ciências Humanas como nas ciências da Natureza, surgem frequentemente quadros multidimensionais onde estão registados valores numéricos (ou atributos qualitativos), resultantes de um conjunto de medidas (ou observações).

Se os quadros em questão são de grandes dimensões, a interpretação dos valores brutos torna-se uma tarefa impraticável, e é então necessário recorrer a técnicas de Análise de Dados para sumarizar a informação de partida.

Na Fig. 1.1 está esquematizado o modelo genérico dos quadros de partida que são o *input* para qualquer método de Análise de Dados. Trata-se de uma matriz Q de dimensões $n \times p$ (n linhas por p colunas) onde, na intersecção da linha i (pertencente ao conjunto I de cardinal n) com a coluna j (pertencente ao conjunto J de cardinal p) se encontra um valor numérico ou um atributo qualitativo $K(i,j)$ que relaciona de algum modo a linha i com a coluna j.

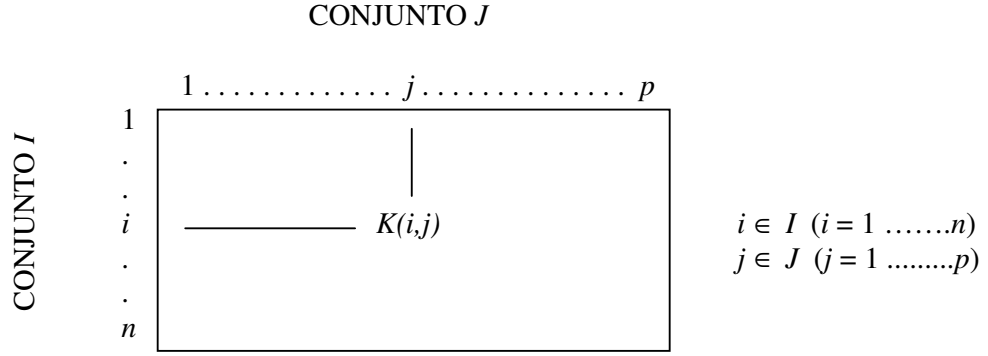


Fig. 1.1 - Quadro de partida para a ANÁLISE DE DADOS (Q).

O significado particular das linhas e colunas (ou seja, a natureza dos conjuntos I e J), bem como o tipo de valores $K(i,j)$ que surgem no quadro Q, condicionam, em grande parte, a escolha da técnica a utilizar no tratamento dos dados de partida.

Assim, por exemplo, as linhas podem ser tomadas como indivíduos (pertencentes a um conjunto *I* de elemento genérico *i*) onde foram registadas *p* propriedades (pertencentes a um conjunto *J*, de elemento genérico *j*); *K(i,j)* tem então o significado do valor que toma a propriedade *j* no indivíduo *i*. Se os valores *K(i,j)* forem variáveis quantitativas (contínuas e mensuráveis), o quadro *Q* denomina-se QUADRO DE MEDIDAS; este tipo de quadro é o *input* característico de um método factorial de Análise de Dados designado por ANÁLISE EM COMPONENTES PRINCIPAIS (ACP).

A ACP foi o primeiro método factorial que suscitou um tratamento matemático rigoroso. De facto, após trabalhos de diferentes investigadores no domínio da psicologia quantitativa (em que se pretendia encontrar os "factores latentes" - tais como "inteligência", "imaginação", "criatividade" - subjacentes aos resultados de uma bateria de testes incidindo sobre um conjunto de indivíduos), Hotteling formulou nos anos 30 a solução do problema da ACP, a partir da diagonalização* de uma matriz de similitude ou de distância que relaciona entre si os resultados dos diferentes testes.

Apresentam-se nas Figs. 1.2 e 1.3 exemplos de quadros de medidas que podem ser tratados através da Análise em Componentes Principais.

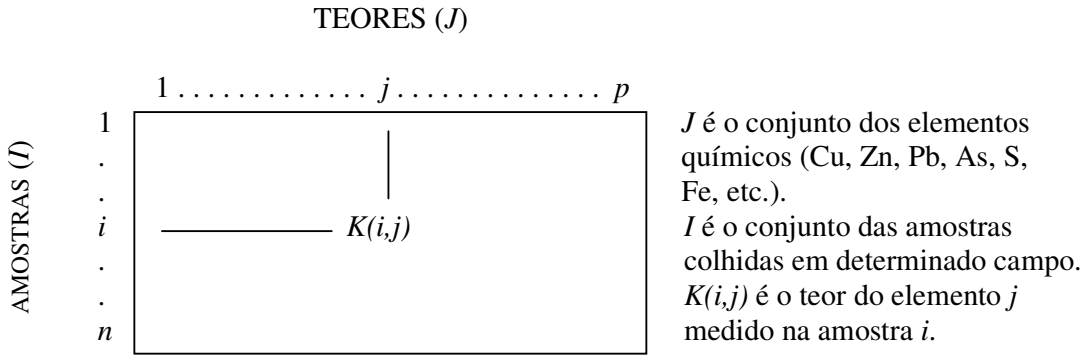


Fig. 1.2 - Exemplo de quadro de partida para a ACP (aplicação em geologia).

* O algoritmo de diagonalização de matrizes permite a sua decomposição numa soma de vectores próprios (factores) ponderados por escalares (valores próprios). Esse algoritmo é facilmente programável e encontra-se em qualquer biblioteca de *software*.

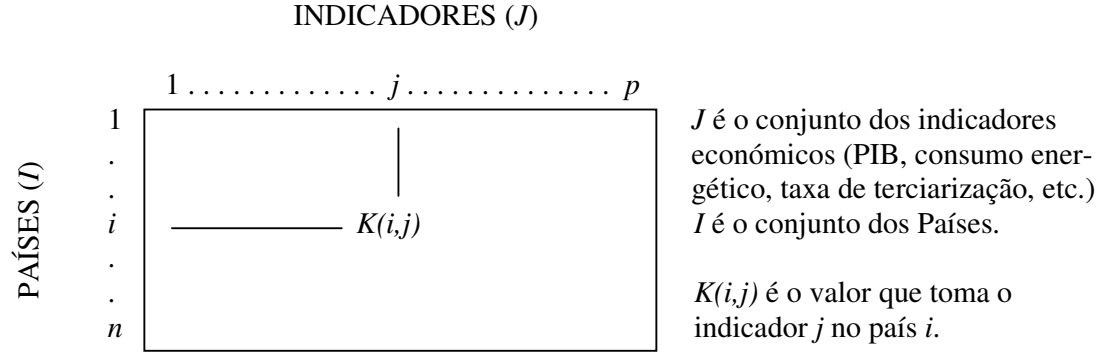


Fig. 1.3 - Exemplo de quadro de partida para a ACP (aplicação em economia).

