Faculdade de Engenharia da Universidade do Porto

Análise Conjunta de Quadros de Dados: Comparação de Alguns Métodos

Patrícia Carvalho Carvalhido
Licenciada em Matemática (ramo educacional) pela
Faculdade de Ciências da Universidade do Porto

Dissertação submetida para satisfação parcial dos
requisitos do grau de mestre em
Estatística Aplicada e Modelação

Dissertação realizada sob a supervisão da
Professora Doutora Adelaide Figueiredo,
da Faculdade de Economia da Universidade do Porto

Porto, Maio de 2005

Resumo

Na Análise Multivariada de Dados é habitual dispor-se de T quadros de dados quantitativos obtidos em diferentes "ocasiões", em que cada um é constituído por n indivíduos (linhas) e p variáveis (colunas). Estes quadros não necessitam de ser cronológicos, isto é, a terceira dimensão pode não ser o tempo. Neste caso, cada quadro representa uma entidade.

Pretende-se a comparação global dos quadros de dados, bem como o estudo da eventual existência de uma estrutura comum a estes. Se os dados forem temporais interessa analisar as tendências evolutivas dos indivíduos ou das variáveis.

O objectivo deste trabalho é o estudo e a comparação de alguns métodos de Análise Conjunta de Quadros de Dados: a metodologia STATIS ("Structuration de Tableaux À Trois Indices de la Statistique"), a Análise Factorial Múltipla e a Dupla Análise em Componentes Principais, tendo por base a Análise em Componentes Principais.

Este estudo é complementado com uma aplicação destes métodos a dados reais.

Abstract

In Multivariate Data Analysis it is usual to have T quantitative data tables, which are obtained in different "occasions". Each table is composed of n individuals (rows) and p variables (columns). These tables don't need to be chronological, which means that the third dimension may not be time. In this case each table represents an entity.

The goal is to compare these tables, as well as investigate if there is a common structure between them. If the data are chronological, the individuals' behaviour, as well as the variables' should be analysed.

The expected goal behind this essay is to study and compare three methods of Three-Way Data Analysis, such as: STATIS methodology ("Structuration de Tableaux À Trois Indices de la Statistique"), Multiple Factor Analysis and Double Principal Component Analysis. The main tool used to implement these methods is Principal Component Analysis.

This essay is concluded with an application of these methods to real data.

Agradecimentos

O trabalho apresentado nesta dissertação é o culminar de dois anos no Mestrado em Estatística Aplicada e Modelação da Faculdade de Engenharia da Universidade do Porto (2002/2004). Durante este período tive a oportunidade de conhecer professores exemplares, quer a nível científico, quer a nível pessoal, que marcaram o meu percurso neste Mestrado. A facilidade de acesso a livros, artigos científicos e equipamento informático disponibilizados por esta Faculdade deram-me a possibilidade de viver uma realidade académica desconhecida até então. O elevado espírito de camaradagem e até mesmo de amizade entre os colegas de Mestrado nunca hão-de ser esquecidos.

Gostaria de expressar o meu reconhecimento a todos os que me ajudaram na realização deste projecto.

À Professora Adelaide Figueiredo, minha orientadora, pelo tema sugerido, pela simpatia e disponibilidade prestadas.

À Professora Fernanda Sousa pelas palavras amigas e encorajadoras que sempre teve para comigo.

À Professora Teresa Arede e à Professora Maria do Carmo Coimbra pela sua ajuda e disponibilidade.

Ao Professor Santos Marques pelo incentivo, dedicação e encorajamento ao longo dos últimos seis anos.

À Solange pela infinita paciência e dedicação.

À Xana Lisboa pela sua ajuda preciosa nas traduções.

Ao Jorge, pela sua dedicação e amizade, troca de opiniões, sugestões e muita ajuda com o LaTeX.

À minha família, pais e irmão, e ao João pelo carinho, compreensão e incentivo...

Índice

Índice de Figuras

Índice de Tabelas

Índice de Abreviaturas

ACP	Análise em Componentes Principais
ACT	Analyse Conjointe de Tableaux
AFM	Análise Factorial Múltipla
DACP	Dupla Análise em Componentes Principais
HS	Hilbert-Schmidt
RV	Coeficiente de correlação vectorial entre objectos
STATIS	Structuration de Tableaux à Trois Indices de la Statistique

Índice de Notações

X	quadro de dados de dimensão $n \times p$
X'	transposta da matriz X
$\boldsymbol{x^j}$	variável de dimensão n
$\boldsymbol{x_i}$	indivíduo de dimensão p
I_n	matriz identidade de ordem n
$\boldsymbol{g}$	centro de gravidade
$\mathbf{1}_n$	vector com todas as componentes iguais a 1, de dimensão n
Y	quadro de dados centrados
V	matriz de variâncias e covariâncias
Q_{1/s^2}	matriz diagonal dos inversos das variâncias entre variáveis
$Q_{1/s}$	matriz diagonal dos inversos dos desvios padrões entre variáveis
Z	quadro de dados centrados e reduzidos
$r(\boldsymbol{x^j}, \boldsymbol{x^k})$	coeficiente de correlação linear entre $\boldsymbol{x^j}$ e $\boldsymbol{x^k}$
R	matriz de correlações entre as variáveis
F	espaço dos indivíduos, de dimensão p
N_I	nuvem dos indivíduos associada ao quadro X
Q	métrica associada ao produto escalar entre indivíduos
$\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle_Q$	produto escalar no espaço F
$d(\boldsymbol{x_i}, \boldsymbol{x_j})$	distância entre os indivíduos $\boldsymbol{x_i}$ e $\boldsymbol{x_j}$
W	matriz dos produtos escalares entre indivíduos
$\mathcal{I}_{\boldsymbol{g}}$	inércia total da nuvem de indivíduos
$\mathcal{I}_{\boldsymbol{a}}$	inércia da nuvem de indivíduos relativamente ao ponto $\boldsymbol{a}$
E	espaço das variáveis, de dimensão n
N_J	nuvem das variáveis associada ao quadro X
D	métrica associada ao produto escalar entre variáveis
$\boldsymbol{u_k}$	eixo principal de dimensão p sobre o qual se projecta um indivíduo

$F_{\boldsymbol{u_k}}(i)$	projecção do indivíduo $\boldsymbol{x_i}$ sobre o eixo $\boldsymbol{u_k}$
$\boldsymbol{F_u(i)}$	vector das projecções do indivíduo $\boldsymbol{x_i}$ sobre os eixos $\boldsymbol{u_1}, \ldots, \boldsymbol{u_q}$
$\mathcal{D}$	deformação em projecção
$\boldsymbol{F_{u_k}}$	componente principal do espaço E
$\boldsymbol{z_k}$	factor principal
$\boldsymbol{v_k}$	eixo principal de dimensão n sobre o qual se projecta uma variável
$E_{\boldsymbol{v_k}}(j)$	projecção da variável $\boldsymbol{x^j}$ sobre o eixo $\boldsymbol{v_k}$
$\boldsymbol{E_v(j)}$	vector das projecções da variável $\boldsymbol{x^j}$ sobre os eixos $\boldsymbol{v_1}, \ldots, \boldsymbol{v_r}$
F^*	espaço dual de F
E^*	espaço dual de E
N_I^*	nuvem dual de N_I
N_J^*	nuvem dual de N_J
$\boldsymbol{u_k^*}$	eixos principais de N_I^*
$\boldsymbol{v_k^*}$	eixos principais de N_J^*
CTA_i^k	contribuição absoluta do indivíduo $\boldsymbol{x_i}$ para a formação de $\boldsymbol{u_k}$
CTA_k^j	contribuição absoluta da variável $\boldsymbol{x^j}$ para a formação de $\boldsymbol{F_{u_k}}$
$(\boldsymbol{u_k}, \boldsymbol{u_l})$	plano formado pelos eixos principais $\boldsymbol{u_k}$ e $\boldsymbol{u_l}$
$\rho_i^{k,l}$	contribuição relativa do plano $(\boldsymbol{u_k}, \boldsymbol{u_l})$ ao indivíduo $\boldsymbol{x_i}$
ρ_i^k	contribuição relativa do eixo principal $\boldsymbol{u_k}$ ao indivíduo $\boldsymbol{x_i}$
γ_k^j	contribuição relativa da componente principal $\boldsymbol{F_{u_k}}$ à variável $\boldsymbol{x^j}$
X_t	quadro de dados no instante t $(t = 1, \ldots, T)$
Q_t	métrica associada ao produto escalar entre indivíduos de X_t
D_t	métrica associada ao produto escalar entre variáveis de X_t
$(\boldsymbol{x^j})^t$	j-ésima variável do quadro X_t
$(\boldsymbol{x_i})^t$	i-ésimo indivíduo do quadro X_t
$\mathcal{W}_t$	objecto que corresponde à matriz dos produtos escalares entre indivíduos do quadro X_t
$\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS}$	produto escalar de Hilbert-Schmidt entre $\mathcal{W}_t$ e $\mathcal{W}_{t'}$
$d_{HS}(\mathcal{W}_t, \mathcal{W}_{t'})$	distância entre os objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$
$\|\mathcal{W}_t\|_{HS}$	norma do objecto $\mathcal{W}_t$
S	matriz dos produtos escalares entre objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$
$\widetilde{S}$	matriz dos produtos escalares entre objectos normados $\mathcal{W}_t/\|\mathcal{W}_t\|_{HS}$ e $\mathcal{W}_{t'}/\|\mathcal{W}_{t'}\|_{HS}$
$RV(t, t')$	coeficiente de correlação vectorial entre $\mathcal{W}_t$ e $\mathcal{W}_{t'}$
Δ	métrica associada ao produto escalar entre objectos
$\mathcal{W}$	matriz compromisso dos objectos $\mathcal{W}_1, \ldots, \mathcal{W}_T$

$\mathcal{V}_t$	objecto que corresponde à matriz de variâncias e covariâncias do quadro X_t
$\mathcal{Z}$	matriz dos produtos escalares entre os objectos $\mathcal{V}_t$ e $\mathcal{V}_{t'}$
$\widetilde{\mathcal{Z}}$	matriz dos produtos escalares entre objectos normados $\mathcal{V}_t/\|\mathcal{V}_t\|_{HS}$ e $\mathcal{V}_{t'}/\|\mathcal{V}_{t'}\|_{HS}$
$\mathcal{V}$	matriz compromisso dos objectos $\mathcal{V}_1, \ldots, \mathcal{V}_T$
$\mathcal{N}_I$	nuvem dos indivíduos definida pelos quadros $X_1, \ldots, X_T$
$\widetilde{\boldsymbol{x}_i}$	vector de $\mathbb{R}^{\sum_{t=1}^{T} p_t}$ representativo do i-ésimo indivíduo definido pelos quadros $X_1, \ldots, X_T$
$\mathcal{N}_I^t$	nuvem dos indivíduos definida pelo quadro de dados X_t
$\mathcal{N}_i^T$	nuvem das T imagens do i-ésimo indivíduo
$\mathcal{N}_I^*$	nuvem dos centros de gravidade das nuvens $\mathcal{N}_i^T$
$\mathcal{X}$	matriz da justaposição dos quadros $X_1, \ldots, X_T$
$\mathcal{Q}$	métrica associada ao produto escalar entre indivíduos da matriz $\mathcal{X}$
$\mathcal{N}_J^t$	nuvem dos indivíduos definida pelo quadro de dados X_t
$\mathcal{N}_J$	nuvem dos grupos de variáveis definida pelos quadros $X_1, \ldots, X_T$
$\eta_g^2(X_t)$	quadrado da norma do objecto representativo do quadro X_t utilizado na AFM
$\mathcal{L}_g(\boldsymbol{u}, X_t)$	medida de ligação entre a variável $\boldsymbol{u}$ e as variáveis do quadro X_t
$\mathcal{L}_g(X_t, X_{t'})$	medida de ligação entre as variáveis dos quadros X_t e $X_{t'}$
$\boldsymbol{g}_t$	centro de gravidade do quadro X_t
G	a matriz de dimensão $T \times p$, formada pelos centros de gravidade das nuvens $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$
$\Phi(., \tau)$	indicador da perda média de inércia das nuvens $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$
$f_1(\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_q})$	índice de qualidade de representação das nuvens $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$

Introdução

Os métodos de Análise de Dados demonstraram nos últimos vinte anos a sua eficácia no estudo de grandes quantidades de informação.

A Análise de Dados começou a dar os primeiros passos com Pearson [36], em 1901, através do estudo de linhas e planos de melhor ajustamento a um conjunto de pontos no espaço. Spearman [43] lançou em 1904 os fundamentos da análise factorial, através da psicologia. Só mais tarde é que Hotelling [21] desenvolveu os estudos de Pearson e Spearman, lançando em 1933 os fundamentos da Análise em Componentes Principais.

A Análise de Dados começa então a ser utilizada em diversas áreas, nomeadamente na psicologia, economia e biologia, com o contributo da equipa de Benzécri [1]. O desenvolvimento da informática acelera este processo, permitindo o tratamento de volumosas quantidades de informação, que sem o auxílio dos computadores tornar-se-ia muito moroso ou até mesmo impraticável.

Por volta de 1967, Harman [20] e Morrison [32] contribuem com desenvolvimentos na área da Análise Multivariada, nomeadamente na sua ligação com a Análise em Componentes Principais, a Análise Factorial e a Análise Discriminante.

É nas décadas de 70 e 80 que Escoufier [13], Bouroche [2], L'Hermier des Plantes [30], Robert e Escoufier [38], Jaffrenou [22], Foucart [16] e Escofier e Pagès [10], entre outros, começam a desenvolver os seus estudos nos métodos de tratamento de tabelas multidimensionais. Em 1981, Glaçon [17] compara vários métodos que envolvam o estudo simultâneo de vários quadros de dados.

A escola anglo-saxónica também desenvolve numerosos trabalhos sobre a Análise Conjunta de Quadros de Dados ("Three-Way Methods" ou "Multidimensional Scaling"). Kiers [25] e Kroonenberg [26] fazem um estudo detalhado de alguns métodos nas suas obras.

O conteúdo da presente dissertação insere-se na área da Análise Multivariada de Dados, nomeadamente na Análise Conjunta de Quadros de Dados, tendo como

objectivo descrever a metodologia STATIS (método STATIS e STATIS dual), a Análise Factorial Múltipla (AFM) e a Dupla Análise em Componentes Principais (DACP).

O **capítulo 1** consiste na descrição da Análise em Componentes Principais, uma vez que esta técnica serve de base aos métodos que irão ser posteriormente estudados. Procura-se ainda estabelecer a dualidade entre o espaço dos indivíduos e o espaço das variáveis.

O **capítulo 2** é dedicado à descrição da metodologia STATIS-ACT ("Analyse Conjointe de Tableaux"). Esta metodologia permite a exploração simultânea de vários quadros de dados quantitativos. Se os quadros forem recolhidos em diferentes "instantes" sobre os mesmos indivíduos, aplicar-se-á o método STATIS. No caso de os grupos de variáveis serem os mesmos para todos os quadros, o método mais adequado é o STATIS dual. Ambos os métodos devem ser aplicados quando todos os quadros de dados possuírem os mesmos indivíduos e as mesmas variáveis. A metodologia STATIS considera a distância euclidiana Hilbert--Schmidt entre quadros de dados como medida de semelhança entre estes. Cada um dos quadros representa um "instante" temporal ou então uma determinada entidade.

O **capítulo 3** é dedicado ao estudo da Análise Factorial Múltipla. Este método é semelhante ao método STATIS, embora também possa ser aplicado a dados qualitativos. Os grupos de variáveis podem diferir ao longo dos quadros e os respectivos coeficientes de ponderação são os inversos dos primeiros valores próprios associados à ACP de cada quadro, no sentido de ponderar o papel desempenhado pelos quadros durante a análise em causa.

O **capítulo 4** destina-se ao estudo da Dupla Análise em Componentes Principais. Este é um método bastante simples de ser executado, uma vez que tem por base essencial a Análise em Componentes Principais, no entanto, o seu campo de aplicação torna-se bastante restrito uma vez que está limitado a quadros que cruzem os mesmos indivíduos com as mesmas variáveis ao longo do tempo. Se os quadros não forem cronológicos a interpretação tornar-se-á bastante mais difícil.

Posteriormente, no **capítulo 5** far-se-á uma comparação destes métodos,

evidenciando vantagens e limitações de cada um deles.

No **capítulo 6**, o estudo é complementado com uma aplicação dos quatro métodos citados a indicadores de desenvolvimento em países europeus. Os dados são provenientes do World Bank Group [46]. Os programas utilizados nestas análises são o Matlab (versão 6.5) e o SPAD ("Système Pour l' Analyse des Données"), na versão 5.5. Este último foi desenvolvido pela escola francesa e destina-se ao tratamento de dados univariado, bivariado e multivariado. No que respeita à Análise Conjunta de Quadros de Dados, este programa trata a metodologia STATIS (método STATIS e STATIS dual) e a Análise Factorial Múltipla. Todos os resultados fornecidos pelo SPAD foram confirmados através do Matlab, com o objectivo de descobrir como funcionavam os algoritmos deste programa. Além disso, todos os resultados que não eram fornecidos pelo SPAD foram calculados no Matlab. Nesta aplicação, procurou-se cruzar a informação fornecida pelos quatro métodos no sentido de tentar averiguar se estes forneciam resultados mais ou menos semelhantes.

Capítulo 1

Análise em Componentes Principais

1.1 Introdução

A Análise em Componentes Principais (ACP) foi inicialmente introduzida por Pearson [36] em 1901 e formalizada por Hotelling [21] em 1933. É um método estatístico multivariado, aplicável a variáveis do tipo quantitativo, que consiste em transformar um conjunto de variáveis correlacionadas entre si num conjunto de variáveis não correlacionadas, denominadas componentes principais. Estas são combinações lineares das variáveis do conjunto inicial, calculadas por ordem decrescente de importância, sendo a primeira componente principal a variável que retém maior informação do quadro de dados, seguida da segunda componente principal, e assim sucessivamente.

O objectivo desta técnica é reter as primeiras componentes, obtendo-se um novo quadro de dados mais reduzido com o mínimo de perda de informação.

1.2 Descrição dos dados e suas características

O **quadro de dados** X consta de n indivíduos e p variáveis:

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^p \end{pmatrix}.$$

x_i^j é o valor que o i-ésimo indivíduo assume na j-ésima variável ($i = 1, ..., n$ e $j = 1, ..., p$), a j-ésima variável é o vector de $\mathbb{R}^n$

$$\boldsymbol{x^j} = \begin{pmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{pmatrix}$$

e o i-ésimo indivíduo o vector de $\mathbb{R}^p$

$$\boldsymbol{x_i}' = \begin{pmatrix} x_i^1 & x_i^2 & \dots & x_i^p \end{pmatrix}.$$

Seja D a matriz diagonal positiva dos pesos atribuídos aos n indivíduos

$$D = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_n \end{pmatrix} \tag{1.1}$$

com $0 < p_i < 1$ e $\sum_{i=1}^n p_i = 1$.

Estes pesos são atribuídos consoante a importância relativa de cada indivíduo. É usual os indivíduos terem todos a mesma importância, logo $D = \frac{1}{n} I_n$, com I_n a matriz identidade de ordem n.

O **centro de gravidade** ou **baricentro** é o vector das médias aritméticas de cada variável definido por

$$\boldsymbol{g}' = \begin{pmatrix} \bar{x}^1 & \bar{x}^2 & \dots & \bar{x}^p \end{pmatrix}$$

com $\bar{x}^j = \sum_{i=1}^n p_i x_i^j$. Na forma matricial tem-se que

$$\boldsymbol{g} = X' D \mathbf{1}_n,$$

em que $\mathbf{1}_n$ é o vector de $\mathbb{R}^n$ com todas as componentes iguais a 1.

Seja Y o **quadro de dados centrado** associado a X com $y_i^j = x_i^j - \bar{x}^j$. Então

$$Y = X - \mathbf{1}_n \boldsymbol{g}' = X - \mathbf{1}_n \left(X' D \mathbf{1}_n\right)' = X - \mathbf{1}_n \mathbf{1}_n' D' X = \left(I_n - \mathbf{1}_n \mathbf{1}_n' D\right) X.$$

No caso particular em que $D = \frac{1}{n} I_n$ tem-se que

$$Y = \left(I_n - \frac{\mathbf{1}_n \mathbf{1}_n'}{n}\right) X. \tag{1.2}$$

A covariância entre as variáveis $\boldsymbol{x^j}$ e $\boldsymbol{x^k}$ é expressa por:

$$s_{jk} = s\left(\boldsymbol{x^j}, \boldsymbol{x^k}\right) = \sum_{i=1}^{n} p_i \left(x_i^j - \bar{x}^j\right)\left(x_i^k - \bar{x}^k\right), \tag{1.3}$$

e $s_j^2 = s^2(\boldsymbol{x^j})$ $(j = 1, \ldots, p)$, logo a **matriz de variâncias e covariâncias** pode exprimir-se em termos matriciais na forma:

$$V = \begin{pmatrix} s_1^2 & s_{12} & \ldots & s_{1p} \\ s_{21} & s_2^2 & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_p^2 \end{pmatrix} = Y'DY = X'DX - \boldsymbol{gg'}. \tag{1.4}$$

V é uma matriz simétrica, definida positiva.

Seja Q_{1/s^2} a matriz diagonal dos inversos das variâncias das variáveis,

$$Q_{1/s^2} = \begin{pmatrix} 1/s_1^2 & 0 & \ldots & 0 \\ 0 & 1/s_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/s_p^2 \end{pmatrix} \tag{1.5}$$

e $Q_{1/s}$ a matriz diagonal dos inversos dos desvios padrões das variáveis.

O **quadro de dados centrados e reduzidos** (isto é, tendo $\frac{x_i^j - \bar{x}^j}{s_j}$ na posição de z_i^j) Z é tal que

$$Z = YQ_{1/s}. \tag{1.6}$$

A **matriz de correlações** entre as p variáveis em que

$$r_{jk} = r(\boldsymbol{x^j}, \boldsymbol{x^k}) = \frac{s_{jk}}{s_j \cdot s_k}$$

é representada por

$$R = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}.$$

Então

$$R = Q_{1/s}VQ_{1/s} = Q_{1/s}Y'DYQ_{1/s} = Z'DZ \tag{1.7}$$

e R coincide com V quando os dados iniciais já estão centrados e reduzidos pois $s_j^2 = 1$ $(j = 1, \ldots, p)$ logo $Q_{1/s} = I_p$.

