3. ANÁLISE EM COMPONENTES PRINCIPAIS

INTRODUÇÃO

É frequente pretender descrever um conjunto de dados constituído por n indivíduos caracterizados por p variáveis do tipo medidas quantitativas.

Este género de dados conduz a um quadro de partida dissimétrico Q cujo termo geral q_{ij} representa o valor tomado pela j-ésima variável no indivíduo i. As variáveis podem apresentar unidades de medida diferentes (algumas podem estar expressas em quilogramas e outras em gramas) com valores médios bastante distintos, sugerindo que os dados sejam centrados. O quadro de partida Q transforma-se assim num quadro X cujo termo geral é

$$x_{ij} = \frac{1}{\sqrt{n}}\left(q_{ij} - \overline{q}_j\right)$$

onde $\overline{q}_j$ é a média aritmética dos valores tomados pela variável j.

A Análise em Componentes Principais é um caso particular da Análise Geral do quadro X, descrita no capítulo anterior. A matriz a diagonalizar X^TX é a matriz variância-covariância.

Frequentemente é ainda necessária uma modificação suplementar do quadro de partida, quando a dispersão das variáveis é muito diferente ou quando as variáveis diferem quanto à sua natureza sendo expressas em unidades de medida não comparáveis. Este problema pode ser ultrapassado reduzindo as variáveis, ou seja, tornando-as adimensionais com média nula e variância unitária. O termo geral do quadro X, neste caso, é dado por

$$x_{ij} = \frac{1}{\sqrt{n}} \frac{q_{ij} - \overline{q}_j}{s_j}$$

onde s_j é o desvio padrão da variável j.

A matriz $X^T X$ transforma-se na matriz das correlações experimentais. A Análise em Componentes Principais neste caso é apelidada de normada.

ANÁLISE EM R^p

A transformação

$$x_{ij} = \frac{1}{\sqrt{n}} \left(q_{ij} - \overline{q}_j \right)$$

traduz-se numa translação da nuvem, de modo a fazer coincidir o centro de gravidade com a origem. A influência do nível geral de cada variável é, assim, eliminada. O coeficiente $\frac{1}{\sqrt{n}}$ tem por objectivo fazer coincidir a matriz $X^T X$ com a matriz variância-covariância.

O quociente pelo desvio padrão s_j provoca a redução do efeito das variáveis muito dispersas sobre as distâncias entre indivíduos:

$$d^2(i,i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \frac{1}{n} \sum_{j=1}^{p} \left(\frac{q_{ij} - q_{i'j}}{s_j} \right)^2$$

Assim cada variável terá uma contribuição análoga na determinação das distâncias.

Resumindo, a análise da nuvem de pontos em R^p conduz à translação da origem para o centro de gravidade e à transformação das escalas dos diferentes eixos. A análise do

quadro transformado traduz-se na pesquisa dos vectores próprios u_j da matriz de correlação experimental $R=X^T X$.

As coordenadas dos indivíduos nos eixos factoriais são dadas pelos produtos escalares seguintes:

$$W = X\,u$$

ANÁLISE EM R^n

A divisão por $s_j\sqrt{n}$, que em R^p se traduzia numa mudança de escala dos eixos, conduz a uma deformação da nuvem em R^n; cada variável passa a estar posicionada à distância unitária da origem.

$$d^2(j,o) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_{ij}-\overline{q}_j}{s_j}\right)^2 = 1$$

As variáveis estão posicionadas sobre uma hiperesfera de raio 1 centrada na origem. A distância entre dois pontos *j* e *j*' é dada por:

$$d^2(j,j') = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_{ij}-\overline{q}_j}{s_j} - \frac{q_{ij'}-\overline{q}_{j'}}{s_{j'}}\right)^2$$

$$d^2(j,j') = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_{ij}-\overline{q}_j}{s_j}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_{ij'}-\overline{q}_{j'}}{s_{j'}}\right)^2 - 2\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_{ij}-\overline{q}_j}{s_j}\,\frac{q_{ij'}-\overline{q}_{j'}}{s_{j'}}\right)$$

$$d^2(j,j') = 2(1-r_{jj'})$$

em que $r_{jj'}$ é o coeficiente de correlação entre as variáveis *j* e *j*'.