Como *output* da ACP obtém-se, para o quadro da Fig. 1.2, as relações entre os teores nos diferentes elementos (quais os teores mais correlacionados, quais os que são independentes, etc.), bem como o sistema de similitudes e oposições entre as amostras. Do mesmo modo, para o quadro da Fig. 1.3, obtém-se a correlação dos indicadores entre si e também as similitudes e oposições entre os países, do ponto de vista do conjunto *J* de indicadores escolhidos para caracterizar cada país.

Um outro tipo de método factorial de Análise de Dados foi desenvolvido por J.P. Benzécri no início dos anos 60. Trata-se da ANÁLISE FACTORIAL DAS CORRESPONDÊNCIAS (AFC), a qual se aplica também a um quadro de partida cujo modelo foi apresentado na Fig. 1.1 (QUADRO *Q*). Esse método, que tem alguns aspectos em comum com a ACP (designadamente a diagonalização de uma matriz distância para a pesquisa de factores), apresenta no entanto a particularidade de conferir um estatuto simétrico às linhas e colunas do quadro *Q*, permitindo assim a projecção simultânea dessas linhas e colunas num espaço de dimensão reduzida. Pode portanto visualizar-se, não só o sistema de relações no interior de cada um dos conjuntos *I* e *J*, mas também a estrutura de $I \times J$, o que é uma vantagem significativa relativamente à ACP clássica. Outra vantagem da AFC resulta do facto de ser possível substituir linhas ou colunas com perfil semelhante pela sua soma, sem alterar a projecção dos outros elementos de *I* ou *J* (princípio de equivalência distribucional). A primeira aplicação da Análise das Correspondências foi efectuada por Benzécri em 1962 a dados no domínio

da linguística, tratando uma tabela de contingência onde, no cruzamento da linha *i* com a coluna *j*, se encontra a frequência absoluta de co-ocorrências da palavra *i* ligada (através de uma sintaxe) a uma palavra *j*, para um dado *corpus* (texto submetido a análise). Apresenta-se na Fig. 1.4 um *input* típico da AFC, exemplificando uma tabela de contingência no domínio da linguística.

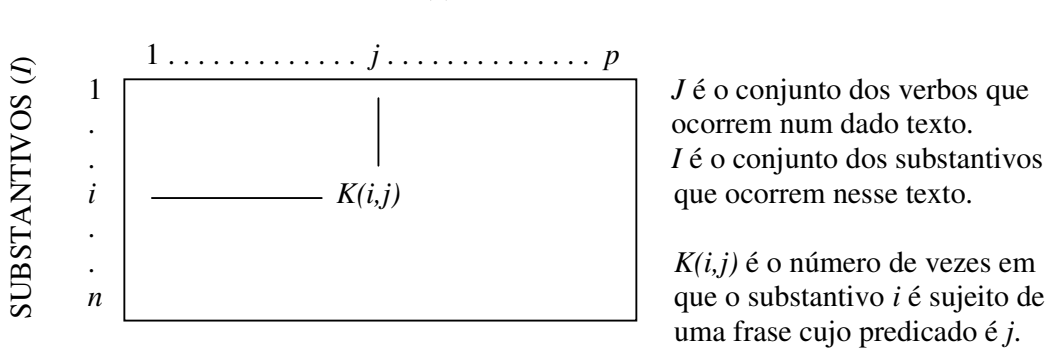


Fig. 1.4 - Quadro de partida para a AFC (aplicação em linguística).

O formalismo da AFC, embora tenha sido desenvolvido no âmbito das tabelas de contingência (onde se registam as frequências absolutas de co-ocorrência de um conjunto de acontecimentos, classificados segundo dois critérios), pode no entanto ser aplicado a qualquer quadro de números positivos. Contudo, na fase de interpretação, só é possível tirar clara vantagem da projecção simultânea das linhas e colunas se a estas puder ser conferido um estatuto simétrico (possibilidade de transpor a matriz sem alterar a análise) e se, os dados de partida tiverem um significado claro, em termos de frequências absolutas ou relativas. Diferentes tipos de quadros de partida têm vindo a ser tratados através da AFC, conduzindo a resultados estáveis e interpretáveis na prática. A diversificação do *input* tem acompanhado de facto o alargamento dos domínios de aplicação do método (o qual abarca hoje praticamente todos os ramos das Ciências do Homem e da Natureza - vd. BIBLIOGRAFIA). Assim, pode dizer-se que a AFC se aplica rigorosamente aos seguintes quadros de partida:

- tabelas de contingência (podendo justapor-se conjuntos de tabelas).

- quadros de notas ou *scores* desdobrados
- quadros contendo quantidades homogéneas e somáveis
- quadros de descrição lógica (presença - ausência).

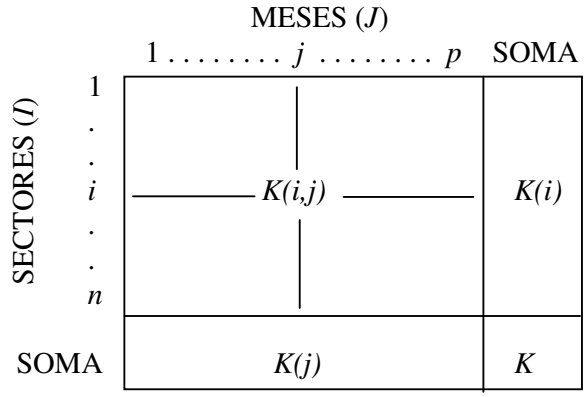
Apresenta-se na Fig. 1.5 um exemplo de quadros de notas desdobradas onde se encontram as notas atribuídas (numa escala 1 - 5) em 4 temas distintos a um conjunto de 3 indivíduos. É de notar que cada tema foi desdobrado em duas colunas (a coluna (+) contém a nota do indivíduo no tema e a coluna (-) contém o complemento para o máximo). Assim o quadro da Fig. 1.5 pode ser considerado como a justaposição de 4 tabelas de contingência. Para uma tabela particular (cuja frequência marginal em linha é constante e igual a 5), a coluna (+) pode ser relacionada com a frequência de "aprovação" do indivíduo num júri de 5 elementos e a coluna (-) como a frequência de "reprovação" (ou o número de subtemas a que o indivíduo respondeu corresponde a (+) e o número a que não respondeu corresponde a (-)).

	TEMA 1		TEMA 2		TEMA 3		TEMA 4		SOMA
	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)	
Indivíduo 1	5	0	4	1	4	1	0	5	20
Indivíduo 2	0	5	4	1	3	2	5	0	20
Indivíduo 3	1	4	0	5	2	3	5	0	20
SOMA	6	9	8	7	9	6	10	5	60
	15		15		15		15		
	Tabela Cont. 1		Tabela Cont. 2		Tabela Cont. 3		Tabela Cont. 4		

Twenty-two points, plus triple-word-score, plus fifty points for using all my letters. Game's over. I'm outta here.

Fig. 1.5 - Exemplo de quadro de notas desdobradas e sua equivalência a uma justaposição de tabelas de contingência.