1.3 Espaço dos indivíduos

Seja F o espaço dos indivíduos de dimensão p que contém cada um dos n indivíduos $\boldsymbol{x_i}' = (x_i^1, x_i^2, \ldots, x_i^p)$, $i = 1, \ldots, n$. Este conjunto forma uma nuvem de pontos em F, designada por N_I, em que o centro de gravidade $\boldsymbol{g}$ coincide com a origem dos eixos O se os dados estiverem centrados (figura 1.1).

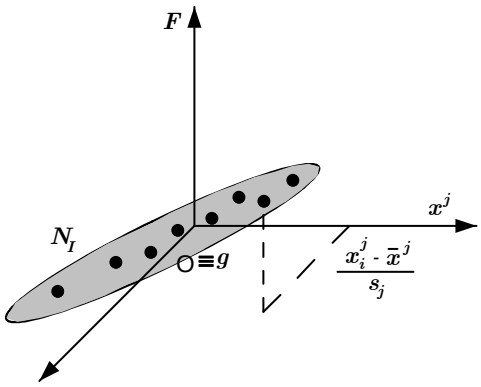


Figura 1.1 Nuvem de indivíduos no espaço F.

A análise do quadro de dados pela óptica dos indivíduos equivale à visualização do mesmo segundo as linhas.

O espaço F deve estar munido de uma métrica que permita determinar as distâncias entre quaisquer indivíduos.

Dada Q, matriz $p \times p$, simétrica, definida positiva, define-se o **produto escalar** em F da seguinte forma:

$$\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle_Q = \boldsymbol{x_i}' Q \boldsymbol{x_j}.$$

Então $\|\boldsymbol{x_i}\|_Q = \sqrt{\langle \boldsymbol{x_i}, \boldsymbol{x_i} \rangle_Q}$ e a **distância entre dois indivíduos** $\boldsymbol{x_i}$ e $\boldsymbol{x_j}$ é dada por:

$$d\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) = \|\boldsymbol{x_i} - \boldsymbol{x_j}\|_Q = \sqrt{\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)' Q \left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)}.$$

A matriz Q é a **métrica** associada ao produto escalar entre indivíduos e a **matriz dos produtos escalares entre indivíduos** é designada por $W = XQX'$ de termo geral:

$$w_{i,j} = \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle_Q = \boldsymbol{x_i}' Q \boldsymbol{x_j}. \tag{1.8}$$

Em ACP as métricas mais utilizadas são:

- $Q = I_p$, que corresponde a utilizar o produto escalar euclidiano:

$$\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle_{I_p} = \boldsymbol{x_i}' I_p \boldsymbol{x_j} = \boldsymbol{x_i}' \boldsymbol{x_j} = \sum_{k=1}^{p} x_i^k x_j^k;$$

- $Q = Q_{1/s^2}$ definida em (1.5). Com esta métrica a distância entre dois indivíduos não depende das unidades de medida.

Propriedade 1.1 *Seja Q matriz $p \times p$ simétrica definida positiva. Como Q pode ser decomposta na forma $T'T$ em que T é uma matriz triangular (decomposição de Cholesky), vem que:*

$$\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle_Q = \boldsymbol{x_i}' Q \boldsymbol{x_j} = \boldsymbol{x_i}' T' T \boldsymbol{x_j} = (T\boldsymbol{x_i})' T \boldsymbol{x_j} = (T\boldsymbol{x_i})' I_p T \boldsymbol{x_j} = \langle T\boldsymbol{x_i}, T\boldsymbol{x_j} \rangle_{I_p}.$$

Isto é, tudo se passa como se se tivesse utilizado a métrica $Q = I_p$ sobre os dados transformados, ou seja, sobre XT'. No caso $Q = Q_{1/s^2}$ verifica-se que $T = Q_{1/s}$ e portanto:

$$\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle_{Q_{1/s^2}} = \langle Q_{1/s} \boldsymbol{x_i}, Q_{1/s} \boldsymbol{x_j} \rangle_{I_p}.$$

Assim utiliza-se a métrica I_p sobre $XQ'_{1/s} = XQ_{1/s}$. Se a matriz X já estiver centrada então por (1.6), $XQ_{1/s}$ corresponde ao quadro de dados centrado e reduzido.

Se no início se proceder à centragem e redução do quadro de dados X (calculando Y a partir de (1.2) e Z em (1.6)), utiliza-se a métrica I_p; se só se centrar, utiliza-se a métrica Q_{1/s^2} definida em (1.5). A distância entre dois indivíduos tanto por um procedimento como pelo outro é igual e não é mais do que a distância euclidiana usual, logo quanto mais próximos estiverem entre si maior será o grau de semelhança entre eles.

Note-se que a centragem dos dados não modifica a problemática inerente a este tipo de análise, apenas faz com que a origem dos eixos coincida com o baricentro $\boldsymbol{g}$. Se as unidades de medida em que as variáveis estão expressas forem muito diferentes, convém então reduzir os dados, ou seja, dividir cada uma das observações da coluna j pelo desvio-padrão s_j. Desta forma, todas as variáveis terão a mesma variabilidade e, consequentemente, a mesma influência no cálculo das distâncias entre indivíduos.

Define-se **inércia total da nuvem de indivíduos** como a média ponderada dos quadrados das distâncias dos indivíduos ao centro de gravidade da nuvem:

$$\mathcal{I}_{\boldsymbol{g}} = \sum_{i=1}^{n} p_i \| (\boldsymbol{x_i} - \boldsymbol{g}) \|_Q^2 = \sum_{i=1}^{n} p_i (\boldsymbol{x_i} - \boldsymbol{g})' Q (\boldsymbol{x_i} - \boldsymbol{g}). \tag{1.9}$$

Naturalmente, se os dados estão centrados então $\boldsymbol{g} = \mathbf{0}$.

A **inércia da nuvem de indivíduos relativamente ao ponto a** é expressa por

$$\mathcal{I}_{\boldsymbol{a}} = \sum_{i=1}^{n} p_i \| (\boldsymbol{x_i} - \boldsymbol{a}) \|_Q^2. \tag{1.10}$$

Surge então o **Teorema de Huyghens**:

Teorema 1.1

$$\mathcal{I}_{\boldsymbol{a}} = \mathcal{I}_{\boldsymbol{g}} + \|\boldsymbol{g} - \boldsymbol{a}\|_Q^2. \tag{1.11}$$

Note-se que:

- $\mathcal{I}_{\boldsymbol{g}} = tr(QV) = tr(VQ)$ (onde $tr(A)$ designa o traço da matriz A);

$$\begin{aligned} tr(QV) &= tr(QY'DY) \\ &= tr\left(Q \sum_{i=1}^{n} p_i \boldsymbol{y_i} \boldsymbol{y_i}' \right) \\ &= tr\left(\sum_{i=1}^{n} Q p_i \boldsymbol{y_i} \boldsymbol{y_i}' \right) \\ &= tr\left(\sum_{i=1}^{n} p_i \underbrace{Q\boldsymbol{y_i}}_{(p\times 1)} \underbrace{\boldsymbol{y_i}'}_{(1\times p)} \right) \\ &= tr\left(\underbrace{\sum_{i=1}^{n} p_i \boldsymbol{y_i}' Q \boldsymbol{y_i}}_{escalar} \right) \\ &= \sum_{i=1}^{n} p_i \boldsymbol{y_i}' Q \boldsymbol{y_i} \overset{(1.9)}{=} \mathcal{I}_{\boldsymbol{g}} \quad \blacksquare \end{aligned}$$

- $\mathcal{I}_{\boldsymbol{g}} = tr(WD) = tr(DW)$ se $\boldsymbol{g} = \mathbf{0}$;

$$\begin{aligned} \mathcal{I}_{\boldsymbol{g}} &\overset{(1.3)}{=} tr(QV) \\ &\overset{(1.4)}{=} tr(QX'DX) \\ &= tr(DXQX') \\ &\overset{(1.8)}{=} tr(DW) \\ &= tr(WD) \quad \blacksquare \end{aligned}$$

- Se $Q = I$, então $\mathcal{I}_{\boldsymbol{g}} = tr(QV) = trV$ (ou seja, é a soma das variâncias das p variáveis);
- Se $Q = Q_{1/s^2}$, então $\mathcal{I}_{\boldsymbol{g}} = tr\,(R) = p$.

$$\begin{aligned} \mathcal{I}_{\boldsymbol{g}} &= tr\left(Q_{1/s^2}V\right) \\ &= tr\left(Q_{1/s}VQ_{1/s}\right) \\ &\overset{(1.7)}{=} tr(R) = p \quad \blacksquare \end{aligned}$$

1.4 Espaço das variáveis

Seja E o espaço das variáveis, de dimensão n, que contém cada uma das p variáveis $\boldsymbol{x^j} = (x_1^j, x_2^j, \ldots, x_n^j)'$, com $j = 1, \ldots, p$. Este conjunto forma uma nuvem de pontos em E, designada por N_J.

A análise do quadro de dados pela óptica das variáveis equivale à visualização do mesmo segundo as colunas.

A métrica utilizada neste espaço será a matriz dos pesos D definida em (1.1), $n \times n$, simétrica, definida positiva. Esta escolha não deixa muitas dúvidas, pelas seguintes razões:

- *O produto escalar de duas variáveis centradas é igual à sua covariância.*

$$\langle \boldsymbol{x^j}, \boldsymbol{x^k} \rangle_D = \left(\boldsymbol{x^j}\right)' D\boldsymbol{x^k} = \sum_{i=1}^{n} p_i x_i^j x_i^k = s\left(\boldsymbol{x^j}, \boldsymbol{x^k}\right) \tag{1.12}$$

- *O comprimento de uma variável centrada é igual ao seu desvio-padrão.*

$$\|\boldsymbol{x^j}\|_D = \sqrt{\langle \boldsymbol{x^j}, \boldsymbol{x^j} \rangle_D} = \sqrt{s\left(\boldsymbol{x^j}, \boldsymbol{x^j}\right)} = \sqrt{s^2\left(\boldsymbol{x^j}\right)} = s\left(\boldsymbol{x^j}\right)$$

- *O co-seno do ângulo entre duas variáveis centradas é igual ao coeficiente de correlação linear entre elas.*

$$\cos\left(\boldsymbol{x^j}, \boldsymbol{x^k}\right) = \frac{\langle \boldsymbol{x^j}, \boldsymbol{x^k} \rangle_D}{\|\boldsymbol{x^j}\|_D \|\boldsymbol{x^k}\|_D} = \frac{s\left(\boldsymbol{x^j}, \boldsymbol{x^k}\right)}{s\left(\boldsymbol{x^j}\right) s\left(\boldsymbol{x^k}\right)} = r\left(\boldsymbol{x^j}, \boldsymbol{x^k}\right)$$

Segundo a métrica utilizada no espaço E, cada variável pode ser representada como um vector de dimensão n, em que o conjunto das extremidades de cada vector forma a nuvem N_J. Se os dados estiverem centrados e reduzidos (**ACP**

normada) cada variável tem norma igual a 1, logo a nuvem N_J encontra-se dentro de uma hiperesfera de raio 1 (aqui representada na figura 1.2, limitada a $\mathbb{R}^3$). Por exemplo, a j-ésima variável é um vector de $\mathbb{R}^n$ cujas componentes são $\frac{x_i^j - \bar{x}^j}{s_j}$ com $i = 1, \ldots, n$.

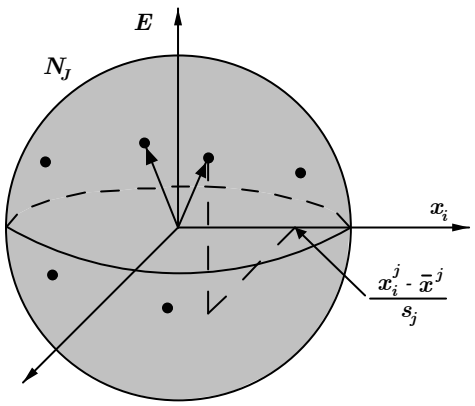


Figura 1.2 Nuvem de variáveis no espaço E.

Se os dados estiverem unicamente centrados o comprimento das variáveis é, como já foi visto, igual ao seu desvio-padrão e está-se em presença de uma **ACP não normada**.

Um tripleto (X, Q, D) pode ser caracterizado por dois objectos diferentes:

- $XQX' = W$, já definido em (1.8);
- $X'DX = V$, já definido em (1.4), se os dados estiverem centrados.

1.5 Projecção dos indivíduos num subespaço

Sendo $p >> 3$, o estudo directo da nuvem N_I torna-se impossível dada a limitação visual para espaços de dimensão superior a 3. Daí o interesse da ACP em fornecer imagens planas aproximadas o melhor possível da nuvem N_I situada num espaço de maior dimensão. À redução da dimensão do espaço dos indivíduos, de p para um certo número q, está naturalmente associada uma perda de informação, que se pretende minimizar.

É então necessário encontrar uma base $\{\boldsymbol{u_k}; k = 1, \ldots, q\}$ de vectores de $\mathbb{R}^p$, chamados **eixos principais**, Q-ortonormados, que definem os planos principais sobre os quais se projecta a nuvem N_I.

Admitindo a existência desta base, seja $F_{\boldsymbol{u_k}}(i)$ a projecção do indivíduo $\boldsymbol{x_i}$ sobre o eixo $\boldsymbol{u_k}$:

$$F_{\boldsymbol{u_k}}(i) = \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q \tag{1.13}$$

e $\boldsymbol{F_u(i)}$ o vector das projecções do indivíduo $\boldsymbol{x_i}$ sobre os eixos $\boldsymbol{u_1}, \ldots, \boldsymbol{u_q}$:

$$\boldsymbol{F_u(i)} = \sum_{k=1}^{q} F_{\boldsymbol{u_k}}(i) \cdot \boldsymbol{u_k} = \sum_{k=1}^{q} \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q \cdot \boldsymbol{u_k}$$

e $\boldsymbol{h}$ o centro de gravidade da nuvem projectada sobre estes mesmos eixos, isto é, $\boldsymbol{F_u(g)} = \boldsymbol{h}$.

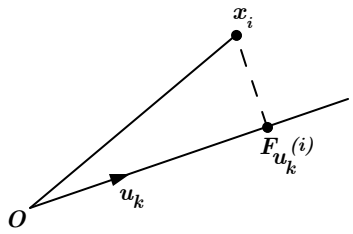


Figura 1.3 Projecção do indivíduo $\boldsymbol{x_i}$ no eixo $\boldsymbol{u_k}$.

A **deformação em projecção** é definida como

$$\mathcal{D} = \sum_{i=1}^{n} p_i \|\boldsymbol{x_i} - \boldsymbol{F_u(i)}\|_Q^2. \tag{1.14}$$

Tem-se que

$$\mathcal{D} = \mathcal{I}_{\boldsymbol{g}} + \|\boldsymbol{g} - \boldsymbol{h}\|_Q^2 - \sum_{i=1}^{n} p_i \|\boldsymbol{F_u(i)} - \boldsymbol{h}\|_Q^2$$

como se mostra facilmente:

$$\begin{aligned}
\mathcal{D} &\overset{(1.14)}{=} \sum_{i=1}^{n} p_i \|\boldsymbol{x_i} - \boldsymbol{F_u(i)}\|_Q^2 \\
&= \sum_{i=1}^{n} p_i \|\boldsymbol{x_i} - \boldsymbol{h}\|_Q^2 - \sum_{i=1}^{n} p_i \|\boldsymbol{F_u(i)} - \boldsymbol{h}\|_Q^2 \quad \text{(Teorema de Pitágoras)} \\
&\overset{(1.10)}{=} \mathcal{I}_{\boldsymbol{h}} - \sum_{i=1}^{n} p_i \|\boldsymbol{F_u(i)} - \boldsymbol{h}\|_Q^2 \\
&\overset{(1.11)}{=} \mathcal{I}_{\boldsymbol{g}} + \|\boldsymbol{g} - \boldsymbol{h}\|_Q^2 - \sum_{i=1}^{n} p_i \|\boldsymbol{F_u(i)} - \boldsymbol{h}\|_Q^2. \quad \blacksquare
\end{aligned}$$

A deformação em projecção é mínima quando $\boldsymbol{g} = \boldsymbol{h}$ e quando a inércia da nuvem projectada, $\sum_{i=1}^{n} p_i \|\boldsymbol{F_u(i)} - \boldsymbol{h}\|_Q^2$, for máxima.

Admitindo que os dados estão centrados, a inércia da nuvem projectada sobre os eixos $\boldsymbol{u_1}, \ldots, \boldsymbol{u_q}$ é:

$$\begin{aligned}
\sum_{i=1}^{n} p_i \|\boldsymbol{F_u(i)}\|_Q^2 &= \sum_{i=1}^{n} p_i \left\| \sum_{k=1}^{q} \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q \cdot \boldsymbol{u_k} \right\|_Q^2 \\
&= \sum_{i=1}^{n} p_i \left\langle \sum_{k=1}^{q} \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q \cdot \boldsymbol{u_k}, \sum_{k=1}^{q} \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q \cdot \boldsymbol{u_k} \right\rangle_Q \\
&= \sum_{i=1}^{n} p_i \sum_{k=1}^{q} \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q^2 \\
&= \sum_{k=1}^{q} \sum_{i=1}^{n} p_i \langle \boldsymbol{u_k}, \boldsymbol{x_i} \rangle_Q \cdot \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q \\
&= \sum_{k=1}^{q} \sum_{i=1}^{n} p_i \boldsymbol{u_k}' Q \boldsymbol{x_i} \cdot \boldsymbol{x_i}' Q \boldsymbol{u_k} \\
&= \sum_{k=1}^{q} \boldsymbol{u_k}' Q \left(\sum_{i=1}^{n} p_i \boldsymbol{x_i} \boldsymbol{x_i}' \right) Q \boldsymbol{u_k} \\
&\overset{(1.4)}{=} \sum_{k=1}^{q} \boldsymbol{u_k}' Q V Q \boldsymbol{u_k}. \qquad (1.15)
\end{aligned}$$

A matriz VQ é Q-simétrica, isto é, para todo o par de vectores $\boldsymbol{u_k}$ e $\boldsymbol{u_l}$ de $\mathbb{R}^p$ tem-se que:

$$\langle \boldsymbol{u_k}, VQ\boldsymbol{u_l} \rangle_Q = \boldsymbol{u_k}' Q V Q \boldsymbol{u_l} = \langle VQ\boldsymbol{u_k}, \boldsymbol{u_l} \rangle_Q,$$

logo, por um resultado de Álgebra Linear, VQ é diagonalizável e admite uma base Q-ortonormada de vectores próprios associados aos valores próprios $\lambda_1, \ldots, \lambda_q, \ldots, \lambda_p$. Por esta razão e pelo facto de os eixos principais serem Q-ortonormados, a inércia da nuvem projectada sobre o eixo $\boldsymbol{u_k}$, $\boldsymbol{u_k}' Q V Q \boldsymbol{u_k}$, é máxima para o maior valor próprio de VQ. Sejam então $\{\boldsymbol{u_k}; k = 1, \ldots, q\}$, os vectores próprios de VQ.

Os eixos principais também são V^{-1}-ortogonais. De facto, uma vez que são Q-ortogonais tem-se que

$$\langle \boldsymbol{u_k}, \boldsymbol{u_l} \rangle_Q = 0 \Longleftrightarrow \boldsymbol{u_k}' Q \boldsymbol{u_l} = 0 \quad (k \neq l) \qquad (1.16)$$

Mas como estes eixos são vectores próprios de VQ (a menos do sinal),

$$VQ\boldsymbol{u_l} = \lambda_l \boldsymbol{u_l}, \qquad (1.17)$$

multiplicando por V^{-1} em ambos os membros

$$Q\boldsymbol{u_l} = \lambda_l V^{-1}\boldsymbol{u_l},$$

logo, substituindo em (1.16) tem-se que

$$\boldsymbol{u_k}'\lambda_l V^{-1}\boldsymbol{u_l} = 0 \iff \boldsymbol{u_k}'V^{-1}\boldsymbol{u_l} = 0 \iff \langle \boldsymbol{u_k}, \boldsymbol{u_l}\rangle_{V^{-1}} = 0 \quad \blacksquare$$

Pela definição de inércia total da nuvem de pontos dada em (1.9) e pelo facto de $\lambda_1, \ldots, \lambda_p$ serem os valores próprios de VQ tem-se que:

$$\mathcal{I}_{\boldsymbol{g}} = tr\,(VQ) = \sum_{k=1}^{p} \lambda_k, \tag{1.18}$$

e, por (1.15) a inércia da nuvem projectada é

$$\sum_{i=1}^{n} p_i \|\boldsymbol{F_u(i)}\|_Q^2 = \sum_{k=1}^{q} \boldsymbol{u_k}'QVQ\boldsymbol{u_k} \overset{(1.17)}{=} \sum_{k=1}^{q} \lambda_k.$$

As projecções dos indivíduos $\boldsymbol{x_i}$ no subespaço de dimensão q formam uma nuvem em que cada ponto possui q coordenadas. As coordenadas dessas projecções no eixo $\boldsymbol{u_k}$ formam o vector $\boldsymbol{F_{u_k}}$ que pertence ao espaço E (ou seja, é um elemento de $\mathbb{R}^n$) e que se designa por **componente principal**.

Surge assim o "novo" quadro de dados $n \times q$:

$$\begin{pmatrix} \langle \boldsymbol{x_1}, \boldsymbol{u_1}\rangle_Q & \langle \boldsymbol{x_1}, \boldsymbol{u_2}\rangle_Q & \ldots & \langle \boldsymbol{x_1}, \boldsymbol{u_q}\rangle_Q \\ \langle \boldsymbol{x_2}, \boldsymbol{u_1}\rangle_Q & \langle \boldsymbol{x_2}, \boldsymbol{u_2}\rangle_Q & \ldots & \langle \boldsymbol{x_2}, \boldsymbol{u_q}\rangle_Q \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{x_n}, \boldsymbol{u_1}\rangle_Q & \langle \boldsymbol{x_n}, \boldsymbol{u_2}\rangle_Q & \ldots & \langle \boldsymbol{x_n}, \boldsymbol{u_q}\rangle_Q \end{pmatrix}$$

em que

$$\boldsymbol{F_{u_k}} = \Big(\langle \boldsymbol{x_1}, \boldsymbol{u_k}\rangle_Q \;\; \langle \boldsymbol{x_2}, \boldsymbol{u_k}\rangle_Q \;\; \ldots \;\; \langle \boldsymbol{x_n}, \boldsymbol{u_k}\rangle_Q\Big)' = XQ\boldsymbol{u_k} \tag{1.19}$$

com $k = 1, \ldots, q$.