Assim, as proximidades entre as variáveis podem ser interpretadas em termos das suas correlações: os pontos estão próximos se apresentam correlação fortemente positiva $(r_{jj'} \approx 1)$ e muito afastados se ela é fortemente negativa $(r_{jj'} \approx -1)$. Distâncias intermédias correspondem a variáveis independentes $(r_{jj'} \approx 0)$.

As coordenadas das variáveis nos eixos factoriais são dadas por

$$F = X^T \, v$$

As coordenadas das variáveis num eixo são os coeficientes de correlação das variáveis com o eixo. Com efeito, a coordenada $f_{j\alpha}$ de uma variável *j* num eixo α, é dada por

$$f_{j\alpha} = \sum_{i=1}^{n} x_{ij} \, v_{i\alpha}$$

Atendendo à transformação utilizada e ao facto de por construção os vectores *v* serem de média nula e variância unitária, a coordenada da variável *j* no eixo α é dada por

$$f_{j\alpha} = \sum_{i=1}^{n} \frac{(q_{ij} - \bar{q}_j)(q_{ij} - \bar{q}_{j'})}{s_j} v_{i\alpha}$$

$$f_{j\alpha} = r_{j\alpha}$$

em que $r_{j\alpha}$ é o coeficiente de correlação entre a variável *j* e a componente principal α.

ALGORITMO DE ANÁLISE EM COMPONENTES PRINCIPAIS

Descreve-se seguidamente o algoritmo da Análise em Componentes Principais normada:

1 - Transformação da matriz dos dados originais. O quadro *Q* é transformado noutra matriz *X* através da operação de redução das variáveis iniciais:

$$x_{ij} = \frac{1}{\sqrt{n}} \frac{q_{ij} - \overline{q}_j}{s_j}$$

2 - Cálculo da matriz de correlações *R*, cujo elemento genérico é dado por:

$$r_{jj'} = \sum_{i=1}^{n} x_{ij}\ x_{ij'} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(q_{ij} - \overline{q}_j\right)\left(q_{ij'} - \overline{q}_{j'}\right)}{s_j\ s_{j'}}$$

3 - Diagonalização da matriz de correlações de que resultam *p* valores próprios λ_α e *p* vectores próprios u_α:

4 - Cálculo das coordenadas das variáveis nos eixos factoriais, dadas por:

$$f_{j\alpha} = \sum_{i=1}^{n} x_{ij}\ v_{i\alpha}$$

5 - Cálculo das projecções dos indivíduos nos eixos factoriais, dadas por:

$$w_{i\alpha} = \sum_{j=1}^{p} x_{ij}\ u_{j\alpha}$$

6 - Selecção da dimensão do sub-espaço, cuja inércia acumulada explique uma percentagem suficiente da inércia total, de acordo com critérios que se analisam adiante.

7 - Projecção eventual de indivíduos e variáveis em suplementar.

8 -Interpretação dos resultados.

REGRAS DE INTERPRETAÇÃO

Os eixos factoriais resultantes de uma Análise em Componentes Principais constituem uma nova base hierarquizada do espaço engendrado pelos dados, cuja inércia total é dada por:

$$I_g = tr(X^T X) = \sum_{\alpha=1}^{p} \lambda_\alpha$$

Cada eixo é responsável por uma determinada percentagem da inércia da nuvem, dada por:

$$100\frac{\lambda_\alpha}{I_g}$$

O objectivo fundamental da Análise em Componentes Principais é a redução da dimensão dos espaços em jogo. Uma forma cómoda de visualizar a nuvem será projectá-la nos planos definidos pelos eixos factoriais que representem, em conjunto, uma percentagem de inércia considerada suficiente. Existem vários critérios para encontrar o número p_r de eixos a reter, balançando a redução da dimensão do espaço com a necessidade de explicar uma proporção importante da variância total. Apresentam-se a seguir os mais utilizados (isoladamente ou combinados):

1. Numa nuvem esférica, sem alongamentos preferenciais, os valores próprios resultantes de uma análise normada são todos iguais:

   $$\lambda_\alpha = 1 \qquad \forall \alpha \in \{1,..., p\}$$

   Então, pode-se escolher p_r como o número de eixos α, tal que $\lambda_\alpha \geq 1$.