Apresenta-se na Fig. 1.6 um exemplo de quadro de quantidades homogéneas e somáveis. Trata-se de uma tabela que cruza todos os sectores industriais (têxtil, química, etc.) com os meses de um dado ano. No cruzamento da linha *i* com a coluna *j* - *K(i,j)* - encontra-se o valor acrescentado do sector *i* durante o mês *j*. É evidente que este quadro se pode assimilar a uma tabela de contingência: a soma em linha - *K(i)* - é o valor acrescentado desse sector durante o ano e a soma em coluna - *K(j)* - é o valor acrescentado da indústria durante o mês; a soma total - *K* - é o valor acrescentado na indústria, admitindo que todos os sectores de actividade estão representados no quadro da Fig. 1.6; $\frac{K(i)}{K}$ é a parte do valor acrescentado total pelo qual o sector *i* é responsável; $\frac{K(j)}{K}$ é a parte do valor acrescentado total obtido no mês *j*.



I é o conjunto dos sectores industriais.

J é o conjunto dos meses.

K(i,j) é o valor acrescentado do sector *i* durante o mês *j*.

Fig. 1.6 - Exemplo do quadro de quantidades homogéneas e somáveis.

Apresenta-se na Fig. 1.7 um exemplo de um quadro de descrição lógica (presença - ausência). Um conjunto de indivíduos é interrogado sobre uma série de questões, admitindo-se apenas a resposta SIM (presença - codificada por 1) ou NÃO (ausência - codificada por 0) a cada uma das modalidades das perguntas (codificação disjuntiva completa). É fácil de ver que este tipo de quadro se reduz também à justaposição de tabelas de contingência (uma tabela por cada pergunta).

PERGUNTAS	SEXO		IDADE			RESIDÊNCIA			APROVA O NUCLEAR	
MODALIDADES	M	F	20/30	30/40	>40	Lisboa	Porto	Província	Sim	Não
Indivíduo 1	1	0	1	0	0	1	0	0	0	1
Indivíduo 2	1	0	0	0	1	0	1	0	1	0
Indivíduo 3	0	1	0	1	0	0	0	1	0	1

Fig. 1.7 - Exemplo de quadro de descrição lógica (presença - ausência) codificando os resultados de um inquérito.

Para além da vantagem da projecção simultânea e da "universalidade" da sua aplicação prática a diferentes domínios e a vários tipos de quadros de partida, a Análise das Correspondências permite ainda tirar partido da projecção de linhas ou colunas em suplementar e da reconstituição aproximada do quadro de partida com base nas projecções das linhas e colunas nos factores.

Assim, quando um indivíduo (ou propriedade) é projectado em suplementar, este (ou esta) não contribui para a construção dos eixos, sendo apenas posicionado(a) no espaço factorial definido pelos outros indivíduos (ou propriedades), ditos principais. Por exemplo, na Fig. 1.7 existem claramente 2 tipos de perguntas - as três primeiras caracterizam estruturalmente o indivíduo e a quarta refere-se à sua opinião sobre o tema do inquérito. Pode ser interessante projectar as respostas à pergunta de opinião sobre o espaço factorial que caracteriza os indivíduos do ponto de vista do sexo, idade e residência. Então diagonalizar-se-á apenas a matriz referente às 3 primeiras perguntas e, sobre os factores assim obtidos, projectam-se as respostas à pergunta de opinião, obtendo-se deste modo uma relação entre essa pergunta e as características dos indivíduos.

A reconstituição aproximada do quadro de partida com base nas projecções das linhas e colunas num pequeno número de eixos factoriais permite, em certos casos, estimar elementos do quadro

que, por qualquer razão, não estão disponíveis. Essa estimação é efectuada com base na estrutura revelada pelos elementos presentes no quadro de partida.

***

Pelas razões aduzidas anteriormente e ainda porque a AFC se conjuga harmoniosamente com outros métodos de ANÁLISE DE DADOS (aplicados a jusante) cujo objectivo é mais "ambicioso" do que a "simples" descrição estrutural do quadro de partida, este tipo de análise tem potencialidades que podem ser exploradas com êxito em diferentes domínios de aplicação. Assim a AFC "filtra", de certo modo, os dados de partida, preparando-os para tratamentos mais poderosos (por exemplo, classificação, discriminação e regressão). As estruturas reveladas pela AFC podem eventualmente sugerir ou guiar análises complementares, as quais, quando aplicadas aos dados brutos, não conduzem a resultados interpretáveis.

OBJECTIVOS E METODOLOGIA DA ANÁLISE DE DADOS

Quanto aos objectivos a atingir, os métodos de ANÁLISE DE DADOS podem segmentar-se em duas categorias:

1. **Métodos Descritivos** - em que o objectivo é a descrição estrutural do quadro de partida **(ANÁLISE EM COMPONENTES PRINCIPAIS E ANÁLISE FACTORIAL DAS CORRESPONDÊNCIAS)**.

2. **Métodos Explicativos** - em que o objectivo é a modelização do fenómeno descrito pelos métodos anteriores, segundo diferentes perspectivas, nomeadamente:

   2.1. Classificação Automática - criação de tipologias;

2.2. Discriminação - afectação de indivíduos anónimos a grupos pré-estabelecidos;

2.3. Regressão - estabelecimento de relações entre as variáveis.

O objectivo dos métodos descritivos consiste essencialmente na pesquisa do sistema de relações entre as linhas e colunas do quadro de partida. Esse sistema de relações não é aparente nos dados brutos e só uma redução na dimensionalidade do espaço permite visualizar as oposições e similitudes existentes entre os elementos submetidos a análise. A ideia básica é projectar indivíduos e propriedades em gráficos planos (a duas dimensões) definidos por um pequeno número de eixos, minimizando contudo a perda de informação (deformando o menos possível as relações geométricas entre os pontos que representam os dados de partida).

O procedimento típico dos métodos descritivos, que aqui se delineou nos seus aspectos fundamentais, não pode ser concretizado com base exclusivamente em técnicas do âmbito da estatística clássica. Esta traz consigo um arsenal de hipóteses (por exemplo, a multi-normalidade) e de testes de significância associados a uma certa lei de probabilidade que não respondem às questões práticas e operacionais que são tratadas pela Análise de Dados.

Por exemplo, a estatística clássica responde à questão de saber se, para um certo risco, há ou não independência entre as linhas e colunas de uma tabela de contingência (teste do χ^2). Essa resposta tem pouco interesse, do ponto de vista da Análise de Dados, visto que, se houver independência não há nada a fazer (o quadro não tem estrutura) e, se houver estrutura, o que importa saber é em que direcção os dados se afastam da independência (e qual a forma das relações existentes para as modalidades que se cruzam na tabela de contingência).

No entanto, certos resultados da estatística clássica (em especial no domínio da regressão) e da estatística não paramétrica (para sintetizar as tabelas de partida e comparar entre si os resultados da Análise de Dados) são utilizados nos métodos descritivos (e *a fortiori* nos métodos explicativos).