As componentes principais possuem algumas propriedades que se seguem.

- Cada componente principal pode ser expressa como combinação linear das p variáveis $\boldsymbol{x^1}, \boldsymbol{x^2}, \ldots, \boldsymbol{x^p}$ (centradas).

  $$\boldsymbol{F_{u_k}} \overset{(1.19)}{=} \sum_{j=1}^{p} q_j \cdot \boldsymbol{u_k}^j \cdot \boldsymbol{x^j},$$

  em que q_j é o j-ésimo elemento da diagonal principal da matriz (métrica) Q definida na secção 1.3.

- As componentes principais são variáveis de média zero.

$$\begin{aligned}
\boldsymbol{F_{u_k}}' D\mathbf{1}_n &\overset{(1.19)}{=} \left(XQ\boldsymbol{u_k}\right)' D\mathbf{1}_n \\
&= \boldsymbol{u_k}' Q' X' D\mathbf{1}_n \\
&= \boldsymbol{u_k}' Q \left(X' D\mathbf{1}_n\right) \quad (Q \text{ é simétrica})
\end{aligned}$$

  Ora como os dados estão centrados $\boldsymbol{g} = X' D\mathbf{1}_n = 0$ logo $\boldsymbol{F_{u_k}}' D\mathbf{1} = 0$. ■

- As componentes principais têm variância λ_k e são não correlacionadas entre si.

$$\begin{aligned}
s\left(\boldsymbol{F_{u_k}}, \boldsymbol{F_{u_l}}\right) &\overset{(1.12)}{=} \boldsymbol{F_{u_k}}' D \boldsymbol{F_{u_l}} \\
&\overset{(1.19)}{=} \left(XQ\boldsymbol{u_k}\right)' D \left(XQ\boldsymbol{u_l}\right) \\
&= \boldsymbol{u_k}' Q X' D X Q \boldsymbol{u_l} \\
&\overset{(1.4)}{=} \boldsymbol{u_k}' Q V Q \boldsymbol{u_l} \\
&\overset{(1.17)}{=} \boldsymbol{u_k}' Q \lambda_l \boldsymbol{u_l} \\
&= \lambda_l \left(\boldsymbol{u_k}' Q \boldsymbol{u_l}\right) \\
&= \lambda_l \cdot \begin{cases} 0 & \text{se } k \neq l \\ 1 & \text{se } k = l \end{cases} \quad \blacksquare
\end{aligned} \tag{1.20}$$

- As componentes principais podem ser obtidas através da diagonalização da matriz WD.

$$\begin{aligned}
VQ\boldsymbol{u_k} &= \lambda_k \boldsymbol{u_k} \\
\left(XQ\right)\left(VQ\boldsymbol{u_k}\right) &= \lambda_k \left(XQ\right)\boldsymbol{u_k} \\
XQX'D\boldsymbol{F_{u_k}} &\overset{(1.19)}{=} \lambda_k \boldsymbol{F_{u_k}} \\
WD\boldsymbol{F_{u_k}} &= \lambda_k \boldsymbol{F_{u_k}} \quad \blacksquare
\end{aligned}$$

Ao eixo principal $\boldsymbol{u_k}$ está associada a forma linear $\boldsymbol{z_k}$, tal que

$$\boldsymbol{z_k} = Q\boldsymbol{u_k}, \tag{1.21}$$

designada por **factor principal**.

Os factores principais também possuem algumas propriedades que importa salientar.

- Os factores principais são Q^{-1}-ortonormados.

$$\begin{aligned}
\langle \boldsymbol{z_k}, \boldsymbol{z_l} \rangle_{Q^{-1}} &= \boldsymbol{z_k}' Q^{-1} \boldsymbol{z_l} \\
&\overset{(1.21)}{=} (\boldsymbol{u_k}' Q') \, Q^{-1} \, (Q \boldsymbol{u_l}) \\
&= (\boldsymbol{u_k}' Q) \, Q^{-1} \, (Q \boldsymbol{u_l}) \qquad (Q \text{ é simétrica}) \\
&= \boldsymbol{u_k}' Q \boldsymbol{u_l} \\
&= \langle \boldsymbol{u_k}, \boldsymbol{u_l} \rangle_Q \\
&= \begin{cases} 0 & \text{se } k \neq l \\ 1 & \text{se } k = l \end{cases} \quad \text{(os eixos principais são } Q\text{-ortonormados)} \quad \blacksquare
\end{aligned}$$

- Os factores principais são V-ortogonais.

$$\begin{aligned}
\langle \boldsymbol{z_k}, \boldsymbol{z_l} \rangle_V &= \boldsymbol{z_k}' V \boldsymbol{z_l} \\
&\overset{(1.21)}{=} (\boldsymbol{u_k}' Q') \, V \, (Q \boldsymbol{u_l}) \\
&= (\boldsymbol{u_k}' Q) \, V \, (Q \boldsymbol{u_l}) \qquad (Q \text{ é simétrica}) \\
&= \boldsymbol{u_k}' Q \, (V Q \boldsymbol{u_l}) \\
&= \boldsymbol{u_k}' Q \, (\lambda_l \boldsymbol{u_l}) \quad \text{(os eixos principais são vectores próprios de } VQ) \\
&= \lambda_l \langle \boldsymbol{u_k}, \boldsymbol{u_l} \rangle_Q \\
&= \lambda_l \cdot \begin{cases} 0 & \text{se } k \neq l \\ 1 & \text{se } k = l \end{cases} \quad \text{(os eixos principais são } Q\text{-ortonormados)} \quad \blacksquare
\end{aligned}$$

- Os factores principais são vectores próprios da matriz QV associados aos valores próprios $\lambda_1, \ldots, \lambda_q$.

$$\begin{aligned}
QV \boldsymbol{z_k} &\overset{(1.21)}{=} QVQ \boldsymbol{u_k} \\
&= Q \lambda_k \boldsymbol{u_k} \qquad \text{(os eixos principais são vectores próprios de } VQ) \\
&\overset{(1.21)}{=} \lambda_k \boldsymbol{z_k} \quad \blacksquare
\end{aligned}$$

Neste momento é importante distinguir duas situações:

- a diagonalização da matriz de variâncias e covariâncias V (considerando $Q = I_p$) equivale a uma **ACP não normada**, que deve ser aplicada quando as variáveis tiverem as mesmas unidades de medida;

- a diagonalização da matriz de correlações R ou da matriz $V \cdot Q_{1/s^2}$ (se os dados estiverem unicamente centrados) corresponde a uma **ACP normada**, que deve ser aplicada no caso de as variáveis não serem da mesma natureza, ou seja, não possuírem as mesmas unidades de medida.

1.6 Projecção das variáveis num subespaço

A projecção das variáveis num subespaço de $\mathbb{R}^n$ é análoga à projecção dos indivíduos num subespaço de $\mathbb{R}^p$. Seja $\{\boldsymbol{v_k}; k = 1, \ldots, r\}$ uma base de vectores de $\mathbb{R}^n$, D-ortonormados, que definem o subespaço (de dimensão r) sobre o qual se vão projectar as variáveis. O critério aplicado na secção anterior para minimizar a deformação em projecção (ou maximizar a inércia da nuvem projectada) também se aplica, mas com um significado diferente, pois a nuvem N_J não é centrada e todos os pontos desta estão situados na hiperesfera de raio unitário. Desta forma interessa considerar os ângulos formados pelos vectores que representam as variáveis e não a distância entre os pontos da nuvem, tal como está definido em 1.4.

Seja $E_{\boldsymbol{v_k}}(j)$ a projecção da variável $\boldsymbol{x^j}$ sobre o eixo $\boldsymbol{v_k}$ tal que

$$E_{\boldsymbol{v_k}}(j) = \langle \boldsymbol{x^j}, \boldsymbol{v_k} \rangle_D \tag{1.22}$$

e $\boldsymbol{E_v(j)}$ o vector das projecções da variável $\boldsymbol{x^j}$ sobre os eixos $\boldsymbol{v_1}, \ldots, \boldsymbol{v_r}$ tal que

$$\boldsymbol{E_v(j)} = \sum_{k=1}^{r} E_{\boldsymbol{v_k}}(j) \cdot \boldsymbol{v_k} = \sum_{k=1}^{r} \langle \boldsymbol{x^j}, \boldsymbol{v_k} \rangle_D \cdot \boldsymbol{v_k}.$$

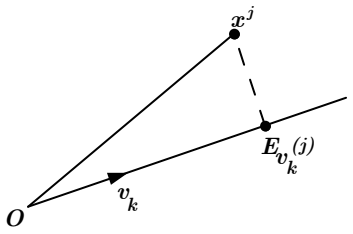


Figura 1.4 Projecção da variável $\boldsymbol{x^j}$ no eixo $\boldsymbol{v_k}$.

A inércia da nuvem projectada nos eixos $\boldsymbol{v_1}, \ldots, \boldsymbol{v_r}$ é deduzida de forma análoga à expressão (1.15):

$$\sum_{j=1}^{p} \tfrac{1}{s_j^2} \|\boldsymbol{E_v(j)}\|_D^2 = \sum_{k=1}^{r} \boldsymbol{v_k}' DWD\boldsymbol{v_k}.$$

A matriz WD é D-simétrica, logo admite uma base D-ortonormada de vectores próprios associada aos valores próprios $\delta_1, \ldots, \delta_r, \ldots, \delta_n$. Conclui-se então que a inércia da nuvem projectada sobre o eixo $\boldsymbol{v_k}$, $\boldsymbol{v_k}' DWD\boldsymbol{v_k}$, é máxima

para o maior valor próprio de WD, uma vez que $\boldsymbol{v_k}'D\boldsymbol{v_k} = 1$ e que os vectores próprios de WD são $\boldsymbol{v_k}$ (a menos do sinal).

As coordenadas das projecções das variáveis $\boldsymbol{x^1}, \ldots, \boldsymbol{x^p}$ no eixo $\boldsymbol{v_k}$ formam o vector $\boldsymbol{E_{v_k}}$ do espaço F:

$$\boldsymbol{E_{v_k}} = X'D\boldsymbol{v_k}.$$

1.7 Dualidade e relações de transição

A determinação dos eixos $\boldsymbol{v_k}$ da nuvem das variáveis é idêntica à determinação dos eixos $\boldsymbol{u_k}$ da nuvem dos indivíduos. Basta, nos resultados relativos à projecção dos indivíduos num subespaço, substituir a matriz X pela matriz transposta X' e a matriz Q pela matriz D.

Viu-se na secção 1.5 que $WD\boldsymbol{F_{u_k}} = \lambda_k \boldsymbol{F_{u_k}}$ e em 1.6 que $WD\boldsymbol{v_k} = \delta_k \boldsymbol{v_k}$. A comparação destas duas equações permite concluir que:

- $\lambda_k = \delta_k$ e que as inércias das nuvens projectadas N_I e N_J sobre os respectivos eixos principais são iguais;

- as matrizes VQ e WD não têm a mesma dimensão, logo, os valores próprios não comuns às duas matrizes são nulos (Escofier e Pagès [12]);

- $\boldsymbol{F_{u_k}}$ e $\boldsymbol{v_k}$ são vectores colineares de $\mathbb{R}^n$ associados ao mesmo valor próprio e, como $\|\boldsymbol{v_k}\|_D = 1$ e $\|\boldsymbol{F_{u_k}}\|_D = \sqrt{\lambda_k}$ tem-se que

  $$\boldsymbol{F_{u_k}} = \sqrt{\lambda_k}\boldsymbol{v_k},$$

  e a relação dual desta é

  $$\boldsymbol{E_{v_k}} = \sqrt{\lambda_k}\boldsymbol{u_k}. \tag{1.23}$$

As projecções de N_I sobre $\boldsymbol{u_k}$ são as coordenadas de $\boldsymbol{F_{u_k}}$, colinear com $\boldsymbol{v_k}$ e as projecções de N_J sobre $\boldsymbol{v_k}$ são as coordenadas de $\boldsymbol{E_{v_k}}$, colinear com $\boldsymbol{u_k}$. Estas relações de dualidade encontram-se esquematizadas na tabela 1.1 e na figura 1.5.

Tabela 1.1 Relações de dualidade entre o espaço dos indivíduos e o das variáveis.

Nuvens	N_I	N_J
Espaço	F	E
Métrica	Q	D
Eixos Principais	$\boldsymbol{u_k}$	$\boldsymbol{v_k}$
Equação	$X'DXQ\boldsymbol{u_k} = \lambda_k \boldsymbol{u_k}$	$XQX'D\boldsymbol{v_k} = \lambda_k \boldsymbol{v_k}$
Norma	$\|\boldsymbol{u_k}\|_Q = 1$	$\|\boldsymbol{v_k}\|_D = 1$
Ortogonalidade	$\langle \boldsymbol{u_i}, \boldsymbol{u_j} \rangle_Q = 0 \ (i \neq j)$	$\langle \boldsymbol{v_i}, \boldsymbol{v_j} \rangle_D = 0 \ (i \neq j)$
Componentes Principais	$\boldsymbol{F_{u_k}} = XQ\boldsymbol{u_k}$	$\boldsymbol{E_{v_k}} = X'D\boldsymbol{v_k}$
Equação	$XQX'D\boldsymbol{F_{u_k}} = \lambda_k \boldsymbol{F_{u_k}}$	$X'DXQ\boldsymbol{E_{v_k}} = \lambda_k \boldsymbol{E_{v_k}}$
Norma	$\|\boldsymbol{F_{u_k}}\|_D = \sqrt{\lambda_k}$	$\|\boldsymbol{E_{v_k}}\|_Q = \sqrt{\lambda_k}$
Ortogonalidade	$\langle \boldsymbol{F_{u_k}}, \boldsymbol{F_{u_l}} \rangle_D = 0 \ (k \neq l)$	$\langle \boldsymbol{E_{v_k}}, \boldsymbol{E_{v_l}} \rangle_Q = 0 \ (k \neq l)$

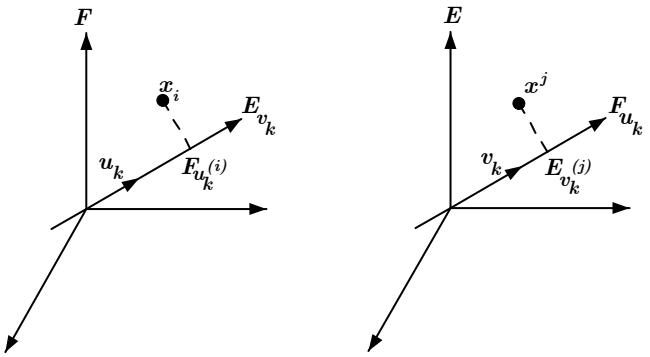


Figura 1.5 Relações de dualidade entre eixos principais e componentes principais.

A matriz XQ define uma aplicação de F em E e a matriz $X'D$ define uma aplicação de E em F. Estas matrizes interligam os eixos e as componentes principais das duas nuvens. A matriz Q define um isomorfismo de F em F^*, espaço dual do espaço dos indivíduos. Se F^* for munido da matriz Q^{-1}, a aplicação Q é um isomorfismo de espaços euclidianos, uma vez que as distâncias e as formas são conservadas. De forma análoga, a matriz D define um isomorfismo de E em E^*, espaço dual do espaço das variáveis.

A matriz X define então uma aplicação de F^* em E e a matriz X' define uma aplicação de E^* em F.

O esquema de dualidade introduzido por Cailliez e Pagès [4] encontra-se

representado na figura 1.6, estabelecendo as relações entre as diversas matrizes que permitem passar dos eixos de uma nuvem para os eixos da outra.

Figura 1.6 Esquema de dualidade da ACP.

Note-se que os eixos principais $\boldsymbol{u_k^*}$ da nuvem N_I^*, dual de N_I, verificam a relação

$$QX'DX\boldsymbol{u_k^*} = \lambda_k \boldsymbol{u_k^*},$$

e os eixos principais $\boldsymbol{v_k^*}$ de N_J^*, dual de N_J, verificam a relação

$$DXQX'\boldsymbol{v_k^*} = \lambda_k \boldsymbol{v_k^*}.$$

Importa agora estabelecer as relações de transição entre as componentes principais $\boldsymbol{F_{u_k}}$ e $\boldsymbol{E_{v_k}}$. A primeira relação (1.24) exprime a projecção do indivíduo $\boldsymbol{x_i}$ sobre o eixo $\boldsymbol{u_k}$ como uma combinação linear das projecções de todas as variáveis. A segunda relação (1.25), dual da primeira, exprime a projecção da variável $\boldsymbol{x^j}$ sobre o eixo $\boldsymbol{v_k}$ como uma combinação linear das projecções de todos os indivíduos.

$$\begin{aligned} F_{\boldsymbol{u_k}}(i) &\overset{(1.13)}{=} \langle \boldsymbol{x_i}, \boldsymbol{u_k} \rangle_Q \\ &= \boldsymbol{x_i'} Q \boldsymbol{u_k} \\ &= \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{p} \frac{x_i^j - \bar{x}^j}{s_j} \cdot E_{\boldsymbol{v_k}}(j) \end{aligned} \tag{1.24}$$

$$\begin{aligned} E_{\boldsymbol{v_k}}(j) &\overset{(1.22)}{=} \langle \boldsymbol{x^j}, \boldsymbol{v_k} \rangle_D \\ &= (\boldsymbol{x^j})' D \boldsymbol{v_k} \\ &= \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} \frac{1}{p_i} \cdot \frac{x_i^j - \bar{x}^j}{s_j} \cdot F_{\boldsymbol{u_k}}(i) \end{aligned} \tag{1.25}$$

1.8 Qualidade e interpretação dos resultados

1.8.1 Fórmulas de reconstituição

Já se viu em (1.19) que $\boldsymbol{F_{u_k}} = XQ\boldsymbol{u_k}$. Ora, multiplicando ambos os membros por $\boldsymbol{u_k}'$ e somando de $k = 1, \ldots, p$, vem que

$$\sum_{k=1}^{p} \boldsymbol{F_{u_k}} \boldsymbol{u_k}' = XQ \sum_{k=1}^{p} \boldsymbol{u_k} \boldsymbol{u_k}'. \tag{1.26}$$

Por um resultado de Álgebra Linear[1] tem-se que a expressão (1.26) é equivalente a

$$XQQ^{-1} = X.$$

Deste modo pode-se obter uma reconstituição aproximada do quadro de dados centrado, utilizando apenas as q primeiras componentes principais:

$$X \simeq \sum_{k=1}^{q} \boldsymbol{F_{u_k}} \boldsymbol{u_k}'.$$

Para obter uma reconstituição da matriz de variâncias e covariâncias, considere-se:

$$\begin{aligned} V &\overset{(1.4)}{=} X'DX \\ &\overset{(1.26)}{=} \left(\sum_{k=1}^{p} \boldsymbol{F_{u_k}} \boldsymbol{u_k}' \right)' D \left(\sum_{k=1}^{p} \boldsymbol{F_{u_k}} \boldsymbol{u_k}' \right) \\ &= \left(\boldsymbol{u_1} \boldsymbol{F_{u_1}}' + \ldots + \boldsymbol{u_p} \boldsymbol{F_{u_p}}' \right) D \left(\boldsymbol{F_{u_1}} \boldsymbol{u_1}' + \ldots + \boldsymbol{F_{u_p}} \boldsymbol{u_p}' \right) \\ &= \boldsymbol{u_1} \left[\boldsymbol{F_{u_1}}' D \boldsymbol{F_{u_1}} \right] \boldsymbol{u_1}' + \ldots + \boldsymbol{u_p} \left[\boldsymbol{F_{u_p}}' D \boldsymbol{F_{u_p}} \right] \boldsymbol{u_p}' \end{aligned} \tag{1.27}$$

uma vez que

$$\begin{aligned} \boldsymbol{u_k} \left[\boldsymbol{F_{u_k}}' D \boldsymbol{F_{u_l}} \right] \boldsymbol{u_l}' &= \boldsymbol{u_k} \cdot \langle \boldsymbol{F_{u_k}}, \boldsymbol{F_{u_l}} \rangle_D \cdot \boldsymbol{u_l}' \\ &= \boldsymbol{u_k} \cdot s\left(\boldsymbol{F_{u_k}}, \boldsymbol{F_{u_l}} \right) \cdot \boldsymbol{u_l}' \overset{(1.20)}{=} 0 \end{aligned}$$

para $k \neq l$ já que as componentes principais são não correlacionadas.

Voltando à expressão (1.27), tem-se que:

$$V = \lambda_1 \boldsymbol{u_1} \boldsymbol{u_1}' + \ldots + \lambda_p \boldsymbol{u_p} \boldsymbol{u_p}'$$

uma vez que

$$\boldsymbol{F_{u_k}}' D \boldsymbol{F_{u_k}} = s\left(\boldsymbol{F_{u_k}}, \boldsymbol{F_{u_k}} \right) \overset{(1.20)}{=} \lambda_k.$$

[1] Se $\{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_n}\}$ é uma base Q-ortonormada, então $\sum_{k=1}^{n} \boldsymbol{u_k} \boldsymbol{u_k}' = Q^{-1}$

Utilizando as q primeiras componentes principais, obtém-se uma reconstituição aproximada da matriz V:

$$V \simeq \sum_{k=1}^{q} \lambda_k \boldsymbol{u_k} \boldsymbol{u_k}'.$$

1.8.2 Medidas de qualidade

Uma vez que, por (1.18), $\mathcal{I}_{\boldsymbol{g}} = tr\,(VQ) = \sum_{k=1}^{p} \lambda_k$, tem-se que

$$\frac{\lambda_k}{\sum_{k=1}^{p} \lambda_k} \tag{1.28}$$

é a proporção de inércia explicada pelo eixo principal $\boldsymbol{u_k}$ e

$$\frac{\sum_{k=1}^{q} \lambda_k}{tr\,(VQ)} \tag{1.29}$$

é a proporção de inércia explicada pelos q primeiros eixos principais. Estas últimas quantidades são indicadores de qualidade da representação da nuvem N_I em $\boldsymbol{u_k}$ e no subespaço de projecção (de dimensão q), respectivamente.