2. Seja τ uma percentagem da inércia total fixada previamente, normalmente 80%. Então p_r é o número de eixos tal que:

$$100 \sum_{j=1}^{p_r} \frac{\lambda_g}{I_g} \geq \tau$$

3. Seja a curva (*scree plot*) que relaciona o número de ordem de cada eixo com o valor próprio que lhe está associado. Se essa curva evidenciar uma estabilização dos valores próprios, pode-se reter apenas os eixos com números de ordem superiores áquele que inicia a estabilização (Fig. 3.1).

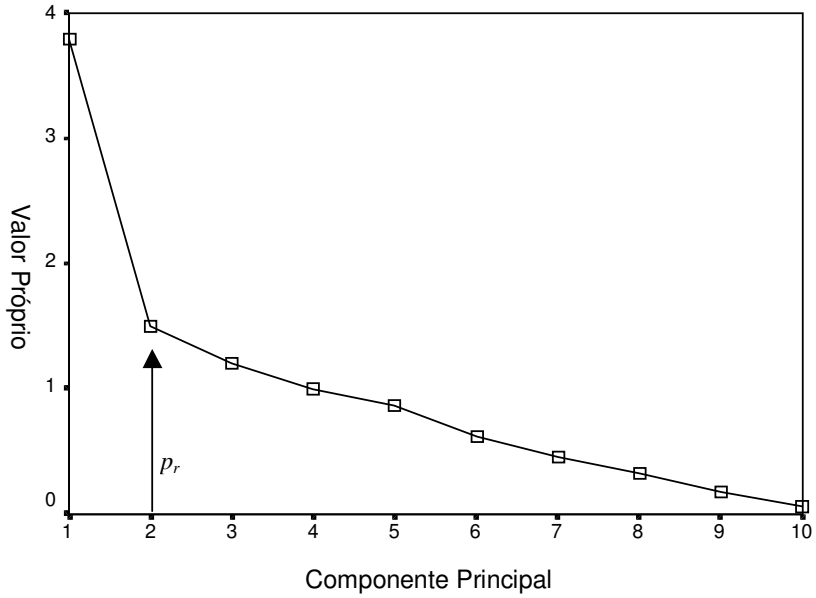

Fig. 3.1 - Distribuição dos valores próprios.

Pode acontecer que existam variáveis mal explicadas nos eixos retidos que apresentem correlações elevadas com eixos não seleccionados, utilizando os critérios referidos anteriormente. Neste caso é aconselhável reter também estes eixos.

A sobreposição das projecções das duas nuvens no mesmo plano torna mais expressiva a interpretação, desde que se tomem algumas precauções. As nuvens têm significados

diferentes pelo que a interpretação das variáveis e dos indivíduos devem ser efectuadas independentemente. As proximidades entre um indivíduo e uma variável não têm um significado matemático muito preciso. No entanto a interpretação dos eixos factoriais, baseada nas correlações que apresentam com as variáveis, permitem relacionar as duas nuvens de uma forma indirecta.

Antes de analisar a posição relativa dos indivíduos ou das variáveis há que verificar a respectiva **qualidade de representação** no plano considerado. A proximidade das projecções não corresponde necessariamente a uma proximidade real (Fig. 3.2).

No caso I as projecções estão próximas embora os indivíduos x_1 e x_2 estejam bastante afastados. Os ângulos θ_1 e θ_2 são grandes. No caso II os ângulos são pequenos, os indivíduos x_1 e x_2 estão próximos das suas projecções, e portanto estão próximos entre si.

O coseno do ângulo formado pelo vector x_i que dá a posição do indivíduo com o plano considerado é uma boa medida da qualidade de representação desse indivíduo.