Também algumas técnicas no âmbito da Investigação Operacional (como a simulação e a optimização) são por vezes um complemento precioso dos métodos aqui considerados.

Mas a base matemática da Análise de Dados é fundamentalmente a ÁLGEBRA LINEAR aplicada a conceitos geométricos como centro de gravidade, espaços euclidianos, distâncias, projecções, inércia, etc..

O objectivo dos métodos descritivos é encontrar, com um mínimo de hipóteses *a priori*, uma representação aproximada do quadro de partida que garanta o máximo de conformidade geométrica com os dados.

* * *

Vejamos agora o procedimento genérico que permite encontrar a estrutura de um quadro de partida que cruza n indivíduos com p propriedades.

O conjunto I dos indivíduos pode ser tomado como uma nuvem em R^p; cada indivíduo dessa nuvem é caracterizado por p coordenadas que o posicionam num espaço a p dimensões (essas coordenadas são os valores numéricos - eventualmente codificados - que estão registados no quadro de partida para a linha correspondente ao indivíduo).

Evidentemente que se p for superior a 3, não é possível visualizar os dados brutos e portanto é trabalhoso (ou até impossível) detectar visualmente a estrutura de relações entre os indivíduos. Sem uma representação gráfica adequada, é impraticável encontrar grupos de indivíduos "semelhantes" ou "opostos" do ponto de vista das suas propriedades, ou pesquisar gradações ou seriações na posição geométrica dos indivíduos em R^p.

J. Bertin imaginou um método manual de permutação das linhas e colunas da matriz dos dados, com vista a detectar a sua estrutura. Cada elemento da matriz é representado por uma tonalidade que dá a intensidade do valor da variável para esse elemento (se forem só duas tonalidades,

representa-se a negro os elementos que tiverem valores superiores à média, por exemplo). Por tentativas, vai-se trocando a posição dos indivíduos e das variáveis até obter uma estrutura do tipo da que se apresenta na Fig. 1.8 (escalograma), onde é claramente possível detectar grupos de indivíduos caracterizados por uma certa intensidade das propriedades, e em que tanto os indivíduos como as propriedades revelam uma gradação clara.

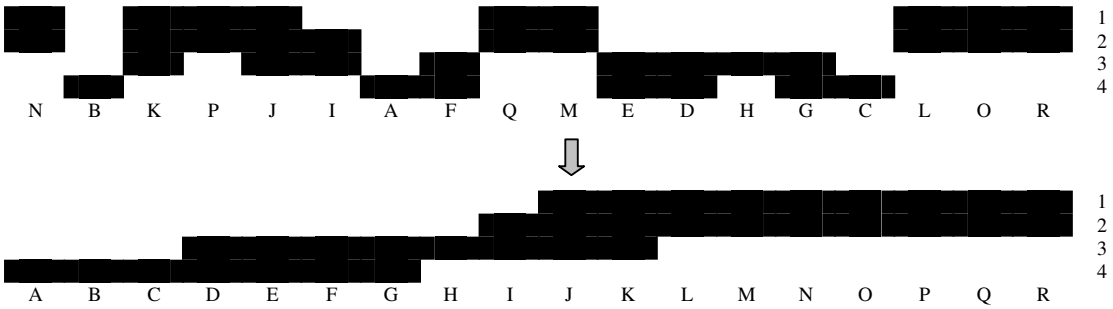


Fig. 1.8 - Escalograma permitindo visualizar a estrutura do quadro de partida sem redução da dimensionalidade do espaço.

Este método, embora graficamente sugestivo (J. Bertin é especialista em semiologia gráfica), é evidentemente impraticável, mesmo para matrizes de pequenas dimensões (20×20), e implica um trabalho moroso e repetitivo. Os mesmos resultados podem ser obtidos através dos métodos factoriais se as linhas e colunas forem ordenadas segundo as suas projecções no 1º factor.

Para reduzir a dimensionalidade do espaço, a ideia de base dos métodos factoriais é pesquisar o conjunto de rectas (no espaço R^p, para posicionar os indivíduos) que "melhor" se ajusta à nuvem inicial, de acordo com um certo critério de distância.

Para encontrar a primeira dessas rectas usa-se um procedimento que é a generalização da regressão linear simples - procura-se a recta em R^p, tal que a soma dos quadrados das distâncias* de cada ponto à recta seja mínima.

As projecções dos *n* indivíduos nessa recta constituem a melhor representação possível da nuvem inicial num espaço unidimensional. Essa recta passa pelo centro de gravidade da nuvem e dá a sua direcção de máximo alongamento, designando-se por **1º Eixo de Inércia, 1º Factor, 1º Eixo Factorial** (ou ainda, o **vector próprio** correspondente ao maior valor próprio da matriz de similitude entre as propriedades).

Prosseguindo o algoritmo, procura-se um espaço ortogonal ao 1º factor (pretende-se que o segundo factor seja ortogonal ao primeiro). Nesse espaço (cuja dimensão é já *p*-1), pesquisa-se a direcção de maior alongamento, projectando a nuvem inicial nesse espaço e, para essas projecções, minimizam-se as suas distâncias a uma recta. Tal recta é o 2º Eixo Factorial e dá a direcção de alongamento de 2ª ordem (o maior alongamento depois de descartado o 1º Factor).

A projecção da nuvem inicial no plano definido pelos dois primeiros factores ortogonais constitui a melhor aproximação possível dessa nuvem, quando interceptada por um plano (a posição relativa dos indivíduos pode ser analisada num gráfico cartesiano habitual situado nesse plano e cujos eixos são os dois primeiros factores).

Reiterando o processo, obtém-se *p* Eixos Factoriais ortogonais, classificados por ordem decrescente da sua "importância" na explicação da forma da nuvem inicial. A importância de cada factor (designada por valor próprio) é medida pelo quociente entre a soma dos quadrados das distâncias ao centro de gravidade das projecções nesse factor e a soma dos quadrados das distâncias ao

---

* Em Análise Factorial das Correspondências, usa-se uma distância ajustada a variáveis qualitativas – trata-se da distância do χ^2.

mesmo centro de gravidade, para a nuvem inicial (em linguagem estatística é a relação entre a variância das projecções no eixo e a variância total).

Se um número pequeno de eixos factoriais (1, 2 ou 3) for suficiente para reproduzir com certa fidelidade a forma da nuvem inicial (o que tem de ser julgado com base na sucessão dos valores próprios da matriz de similitude[*], pode encontrar-se o sistema de relações entre os indivíduos através da análise de gráficos planos cujos eixos são os factores retidos.

Em ACP faz-se seguidamente uma análise paralela em R^n (projectando as propriedades no espaço dos indivíduos e reduzindo a sua dimensionalidade). Em AFC projectam-se simultaneamente os indivíduos e propriedades no mesmo espaço, mercê de um tratamento simétrico das linhas e colunas.