Anteriormente já se mostrou em (1.20) que a variância da k-ésima componente principal é igual ao valor próprio λ_k:

$$s^2\left(\boldsymbol{F}_{\boldsymbol{u_k}}\right) \overset{(1.3)}{=} \sum_{i=1}^{n} p_i \left(F_{\boldsymbol{u_k}}(i)\right)^2 = \lambda_k.$$

Por definição, $F_{\boldsymbol{u_k}}(i)$ é a projecção do indivíduo $\boldsymbol{x_i}$ sobre o eixo principal $\boldsymbol{u_k}$. Portanto, a **contribuição absoluta do indivíduo $\boldsymbol{x_i}$ para a formação do eixo principal $\boldsymbol{u_k}$** é:

$$CTA_i^k = \frac{p_i \left(F_{\boldsymbol{u_k}}(i)\right)^2}{\lambda_k}.$$

Note-se que:

- $\sum_{i=1}^{n} CTA_i^k = \sum_{i=1}^{n} \frac{p_i \left(F_{\boldsymbol{u_k}}(i)\right)^2}{\lambda_k} = 1$;
- a contribuição absoluta de um indivíduo para a formação do eixo principal $\boldsymbol{u_k}$, isto é, para a variância explicada pelo eixo $\boldsymbol{u_k}$, permite evidenciar os indivíduos que apresentam características relacionadas com o fenómeno traduzido pela componente principal que lhe corresponde;
- de uma forma geral, considera-se a contribuição do indivíduo $\boldsymbol{x_i}$ para a formação do eixo principal $\boldsymbol{u_k}$ relevante se esta exceder o seu peso (p_i) sobre a amostra (Saporta [40]);

- se um indivíduo tiver uma contribuição excessiva para a formação dos primeiros eixos isso poderá ser um factor de instabilidade na ACP. Esse indivíduo deverá então ser considerado **suplementar**, isto é, não deverá intervir na determinação dos elementos principais, sendo posteriormente incluído nas representações gráficas.

Para medir a qualidade de representação de um indivíduo $\boldsymbol{x_i}$ no plano principal $(\boldsymbol{u_k}, \boldsymbol{u_l})$, formado pelos eixos principais $\boldsymbol{u_k}$ e $\boldsymbol{u_l}$, é necessário considerar o ângulo α que este indivíduo forma com esse plano principal (figura 1.7).

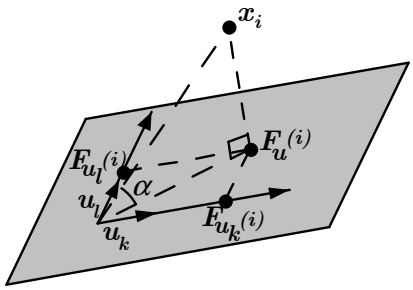


Figura 1.7 Projecção do indivíduo $\boldsymbol{x_i}$ no plano $(\boldsymbol{u_k}, \boldsymbol{u_l})$.

Tem-se que

$$\cos^2 \alpha = \frac{\|\boldsymbol{F_u(i)}\|_Q^2}{\|\boldsymbol{x_i}\|_Q^2} = \frac{(F_{\boldsymbol{u_k}}(i))^2 + (F_{\boldsymbol{u_l}}(i))^2}{\sum_{j=1}^p q_j \left(x_i^j\right)^2}. \tag{1.30}$$

Esta quantidade é denominada por **contribuição relativa do plano** $(\boldsymbol{u_k}, \boldsymbol{u_l})$ **ao indivíduo** $\boldsymbol{x_i}$ e é designada por $\rho_i^{k,l}$.

Note-se que quanto mais próximo $\boldsymbol{x_i}$ estiver do plano, mais próximo estará de $\boldsymbol{F_u(i)}$, menor será a amplitude do ângulo α, ou seja, mais próximo estará $\cos^2 \alpha$ de 1 e desta forma, melhor será a qualidade da sua representação.

Define-se a **contribuição relativa do eixo principal** $\boldsymbol{u_k}$ **ao indivíduo** $\boldsymbol{x_i}$, ρ_i^k, como

$$\rho_i^k = \frac{(F_{\boldsymbol{u_k}}(i))^2}{\|\boldsymbol{x_i}\|_Q^2} = \frac{(F_{\boldsymbol{u_k}}(i))^2}{\sum_{j=1}^p q_j \left(x_i^j\right)^2}.$$

Esta contribuição traduz a proximidade entre o indivíduo $\boldsymbol{x_i}$ e o eixo principal $\boldsymbol{u_k}$, logo quanto mais próximo de 1 for ρ_i^k melhor será a qualidade de representação do indivíduo $\boldsymbol{x_i}$ no eixo $\boldsymbol{u_k}$. Note-se que

$$\rho_i^{k,l} = \rho_i^k + \rho_i^l.$$

De forma análoga, seja θ o ângulo entre a variável $\boldsymbol{x^j}$ e a sua projecção no plano principal $(\boldsymbol{v_k}, \boldsymbol{v_l})$ (figura 1.8) e

$$\cos^2\theta = \frac{\|\boldsymbol{E_v(j)}\|_D^2}{\|\boldsymbol{x^j}\|_D^2} = \frac{\left(E_{\boldsymbol{v_k}}(j)\right)^2 + \left(E_{\boldsymbol{v_k}}(j)\right)^2}{\sum_{i=1}^n p_i\left(x_i^j\right)^2} \overset{(1.23)}{=} \frac{\left(\sqrt{\lambda_k}u_k^j\right)^2 + \left(\sqrt{\lambda_l}u_l^j\right)^2}{s^2\left(\boldsymbol{x^j}\right)},$$

sendo u_k^j o j-ésimo elemento do vector $\boldsymbol{u_k}$.

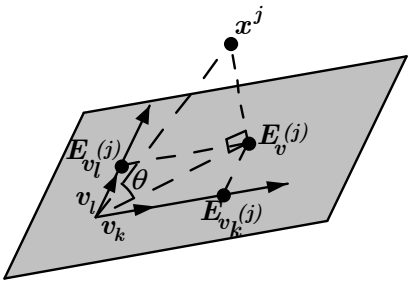


Figura 1.8 Projecção da variável $\boldsymbol{x^j}$ no plano $(\boldsymbol{v_k}, \boldsymbol{v_l})$.

Define-se a **contribuição relativa da componente principal $\boldsymbol{F_{u_k}}$ à variável $\boldsymbol{x^j}$** como:

$$\gamma_k^j = \frac{\left(\sqrt{\lambda_k}u_k^j\right)^2}{s^2\left(\boldsymbol{x^j}\right)}.$$

Sendo as componentes principais $\boldsymbol{F_{u_k}}$ "novas" variáveis, é importante saber interpretá-las. Uma forma de o fazer é determinar o grau de correlação linear entre cada componente principal e cada uma das variáveis iniciais $\boldsymbol{x^j}$:

$$\begin{aligned} r(\boldsymbol{x^j}, \boldsymbol{F_{u_k}}) &= \frac{s(\boldsymbol{x^j}, \boldsymbol{F_{u_k}})}{s(\boldsymbol{x^j})s(\boldsymbol{F_{u_k}})} \\ &\overset{(1.20)}{=} \frac{(\boldsymbol{x^j})' D \boldsymbol{F_{u_k}}}{s(\boldsymbol{x^j})\sqrt{\lambda_k}} \\ &= \frac{\lambda_k u_k^j}{s(\boldsymbol{x^j})\sqrt{\lambda_k}} \\ &= \frac{\sqrt{\lambda_k}u_k^j}{s(\boldsymbol{x^j})}. \end{aligned} \tag{1.31}$$

Note-se que

$$\gamma_k^j = \left[r(\boldsymbol{x^j}, \boldsymbol{F_{u_k}})\right]^2.$$

Da seguinte relação

$$\|\boldsymbol{u_k}\|_Q^2 = 1 \Longleftrightarrow \boldsymbol{u_k}'Q\boldsymbol{u_k} = 1 \Longleftrightarrow \sum_{j=1}^p q_j (u_k^j)^2 = 1,$$

surge a **contribuição absoluta da variável $\boldsymbol{x^j}$ para a formação da componente principal $\boldsymbol{F_{u_k}}$**:

$$CTA_k^j = q_j (u_k^j)^2.$$

Tem-se então que as contribuições das variáveis são proporcionais:

$$\gamma_k^j = \lambda_k \cdot CTA_k^j,$$

com λ_k o factor de proporcionalidade.

1.8.3 Círculo de correlações

Se os dados estiverem centrados e reduzidos, a representação gráfica das variáveis pode ser feita utilizando o coeficiente de correlação linear entre $\boldsymbol{F_{u_k}}$ e $\boldsymbol{x^j}$:

$$r(\boldsymbol{x^j}, \boldsymbol{F_{u_k}}) \overset{(1.31)}{=} \sqrt{\lambda_k} u_k^j.$$

Considerando os eixos associados a um par de componentes principais $(\boldsymbol{F_{u_k}}, \boldsymbol{F_{u_l}})$, representem-se os pontos

$$\left(r(\boldsymbol{x^j}, \boldsymbol{F_{u_k}}), r(\boldsymbol{x^j}, \boldsymbol{F_{u_l}}) \right) \tag{1.32}$$

em que cada variável $\boldsymbol{x^j}$ é representada pelo ponto de abcissa $r(\boldsymbol{x^j}, \boldsymbol{F_{u_k}})$ e ordenada $r(\boldsymbol{x^j}, \boldsymbol{F_{u_l}})$ numa figura denominada **círculo de correlações**, contida no plano principal $(\boldsymbol{v_k}, \boldsymbol{v_l})$ (figura 1.9).

A qualidade de representação da variável sobre o eixo representado por $\boldsymbol{F_{u_k}}$ só será razoável se a sua representação no círculo de correlações se aproximar da fronteira uma vez que $\gamma_k^j = \left[r\left(\boldsymbol{x^j}, \boldsymbol{F_{u_k}}\right) \right]^2$ e o coeficiente de correlação linear varia entre -1 e 1.

Considere-se o exemplo da figura 1.10. A componente principal $\boldsymbol{F_{u_k}}$ está bastante correlacionada com as variáveis $\boldsymbol{x^1}$ (de forma positiva) e $\boldsymbol{x^2}$ (de forma negativa) e não correlacionada com as variáveis $\boldsymbol{x^3}$ e $\boldsymbol{x^4}$. Por oposição, a componente principal $\boldsymbol{F_{u_l}}$ está bastante correlacionada com as variáveis $\boldsymbol{x^3}$ (de forma positiva) e $\boldsymbol{x^4}$ (de forma negativa) e não correlacionada com as variáveis $\boldsymbol{x^1}$ e $\boldsymbol{x^2}$.

Desta forma cada variável (centrada e reduzida) é representada em $\mathbb{R}^n$ por um ponto da hiperesfera de raio 1 e projectada no plano definido por um par de componentes principais, no interior de um círculo de raio 1.

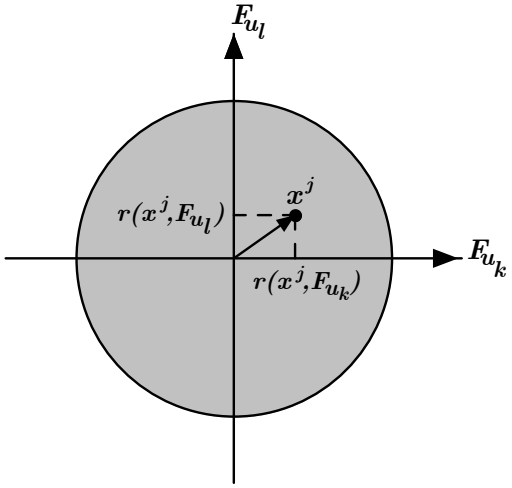


Figura 1.9 Coordenadas da variável x^j no círculo de correlações.

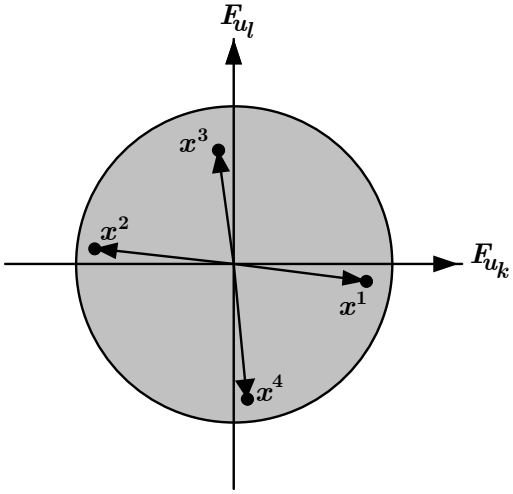


Figura 1.10 Significado do círculo de correlações: exemplo.

1.8.4 Número de eixos a considerar

Uma vez que já se sabe como reduzir a dimensão do espaço dos indivíduos de p para q (com $q << p$), a questão que se coloca agora é qual o número de componentes a reter, isto é, qual vai ser o valor de q?

Já foi visto em (1.29) que

$$\frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

é a percentagem de inércia explicada pelos q primeiros eixos principais e funciona

como um indicador global de qualidade da ACP.

O critério empírico habitualmente usado consiste em determinar o menor valor de q que garanta uma percentagem de variância explicada não inferior a uma quantidade fixada e que usualmente ronda os 80%. Este critério pode levar à retenção de componentes com peso significativo apenas para uma variável.

O scree plot proposto por Cattell [6] (figura 1.11) consiste na representação gráfica da percentagem de variância explicada por cada componente, por ordem decrescente. Quando esta percentagem é reduzida e a curva é quase paralela ao eixo das abcissas, excluem-se as componentes principais seguintes.

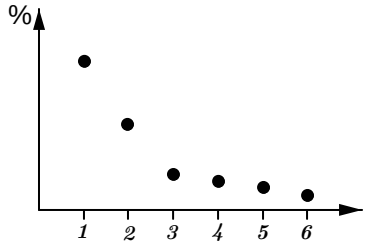


Figura 1.11 Scree plot.

Kaiser [24] propõe outro método para dados centrados e reduzidos que consiste em excluir as componentes principais cujos valores próprios são inferiores a 1.

Mardia *et al.* [31] sugere que para $p \leq 20$ o critério de Kaiser tem tendência a incluir poucas componentes e o scree plot de Cattell tende a incluir muitas.

Outra alternativa possível seria recorrer aos testes de hipóteses propostos por Mardia *et al.* [31] e Sousa [42].

Capítulo 2

Metodologia STATIS

2.1 Introdução

A metodologia STATIS (*"Structuration de Tableaux À Trois Indices de la Statistique"*) foi inicialmente introduzida por Escoufier [13] e L'Hermier des Plantes [30] no Laboratório de Probabilidades e Estatística da Universidade de Montpellier II, por volta de 1976, e posteriormente desenvolvida por Lavit [27], em 1988. Este método não se restringe apenas à análise de um quadro de dados como era o caso da ACP, mas permite a exploração simultânea de vários quadros de dados (quantitativos) recolhidos de uma das seguintes formas:

- T quadros de dados recolhidos em diferentes "ocasiões" sobre os mesmos indivíduos; as variáveis podem diferir ao longo dos quadros; neste caso o método é caracterizado por T estudos (X_t, Q_t, D), $t = 1, \ldots, T$;
- as mesmas variáveis são descritas ao longo de T quadros de dados recolhidos em diferentes "ocasiões" sobre indivíduos que podem diferir ao longo dos quadros; neste caso o método é caracterizado por T estudos (X_t, Q, D_t), $t = 1, \ldots, T$.

A cada uma destas situações corresponde uma estratégia diferente; a primeira evidencia as proximidades entre indivíduos (**método STATIS**) e a segunda privilegia as relações entre variáveis (**método STATIS dual**). Estes dois métodos compreendem as seguintes etapas fundamentais:

I) inter-estrutura: comparação global dos quadros de dados;

II) intra-estrutura: descrição da estrutura comum aos vários quadros de dados através da determinação do compromisso e da respectiva imagem euclidiana;

III) representação das trajectórias dos indivíduos: permite destacar os indivíduos ("STATIS") ou as variáveis ("STATIS dual") responsáveis pelas semelhanças (ou diferenças) entre os quadros; a partir da imagem compromisso traçam-se as trajectórias que descrevem o comportamento evolutivo de cada indivíduo (ou variável).

2.2 Método STATIS

Os T quadros de dados constam de n indivíduos sobre p_t variáveis quantitativas, $t = 1, \ldots, T$.

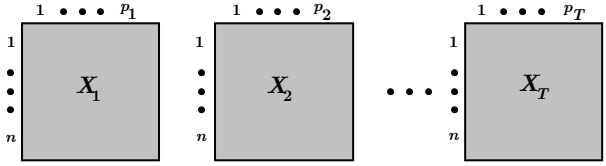


Figura 2.1 T quadros de dados: notação do método STATIS.

No instante t o quadro X_t é a matriz de dimensão $n \times p_t$

$$X_t = \begin{pmatrix} (x_1^1)^t & (x_1^2)^t & \ldots & (x_1^{p_t})^t \\ (x_2^1)^t & (x_2^2)^t & \ldots & (x_2^{p_t})^t \\ \vdots & \vdots & \ddots & \vdots \\ (x_n^1)^t & (x_n^2)^t & \ldots & (x_n^{p_t})^t \end{pmatrix},$$

a j-ésima variável é o vector de $\mathbb{R}^n$

$$(\boldsymbol{x^j})^t = \begin{pmatrix} \left(x_1^j\right)^t \\ \left(x_2^j\right)^t \\ \vdots \\ (x_n^j)^t \end{pmatrix}$$

e o i-ésimo indivíduo é o vector de $\mathbb{R}^{p_t}$

$$\left((\boldsymbol{x_i})^t\right)' = \begin{pmatrix} (x_i^1)^t & (x_i^2)^t & \ldots & (x_i^{p_t})^t \end{pmatrix}.$$

As notações t e T apelam à noção de tempo, no entanto, o método STATIS poder-se-á aplicar a dados não temporais.

2.2.1 Inter-estrutura

Esta primeira etapa consiste na comparação global dos T quadros. Para tal é necessário definir um objecto representativo para cada estudo, uma métrica sobre os objectos representativos de cada estudo e, finalmente, definir uma imagem euclidiana dos objectos representativos associada aos produtos escalares já definidos.

Analogamente ao que foi estudado na secção 1.4, é possível caracterizar um estudo (X_t, Q_t, D), $t = 1, \ldots, T$, por um objecto

$$\mathcal{W}_t = X_t Q_t X_t',$$

em que Q_t é a métrica associada ao espaço dos indivíduos do quadro X_t e D é a métrica associada ao espaço das variáveis definida em (1.1). O objecto $\mathcal{W}_t$ é uma matriz de dimensão $n \times n$ denominada **matriz dos produtos escalares entre indivíduos do quadro** X_t.

Tal como já foi referido, para representar graficamente os T estudos é necessário definir uma distância (métrica) entre estes, para tal é necessário definir um produto escalar entre os objectos:

$$\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS} = tr\left(\mathcal{W}_t D \mathcal{W}_{t'} D\right), \tag{2.1}$$

denominado **produto escalar de Hilbert-Schmidt**, introduzido inicialmente por Escoufier [13], para induzir uma distância (euclidiana) entre os objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$, dada pela seguinte expressão:

$$\begin{aligned} d_{HS}\left(\mathcal{W}_t, \mathcal{W}_{t'}\right) &= \|\mathcal{W}_t - \mathcal{W}_{t'}\|_{HS} \\ &= \sqrt{\langle \mathcal{W}_t - \mathcal{W}_{t'}, \mathcal{W}_t - \mathcal{W}_{t'} \rangle_{HS}} \\ &\overset{(2.1)}{=} \sqrt{tr\left(\left(\mathcal{W}_t - \mathcal{W}_{t'}\right) D\right)^2} \end{aligned}$$

Note-se que a distância entre estes objectos também pode ser assim estabelecida:

$$\begin{aligned} d_{HS}\left(\mathcal{W}_t, \mathcal{W}_{t'}\right) &= \|\mathcal{W}_t - \mathcal{W}_{t'}\|_{HS} \\ &= \sqrt{\langle \mathcal{W}_t - \mathcal{W}_{t'}, \mathcal{W}_t - \mathcal{W}_{t'} \rangle_{HS}} \\ &= \sqrt{\|\mathcal{W}_t\|_{HS}^2 + \|\mathcal{W}_{t'}\|_{HS}^2 - 2\langle \mathcal{W}_t, \mathcal{W}_t \rangle_{HS}}. \end{aligned}$$

A **norma do objecto** $\mathcal{W}_t$ é assim definida:

$$\|\mathcal{W}_t\|_{HS} = \sqrt{\langle \mathcal{W}_t, \mathcal{W}_t \rangle_{HS}} \overset{(2.1)}{=} \sqrt{tr\left(\mathcal{W}_t D \mathcal{W}_t D\right)} = \sqrt{tr\left(\mathcal{W}_t D\right)^2} = \sqrt{\sum_{i=1}^{n} \left(\lambda_i^t\right)^2},$$

com λ_i^t o i-ésimo valor próprio de $\mathcal{W}_t D$.