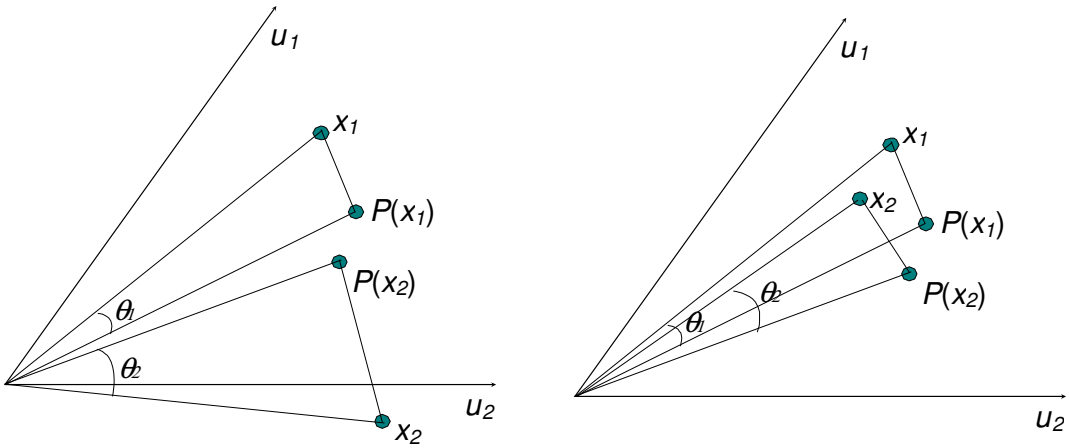

Fig. 3.2 - Projecção ortogonal no plano $u_1 \times u_2$.

Os pontos representativos das variáveis estão na hiperesfera de raio 1. A qualidade de representação de uma variável pode ser avaliada directamente traçando o círculo

unitário: as variáveis posicionadas próximas do plano projectam-se junto à circunferência.

O valor do produto interno dos vectores que unem dois pontos da nuvem em R^n é o coeficiente de correlação entre as variáveis correspondentes (é também o coseno do ângulo entre os dois vectores). Também, como referido antes, as coordenadas das variáveis num eixo são os coeficientes de correlação das variáveis com o eixo.

Assim, a análise das proximidades ou oposições entre variáveis é feita em termos de correlações. No exemplo da Fig. 3.3 estão representadas as projecções de 5 variáveis no plano $u_1\ u_2$ bem como o círculo de correlação.

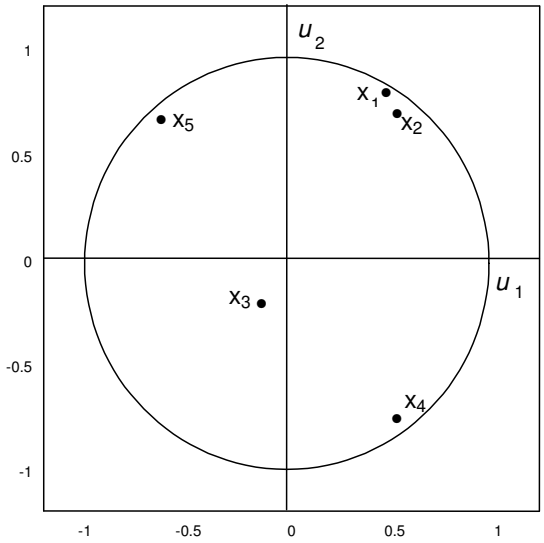

Fig. 3.3 - Círculo de correlação.

As variáveis x_1, x_2, x_4 e x_5 estão bem representadas neste plano, pois encontram-se próximo da circunferência unitária: x_1 e x_2 estão fortemente correlacionadas entre si, mas são independentes das variáveis x_4 e x_5, as quais, entre si apresentam uma correlação negativa forte. Quanto à variável x_3, mal representada neste plano, nada se pode concluir.

Se as coordenadas das variáveis são interpretáveis em termos de correlações, o mesmo não acontece com os indivíduos. A análise da nuvem em R^p faz-se em relação ao centro de gravidade, sendo a distância euclideana a medida que quantifica as relações (proximidades e oposições) entre os pontos.

INDIVÍDUOS E VARIÁVEIS SUPLEMENTARES

Acontece frequentemente que se conhecem os valores das *p* variáveis num conjunto de novos indivíduos. Pode ser interessante posicionar estes novos indivíduos na nuvem já analisada. Noutros casos pode interessar analisar como pontos suplementares os centros de gravidade de indivíduos pertencentes à mesma categoria. Pode ainda acontecer que novas variáveis tenham sido medidas sobre o conjunto dos indivíduos ou então que voluntariamente tenham sido "postas de lado" porque se queria conservar exclusivamente um grupo de características homogéneas.

Em qualquer dos casos anteriores, a interpretação dos factores pode ser enriquecida projectando estas variáveis ilustrativas nos planos principais da nuvem das variáveis activas.