* * *

Os métodos explicativos vão ser aqui tratados como complemento dos métodos descritivos. Vejamos como é possível articular entre si as diferentes técnicas de Análise de Dados, partindo dos métodos descritivos.

Quanto à classificação automática, o objectivo genérico é criar grupos de indivíduos semelhantes (do ponto de vista das respectivas propriedades) ou grupos de propriedades correlacionadas entre si. Estabelecem-se assim tipologias através de um procedimento automático. Para a análise ascendente hierárquica dita em Modo Q (classificação dos indivíduos), pesquisa-se o par de indivíduos mais semelhantes (formado pelos dois indivíduos cuja distância no espaço das propriedades é mínima), constitui-se, com esse par, um grupo e reinicia-se o processo até esgotar o conjunto total de

[*] A soma dos valores próprios correspondentes aos factores descartados deve ser pequena em face da soma de todos os valores próprios.

indivíduos. O resultado final da Classificação Ascendente Hierárquica é um dendrograma (vd. Fig. 1.9).

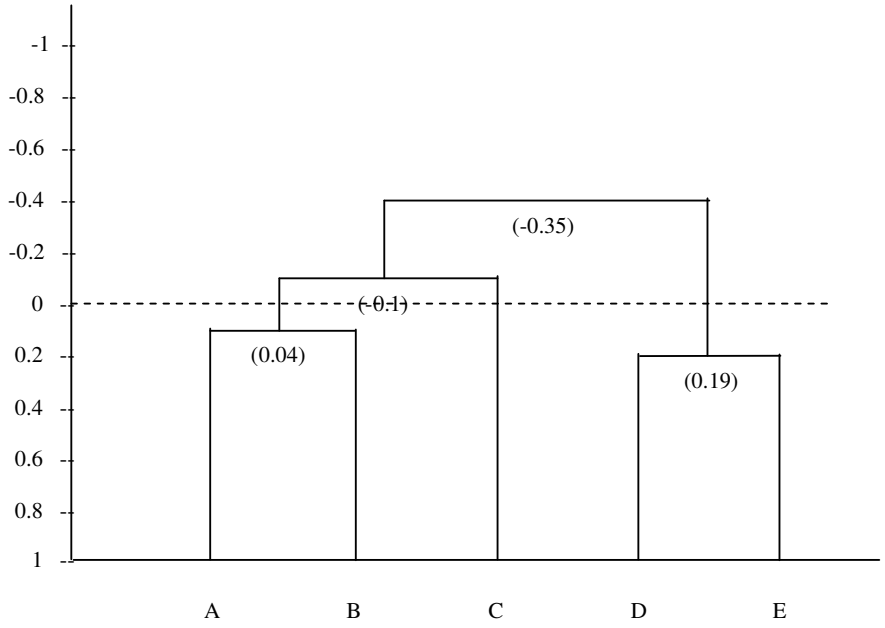
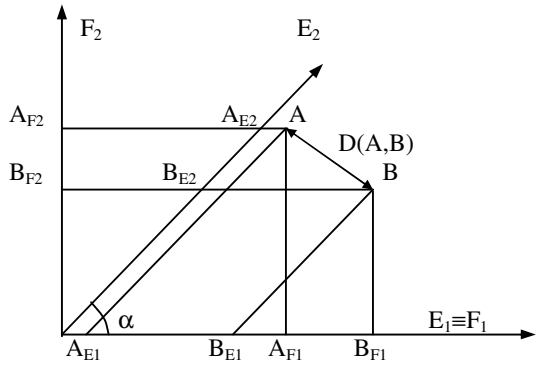


Fig. 1.9 - Exemplo de dendrograma.

A classificação automática pode ser efectuada sobre os dados de partida. Mas se, previamente, estes tiverem sido submetidos a um método factorial, a classificação dos indivíduos no espaço dos factores conduz, em geral, a resultados mais coerentes do ponto de vista formal, visto que as distâncias entre os indivíduos (calculada como é habitual pela soma dos quadrados das diferenças entre coordenadas) foram obtidas num espaço ortogonal (e não no espaço inicial de "eixos oblíquos" definidos pelas propriedades, em geral correlacionadas entre si - vd. Fig. 1.10).



E1, E2 - eixos iniciais (propriedades)
F1, F2 - eixos factoriais

D(A,B) - distância entre os indivíduos A e B

$D^2 (A,B) = (A_{F1} - B_{F1})^2 + (A_{F2} - B_{F2})^2$

$D^2 (A,B) \neq (A_{E1} - B_{E1})^2 + (A_{E2} - B_{E2})^2$

Cos α = coeficiente de correlação entre as propriedades 1 e 2

Fig. 1.10 - Distância num espaço de eixos oblíquos e num espaço de eixos ortogonais.

Também os resultados de uma classificação automática sobre os indivíduos podem ser projectados no espaço dos factores obtidos após uma AFC, permitindo visualizar a relação entre os grupos de indivíduos e enriquecer a interpretação da tipologia pela detecção das propriedades que para ela mais contribuem. Assim os dendrogramas habituais da classificação automática ganham um sentido mais coerente quando os grupos são projectados no espaço dos factores.

* * *

Quanto à análise discriminante clássica, tal como foi formulada por Fisher para variáveis quantitativas, o objectivo é encontrar um eixo discriminante *U* que maximiza o quociente das projecções (sobre *U*) entre a variância inter-classes (de classe para classe) e a variância intra-classes (no interior das classes).

Uma variante da análise discriminante, proposta por Benzécri para variáveis qualitativas, consiste em efectuar a AFC da tabela contingência *TC* da Fig. 1.11, constituída por duas linhas, contendo a soma das frequências em *R* e *S*.

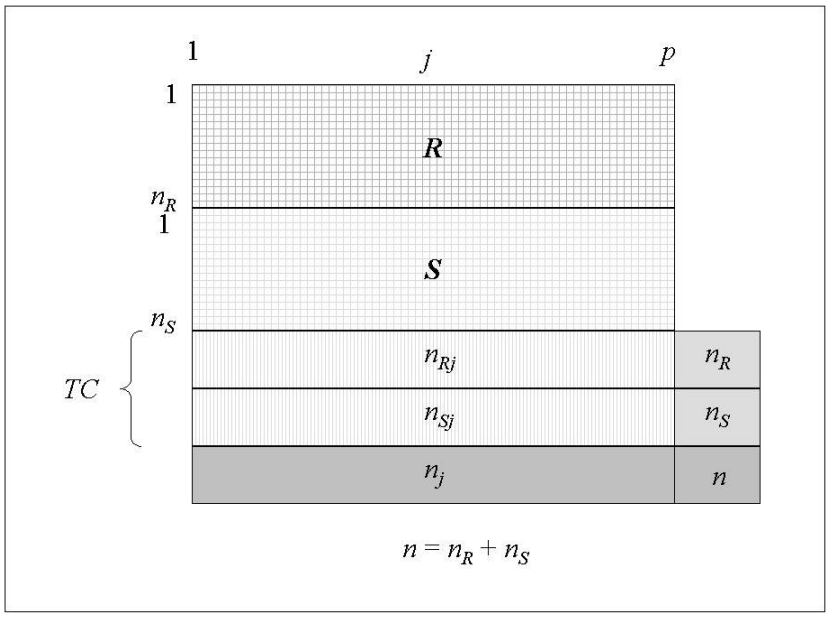


Fig. 1.11 – Tabela de contingência.