A **matriz S dos produtos escalares** entre os objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$, de dimensão $T \times T$, tem como termo geral:

$$S_{tt'} = \langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS}.$$

Analogamente ao que acontecia na ACP, se os objectos $\mathcal{W}_t$ tiverem normas muito diferentes é conveniente normá-los, ou seja, considerar os objectos $\mathcal{W}_t / \|\mathcal{W}_t\|_{HS}$; neste caso, a matriz dos produtos escalares $\widetilde{S}$ tem como termo geral:

$$\widetilde{S}_{tt'} = \frac{\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS}}{\|\mathcal{W}_t\|_{HS} \|\mathcal{W}_{t'}\|_{HS}}.$$

O **coeficiente de correlação vectorial entre os objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$**, ou **coeficiente RV** proposto por Robert e Escoufier [38] é designado por

$$RV(t,t') = \left\langle \frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}}, \frac{\mathcal{W}_{t'}}{\|\mathcal{W}_{t'}\|_{HS}} \right\rangle_{HS} = \frac{S_{tt'}}{\sqrt{S_{tt}}\sqrt{S_{t't'}}} = \widetilde{S}_{tt'}. \tag{2.2}$$

Os coeficientes RV são muito úteis na interpretação da inter-estrutura, uma vez que possuem as seguintes

Propriedades:

- Os coeficientes RV permitem obter a distância entre dois objectos normados:

$$\begin{aligned} d_{HS}\left(\frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}}, \frac{\mathcal{W}_{t'}}{\|\mathcal{W}_{t'}\|_{HS}} \right) &= \left\| \frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}} - \frac{\mathcal{W}_{t'}}{\|\mathcal{W}_{t'}\|_{HS}} \right\|_{HS} \\ &= \sqrt{2 - 2RV(t,t')}. \end{aligned}$$

- Se $RV(t,t') = 1$ a distância acima é nula e

$$\frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}} = \frac{\mathcal{W}_{t'}}{\|\mathcal{W}_{t'}\|_{HS}}.$$

  Isto significa que a imagem euclidiana dos indivíduos do estudo t deduz-se da imagem euclidiana do estudo t' pela homotetia de razão $\|\mathcal{W}_t\|_{HS} / \|\mathcal{W}_{t'}\|_{HS}$.

Mas porquê a utilização do produto escalar de Hilbert-Schmidt? Há duas justificações importantes para tal facto.

- Decompondo $\|\mathcal{W}_t - \mathcal{W}_{t'}\|^2_{HS}$ no espaço dos indivíduos (isto é, com as métricas Q_t e $Q_{t'}$) fica-se com:

$$\|\mathcal{W}_t - \mathcal{W}_{t'}\|^2_{HS} = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \left(\left\langle (\boldsymbol{x_i})^t, (\boldsymbol{x_j})^t \right\rangle_{Q_t} - \left\langle (\boldsymbol{x_i})^{t'}, (\boldsymbol{x_j})^{t'} \right\rangle_{Q_{t'}} \right)^2.$$

Então $\|\mathcal{W}_t - \mathcal{W}_{t'}\|^2_{HS}$ é igual à soma ponderada dos quadrados das diferenças entre os produtos escalares entre indivíduos do quadro X_t e os produtos escalares entre indivíduos do quadro $X_{t'}$.

- Exprimindo $\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS}$ no espaço das variáveis e utilizando a métrica identidade para calcular os produtos escalares entre indivíduos (isto é, $Q_t = I_{p_t}$ e $Q_{t'} = I_{p_{t'}}$) tem-se:

$$\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS} = \sum_{k=1}^{p_t} \sum_{l=1}^{p_{t'}} \left(\left\langle (\boldsymbol{x}^{\boldsymbol{k}})^t, (\boldsymbol{x}^{\boldsymbol{l}})^{t'} \right\rangle_D \right)^2 .$$

  Então $\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS}$ é igual à soma dos quadrados das covariâncias entre as variáveis do quadro X_t e as variáveis do quadro $X_{t'}$.

  Se os objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$ são ortogonais, tem-se que $\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS} = 0$, então $RV(t, t') = 0$ e poder-se-á dizer que as covariâncias entre as variáveis de X_t e as variáveis de $X_{t'}$ são nulas.

Para se construir a imagem euclidiana dos T estudos é necessário afectar cada um destes de um peso designado por π_t. Logo a **matriz dos pesos dos estudos** é:

$$\Delta = \begin{pmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \pi_T \end{pmatrix} .$$

No caso de se atribuir aos estudos a mesma importância tem-se que $\Delta = \frac{1}{T} I_T$.

Se houver algum quadro que deva intervir na imagem euclidiana sem contribuir para a análise e determinação dos eixos, esse quadro deverá ser considerado **suplementar**, ou seja, deverá ser-lhe atribuído um peso nulo.

A partir deste momento basta aplicar uma ACP à matriz S; para tal é necessário calcular os valores próprios e os vectores próprios da matriz $S\Delta$. Sejam então:

- $\tau_1, \tau_2, \dots, \tau_T$ os valores próprios da matriz $S\Delta$ associados aos vectores próprios Δ-ortonormados $\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2} \dots, \boldsymbol{\gamma_T}$, respectivamente;

- $A_1, A_2, \dots, A_T$ os pontos associados a $\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_T$, respectivamente, na imagem euclidiana.

As coordenadas de A_t sobre o i-ésimo eixo são as componentes do vector $\sqrt{\tau_i}\boldsymbol{\gamma_i}$ (de dimensão T) com $t = 1, \ldots, T$, por analogia com a expressão (1.23).

Na prática, a imagem euclidiana restringe-se aos dois primeiros eixos, obtendo-se uma imagem euclidiana plana aproximada dos T estudos, associada aos produtos escalares entre objectos: **o plano principal**. Desta forma, a distância entre os pontos A_t e $A_{t'}$ é a melhor aproximação entre os objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$ no sentido do produto escalar de Hilbert-Schmidt e dois pontos suficientemente próximos no plano principal revelam uma estrutura de indivíduos comum aos quadros correspondentes.

O seguinte teorema, cuja demonstração se encontra em Lavit [27], permite ter uma ideia de como se vão situar os pontos A_t na respectiva imagem euclidiana.

Teorema 2.1 *Uma matriz simétrica com todos os termos positivos admite um vector próprio associado ao maior valor próprio cujas coordenadas têm todas o mesmo sinal.*

Ora, representando os pontos A_t $(t = 1, \ldots, T)$ no plano principal constituído pelos dois primeiros eixos, estes vão-se situar todos no primeiro e quarto quadrantes, admitindo que as coordenadas do primeiro vector próprio são todas positivas (figura 2.2).

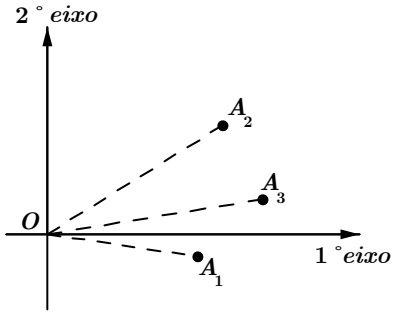


Figura 2.2 Representação dos objectos no plano principal.

Note-se que o coeficiente $RV\,(t, t')$ representa o co-seno do ângulo formado pelos vectores $\overrightarrow{OA_t}$ e $\overrightarrow{OA_{t'}}$:

$$RV\,(t,t') \overset{(2.2)}{=} \frac{\langle \mathcal{W}_t, \mathcal{W}_t \rangle_{HS}}{\|\mathcal{W}_t\|_{HS}\|\mathcal{W}_{t'}\|_{HS}} = \cos\left(\overrightarrow{OA_t}, \overrightarrow{OA_{t'}}\right).$$

Se os objectos forem normados $(\mathcal{W}_t/\|\mathcal{W}_t\|_{HS})$ a matriz dos coeficientes RV coincide com a matriz S.

Caso se pretenda atribuir pesos iguais aos vários estudos diagonalizar-se-á apenas a matriz S. A diagonalização desta matriz permite obter uma representação análoga à representação das variáveis na ACP (como foi exposto no capítulo 1), designada por **imagem euclidiana da inter-estrutura não centrada**.

Outra representação gráfica alternativa será a **imagem euclidiana da inter--estrutura centrada**, em que a matriz S será centrada por linhas e por colunas. Esta representação é diferente e complementar daquela obtida anteriormente e tem como vantagem a visualização das proximidades e oposições entre objectos. Neste caso a matriz em questão é

$$\mathcal{C} = (I_T - \mathbf{1}_T\mathbf{1}'_T\Delta)\, S\, (I_T - \Delta\mathbf{1}_T\mathbf{1}'_T)\,.$$

A diagonalização da matriz $\mathcal{C}\Delta$ ou $\mathcal{C}$ (no caso de os pesos atribuídos aos objectos serem todos iguais) permite a representação gráfica dos objectos análoga à representação dos indivíduos na ACP, conforme exposto no capítulo 1.

2.2.2 Intra-estrutura

O estudo da inter-estrutura evidenciou as semelhanças entre os vários estudos, sem as explicar. Chegou então a altura de procurar um novo quadro que resuma o conjunto dos objectos e que seja da mesma natureza destes. Este quadro é designado por **compromisso** e não é mais do que uma média ponderada dos objectos $\mathcal{W}_t$:

$$\mathcal{W} = \sum_{t=1}^{T} \alpha_t \mathcal{W}_t.$$

Se os objectos forem normados o compromisso é definido por:

$$\mathcal{W} = \sum_{t=1}^{T} \alpha_t \frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}}.$$

A determinação dos coeficientes α_t depende de dois critérios:

- O compromisso $\mathcal{W}$ é o objecto mais correlacionado com os objectos $\mathcal{W}_t$ (no sentido do produto escalar de Hilbert-Schmidt);
- $\mathcal{W}$ deve ser um objecto da mesma natureza que os objectos $\mathcal{W}_t$, isto é,

$$\|\mathcal{W}\|_{HS} = \sum_{t=1}^{T} \pi_t \|\mathcal{W}_t\|_{HS}.$$

Se os objectos forem normados então

$$\|\mathcal{W}\|_{HS} = 1.$$

Seja $\boldsymbol{\gamma_1}$ o vector próprio de $S\Delta$ associado ao maior valor próprio τ_1 designado por

$$\boldsymbol{\gamma_1} = \begin{pmatrix} \gamma_1^1 \\ \gamma_1^2 \\ \vdots \\ \gamma_1^T \end{pmatrix}$$

cujas coordenadas são todas do mesmo sinal (admita-se que são todas positivas) em virtude do teorema 2.1.

Ora os coeficientes α_t são determinados pelas seguintes fórmulas (Lavit [27]):

- no caso dos objectos $\mathcal{W}_t$

$$\alpha_t = \frac{1}{\sqrt{\tau_1}} \left(\sum_{l=1}^{T} \pi_l \|\mathcal{W}_l\|_{HS} \right) \pi_t \gamma_1^t;$$

- no caso dos objectos normados $\mathcal{W}_t / \|\mathcal{W}_t\|_{HS}$

$$\alpha_t = \frac{1}{\sqrt{\tau_1}} \pi_t \gamma_1^t.$$

Então a expressão que define o compromisso é dada nas seguintes fórmulas:

- no caso dos objectos $\mathcal{W}_t$

$$\mathcal{W} = \sum_{t=1}^{T} \left[\frac{1}{\sqrt{\tau_1}} \left(\sum_{l=1}^{T} \pi_l \|\mathcal{W}_l\|_{HS} \right) \pi_t \gamma_1^t \cdot \mathcal{W}_t \right];$$

- no caso dos objectos normados $\mathcal{W}_t / \|\mathcal{W}_t\|_{HS}$

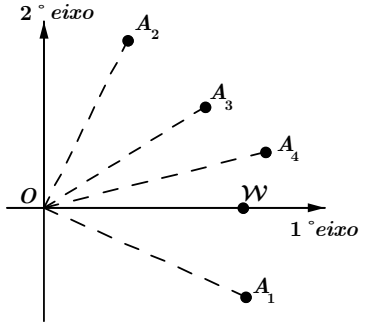
$$\mathcal{W} = \sum_{t=1}^{T} \left[\frac{1}{\sqrt{\tau_1}} \pi_t \gamma_1^t \cdot \frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}} \right].$$

A coordenada do compromisso $\mathcal{W}$ sobre o i-ésimo eixo é obtida por combinação linear das coordenadas $\sqrt{\tau_i}\gamma_i^t$ dos pontos A_t sobre o i-ésimo eixo:

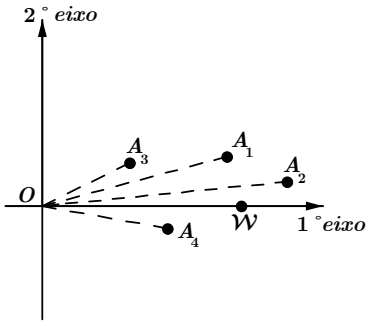
$$\sum_{t=1}^{T} \alpha_t \sqrt{\tau_i} \gamma_i^t.$$

Como os vectores próprios da matriz $S\Delta$ são Δ-ortonormados (uma vez que $S\Delta$ é Δ-simétrica) e por definição de α_t, todas as coordenadas do compromisso serão nulas com excepção da primeira. Desta forma, pode-se concluir que o objecto compromisso $\mathcal{W}$ situar-se-á no primeiro eixo da imagem euclidiana.

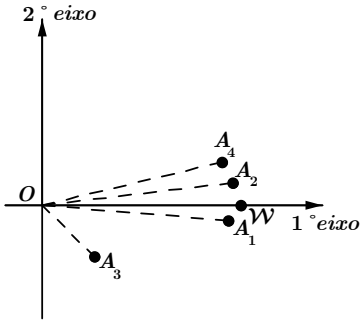
Note-se que só será válida a interpretação da imagem euclidiana dos objectos se os coeficientes RV entre os respectivos estudos forem elevados. As figuras seguintes, também expostas em Lavit [27], ajudam a esta interpretação.



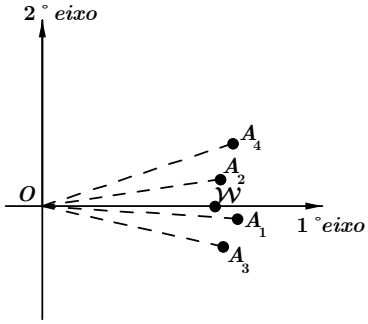
(a) Inexistência de uma estrutura comum



(b) Normas muito diferentes



(c) O objecto $\mathcal{W}_3$ está mal representado



(d) Existência de uma estrutura comum

Figura 2.3 Representação e interpretação dos objectos no plano principal.

Na figura 2.3(a) o objecto compromisso é somente uma média ponderada dos objectos e não reflecte uma estrutura de indivíduos comum aos objectos, uma vez que estes são muito diferentes e, consequentemente, os coeficientes RV são fracos.

As normas dos objectos da figura 2.3(b) são muito diferentes, sendo os objectos de normas mais elevadas que contribuem para a construção do compromisso. Neste caso é conveniente considerar os objectos normados $\mathcal{W}_t/\|\mathcal{W}_t\|_{HS}$.

O objecto $\mathcal{W}_3$ intervém pouco na construção do compromisso da figura 2.3(c), ou seja, o quadro X_3 possui uma estrutura diferente da dos restantes. Uma alternativa possível será considerá-lo suplementar.

Os objectos considerados na figura 2.3(d) têm normas muito aproximadas e coeficientes RV elevados. Neste caso, existe uma estrutura de indivíduos comum aos vários quadros e o compromisso traduz correctamente esta estrutura.

Neste momento reúnem-se as condições para efectuar dois tipos de representação:

- representação da nuvem dos n indivíduos caracterizados pelos T quadros de forma a obter a imagem euclidiana compromisso;
- representação das correlações das variáveis dos diversos quadros com os eixos do compromisso, visando a interpretação destes eixos e das posições dos indivíduos no respectivo plano compromisso.

O compromisso é designado pela matriz $\mathcal{W}$ de dimensão $n \times n$, que é centrada pelos pesos dos indivíduos. Com efeito, sendo os quadros $\mathcal{W}_t$ centrados e fixando a coluna j, tem-se que:

$$\begin{aligned}\sum_{i=1}^{n} p_i \mathcal{W}_{i,j} &= \sum_{i=1}^{n} p_i \sum_{t=1}^{T} \alpha_t \left(\mathcal{W}_t\right)_{i,j} \\ &= \sum_{t=1}^{T} \alpha_t \sum_{i=1}^{n} p_i \left(\mathcal{W}_t\right)_{i,j} = 0. \quad \blacksquare\end{aligned}$$

Aplicando uma ACP à nuvem de indivíduos da matriz $\mathcal{W}$, obtém-se a imagem euclidiana do compromisso. Sejam então:

- $\mu_1, \mu_2, \ldots, \mu_n$ os valores próprios da matriz $\mathcal{W}D$ associados aos vectores próprios (D-ortonormados) $\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2} \ldots, \boldsymbol{\varepsilon_n}$, respectivamente;
- $B_1, B_2, \ldots, B_n$ os pontos associados aos indivíduos, na imagem euclidiana do compromisso.

As coordenadas destes pontos sobre o k-ésimo eixo são as componentes do vector $\sqrt{\mu_k}\boldsymbol{\varepsilon_k}$ (de dimensão n) com $k = 1, \ldots, n$.

Se o estudo da inter-estrutura evidenciou a existência de uma estrutura de indivíduos comum aos quadros (como na figura 2.3(d)), será conveniente efectuar a representação da imagem euclidiana do compromisso aproximada, ou seja, restringindo aos dois ou três primeiros eixos, segundo a percentagem de inércia explicada por estes.

A distância d_{B_i,B_j} entre os pontos B_i e B_j nesta imagem euclidiana corresponde à distância entre os indivíduos i e j no período $[1, T]$ e deduz-se

das distâncias entre os indivíduos i e j em cada estudo:

$$d^2_{B_i,B_j} = \sum_{t=1}^{T} \alpha_t \left\| (\boldsymbol{x_i})^t - (\boldsymbol{x_j})^t \right\|^2_{Q_t}.$$

Note-se que a imagem euclidiana do compromisso obtida é equivalente à que se teria obtido implementando uma ACP sobre o quadro de dados de dimensão $n \times \sum_{t=1}^{T} p_t$, construído a partir da justaposição dos quadros $\sqrt{\alpha_1}X_1, \ldots, \sqrt{\alpha_T}X_T$, ou seja, dispondo estes mesmos sob a forma de colunas (figura 2.4).

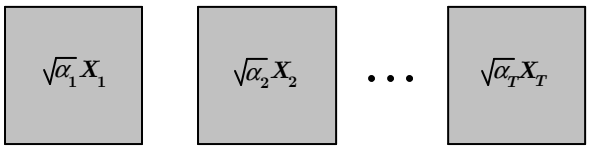


Figura 2.4 Justaposição dos quadros no método STATIS.

Sendo o compromisso centrado pelos pesos dos indivíduos, a imagem euclidiana do compromisso também o é, ou seja, o ponto de intersecção dos eixos corresponde ao centro de gravidade dos pontos $B_1, B_2, \ldots, B_n$. Neste caso, é possível interpretar a posição dos indivíduos sobre um eixo qualquer. Analogamente ao que foi feito na ACP, calculam-se as correlações da componente principal do compromisso correspondente ao k-ésimo eixo com as variáveis de cada estudo.

Para uma variável $\left(\boldsymbol{x^j}\right)^t$ centrada e reduzida, a correlação entre esta e o k-ésimo eixo é igual a

$$\left\langle \boldsymbol{\varepsilon_k}, (\boldsymbol{x^j})^t \right\rangle_D = \left((\boldsymbol{x^j})^t\right)' D \boldsymbol{\varepsilon_k},$$

análoga à expressão (1.31).

O gráfico das correlações, em que a variável $\left(\boldsymbol{x^j}\right)^t$ é representada por um ponto cuja coordenada sobre o k-ésimo eixo é igual a $\left\langle \boldsymbol{\varepsilon_k}, (\boldsymbol{x^j})^t \right\rangle_D$, permite visualizar e interpretar as posições compromisso dos indivíduos ao longo dos eixos.

2.2.3 Trajectórias dos indivíduos

A inter-estrutura evidenciou as diferenças entre os objectos e destes com o compromisso. Para explicar estas diferenças a nível individual decompõe-se

$d^2_{HS}(\mathcal{W}_t, \mathcal{W}_{t'})$ na soma

$$\sum_{i=1}^{n} p_i \sum_{j=1}^{n} p_j \left[(\mathcal{W}_t)_{i,j} - (\mathcal{W}_{t'})_{i,j} \right]^2,$$

em contribuições de indivíduos, elemento a elemento:

$$\frac{p_i \sum_{j=1}^{n} p_j \left[(\mathcal{W}_t)_{i,j} - (\mathcal{W}_{t'})_{i,j} \right]^2}{d^2_{HS}(\mathcal{W}_t, \mathcal{W}_{t'})}.$$

Esta decomposição dá origem a uma matriz de dimensão $n \times \left[\frac{1}{2} T (T-1)\right]$ que permite detectar quais os indivíduos que mais contribuem para as oposições entre pares de objectos, conforme expõe Lavit *et al.* [28].

Outra decomposição possível será a da soma dos quadrados das distâncias entre todos os pares de objectos, em contribuições de indivíduos:

$$\frac{p_i \sum_{t=1}^{T} \sum_{t'=1}^{T} \sum_{j=1}^{n} p_j \left[(\mathcal{W}_t)_{i,j} - (\mathcal{W}_{t'})_{i,j} \right]^2}{\sum_{t=1}^{T} \sum_{t'=1}^{T} d^2_{HS}(\mathcal{W}_t, \mathcal{W}_{t'})},$$

dando origem a um vector de dimensão $n \times 1$.

Outra forma de salientar os indivíduos que mais contribuem para as diferenças entre objectos será a representação das **trajectórias**, que se efectua na imagem euclidiana do compromisso e consiste em representar nesta imagem as T nuvens de indivíduos, sendo a t-ésima nuvem definida pelas variáveis do quadro X_t. Desta forma obtém-se uma representação de nT pontos com n trajectórias, cada uma com T pontos. A técnica implementada para esta representação é semelhante à dos pontos suplementares; as diferentes posições de um indivíduo definem a sua trajectória.

As coordenadas dos pontos compromisso $B_1, B_2, \ldots, B_n$ sobre o k-ésimo eixo são as componentes do vector de dimensão n

$$\sqrt{\mu_k}\boldsymbol{\varepsilon_k} = \frac{1}{\sqrt{\mu_k}} \mathcal{W} D \boldsymbol{\varepsilon_k}.$$

Considerando cada objecto como elemento suplementar, as coordenadas dos pontos $B_1^t, B_2^t, \ldots, B_n^t$ sobre o k-ésimo eixo são

$$\frac{1}{\sqrt{\mu_k}} \mathcal{W}_t D \boldsymbol{\varepsilon_k}$$

para $t = 1, \ldots, T$.

Saliente-se que nenhum destes pontos interveio na construção da imagem euclidiana do compromisso, mas todos eles podem ser representados nela.