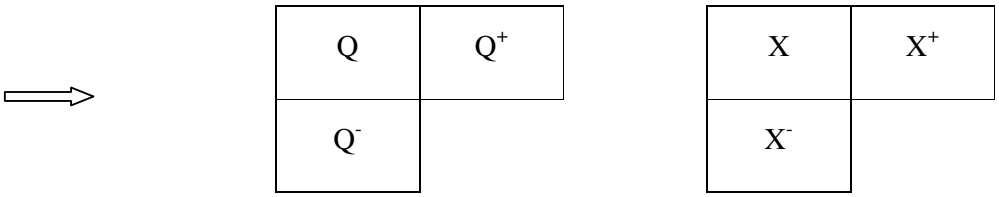

Fig. 3.4 – Exemplo ilustrativo de novos pontos-linha e pontos-coluna associados ao quadro de dados.

Pode acontecer portanto que existam novas linhas e novas colunas bordejando o quadro de dados.

As novas variáveis do quadro Q^+ ficarão posicionadas sobre a esfera de raio 1 de R^n após a transformação.

$$x_{ij}^{+} = \frac{1}{\sqrt{n}} \frac{q_{ij} - \overline{q}_j^{+}}{s_j^{+}}$$

As coordenadas destes pontos num eixo α, é obtida por projecção, isto é, efectuando o produto interno com o vector v_α:

$$(X^{+})^T v_\alpha$$

As novas linhas do quadro Q^- devem ser comparáveis às linhas do quadro analisado:

$$x_{ij}^{-} = \frac{1}{\sqrt{n}} \frac{q_{ij} - q_j}{s_j}$$

As coordenadas dos novos pontos linha são dadas por $X^- u_\alpha$.

EXEMPLOS DE APLICAÇÃO

Os dois casos de aplicação, descritos nos parágrafos seguintes, pretendem exemplificar as potencialidades da Análise em Componentes Principais a dois domínios diferentes: no primeiro exemplo pretende-se uma caracterização do nível de literacia médio em 3 vertentes para 20 países[*]; o segundo caso diz respeito a um conjunto de amostras de um jazigo de sulfuretos caracterizadas por 6 teores químicos.

[*] A Suíça encontra-se dividida nas suas três regiões linguísticas o que se traduz em 22 indivíduos.

1. Nível de literacia médio 3 vertentes para 20 países

Com base no estudo ”Literacia na Era da Informação” (Público, 16 de Junho 2000) foi constituído o quadro de partida para a ACP (Quadro 3.1), o qual consiste nas “notas” médias (numa escala de 1 a 5) obtidas em 20 países para 3 vertentes de literacia – a vertente Documental (DOC) refere-se à capacidade de lidar com informação documental (impressos, mapas, tabelas, horários); a vertente Quantitativa (QUANT) refere-se à capacidade de tratar material escrito envolvendo operações numéricas; a vertente PROSA refere-se à capacidade de compreender e usar a informação em textos corridos.

Quadro 3.1 – Tabela de dados de partida.

PAÍSES	DOC	QUANT	PROSA
Canadá	2.73	2.69	2.71
Alemanha	2.66	2.85	2.47
Irlanda	2.32	2.45	2.42
Holanda	2.76	2.76	2.62
Polónia	1.94	2.06	1.89
Suécia	3.16	3.16	3.07
Suíça (francesa)	2.56	2.75	2.38
Suíça (alemã)	2.53	2.68	2.31
EUA	2.54	2.64	2.62
Austrália	2.59	2.64	2.62
Bélgica (Flandres)	2.66	2.75	2.51
Nova Zelândia	2.51	2.52	2.60
Reino Unido	2.53	2.51	2.47
Chile	1.72	1.79	1.75
República Checa	2.66	3.01	2.32
Dinamarca	2.90	3.01	2.41
Filândia	2.82	2.72	2.79
Hungria	2.12	2.46	1.89
Noruega	3.00	2.97	2.77
Portugal	1.85	2.00	1.91
Eslovénia	1.99	2.15	1.90
Suíça (italiana)	2.48	2.58	2.31

A ACP desta tabela de medidas 22×3 dá origem a 3 eixos, cuja importância é obviamente muito desigual: o eixo 1 explica 92.5% da informação de partida, o eixo 2, 7% e o eixo 3, 0.5%.