Sobre o eixo factorial resultante desta AFC, projectam-se em suplementar os indivíduos dos grupos *R* e *S*, o que permite interpretar a sua separação com base nas propriedades que mais contribuem para essa separação.

Uma vez encontrado o eixo factorial, é então possível afectar indivíduos anónimos aos grupos pré-estabelecidos, desde que se defina uma fronteira no eixo factorial.

Na Fig. 1.12 encontra-se esquematizado o procedimento que permite utilizar a AFC como método discriminante. A posição da fronteira ψ é encontrada por minimizção dos indivíduos mal classificados na zona de coalescência.

Projectando agora quaisquer indivíduos anónimos, eles ficam afectados a *A* se se projectam à esquerda de ψ e a *B* se se projectam à direita de ψ.

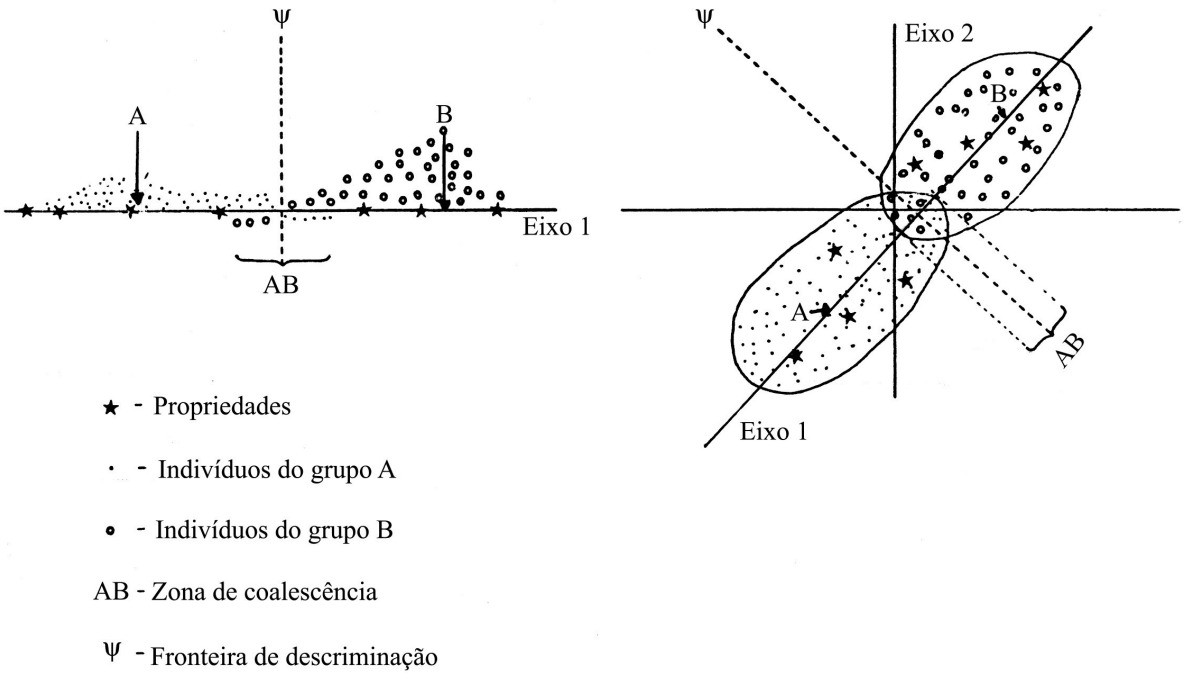


Fig. 1.12 - Discriminação através da Análise Factorial das Correspondências.

Dada a propriedade da projecção simultânea em AFC, ψ pode pôr-se em função das variáveis e estabelecer assim um critério quantitativo para separar os dois grupos. Qualquer novo indivíduo "anónimo" (cuja pertença aos grupos é desconhecida) pode agora ser projectado (através do seu perfil nas mesmas propriedades) no espaço de dimensão reduzida (Eixo 1 ou plano dos Eixos 1, 2) e afectado a um dos grupos pré-estabelecidos de acordo com a posição relativamente a ψ.

* * *

No que diz respeito à regressão múltipla, pretende-se encontrar uma relação entre uma variável privilegiada – dita dependente – e uma série de outras variáveis ditas explicativas ou independentes. Se estas últimas forem inter-correlacionadas (multi-colinearidade), a regressão clássica não permite tirar o melhor partido dos dados (e pode conduzir a soluções absurdas). Então uma ACP (ou AFC) sobre estas variáveis permite encontrar os factores (que são ortogonais por

construção) e é possível posteriormente efectuar a regressão da variável dependente sobre estes factores, eliminando assim a redundância nas variáveis explicativas.

O PROBLEMA DA VALIDAÇÃO DOS RESULTADOS EM ANÁLISE DE DADOS

Embora algumas tentativas tenham sido feitas para estabelecer testes de hipóteses no quadro da estatística multivariada para validar os resultados obtidos pela análise de dados, o ponto de vista exploratório, inspirado nos trabalhos de Tukey (EDA - *Exploratory Data Analysis*), tem prevalecido sobre a atitude formalista porque permite tirar partido do computador como instrumento de rápida experimentação numérica. De facto, a atitude exploratória subjacente a este enfoque da análise de dados tem-se revelado fértil, desde que validada pelo contexto onde se situa o estudo. Esta aproximação tem-se mostrado válida, não só porque não apela a testes que se baseiam em hipóteses estatísticas inverificáveis e irrealistas (como a multigaussianidade), como ainda porque tem a capacidade de levantar hipóteses novas, sugeridas pelo reconhecimento de padrões provenientes da análise (que terão de ser confrontadas com o conhecimento pericial sobre a fenomenologia reflectida nos dados empíricos).

A atitude aqui proposta assenta na complementaridade entre os resultados da análise de dados e a contribuição exterior das disciplinas que estabelecem o contexto preciso onde os dados se inserem. A interpretação faz-se com base nas regularidades encontradas pela análise, tomando em conta as analogias formais sugeridas pela geometria e topologia das nuvens de pontos que resultam dos métodos, e recorrendo ainda a informação "externa" aos dados, fornecida pelo especialista do domínio a que o problema se refere.

No que diz respeito à validação dos resultados obtidos pelos métodos explicativos, para além da sua eventual operacionalidade pragmática, é necessário chamar a atenção para o facto de tais métodos, para penetrarem mais profundamente na realidade, exigirem a utilização incessante de outros

modelos imbricados, pelo que o ganho em previsão resulta, em última instância, de uma série de pressupostos, por vezes inconscientes, mas cujo estatuto tem de ser claramente discutido, no momento da validação.