Se os estudos forem caracterizados pelos objectos normados $\mathcal{W}_t/\|\mathcal{W}_t\|_{HS}$, as coordenadas dos pontos $B_1^t, B_2^t, \ldots, B_n^t$ sobre o k-ésimo eixo são

$$\frac{1}{\sqrt{\mu_k}} \frac{1}{\|\mathcal{W}_t\|_{HS}} \mathcal{W}_t D \boldsymbol{\varepsilon_k}$$

para $t = 1, \ldots, T$.

Propriedade 2.1 *O ponto compromisso B_i é o centro de gravidade dos pontos $B_i^1, B_i^2, \ldots, B_i^T$ ponderados pelos coeficientes $\alpha_1, \alpha_2, \ldots, \alpha_T$; esta propriedade mantém-se em projecção.*

As trajectórias permitem evidenciar quais os indivíduos responsáveis pelos desvios entre os quadros X_t e $X_{t'}$. De facto, a distância entre os pontos B_i^t e $B_i^{t'}$ sobre o k-ésimo eixo é directamente proporcional à distância entre objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$, $\|\mathcal{W}_t - \mathcal{W}_{t'}\|_{HS}$, sendo o coeficiente de dilatação $1/\sqrt{\mu_k}$.

Os coeficientes de dilatação $1/\sqrt{\mu_k}$ dispostos em ordem crescente de k, deformam cada vez mais as projecções dos indivíduos nos eixos. Esta é a razão pela qual a análise das trajectórias se limita aos dois primeiros eixos, ou, se não houver grande diferença entre μ_2 e μ_3, aos três primeiros eixos.

As trajectórias interpretam-se segundo a evolução de um indivíduo fictício cujos valores são as médias das variáveis de um estudo. Se as variáveis estiverem centradas por estudo, o indivíduo médio de cada estudo estará situado na origem da imagem euclidiana do compromisso e a sua trajectória reduzir-se-á a um único ponto.

É possível distinguir dois casos em relação à forma que o sentido das trajectórias pode tomar:

- uma trajectória pouco alargada e em torno da sua posição compromisso corresponde a um indívidiuo com uma evolução muito próxima da evolução média; por outras palavras, para cada variável, o desvio entre o valor desta variável para este indivíduo e o indivíduo fictício médio é regular de um estudo para o outro;

- uma trajectória bastante alargada reflecte uma mudança da estrutura do indivíduo ao longo dos estudos, que difere da evolução média.

Se os pontos têm tendência a agrupar-se por variável, é conveniente elaborar o gráfico das correlações das variáveis com os eixos do compromisso, de forma

a estudar mais detalhadamente as trajectórias dos indivíduos. Deste modo é possível explicar os eixos do compromisso em função das variáveis e interpretar as trajectórias.

Se, pelo contrário, as correlações entre as variáveis de um estudo são fortes, os pontos do gráfico das correlações agrupam-se mais por estudo que por variável. Assim, não é possível descrever os eixos em função das variáveis, nem interpretar as trajectórias.

2.3 Método STATIS dual

Este método é análogo ao método STATIS e aplica-se quando se dispõe de T quadros de dados recolhidos sobre as mesmas variáveis mas em que os T grupos de indivíduos podem ser diferentes ao longo dos quadros.

Os T quadros de dados constam de n_t indivíduos sobre p variáveis $(t = 1, \ldots, T)$.

No instante t o quadro X_t é a matriz de dimensão $n_t \times p$

$$X_t = \begin{pmatrix} (x_1^1)^t & (x_1^2)^t & \ldots & (x_1^p)^t \\ (x_2^1)^t & (x_2^2)^t & \ldots & (x_2^p)^t \\ \vdots & \vdots & \ddots & \vdots \\ (x_{n_t}^1)^t & (x_{n_t}^2)^t & \ldots & (x_{n_t}^p)^t \end{pmatrix},$$

a j-ésima variável é o vector de $\mathbb{R}^{n_t}$

$$(\boldsymbol{x^j})^t = \begin{pmatrix} (x_1^j)^t \\ (x_2^j)^t \\ \vdots \\ (x_{n_t}^j)^t \end{pmatrix}$$

e o i-ésimo indivíduo é o vector de $\mathbb{R}^p$

$$\left((\boldsymbol{x_i})^t\right)' = \begin{pmatrix} (x_i^1)^t & (x_i^2)^t & \ldots & (x_i^p)^t \end{pmatrix}.$$

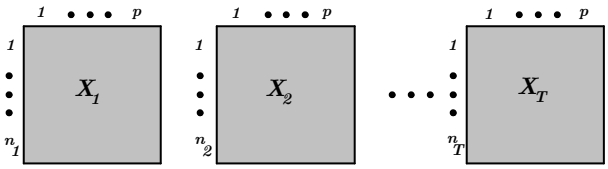


Figura 2.5 T quadros de dados: notação do método STATIS dual.

O estudo t é da forma (X_t, Q, D_t), $t = 1, \ldots, T$, e o objecto representativo associado é a matriz

$$\mathcal{V}_t = X_t' D_t X_t,$$

de dimensão $p \times p$ denominada **matriz de variâncias e covariâncias do quadro** X_t, em que D_t é a métrica associada às variáveis do quadro X_t e Q é a métrica associada aos indivíduos definida na secção 1.3.

O produto escalar utilizado para induzir uma distância entre os objectos $\mathcal{V}_t$ e $\mathcal{V}_{t'}$ também é o produto escalar de Hilbert-Schmidt:

$$\langle \mathcal{V}_t, \mathcal{V}_{t'} \rangle_{HS} = tr\left(\mathcal{V}_t Q \mathcal{V}_{t'} Q\right).$$

A **matriz $\mathcal{Z}$ dos produtos escalares** entre objectos $\mathcal{V}_t$ e $\mathcal{V}_{t'}$, de dimensão $T \times T$, tem como termo geral

$$\mathcal{Z}_{tt'} = \langle \mathcal{V}_t, \mathcal{V}_{t'} \rangle_{HS}.$$

Considerando os objectos normados $\mathcal{V}_t / \|\mathcal{V}_t\|_{HS}$ a matriz dos produtos escalares $\widetilde{\mathcal{Z}}$ tem como termo geral:

$$\widetilde{\mathcal{Z}}_{tt'} = \frac{\langle \mathcal{V}_t, \mathcal{V}_{t'} \rangle_{HS}}{\|\mathcal{V}_t\|_{HS} \|\mathcal{V}_{t'}\|_{HS}}.$$

A diagonalização da matriz $\mathcal{Z}\Delta$ permite a construção da **imagem euclidiana** dos objectos. No caso de os pesos atribuídos aos estudos serem todos iguais diagonalizar-se-á apenas a matriz $\mathcal{Z}$.

No caso da **imagem euclidiana centrada**, a matriz em questão é

$$\mathsf{C} = \left(I_T - \mathbf{1}_T \mathbf{1}_T' \Delta\right) \mathcal{Z} \left(I_T - \Delta \mathbf{1}_T \mathbf{1}_T'\right),$$

e a diagonalização da matriz $\mathsf{C}\Delta$ ou apenas da matriz C (no caso de os pesos atribuídos aos objectos serem todos iguais) permitirá a representação gráfica dos objectos.

O compromisso, designado por $\mathcal{V}$ é construído de forma análoga ao compromisso $\mathcal{W}$ obtido pelo método STATIS:

$$\mathcal{V} = \sum_{t=1}^{T} \beta_t \mathcal{V}_t,$$

e representa a matriz de variâncias e covariâncias entre as variáveis, no período $[1, T]$. Se as variáveis de cada quadro estiverem centradas e reduzidas, os objectos

$\mathcal{V}_t$ serão matrizes de correlações e a matriz compromisso será definida da seguinte forma

$$\mathcal{V} = \frac{\sum_{t=1}^{T} \beta_t \mathcal{V}_t}{\sum_{t=1}^{T} \beta_t},$$

de modo a obter um compromisso da mesma natureza dos objectos $\mathcal{V}_t$.

Os valores próprios e vectores próprios da matriz $\mathcal{V}Q$ fornecem a imagem euclidiana do compromisso aproximada das variáveis. Utilizando a técnica dos pontos suplementares, obtêm-se as trajectórias das variáveis nesta mesma imagem.

A matriz compromisso $\mathcal{V}$ é equivalente àquela que se teria obtido implementando uma ACP sobre o quadro de dados de dimensão $\sum_{t=1}^{T} n_t \times p$, construído a partir da sobreposição dos quadros $\sqrt{\beta_1}X_1, \ldots, \sqrt{\beta_T}X_T$, isto é, dispondo estes mesmos sobre a forma de uma só coluna (figura 2.6).

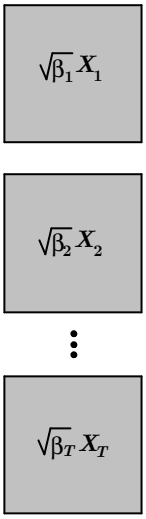


Figura 2.6 Sobreposição dos quadros no método STATIS dual.

Os métodos STATIS e STATIS dual devem ser ambos aplicados se os T quadros de dados possuírem os mesmos indivíduos e as mesmas variáveis.

Capítulo 3

Análise Factorial Múltipla

3.1 Introdução

A Análise Factorial Múltipla (AFM) foi inicialmente estudada por Escofier e Pagès [10] em 1985 e permite analisar uma população de indivíduos caracterizados por vários grupos de variáveis quantitativas ou qualitativas, embora nesta dissertação a abordagem recaia unicamente sobre variáveis quantitativas. Para o estudo das variáveis qualitativas, Escofier e Pagès [12] é uma referência importante.

De forma análoga à descrita pelo método STATIS, há T quadros de dados, $X_1, \ldots, X_T$, recolhidos em diferentes "ocasiões", sobre os mesmos indivíduos em que as variáveis podem diferir ao longo destes quadros. As etapas que constituem a AFM diferem daquelas propostas pelo método STATIS:

I) determinação dos valores próprios associados a cada um dos grupos de variáveis;

II) determinação do **compromisso**, dos eixos da **intra-estrutura** e representação simultânea das nuvens (**trajectórias**);

III) interpretação dos eixos da intra-estrutura segundo as variáveis;

IV) estudo da **inter-estrutura** e interpretação da posição dos quadros de dados;

V) interpretação das posições compromisso e das trajectórias dos indivíduos.

A primeira etapa da AFM consiste em efectuar T Análises em Componentes Principais de cada um dos quadros, determinando os valores e os vectores próprios das matrizes $\mathcal{W}_t D$. Seja λ_1^t o maior valor próprio da matriz $\mathcal{W}_t D$ para

$t = 1, \ldots, T$. Os coeficientes de ponderação utilizados na AFM são os inversos destes valores próprios. Esta ponderação tem como objectivo equilibrar o papel desempenhado pelos quadros durante a análise.

3.2 Intra-estrutura

O espaço $\mathbb{R}^{\sum_{t=1}^{T} p_t}$ contém as representações dos indivíduos e a partir deste há dois tipos de representação:

- representação da nuvem dos indivíduos, caracterizadas pelo conjunto de variáveis (compromisso);
- representação simultânea das T nuvens caracterizadas por cada grupo de variáveis (trajectórias dos indivíduos).

Seja $\mathcal{N}_I = \{\widetilde{\boldsymbol{x_i}} : i = 1, \ldots, n\}$ a nuvem dos indivíduos definida pelos vários quadros de dados $X_1, \ldots, X_T$ em $\mathbb{R}^{\sum_{t=1}^{T} p_t}$, com

$$\widetilde{\boldsymbol{x_i}} = \begin{pmatrix} (\boldsymbol{x_i})^1 \\ \vdots \\ (\boldsymbol{x_i})^T \end{pmatrix} = \begin{pmatrix} (x_i^1)^1 \\ \vdots \\ (x_i^{p_1})^1 \\ \vdots \\ (x_i^1)^T \\ \vdots \\ (x_i^{p_T})^T \end{pmatrix},$$

vector de $\mathbb{R}^{\sum_{t=1}^{T} p_t}$ e seja $\mathcal{N}_I^t = \{(\boldsymbol{x_i})^t : i = 1 \ldots, n\}$ a nuvem dos indivíduos definida pelo quadro de dados X_t. Para representar as nuvens $\mathcal{N}_I^t$, $t = 1, \ldots, T$, no espaço $\mathbb{R}^{\sum_{t=1}^{T} p_t}$, basta reparar que este espaço se pode decompor como soma directa de T subespaços isomórfos aos espaços $\mathbb{R}^{p_t}$:

$$\mathbb{R}^{\sum_{t=1}^{T} p_t} = \bigoplus_{t=1}^{T} \mathbb{R}^{p_t}. \tag{3.1}$$

As coordenadas dos pontos da nuvem $\mathcal{N}_I^t$ no espaço $\mathbb{R}^{\sum_{t=1}^{T} p_t}$ estão contidas num quadro $\widetilde{X_t}$ de dimensão $n \times \sum_{t=1}^{T} p_t$, no qual X_t é completado por zeros (figura 3.1).

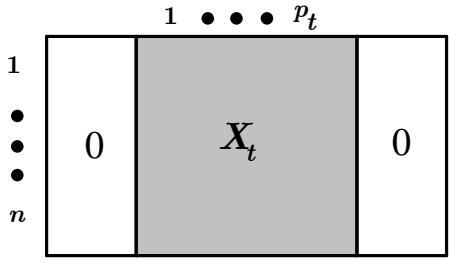


Figura 3.1 O quadro de dados $\widetilde{X_t}$.

Seja $\mathcal{N}_i^T = \{(\boldsymbol{x_i})^t : t = 1 \ldots, T\}$ a nuvem do espaço $\mathbb{R}^{\sum_{t=1}^T p_t}$ que fornece as T imagens do i-ésimo indivíduo e $\mathcal{N}_I^* = \{\boldsymbol{x_i}^* : i = 1, \ldots, n\}$ a nuvem dos centros de gravidade das nuvens $\mathcal{N}_i^T$, com $\boldsymbol{x_i}^* = \frac{1}{T}\sum_{t=1}^T (\boldsymbol{x_i})^t$. Com efeito, tendo em conta a relação (3.1), a nuvem $\mathcal{N}_I^*$ deduz-se da nuvem $\mathcal{N}_I$ por uma homotetia de razão $1/T$. Logo as nuvens $\mathcal{N}_I^*$ e $\mathcal{N}_I$ dizem-se **homotéticas**. Note-se que:

$$\bigcup_{t=1}^{T} \mathcal{N}_I^t = \bigcup_{i=1}^{n} \mathcal{N}_i^T.$$

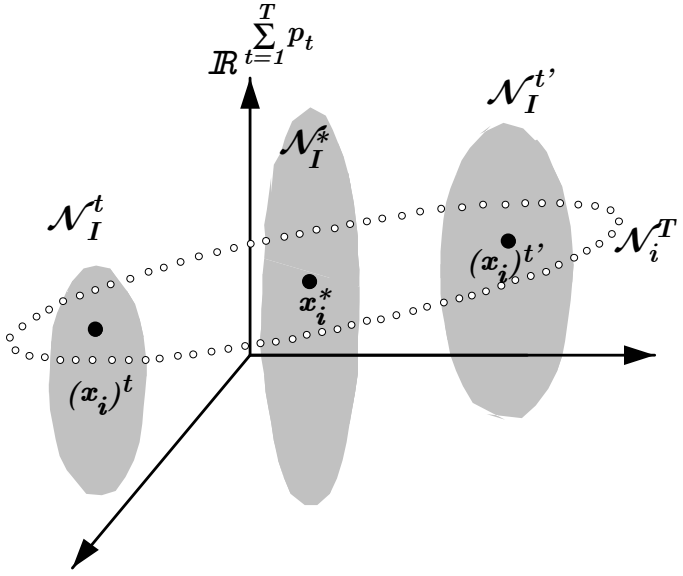


Figura 3.2 Relação entre as nuvens $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$ e $\mathcal{N}_i^T$ $(i = 1, \ldots, n)$.

3.2.1 AFM em $\mathbb{R}^{\sum_{t=1}^{T} p_t}$: representação dos indivíduos

A AFM tem a propriedade de equilibrar a influência dos diferentes grupos de variáveis dando a cada variável um peso, que deve ser o mesmo para todas as variáveis de um mesmo quadro, de forma a não modificar a forma das nuvens $\mathcal{N}_I^t$. O peso atribuído a cada uma das variáveis de um quadro X_t é igual ao inverso da inércia da primeira componente principal associada a este quadro. Em vez de se utilizar a métrica Q_t, utiliza-se a métrica $(1/\lambda_1^t) \cdot Q_t$, que dá uma ponderação diferenciada para os T grupos de variáveis e normaliza as diversas nuvens, uma vez que a inércia da primeira componente principal de cada grupo de variáveis é igual a 1.

A representação dos indivíduos obtém-se projectando a nuvem $\mathcal{N}_I$ ou a nuvem $\mathcal{N}_I^*$ (uma vez que estas são homotéticas) sobre um espaço de dimensão menor, de tal forma que a sua projecção se assemelhe o mais possível à nuvem $\mathcal{N}_I$. Para tal é necessário efectuar uma ACP ao quadro de dados formado pela justaposição dos quadros $X_1, \ldots, X_T$, notado por $\mathcal{X}$, utilizando a métrica $\mathcal{Q}$ definida por

$$\mathcal{Q} = \begin{pmatrix} \boxed{\frac{1}{\lambda_1^1} \cdot Q_1} & 0 & \ldots & 0 \\ 0 & \boxed{\frac{1}{\lambda_1^2} \cdot Q_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boxed{\frac{1}{\lambda_1^T} \cdot Q_T} \end{pmatrix},$$

de dimensão $\sum_{t=1}^{T} p_t \times \sum_{t=1}^{T} p_t$, ou, de forma equivalente, efectuar uma ACP da justaposição dos quadros

$$\frac{X_1}{\sqrt{\lambda_1^1}}, \ldots, \frac{X_T}{\sqrt{\lambda_1^T}},$$

a partir da métrica

$$\mathcal{Q} = \begin{pmatrix} \boxed{Q_1} & 0 & \ldots & 0 \\ 0 & \boxed{Q_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boxed{Q_T} \end{pmatrix}.$$

A partir desta ACP ponderada é possível obter a imagem euclidiana do compromisso, que resume a informação contida nos T quadros num só. Efectuando uma ACP ponderada à nuvem $\mathcal{N}_I^*$, a imagem euclidiana é idêntica, uma vez que, as nuvens $\mathcal{N}_I$ e $\mathcal{N}_I^*$ são homotéticas. A expressão geral do

compromisso é:

$$\mathcal{W} = \mathcal{X}\mathcal{Q}\mathcal{X}' = \sum_{t=1}^{T} \frac{\mathcal{W}_t}{\lambda_1^t}.$$

Para interpretar as posições compromisso, determinam-se as correlações das variáveis com os eixos do compromisso, tal como no método STATIS.

3.2.2 Representação simultânea das T nuvens $\mathcal{N}_I^t$

Para efectuar a representação simultânea das T nuvens de forma a construir as trajectórias dos indivíduos é necessário projectá-las num subespaço de $\mathbb{R}^{\sum_{t=1}^{T} p_t} = \bigoplus_{t=1}^{T} \mathbb{R}^{p_t}$, de forma a permitir a comparação simultânea de um mesmo indivíduo nas diversas nuvens $\mathcal{N}_I^1, \ldots, \mathcal{N}_I^T$. Para que isto aconteça, a escolha deste subespaço deve satisfazer duas condições essenciais:

I) Cada nuvem $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$ deve estar bem representada;

II) As representações das nuvens $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$ devem ser semelhantes.

No que diz respeito à condição **I**, deve-se fazer uma projecção ortogonal das respectivas nuvens no subespaço, sendo a qualidade de representação destas medida pela soma das inércias das nuvens projectadas. Pretende-se então maximizar a inércia da união das nuvens $\mathcal{N}_I^t$: $\bigcup_{t=1}^{T} \mathcal{N}_I^t$. Seja $\boldsymbol{x^*}$ o centro de gravidade de $\bigcup_{t=1}^{T} \mathcal{N}_I^t$, tal que:

$$\boldsymbol{x^*} = \frac{1}{n \cdot T} \sum_{i=1}^{n} \sum_{t=1}^{T} (\boldsymbol{x_i})^t.$$

Não é possível comparar as posições de um indivíduo nas diversas nuvens se as representações deste nas mesmas forem muito diferentes. Para assegurar a condição **II** é necessário que os pontos que representam o mesmo indivíduo nos vários quadros, $(\boldsymbol{x_i})^1, \ldots, (\boldsymbol{x_i})^T$, estejam próximos uns dos outros. Por outras palavras, é necessário minimizar a inércia das nuvens $\mathcal{N}_i^T$ em torno dos respectivos centros de gravidade $\boldsymbol{x_i}^*$. As condições **I** e **II** são incompatíveis, uma vez que a qualidade de representação das nuvens e a semelhança entre estas representações não podem ser optimizadas em simultâneo.

Segundo o Teorema de Huyghens, já descrito pela expressão (1.11) e aplicado neste contexto, a **inércia total** de $\bigcup_{t=1}^{T} \mathcal{N}_I^t$ decompõe-se na soma da inércia da nuvem $\mathcal{N}_I^*$ (**inércia inter**) com a inércia das nuvens $\mathcal{N}_i^T$ (**inércia intra**), como exemplifica a figura 3.3.

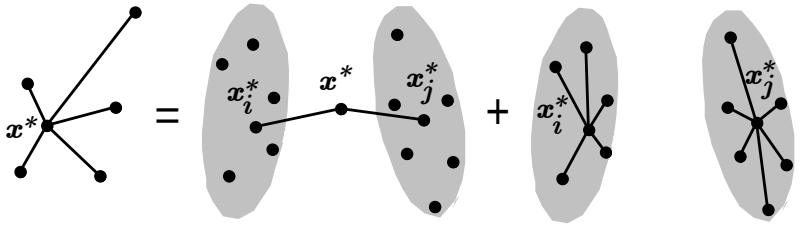


Figura 3.3 Inércia total= Inércia inter + Inércia intra.

O teorema de Huyghens sugere que se minimize a inércia intra e maximize a inércia total, tornando a inércia inter máxima.