A matriz de correlação entre as 3 variáveis é dada no Quadro 3.2, onde se pode verificar que a menor correlação ocorre entre a vertente quantitativa (QUANT) e a vertente textual (PROSA).

Quadro 3.2 – Matriz de correlação entre as variáveis.

	DOC	QUANT	PROSA
DOC	1.00	0.95	0.92
QUANT	0.95	1.00	0.80
PROSA	0.92	0.80	1.00

As coordenadas das variáveis e indivíduos nos eixos encontra-se na Quadro 3.3.

Quadro 3.3 – Coordenadas dos países e das vertentes de literacia nos 3 eixos.

PAÍSES	**Eixo 1**	**Eixo 2**	**Eixo3**
Canadá	0.6072	-0.2699	-0.0270
Alemanha	0.4657	0.2086	0.0561
Irlanda	-0.2929	-0.2181	0.1169
Holanda	0.6169	-0.0718	-0.0583
Polónia	-1.5429	-0.0203	-0.0444
Suécia	1.8182	-0.1527	0.0394
Suíça (francesa)	0.1871	0.1998	0.0448
Suíça (alemã)	0.0230	0.2080	-0.0176
EUA	0.2927	-0.2348	0.1174
Austrália	0.3397	-0.2282	0.0544
Bélgica (Flandres)	0.4067	0.0456	-0.0051
Nova Zelândia	0.1280	-0.3468	0.0420
Reino Unido	0.0116	-0.1878	-0.0678
Chile	-2.1486	-0.1701	-0.0776
República Checa	0.4769	0.5804	0.1037
Dinamarca	0.7898	0.4963	-0.1461
Filândia	0.7985	-0.3276	-0.0683
Hungria	-0.9832	0.4497	0.0672
Noruega	1.1925	0.0009	-0.0953
Portugal	-1.6668	-0.1249	0.0300
Eslovénia	-1.3984	0.0738	-0.0254
Suíça (italiana)	-0.1217	0.0899	-0.0392
VERTENTES	**Eixo 1**	**Eixo 2**	**Eixo3**
DOC	0.9945	0.0379	-0.0975
QUANT	0.9537	0.2947	0.0606
PROSA	0.9398	-0.3392	0.0417

Cruzando o eixo 1 e o eixo 2 obtém-se a Fig. 3.5, cuja interpretação é imediata: o eixo 1 é um eixo de escala, explicado pelas 3 vertentes da literacia, que ordena globalmente os países em relação ao seu nível médio de literacia.

Quanto ao eixo 2, embora explique apenas 7% de informação de partida, separa a vertente QUANTITATIVA (de coordenada +0.3) da vertente PROSA (de coordenada -0.3).