Pelo contrário, os métodos descritivos, estando mais próximos da realidade empírica, são mais facilmente controláveis, podendo, em certos casos, com economia de hipóteses, atingir resultados de valor pragmático imediato.

ARTICULAÇÃO DAS DIFERENTES FASES E TÉCNICAS DE ANÁLISE DE DADOS

Perante um conjunto de dados brutos, e tendo uma ideia do que se pode esperar do estudo a empreender, o especialista de um certo domínio de aplicação tem de definir uma estratégia de articulação das diferentes técnicas de Análise de Dados que lhe permita atingir os seus objectivos (os quais podem não estar rigorosamente definidos *a priori*, mas que vão sendo mais precisos à medida que a análise avança).

Essa estratégia comporta diferentes fases sequenciais ou em retroacção, implicando muitas vezes escolha e ensaio de métodos alternativos ou complementares, selecção de variáveis e respectiva (re)codificação, regresso ao quadro de partida, etc. Apresenta-se na Fig. 1.13 um diagrama onde se estabelecem as relações entre as diferentes fases de uma ANÁLISE DE DADOS.

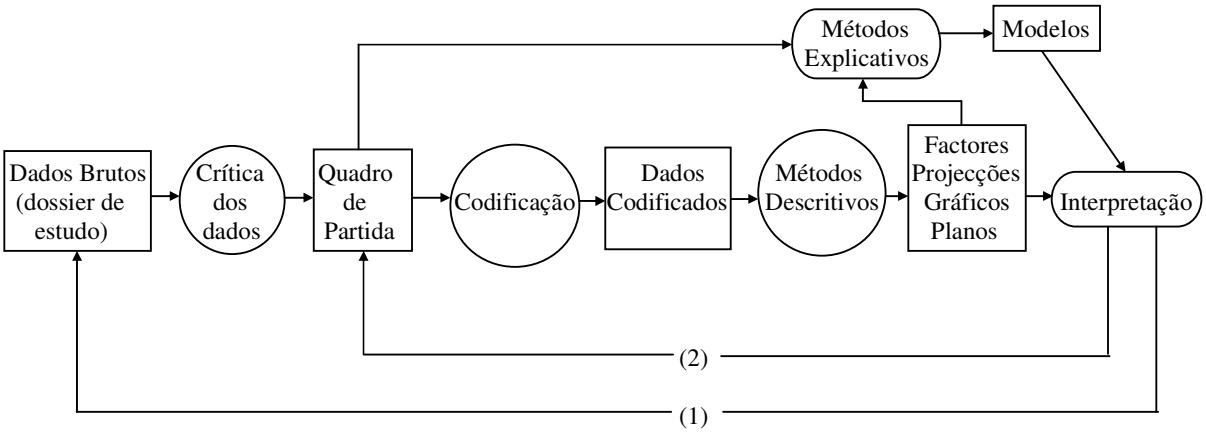


Fig. 1.13 - Diagrama de articulação das fases de uma análise de dados.

Partindo do dossier do estudo (DADOS BRUTOS), é necessário, em geral, efectuar uma CRÍTICA DOS DADOS, do ponto de vista semântico (significado das linhas e colunas da matriz, representatividade do conjunto dos indivíduos e propriedades, presença de valores aberrantes ou ausência de certos elementos do quadro, etc.) e do ponto de vista estatístico (análises uni e bi-dimensionais sobre histogramas e tabelas de contingência, cálculo de médias, variâncias e outras medidas que sumarizem a distribuição de cada variável ou de pares de variáveis - coeficientes de correlação, χ^2 das tabelas de contingência, etc.). Terminada esta fase, pode construir-se o QUADRO DE PARTIDA, que deve conter apenas os dados significativos para o estudo em questão.

Esse quadro de partida pode agora ser CODIFICADO para servir de *input* aos MÉTODOS DESCRITIVOS. A AFC, por exemplo, pode admitir diferentes tipos de codificação, e há uma ligação estreita entre cada tipo de codificação utilizada e as regras sintácticas de interpretação dos resultados.

O objectivo da codificação é assegurar uma homogeneidade das variáveis a serem submetidas a análise, podendo ainda fazer ressaltar estruturas não aparentes nos dados brutos. No caso mais geral, os dados podem conter variáveis de diferente natureza:

- Variáveis **nominais** (só admitem a igualdade) - Ex.: categoria sócio-profissional, sexo, litologia.

- Variáveis **ordinais** (só admitem a igualdade e a relação de ordem) - Ex.: grupos de idade, grau de metamorfismo, sabor, classe social.

- Variáveis **mensuráveis** (admitem a relação de ordem e a soma) - Ex.: teores em suporte constante, comprimentos, rendimentos familiares.

Antes de aplicar um método descritivo, é necessário assegurar a homogeneidade das variáveis, codificando os dados de partida. Como foi anteriormente referido, a ACP só se aplica a variáveis mensuráveis. A AFC pode ser aplicada a todos os tipos de variáveis, se estas forem convenientemente codificadas (por exemplo, passar as variáveis mensuráveis a ordinais por divisão em classes).

É de notar ainda que a codificação deve ser efectuada em colaboração com o especialista do domínio do estudo, visto que, alterando a codificação (até como consequência de uma primeira interpretação - vd. retroacção (2) da Fig. 1.13) podem surgir estruturas diferentes, mais facilmente interpretáveis e com maior significado prático. Por exemplo, a variável idade pode ser tomada como mensurável (a idade exacta do indivíduo), como ordinal (jovens, adultos, idosos) ou como nominal (se, em resultado da interpretação, for conveniente, por exemplo, agrupar os jovens e os idosos). O modo como cada variável deve ser codificada depende pois, não só de todas as outras (para assegurar a homogeneidade), mas também dos resultados da interpretação.

Os dados codificados são então submetidos a um MÉTODO DESCRITIVO (Fig. 1.13), o qual conduz a uma série de *outputs* (FACTORES, PROJECÇÕES, GRÁFICOS PLANOS), que devem seguidamente ser interpretados. Essa INTERPRETAÇÃO pode levar a retomar o dossier do estudo (retroacção (1) da Fig. 1.13) e/ou reformular o quadro de partida (retroacção (2) da Fig. 1.13),

ensaiando nova codificação, segmentando-o em blocos, eliminando ou projectando em suplementar indivíduos ou variáveis, etc..

Para aplicar os métodos explicativos pode evidentemente utilizar-se como *input* o quadro de partida, se a crítica de dados tiver fornecido indicações suficientes para guiar ou validar a aplicação de tais métodos. Se, pelo contrário, o conhecimento existente sobre o quadro de partida for insuficiente, é aconselhável aplicar os métodos explicativos sobre os resultados obtidos através dos métodos descritivos.