O subespaço de $\mathbb{R}^{\sum_{t=1}^{T} p_t}$ sobre o qual a projecção de $\bigcup_{t=1}^{T} \mathcal{N}_I^t$ tem inércia inter máxima é formado pelos primeiros eixos de inércia da nuvem $\mathcal{N}_I^*$ ou da nuvem $\mathcal{N}_I$, uma vez que estas nuvens são homotéticas. Assim, o subespaço procurado obter-se-á efectuando uma ACP ao quadro $\mathcal{X}$. As coordenadas dos pontos de $\mathcal{N}_I^t$ estão contidas no quadro $\widetilde{X_t}$, logo, considerando cada $\widetilde{X_t}$ como elemento suplementar, obtém-se a representação simultânea das nuvens $\mathcal{N}_I^t$ e, desta forma, as trajectórias dos indivíduos são obtidas de forma semelhante àquela utilizada no método STATIS.

A projecção das nuvens

A nuvem $\mathcal{N}_I^t$, pertencente ao subespaço $\mathbb{R}^{p_t}$, é então projectada sobre um vector $\boldsymbol{u_k}$ de $\mathbb{R}^{\sum_{t=1}^{T} p_t}$, que não pertence a $\mathbb{R}^{p_t}$. A projecção de $\mathcal{N}_I^t$ sobre $\boldsymbol{u_k}$ consiste na projecção sobre um vector $\boldsymbol{u_k}^t$, seguida de uma projecção sobre $\boldsymbol{u_k}$, multiplicando as suas componentes por $\cos\theta_k^t$, sendo este o ângulo formado pelos vectores $\boldsymbol{u_k}$ e $\boldsymbol{u_k}^t$ (figura 3.4).

Esta sucessão de projecções leva a questionar se não seria mais vantajoso conservar as projecções sobre $\boldsymbol{u_k}^t$ na representação simultânea das nuvens. Ora, isto faria com que se representassem as nuvens em espaços munidos de métricas diferentes, impossibilitando a comparação entre estas. Para além disso, os eixos $\boldsymbol{u_k}$ são ortogonais em $\mathbb{R}^{\sum_{t=1}^{T} p_t}$, já não se passando o mesmo com os eixos $\boldsymbol{u_k}^t$.

Qualidade da representação de cada nuvem

A qualidade de representação de cada nuvem $\mathcal{N}_I^t$ mede-se de forma já conhecida e abordada em 1.8.2, através da razão entre a inércia projectada e a inércia total da nuvem. Mas, na AFM, esta medida de qualidade é pouco fiável uma vez que o vector $\boldsymbol{u_k}$ do espaço $\mathbb{R}^{\sum_{t=1}^{T} p_t}$, sobre o qual a nuvem $\mathcal{N}_I^t$ é

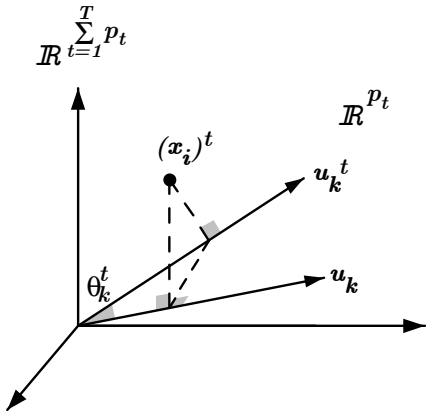


Figura 3.4 Projecção do indivíduo $(\boldsymbol{x_i})^t$ no espaço $\mathbb{R}^{\sum_{t=1}^{T} p_t}$.

projectada, não pertence ao subespaço $\mathbb{R}^{p_t}$ no qual esta nuvem se situa (figura 3.4). Tal como se verificou em 1.8.2, tem-se que $\cos^2 \theta_k^t$ é uma medida de qualidade de representação do indivíduo $(\boldsymbol{x_i})^t$ projectado no eixo $\boldsymbol{u_k}$.

Validade da representação simultânea

Sendo o objectivo desta análise minimizar a inércia intra das nuvens $\mathcal{N}_i^T$ em torno dos $\boldsymbol{x_i}^*$, de forma a que os pontos $(\boldsymbol{x_i})^t$, $t = 1, \ldots, T$, que representam o i-ésimo indivíduo, estejam o mais próximo possível uns dos outros. Pode-se então tomar como medida de semelhança entre as projecções das nuvens $\mathcal{N}_I^t$ sobre um determinado eixo, a inércia intra. Mas este valor só terá significado se ponderado pela inércia total, logo interessa calcular para cada eixo a razão:

$$\frac{\text{inércia inter}}{\text{inércia total}} = 1 - \frac{\text{inércia intra}}{\text{inércia total}}.$$

Esta razão não é a quantidade directamente minimizada, logo não decresce necessariamente com a ordem dos eixos, contrariamente ao método STATIS. No entanto, representa um indicador global da qualidade da representação simultânea das nuvens $\mathcal{N}_I^t$.

Com uma razão [inércia inter/inércia total] próxima de 1, há uma maior estabilidade entre as nuvens $\mathcal{N}_I^t$, permitindo um estudo pormenorizado das suas diferenças para aquele eixo. Se esta razão estiver próxima de 0, é sinal de que as diferenças de forma são tão grandes que não há interesse em estudar as trajectórias.

3.2.3 AFM em $\mathbb{R}^n$: representação das variáveis

A AFM em $\mathbb{R}^{\sum_{t=1}^{T} p_t}$ permitiu visualizar as posições "médias" de cada indivíduo e as respectivas trajectórias no plano compromisso. A representação simultânea dos T grupos de variáveis no espaço $\mathbb{R}^n$ é obtida através de uma ACP do quadro $\mathcal{X}$ e permite a interpretação da representação da nuvem dos indivíduos e das correlações entre estas variáveis. Esta representação não é mais do que a dual da representação de $\mathcal{N}_I$ em $\mathbb{R}^{\sum_{t=1}^{T} p_t}$.

A inércia projectada de cada nuvem $\mathcal{N}_J^t$, formada pelo grupo de variáveis do quadro X_t, pode ser interpretada como a contribuição deste grupo para a formação dos eixos. A ponderação destes grupos por $1/\lambda_1^t$ equilibra a sua influência no sentido em que a contribuição de cada grupo para a construção de um eixo é, no máximo, 1.

A comparação de grupos de variáveis pode ser efectuada através de um estudo sistemático das correlações entre as primeiras componentes principais de cada grupo (Escofier e Pagès [12]). Esse estudo baseia-se numa ACP não normada das componentes principais de todos os grupos. Sendo as componentes principais do quadro X_t, projecções dos indivíduos sobre uma base ortonormada, as nuvens de indivíduos definidas pela ACP do quadro X_t e pela ACP do quadro definido pelas componentes principais de X_t são idênticas.

Para proceder à comparação das componentes principais dos vários grupos, basta considerá-las como elementos suplementares na análise do quadro completo $\mathcal{X}$. Outra alternativa possível seria efectuar uma ACP das componentes principais, tomando as variáveis como elementos suplementares. Conservando apenas as primeiras componentes de cada grupo, tendo como vantagens a franca diminuição da dimensão dos quadros, bem como da duração dos cálculos, obter-se-iam resultados aproximados em que a qualidade da aproximação depende da inércia máxima da componente principal abandonada.

3.3 Inter-estrutura

3.3.1 AFM em $\mathbb{R}^{n^2}$: representação dos grupos de variáveis

Esta etapa consiste na comparação global dos T grupos de variáveis, através da sua representação num plano. Seja $\mathcal{N}_J$ a nuvem dos grupos de variáveis definida pelos quadros $X_1, \ldots, X_T$. No método STATIS, foi definido um objecto representativo de cada estudo da forma $\mathcal{W}_t = X_t Q_t X_t'$ e o produto escalar

utilizado era o produto de Hilbert-Schmidt:

$$\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS} = tr\left(\mathcal{W}_t D \mathcal{W}_{t'} D\right),$$

em que

$$\mathcal{W}_t = \sum_{j=1}^{p_t} q_j \boldsymbol{x^j} (\boldsymbol{x^j})'. \tag{3.2}$$

Em vez de se escrever $(\boldsymbol{x^j})^t$ optou-se por escrever $\boldsymbol{x^j}$ de forma a não sobrecarregar a notação utilizada.

Cada $\mathcal{W}_t$ é de dimensão $n \times n$, logo este objecto pertence ao espaço vectorial de dimensão n^2 notado por $\mathbb{R}^{n^2}$.

Na AFM, os objectos utilizados vão ser ponderados, da forma $\mathcal{W}_t/\lambda_1^t$. A ponderação por $1/\lambda_1^t$ traduz-se em $\mathbb{R}^{n^2}$ por uma homotetia do vector representativo deste grupo. Daí a nuvem $\mathcal{N}_J$ se assemelhar mais à nuvem das variáveis que à nuvem dos indivíduos.

O quadrado da norma dos objectos representativos de cada grupo de variáveis de X_t, em $\mathbb{R}^{n^2}$ é:

$$\begin{aligned} \eta_g^2(X_t) &= \left\| \frac{\mathcal{W}_t}{\lambda_1^t} \right\|_{HS}^2 \\ &= \left\langle \frac{\mathcal{W}_t}{\lambda_1^t}, \frac{\mathcal{W}_t}{\lambda_1^t} \right\rangle_{HS} \\ &= tr\left(\frac{\mathcal{W}_t}{\lambda_1^t} D \right)^2 \\ &= \sum_{i=1}^{n} \frac{(\lambda_i^t)^2}{(\lambda_1^t)^2}, \end{aligned} \tag{3.3}$$

com λ_i^t o i-ésimo valor próprio de $\mathcal{W}_t D$. Esta medida é sempre superior ou igual a 1, sendo tanto maior quanto maior for número de valores próprios de importância comparável ao primeiro valor próprio, logo depende da estrutura de cada grupo de variáveis. Pode-se então dizer que $\eta_g^2(X_t)$ constitui um índice de dimensionalidade do grupo de variáveis X_t.

Considere-se, neste capítulo, que os objectos $\mathcal{W}_t$ já estão ponderados por $1/\lambda_1^t$.

Interpretação do produto escalar entre dois grupos

O produto escalar entre dois grupos pode-se interpretar como uma medida de ligação entre estes.

I) Cada um dos dois objectos representativos de cada grupo compreende uma única variável.

Sejam $\boldsymbol{u}$ e $\boldsymbol{v}$ duas variáveis centradas e reduzidas, constituindo cada uma delas um grupo, afectadas dos respectivos coeficientes de ponderação. Os elementos de $\mathbb{R}^{n^2}$ associados a estes dois grupos têm norma 1, como se pode verificar a partir da expressão (3.3) e o produto escalar entre estes é o quadrado do coeficiente de correlação entre as variáveis $\boldsymbol{u}$ e $\boldsymbol{v}$:

$$\begin{aligned}
\langle \boldsymbol{u}\boldsymbol{u}', \boldsymbol{v}\boldsymbol{v}' \rangle_{HS} &= tr(\boldsymbol{u}\boldsymbol{u}' D \boldsymbol{v}\boldsymbol{v}' D) \\
&= \sum_{i=1}^{n}\sum_{l=1}^{n} p_i p_l u_i u_l v_i v_l \\
&= \sum_{i=1}^{n} p_i u_i v_i \sum_{l=1}^{n} p_l u_l v_l \\
&= \left(\sum_{i=1}^{n} p_i u_i v_i \right)^2 \\
&= \left[r\left(\boldsymbol{u}, \boldsymbol{v}\right) \right]^2 \\
&= \text{inércia da projecção de } \boldsymbol{v} \text{ sobre } \boldsymbol{u}.
\end{aligned}$$

II) Um dos grupos compreende uma variável e outro compreende diversas variáveis.

Seja $\boldsymbol{u}$ uma variável centrada e reduzida do grupo $X_{t'}$ e $\boldsymbol{x^j}$ o conjunto de variáveis, centradas e reduzidas ao peso q_j, do grupo X_t com $j = 1, \ldots, p_t$. Tem-se que:

$$\begin{aligned}
\langle \mathcal{W}_{t'}, \mathcal{W}_t \rangle_{HS} &= \left\langle \boldsymbol{u}\boldsymbol{u}', \sum_{j=1}^{p_t} q_j \boldsymbol{x^j} (\boldsymbol{x^j})' \right\rangle_{HS} \\
&= \sum_{j=1}^{p_t} q_j \left\langle \boldsymbol{u}\boldsymbol{u}', \boldsymbol{x^j}(\boldsymbol{x^j})' \right\rangle_{HS} \\
&= \sum_{j=1}^{p_t} q_j \left[r(\boldsymbol{u}, \boldsymbol{x^j}) \right]^2 \\
&= \sum_{j=1}^{p_t} \text{inércia da projecção de } \boldsymbol{x^j} \text{ sobre } \boldsymbol{u},
\end{aligned}$$

que constitui uma medida de ligação entre a variável $\boldsymbol{u}$ e o grupo de variáveis de X_t, que se denotará por $\mathcal{L}_g\left(\boldsymbol{u}, X_t\right)$, e que foi introduzida por Carroll [5] na abordagem à Análise Canónica Generalizada.

III) Os dois grupos compreendem diversas variáveis.

Seja $\boldsymbol{x^j}$ o conjunto de variáveis, centradas e reduzidas ao peso q_j do grupo X_t $(j = 1, \ldots, p_t)$ e $\boldsymbol{x^k}$ o conjunto de variáveis, centradas e reduzidas ao peso q_k do grupo $X_{t'}$ $(k = 1, \ldots, p_{t'})$. O produto escalar entre os objectos representativos de X_t e $X_{t'}$ é:

$$\begin{aligned}\langle \mathcal{W}_{t'}, \mathcal{W}_t \rangle_{HS} &= \sum_{j=1}^{p_t} q_j \sum_{k=1}^{p_{t'}} q_k \left\langle \boldsymbol{x^j}(\boldsymbol{x^j})', \boldsymbol{x^k}(\boldsymbol{x^k})' \right\rangle_{HS} \\ &= \sum_{j=1}^{p_t} q_j \cdot \mathcal{L}_g(\boldsymbol{x^j}, X_{t'}) \\ &= \sum_{k=1}^{p_{t'}} q_k \cdot \mathcal{L}_g(\boldsymbol{x^k}, X_t). \end{aligned} \tag{3.4}$$

Esta medida pode-se explicitar em função de dois grupos de variáveis (afectados dos respectivos coeficientes de ponderação) sob a forma:

$$\begin{aligned}\mathcal{L}_g\left(X_t, X_{t'}\right) &= \left\langle \frac{\mathcal{W}_t}{\lambda_1^t}, \frac{\mathcal{W}_{t'}}{\lambda_1^{t'}} \right\rangle_{HS} \\ &\overset{(3.4)}{=} \sum_{j=1}^{p_t} q_j \cdot \mathcal{L}_g(\boldsymbol{x^j}, X_{t'}) \\ &= \sum_{j=1}^{p_t} q_j \sum_{k=1}^{p_{t'}} q_k \left[r(\boldsymbol{x^j}, \boldsymbol{x^k})\right]^2. \end{aligned}$$

Esta expressão constitui uma medida de ligação entre os dois grupos de variáveis e é tanto maior quanto mais ligada estiver cada uma das variáveis de um grupo às variáveis do outro e quanto maior for o número de direcções comuns às direcções de inércia máxima de cada grupo.

Importa salientar que ao grupo formado pela variável centrada e reduzida definida pelo eixo $\boldsymbol{v_k}$ de $\mathbb{R}^n$ está associado o objecto representativo desta,

$$\mathcal{W}_{\boldsymbol{v_k}} = \boldsymbol{v_k}\boldsymbol{v_k}'$$

em $\mathbb{R}^{n^2}$. Estes objectos devem ser de ordem 1, no sentido em que cada um deles está associado a uma única direcção de $\mathbb{R}^n$. A inércia projectada das variáveis do grupo X_t sobre $\boldsymbol{v_k}$ não é mais do que o produto escalar (de Hilbert-Schmidt) entre $\mathcal{W}_t$ e $\boldsymbol{v_k}\boldsymbol{v_k}'$, isto é, a projecção de $\mathcal{W}_t$ sobre $\boldsymbol{v_k}\boldsymbol{v_k}'$ (figura 3.5). Logo, se dois vectores são D-ortogonais em $\mathbb{R}^n$ então os vectores associados, em $\mathbb{R}^{n^2}$, também são D-ortogonais. Desta forma, a representação gráfica dos grupos de

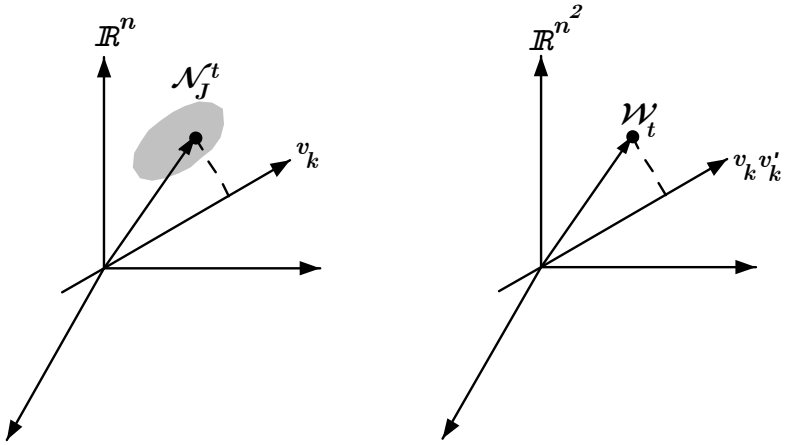


Figura 3.5 Representação dos grupos de variáveis em $\mathbb{R}^n$ e em $\mathbb{R}^{n^2}$.

variáveis pode ser interpretada como uma projecção da nuvem $\mathcal{N}_J$ sobre um conjunto de eixos ortogonais. Para proceder à comparação global dos grupos, devem-se analisar as proximidades das projecções dos objectos $\mathcal{W}_t$ num espaço de dimensão r, bastante inferior à de $\mathbb{R}^{n^2}$.

Os ângulos entre os objectos devem estar bem representados e não é conveniente centrar a nuvem $\mathcal{N}_J$. A projecção desta nuvem sobre os eixos principais é análoga à da nuvem das variáveis na ACP. O único inconveniente deste tipo de análise será o de fornecer um referencial constituído por eixos de difícil interpretação, uma vez que cada um destes eixos não se exprime explicitamente em função dos dados. É por esta razão que se impõe que os elementos $\boldsymbol{v_k v_k}'$ $(k = 1, \ldots, r)$ sejam D-simétricos, de ordem 1. Estes elementos estão associados unicamente a $\boldsymbol{v_k}$ e interpretam-se a partir das correlações deste vector com as variáveis iniciais.

Na ACP, utilizou-se o critério dos mínimos quadrados, segundo o qual se maximizava a soma dos quadrados das projecções dos vectores das nuvens. Devido ao tipo de eixos utilizados na AFM (elementos D-simétricos de ordem 1), a quantidade maximizada é a soma das projecções. Uma vez que as coordenadas dos objectos $\mathcal{W}_t$ sobre os elementos do tipo $\boldsymbol{v_k v_k}'$ são todas positivas tem-se que a soma das projecções de $\mathcal{W}_t$ sobre $\boldsymbol{v_k v_k}'$,

$$\sum_{t=1}^{T} \langle \mathcal{W}_t, \boldsymbol{v_k v_k}' \rangle_{HS} ,$$

é a inércia das variáveis (de todos os grupos) projectadas sobre $\boldsymbol{v_k}$. O conjunto ortonormado de elementos D-simétricos de ordem 1 que maximizam esta soma,

está associado às componentes principais do quadro $\mathcal{X}$. Logo, os cálculos necessários à análise em $\mathbb{R}^{n^2}$ deduzem-se directamente dos resultados da ACP aplicada a $\mathcal{X}$. Os vectores $\boldsymbol{v_k}$ são as componentes principais normadas de $\mathcal{X}$ e a coordenada de $\mathcal{W}_t$ sobre $\boldsymbol{v_k}\boldsymbol{v_k}'$ é a contribuição do grupo X_t à inércia da componente principal $\boldsymbol{v_k}$.

A representação dos grupos de variáveis pode ser encarada como uma ajuda complementar na interpretação dos outros gráficos. A coordenada do objecto $\mathcal{W}_t$ (já ponderado) sobre o eixo $\boldsymbol{v_k}\boldsymbol{v_k}'$ varia entre 0 e 1 e pode ser interpretada como:

- a inércia da projecção da nuvem $\mathcal{N}_J^t$ em $\mathbb{R}^n$ sobre a componente principal $\boldsymbol{v_k}$ do quadro $\mathcal{X}$ (figura 3.5);
- uma medida de ligação entre a componente principal $\boldsymbol{v_k}$ e as variáveis do quadro X_t, já designada por $\mathcal{L}_g(\boldsymbol{v_k}, X_t)$;
- a projecção do grupo X_t no espaço $\mathbb{R}^{n^2}$.

Quando $\mathcal{L}_g(\boldsymbol{v_k}, X_t) = 1$ significa que $\boldsymbol{v_k}$ é a primeira componente principal de X_t. Por outro lado, se $\mathcal{L}_g(\boldsymbol{v_k}, X_t) = 0$, a variável $\boldsymbol{v_k}$ não está correlacionada com nenhuma das variáveis de X_t, e se estiver próxima de 0 indica uma direcção de inércia fraca para $\mathcal{N}_J^t$, mas que pode suscitar alguma ambiguidade. Deve-se então calcular o coeficiente de correlação entre as componentes principais associadas ao quadro completo $\mathcal{X}$ e ao quadro X_t

$$\boldsymbol{F_{u_k}} = \mathcal{X}\mathcal{Q}\boldsymbol{u_k} = \sqrt{\lambda_k}\boldsymbol{v_k} \tag{3.5}$$

e

$$\boldsymbol{F_{u_k}}^t = \widetilde{X_t}\mathcal{Q}\boldsymbol{u_k}, \tag{3.6}$$

com $\boldsymbol{u_k} = \frac{1}{\lambda_k}\mathcal{X}'D\boldsymbol{F_{u_k}}$, respectivamente. Estas componentes estão ligadas através da seguinte relação:

$$\begin{aligned}\boldsymbol{F_{u_k}}^t &\overset{(3.6)}{=} \frac{1}{\lambda_k}\widetilde{X_t}\mathcal{Q}\mathcal{X}'D\boldsymbol{F_{u_k}} \\ &= \frac{1}{\lambda_k}\mathcal{W}_t D\boldsymbol{F_{u_k}} \\ &\overset{(3.5)}{=} \frac{1}{\sqrt{\lambda_k}}\mathcal{W}_t D\boldsymbol{v_k}\end{aligned}$$

com λ_k valor próprio de $\mathcal{W}D$ de ordem k, em que $\mathcal{W} = \sum_{t=1}^{T}\mathcal{W}_t$. O coeficiente de correlação $r(\boldsymbol{F_{u_k}}, \boldsymbol{F_{u_k}}^t)$ é um indicador da presença de $\boldsymbol{F_{u_k}}$ no grupo X_t.