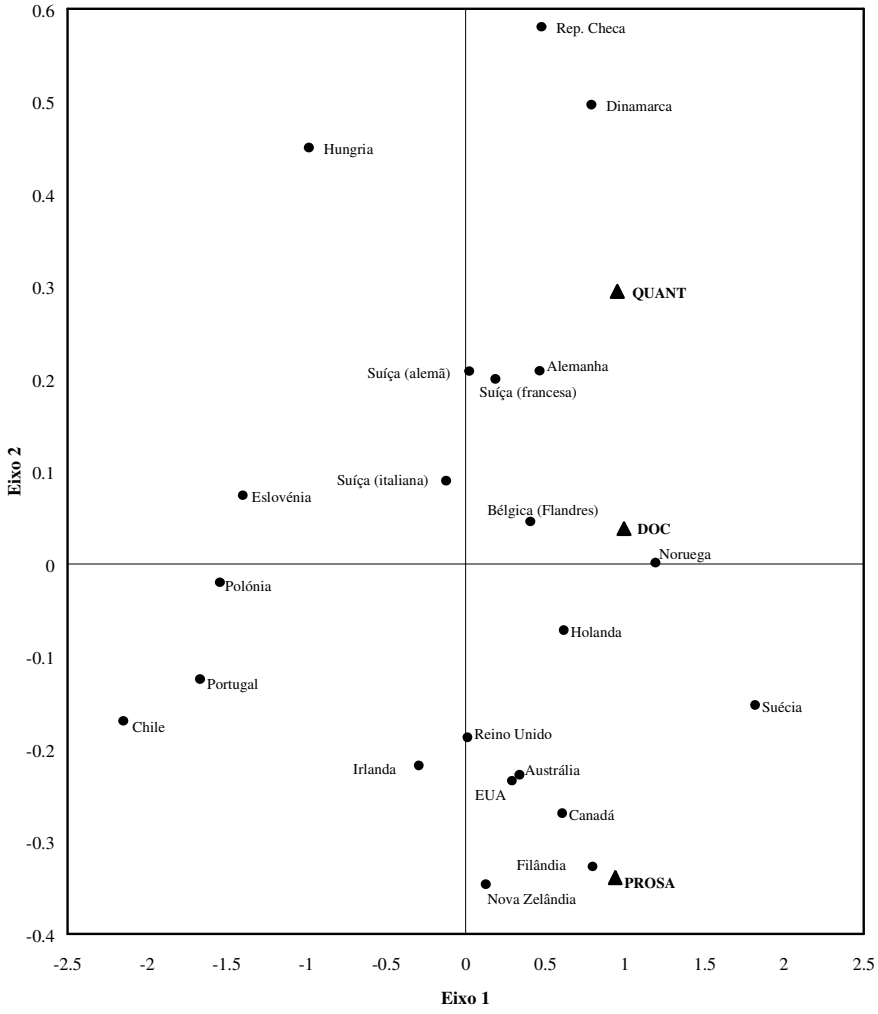

Fig. 3.5 – Projecção das vertentes de literacia e países no plano factorial 1, 2.

Para lá da óbvia ordenação dos países relativamente ao seu nível global de literacia, dada pelas coordenadas no eixo 1, a ACP permite agora sequenciá-los no que diz respeito às 2 vertentes menos correlacionáveis entre si (QUANTITATIVA e PROSA).

Por recordação do Quadro 3.3 relativa às coordenadas no eixo 2, obtém-se a seguinte sequência:

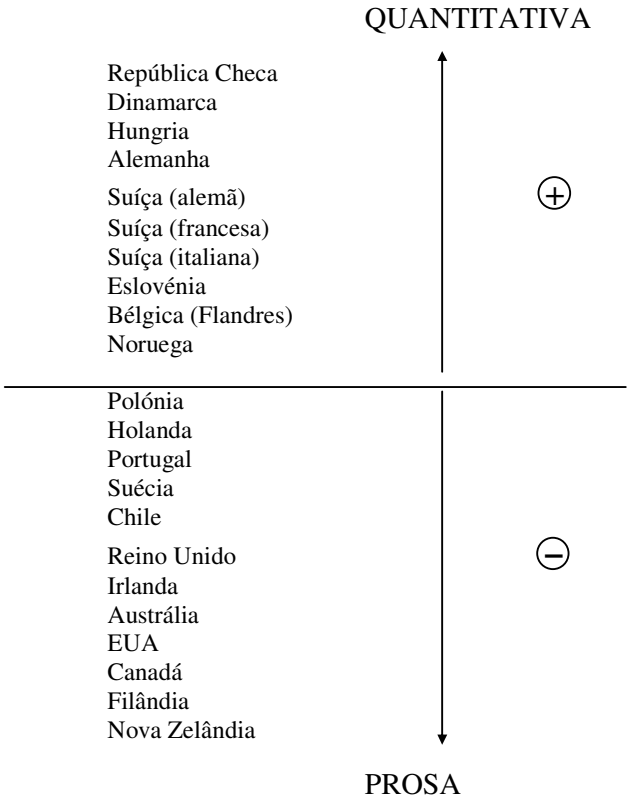

Fig. 3.6 – Sequência de países segundo o eixo 2.

2. Tipologia de um jazigo de sulfuretos

O conjunto de dados de partida é constituído por 172 amostras de um jazigo de sulfuretos, em que se conhecem os teores em Enxofre, Ferro, Cobre, Zinco, Chumbo e Arsénio.

Os eixos factoriais resultantes da Análise em Componentes Principais estão caracterizados no Quadro 3.4.

Quadro 3.4 - Valores próprios e inércia explicada.