EXEMPLO ILUSTRATIVO DA ANÁLISE FACTORIAL DAS CORRESPONDÊNCIAS

Seja o Quadro 1.1 referente à frequência absoluta de atribuição de prémios Nobel por países, por especialidade e por época. O Quadro 1.1 é a justaposição de 2 tabelas de contingência – a tabela A cruza os países com as especialidades e a tabela B cruza os países com as épocas em que os prémios foram atribuídos.

Quadro 1.I – Dados de partida.

PAÍSES	A			B					
	Física	Química	Medicina	1901-1915	1916-1930	1931-1945	1946-1960	1961-1975	1976-1991
EUA	54	36	66	3	3	14	38	41	57
Reino Unido	21	24	23	7	8	11	14	20	8
Alemanha	16	24	13	15	12	11	4	8	3
França	9	7	8	10	3	2	0	5	4
Ex-URSS	7	1	2	2	0	0	4	3	1
Japão	3	1	0	0	0	0	1	2	1
Outros Países	27	19	37	15	15	11	13	13	16
	ESPECIALIDADES			ÉPOCAS					

Podem então aplicar-se duas estratégias em paralelo:

1. Projectar B em suplementar sobre A, obtendo-se, pela análise das variáveis principais, o modo como as especialidades se relacionam com os países. Na Fig. 1.14 (que contém a totalidade da informação do quadro de partida A), verifica-se que o eixo 1 opõe a

(Física+Medicina) à Química. Então os países mais relacionados com a (Física + Medicina) são os EUA, e, com a Química, o Reino Unido e Alemanha. A separação entre a Física e a Medicina é dada pelo eixo 2, estando o Japão e a Ex-URSS ligados à Física e os "Outros Países" à Medicina. É de notar a posição central da França, que se projecta na proximidade da origem. Sobre esta estrutura pode então interpretar-se a projecção em suplementar das épocas: até 1945, há predominância da Alemanha e da Química. Depois de 1945, surgem os EUA, ligados à (Física + Medicina).

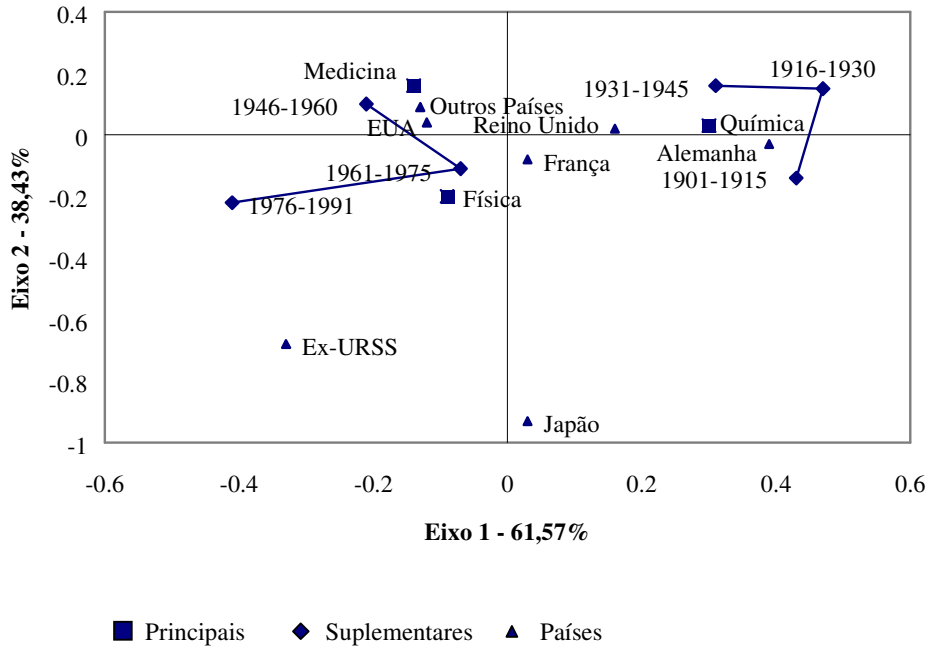


Fig. 1.14 – Projecção de B em suplementar sobre A.

2. Projectar A em suplementar sobre B, obtendo-se o modo como os países se relacionam com as épocas, por interpretação dos eixos criados pelas variáveis principais. Na Fig. 1.15 (que explica 90% da informação contida no quadro de partida), verifica-se uma sequência de épocas ao longo do eixo 1 (excepto para os períodos de 46-60 e 61-75, que aparecem invertidos). Nesta sequência, pode estabelecer-se uma ordenação dos países, segundo a sua projecção no eixo 1, *i. e.*, segundo a época em que houve mais Prémios Nobel atribuídos a cada país:

Alemanha ⇒ França ⇒ Outros ⇒ Reino Unido ⇒ Ex-URSS ⇒ EUA ⇒ Japão

Quanto às variáveis em suplementar, a Química projecta-se antes de 1945 e a (Física + Medicina), depois dessa data.

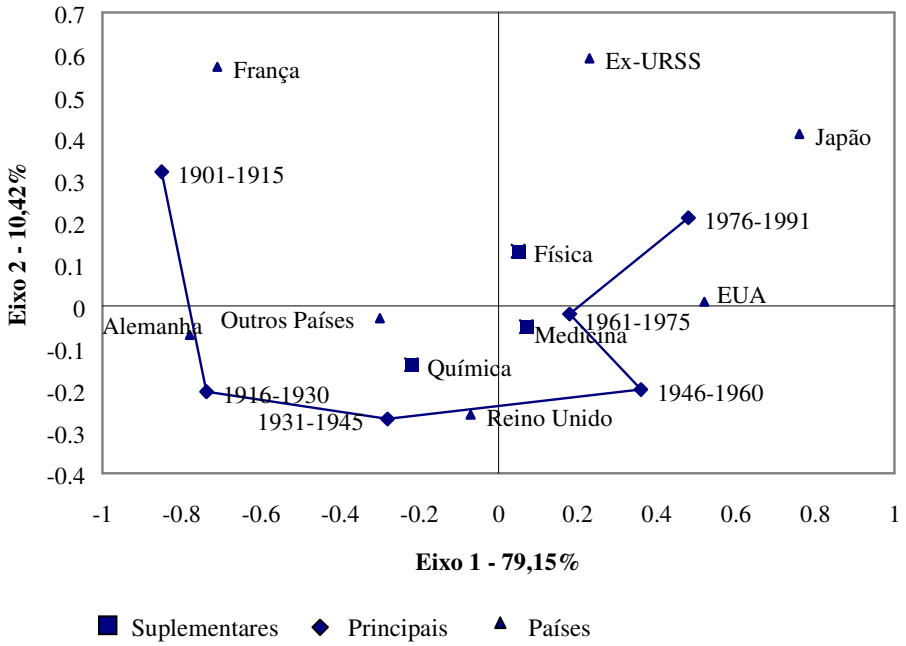


Fig. 1.15 – Projecção de A em suplementar sobre B.