EIXO	VALOR PRÓPRIO	INÉRCIA EXPLICADA (%)	INÉRCIA ACUMULADA (%)
1	2.81808	46.97	46.97
2	1.25763	20.96	67.93
3	0.74283	12.38	80.31
4	0.59967	9.99	90.30
5	0.41603	6.93	97.94
6	0.16573	2.76	100.00

No gráfico da Fig. 3.7 estão representadas as projecções das variáveis no primeiro plano factorial, o qual preserva a estrutura topológica das variáveis. Evidenciam-se nitidamente as elevadas correlações positivas entre o Enxofre e o Ferro, o Zinco e o Chumbo e o Cobre e o Arsénio. O Cobre e o Arsénio estão fracamente correlacionados com os outros pares de variáveis (que apresentam forte correlação negativa entre si).

O primeiro eixo traduz a oposição entre os pares Zn/Pb e S/Fe. O segundo eixo separa o par Cu/As das restantes variáveis.

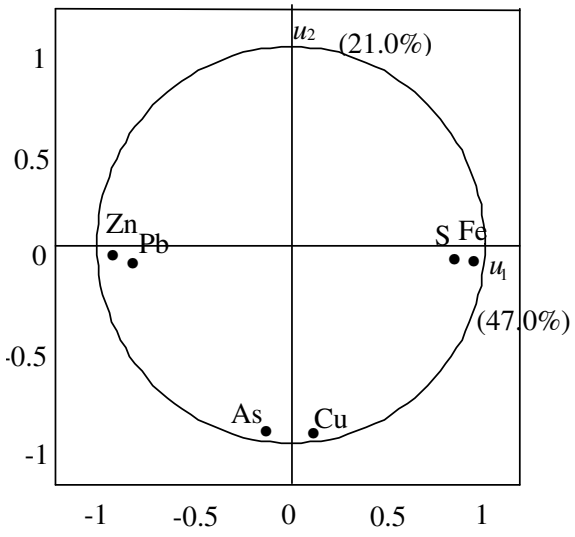

Fig. 3.7 - Projecção dos elementos químicos no primeiro plano factorial.

A projecção dos indivíduos no plano dos dois primeiros eixos factoriais está representada na Fig. 3.8 (só se apresentam as amostras bem explicadas por este plano). A mesma figura permite diferenciar três grupos de amostras: aquelas que são mais ricas em Zn e Pb (grupo A), ou em Cobre e Arsénio (grupo B) e as que são mais pobres nestes quatro elementos (grupo C).

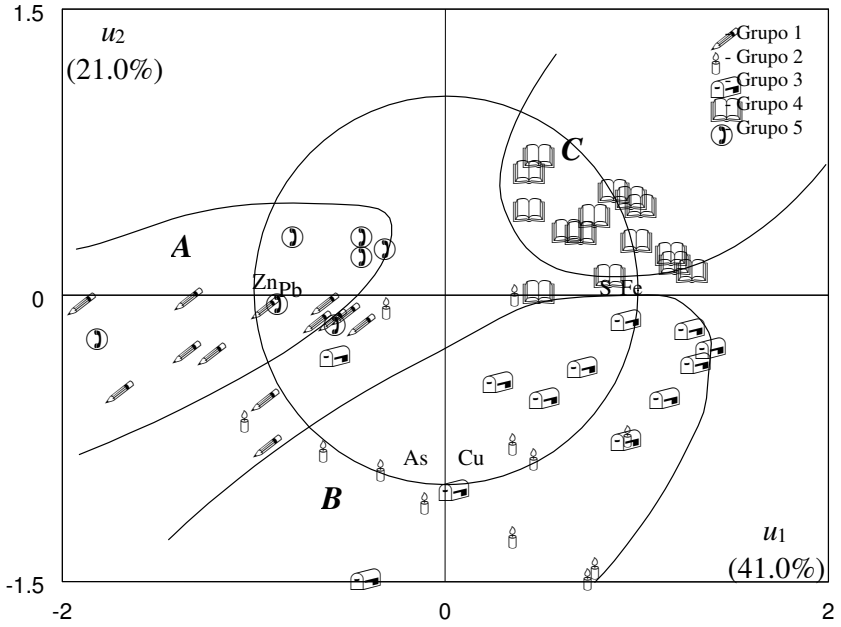

Fig. 3.8 – Projecção das amostras no primeiro plano factorial (os indivíduos estão representados pelos códigos dos grupos a que pertencem – vd. CLASSIFICAÇÃO SOBRE OS FACTORES).