Tem-se então que a imagem compromisso do i-ésimo indivíduo é o baricentro das projecções deste nas nuvens $\mathcal{N}_I^t$:

$$F_{\boldsymbol{u_k}}(i) = \frac{1}{T} \sum_{t=1}^{T} F_{\boldsymbol{u_k}}{}^t(i).$$

A qualidade de representação dos objectos $\mathcal{W}_t$ pode ser avaliada através da razão entre a inércia projectada e a inércia total, que é inferior e raramente igual a 1, mesmo que se aumente o número de eixos até n. Isto deve-se à forma como foram definidos os eixos (D-simétricos e de ordem 1).

Capítulo 4

Dupla Análise em Componentes Principais

4.1 Introdução

A Dupla Análise em Componentes Principais (DACP) foi inicialmente introduzida por Bouroche [2] em 1975 e destina-se à análise de um conjunto de variáveis sobre uma população de indivíduos em instantes diferentes. O objectivo principal da DACP é, tal como do método STATIS, STATIS dual e da AFM, analisar globalmente a evolução das proximidades entre indivíduos assim como relações entre as variáveis. As características dos dados em estudo são as seguintes:

- as variáveis são reais, positivas ou negativas, quantitativas e homogéneas (isto é, têm as mesmas unidades de medida), caso contrário terão que ser centradas e reduzidas;

- os quadros de dados $X_1, \ldots, X_T$ são obtidos cronologicamente; a DACP também pode ser utilizada no caso de os dados não serem temporais, no entanto, a interpretação dos resultados tornar-se-á mais difícil;

- os mesmos indivíduos e as mesmas variáveis são medidos em cada instante; esta restrição pode ser evitada se se analisarem apenas os indivíduos e as variáveis invariantes e aqueles que forem ocasionais serão tratados como suplementares.

A DACP compreende três etapas fundamentais:

I) inter-estrutura: análise da evolução global, através de uma ACP sobre a nuvem dos centros de gravidade de cada quadro X_t $(t = 1, \ldots, T)$;

II) estudo da **deformação das nuvens** de um instante ao outro, através de uma ACP sobre cada uma das nuvens $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$ em torno do respectivo centro de gravidade;

III) intra-estrutura: procura de um espaço comum de representação às várias nuvens, que sintetize da melhor forma possível as diferenças e semelhanças existentes entre estas.

Os T quadros de dados $X_1, \ldots, X_T$ constam de n indivíduos sobre p variáveis.

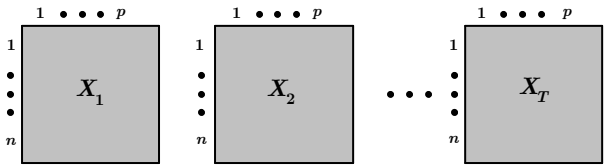


Figura 4.1 T quadros de dados: notação da DACP.

O centro de gravidade do quadro X_t associado à matriz (métrica) D definida em (1.1) é expresso por:

$$\boldsymbol{g}_t = \begin{pmatrix} (\bar{x}^1)^t \\ (\bar{x}^2)^t \\ \vdots \\ (\bar{x}^p)^t \end{pmatrix},$$

onde

$$\left(\bar{x}^j\right)^t = \sum_{i=1}^{n} p_i \left(x_i^j\right)^t,$$

com $j = 1, \ldots, p$ e $t = 1, \ldots, T$. A métrica associada ao espaço das variáveis no instante t é a métrica Q definida no capítulo 1. As restantes notações são análogas àquelas utilizadas na metodologia STATIS e AFM.

4.2 Inter-estrutura

Seja

$$G = \begin{pmatrix} \boldsymbol{g}_1' \\ \vdots \\ \boldsymbol{g}_T' \end{pmatrix} = \begin{pmatrix} (\bar{x}^1)^1 & \dots & (\bar{x}^p)^1 \\ \vdots & \ddots & \vdots \\ (\bar{x}^1)^T & \dots & (\bar{x}^p)^T \end{pmatrix},$$

a matriz de dimensão $T \times p$, formada pelos centros de gravidade das nuvens $\mathcal{N}_I^t$ $(t = 1, \dots, T)$. Esta fase consiste, como já foi referido, numa ACP sobre a matriz G.

O estudo da inter-estrutura na DACP difere ligeiramente da metodologia STATIS e AFM, uma vez que estes dois últimos analisam as semelhanças (ou diferenças) entre os quadros centrados em relação ao respectivo centro de gravidade, enquanto que a DACP estuda a evolução global dos quadros através dos centros de gravidade de cada nuvem $\mathcal{N}_I^t$ $(t = 1, \dots, T)$. Daí a necessidade de os dados serem temporais, caso contrário, a interpretação da inter-estrutura poder-se-á tornar difícil.

4.3 Análise das T nuvens de indivíduos

A segunda etapa consiste em eliminar o fenómeno de evolução global (inter-estrutura) através de uma ACP sobre cada um dos T quadros de dados centrados

$$Y_t = X_t - \mathbf{1}_n \boldsymbol{g}_t' = (I_n - \mathbf{1}_n \mathbf{1}_n' D)\, X_t,$$

de termo geral

$$\left(y_i^j\right)^t = \left(x_i^j\right)^t - \left(\bar{x}^j\right)^t,$$

para $t = 1, \dots, T$.

Embora a análise a cada um dos quadros se torne vantajosa, uma vez que permite a representação gráfica e a avaliação da qualidade de representação já expostas no capítulo 1, poder-se-á tornar bastante morosa, se T for elevado.

As T Análises em Componentes Principais permitem determinar T sistemas de eixos principais

$$\left(\boldsymbol{u_k}^t\right)_{k=1,\dots,q}$$

em cada nuvem $\mathcal{N}_I^t$, $t = 1, \dots, T$. Dispõem-se então de $2T$ sistemas de eixos ortogonais:

- T sistemas de factores principais $(\boldsymbol{z_k}^t)_{k=1,\ldots,q}$ (vectores de $\mathbb{R}^p$), que são os vectores próprios de $Q\mathcal{V}_t$ associados aos q maiores valores próprios $\lambda_1^t, \ldots, \lambda_q^t$, para $t = 1, \ldots, T$;

- T sistemas de componentes principais $\left(\boldsymbol{F_{u_k}}^t\right)_{k=1,\ldots,q}$ (vectores de $\mathbb{R}^n$), que são os vectores próprios de $\mathcal{W}_t D$ associados aos q maiores valores próprios $\lambda_1^t, \ldots, \lambda_q^t$, para $t = 1, \ldots, T$;

É através destes sistemas que se vai determinar um espaço de representação comum aos T quadros de dados.

4.4 Intra-estrutura: critérios para o melhor sistema de eixos

Bouroche [2] propõe a pesquisa de dois sistemas de q vectores ortogonais que resumam da melhor forma possível, segundo quatro critérios, as semelhanças e diferenças entre os vários quadros de dados. A notação para os dois sistemas de eixos pretendidos é a seguinte:

- $(\boldsymbol{\nu_k})_{k=1,\ldots,q}$ para os factores principais;

- $(\boldsymbol{\varsigma_k})_{k=1,\ldots,q}$ para as componentes principais.

É neste espaço de representação comum que será possível representar as trajectórias de cada indivíduo ao longo do período considerado, através da projecção de cada um no sistema de eixos encontrado.

A quantidade de inércia explicada pelo k-ésimo factor principal da nuvem $\mathcal{N}_I^t$ é

$$\langle \boldsymbol{z_k}^t, \boldsymbol{z_k}^t \rangle_{\mathcal{V}_t} = \left(\boldsymbol{z_k}^t\right)' \mathcal{V}_t \boldsymbol{z_k}^t = \lambda_k^t.$$

Seja $\tau \in \{1, \ldots, T\}$ e $\tau \neq t$. A inércia de $\mathcal{N}_I^t$ explicada pelo subespaço formado pelos vectores $(\boldsymbol{z_k}^\tau)_{k=1,\ldots,q}$ é

$$\sum_{k=1}^{q} \left(\boldsymbol{z_k}^\tau\right)' \mathcal{V}_t \left(\boldsymbol{z_k}^\tau\right) \leq \lambda_1^t + \ldots + \lambda_q^t.$$

É de notar que a igualdade se efectua apenas se os subespaços formados pelos sistemas $(\boldsymbol{z_k}^t)_{k=1,\ldots,q}$ e $(\boldsymbol{z_k}^\tau)_{k=1,\ldots,q}$ coincidirem.

Seja o índice $\Phi(t,\tau)$ definido por

$$\Phi(t,\tau) = \frac{\sum_{k=1}^{q} \lambda_k^t - \sum_{k=1}^{q} (\boldsymbol{z_k}^\tau)' \mathcal{V}_t \boldsymbol{z_k}^\tau}{\sum_{k=1}^{q} \lambda_k^t}.$$

Este índice representa a percentagem de perda de inércia da nuvem $\mathcal{N}_I^t$ quando se passa à projecção dos seus indivíduos sobre os q primeiros factores principais da nuvem $\mathcal{N}_I^\tau$, em vez da sua projecção sobre os q primeiros factores principais da nuvem $\mathcal{N}_I^t$. Por outras palavras, a partir do momento em que se projecta a nuvem $\mathcal{N}_I^t$ no subespaço formado pelos vectores $(\boldsymbol{z_k}^\tau)_{k=1,\ldots,q}$, a percentagem de inércia explicada pela nuvem $\mathcal{N}_I^t$ diminui para $\Phi(t,\tau)$.

Note-se que $\Phi(t,\tau) \neq \Phi(\tau,t)$ e $\Phi(t,t) = 0$.

Efectuando a projecção das T nuvens $\mathcal{N}_I^1, \ldots, \mathcal{N}_I^T$ sobre o subespaço formado por $(\boldsymbol{z_k}^\tau)_{k=1,\ldots,q}$ a perda média de inércia é

$$\Phi(.,\tau) = \frac{1}{T} \sum_{t=1}^{T} \Phi(t,\tau).$$

O **primeiro critério** consiste em seleccionar o sistema $(\boldsymbol{z_k}^\tau)_{k=1,\ldots,q}$ tal que:

$$\Phi(.,\tau) = \min_{t=1,\ldots,T} \Phi(.,t).$$

Segundo este critério, os sistemas nos quais se vão representar as trajectórias dos indivíduos são

- $\boldsymbol{\nu_k} = \boldsymbol{z_k}^\tau$, para os factores principais;
- $\boldsymbol{\varsigma_k} = \boldsymbol{F_{u_k}}^\tau$, para as componentes principais,

com $k = 1, \ldots, q$.

Note-se que, de acordo com aquilo que foi exposto no capítulo 1, na relação (1.19), tem-se que

$$\boldsymbol{F_{u_k}}^\tau = Y_\tau \boldsymbol{z_k}^\tau.$$

A inércia da nuvem $\mathcal{N}_I^t$ explicada pelo sistema $(\boldsymbol{\nu_k})_{k=1,\ldots,q}$ é igual a

$$\sum_{k=1}^{q} (\boldsymbol{\nu_k})' \mathcal{V}_t \boldsymbol{\nu_k}.$$

O **segundo critério** consiste em maximizar a inércia de todas as nuvens projectadas

$$\sum_{t=1}^{T}\sum_{k=1}^{q}(\boldsymbol{\nu_k})'\,\mathcal{V}_t\boldsymbol{\nu_k}, \tag{4.1}$$

ou seja, o sistema $(\boldsymbol{\nu_k})_{k=1,\ldots,q}$ não é mais do que o conjunto dos q vectores próprios da matriz $\sum_{t=1}^{T}\mathcal{V}_tQ$ associados aos q maiores valores próprios, de acordo com o exposto na Análise em Componentes Principais.

O segundo critério pode ser comparado com o primeiro que consistia em maximizar a função

$$\Psi(.,\tau)=\frac{1}{T}\sum_{t=1}^{T}\Psi(t,\tau)$$

em que

$$\Psi(t,\tau)=\frac{\sum_{k=1}^{q}(\boldsymbol{z_k}^{\tau})'\mathcal{V}_t\boldsymbol{z_k}^{\tau}}{\sum_{k=1}^{q}\lambda_k^t}$$

é o índice representativo da percentagem de inércia da nuvem $\mathcal{N}_I^t$ explicada pelo sistema $(\boldsymbol{z_k}^{\tau})_{k=1,\ldots,q}$. Ora a solução encontrada é a mesma, quer se maximize $\Psi(.,t)$ ou minimize $\Phi(.,t)$.

Pretende-se então, de entre os sistemas obtidos, um sistema $(\boldsymbol{z_k}^{\tau})_{k=1,\ldots,q}$ tal que

$$\begin{aligned}\Psi(.,\tau)&=\frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{k=1}^{q}(\boldsymbol{z_k}^{\tau})'\mathcal{V}_t\boldsymbol{z_k}^{\tau}}{\sum_{k=1}^{q}\lambda_k^t}\\&=\frac{1}{T}\sum_{k=1}^{q}(\boldsymbol{z_k}^{\tau})'\left[\sum_{t=1}^{T}\frac{\mathcal{V}_t}{\sum_{k=1}^{q}\lambda_k^t}\right]\boldsymbol{z_k}^{\tau}\end{aligned}$$

seja máximo. Comparando esta relação com (4.1), verifica-se que estas não podem ser iguais, uma vez que cada matriz $\mathcal{V}_t$ está a ser normalizada pelo seu traço e

$$\sum_{t=1}^{T}\frac{\mathcal{V}_t}{\sum_{k=1}^{q}\lambda_k^t}\neq\sum_{t=1}^{T}\mathcal{V}_t,$$

o que não significa forçosamente que o sistema encontrado através do primeiro critério possa ser considerado menos eficaz que o sistema obtido através do segundo critério.

Note-se que o segundo critério consiste em efectuar uma ACP sobre a nuvem $\bigcup_{t=1}^{T} \mathcal{N}_I^t$ dos indivíduos centrados, cuja matriz de inércia é $\sum_{t=1}^{T} \mathcal{V}_t$, obtida por sobreposição dos quadros $Y_1, \ldots, Y_T$ (figura 4.2), e que está associada à matriz compromisso do método STATIS dual. Na imagem euclidiana associada representam-se as posições compromisso das variáveis, que são as coordenadas das correlações médias destas com os eixos do compromisso.

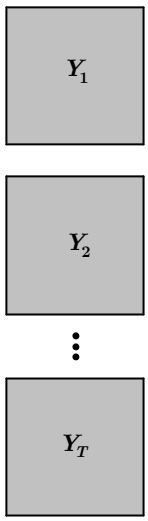


Figura 4.2 Sobreposição dos quadros de dados centrados.

Neste momento é possível determinar as "novas" variáveis $(\boldsymbol{\varsigma_k})_{k=1,\ldots,q}$, projecções dos indivíduos sobre os vectores $(\boldsymbol{\nu_k})_{k=1,\ldots,q}$. Como medidas de qualidade, podem-se determinar os co-senos dos ângulos entre as componentes principais $(\boldsymbol{\varsigma_k})_{k=1,\ldots,q}$ e

- e as variáveis $(\boldsymbol{x^j})^t$, $j = 1, \ldots, p$;
- e as componentes principais ${\boldsymbol{F_{u_k}}}^t$, $k = 1, \ldots, q$.

O **terceiro critério** consiste numa procura sequencial de um sistema de eixos $(\boldsymbol{\nu_k})_{k=1,\ldots,q}$. Seja $({\boldsymbol{z_1}}^t)_{t=1,\ldots,T}$ o conjunto de factores principais associados aos primeiros eixos principais de cada quadro. Inicialmente, pretende-se encontrar $\boldsymbol{\nu_1}$ tal que, em média, o ângulo $({\boldsymbol{z_1}}^t, \boldsymbol{\nu_1})$ seja mínimo. Logo $\boldsymbol{\nu_1}$ é tal que

$$\sum_{t=1}^{T} \cos^2\left({\boldsymbol{z_1}}^t, \boldsymbol{\nu_1}\right)$$

seja máximo. De seguida, pretende-se encontrar $\boldsymbol{\nu_2}$, ortogonal a $\boldsymbol{\nu_1}$, tal que

$$\sum_{t=1}^{T} \cos^2\left(\boldsymbol{z_2}^t, \boldsymbol{\nu_2}\right)$$

seja máximo, e assim sucessivamente até se encontrar $\boldsymbol{\nu_q}$.

Note-se que o sistema de eixos pode ser obtido analiticamente:

(i) pretende-se $\boldsymbol{\nu_1}$ que maximize a função

$$\begin{aligned} f_1(\boldsymbol{\nu_1}) &= \sum_{t=1}^{T} \cos^2\left(\boldsymbol{z_1}^t, \boldsymbol{\nu_1}\right) \\ &= \sum_{t=1}^{T} \boldsymbol{\nu_1}' \boldsymbol{z_1}^t \left(\boldsymbol{z_1}^t\right)' \boldsymbol{\nu_1} \\ &= \boldsymbol{\nu_1}' \left(\sum_{t=1}^{T} \boldsymbol{z_1}^t \left(\boldsymbol{z_1}^t\right)'\right) \boldsymbol{\nu_1}. \end{aligned} \tag{4.2}$$

Seja

$$\mathcal{U}_1 = \left[\boldsymbol{z_1}^1 \ldots \boldsymbol{z_1}^T\right]$$

de dimensão $p \times T$. Tem-se então que

$$f_1(\boldsymbol{\nu_1}) = \boldsymbol{\nu_1}' \mathcal{U}_1 \left(\mathcal{U}_1\right)' \boldsymbol{\nu_1},$$

que se pretende maximizar. $\boldsymbol{\nu_1}$ é o vector próprio normado (pela métrica identidade) da matriz $\mathcal{U}_1 \left(\mathcal{U}_1\right)'$, associado ao maior valor próprio.

(ii) na k-ésima etapa pretende-se $\boldsymbol{\nu_k}$ tal que

$$f_1(\boldsymbol{\nu_k} | \boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_{k-1}}) = \sum_{t=1}^{T} \cos^2\left(\boldsymbol{z_k}^t, \boldsymbol{\nu_k}\right)$$

seja máxima, com $\boldsymbol{\nu_k}$ ortogonal ao subespaço $\mathcal{H}$ formado por $\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_{k-1}}$. Sejam $\widetilde{\boldsymbol{z_k}^t}$ as projecções ortogonais dos vectores $\boldsymbol{z_k}^t$ sobre $\mathcal{H}^\perp$. Tem-se que:

$$\boldsymbol{z_k}^t = \widetilde{\boldsymbol{z_k}^t} + \left(\boldsymbol{z_k}^t - \widetilde{\boldsymbol{z_k}^t}\right)$$

com $\boldsymbol{z_k}^t - \widetilde{\boldsymbol{z_k}^t} \in \mathcal{H}$ e $\widetilde{\boldsymbol{z_k}^t} \in \mathcal{H}^\perp$, logo

$$\begin{aligned} \cos^2\left(\boldsymbol{z_k}^t, \boldsymbol{\nu_k}\right) &= \left\langle \widetilde{\boldsymbol{z_k}^t} + \left(\boldsymbol{z_k}^t - \widetilde{\boldsymbol{z_k}^t}\right), \boldsymbol{\nu_k} \right\rangle^2 \\ &= \left(\boldsymbol{\nu_k}' \cdot \left[\widetilde{\boldsymbol{z_k}^t} + \left(\boldsymbol{z_k}^t - \widetilde{\boldsymbol{z_k}^t}\right)\right]\right)^2, \end{aligned}$$

e como $\boldsymbol{\nu_k} \in \mathcal{H}^{\perp}$ vem que

$$\cos^2\left(\boldsymbol{z_k}^t, \boldsymbol{\nu_k}\right) = \left(\boldsymbol{\nu_k}' \cdot \widetilde{\boldsymbol{z_k}}^t\right)^2.$$

A função que se pretende maximizar é:

$$\begin{aligned} f_1(\boldsymbol{\nu_k}|\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_{k-1}}) &= \sum_{t=1}^{T} \left(\boldsymbol{\nu_k}' \cdot \widetilde{\boldsymbol{z_k}}^t\right)^2 \\ &= \boldsymbol{\nu_k}' \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{z_k}}^t \left(\widetilde{\boldsymbol{z_k}}^t\right)'\right) \boldsymbol{\nu_k} \\ &= \boldsymbol{\nu_k}' \widetilde{\mathcal{U}_k} \left(\widetilde{\mathcal{U}_k}\right)' \boldsymbol{\nu_k}, \end{aligned}$$

com

$$\widetilde{\mathcal{U}_k} = \left[\widetilde{\boldsymbol{z_k}}^1 \ldots \widetilde{\boldsymbol{z_k}}^T\right].$$

$\boldsymbol{\nu_k}$ é o vector próprio normado de $\widetilde{\mathcal{U}_k}\left(\widetilde{\mathcal{U}_k}\right)'$ associado ao maior valor próprio. $\widetilde{\mathcal{U}_k}$ deduz-se facilmente de $\mathcal{U}_k$. Se $\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_{k-1}}$ é um sistema de eixos ortonormado, então

$$\widetilde{\mathcal{U}_k} = \left[\prod_{l=1}^{k-1} \left(I_p - \boldsymbol{\nu_l}\boldsymbol{\nu_l}'\right)\right] \mathcal{U}_k.$$

Este critério dá uma importância decrescente aos eixos, privilegiando os primeiros em detrimento dos últimos, o que faz algum sentido pois o primeiro eixo principal é aquele que tem maior percentagem de inércia explicada, seguido do segundo eixo, e assim sucessivamente, até ao q-ésimo eixo.

É possível comparar o resultado obtido com os dois critérios anteriores, através do cálculo do índice Φ sobre a solução sequencial $(\boldsymbol{\nu_k})_{k=1,\ldots,q}$ obtida:

$$\Phi\left(\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_q}\right) = \frac{1}{T} \sum_{t=1}^{T} \Phi\left(t, \boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_q}\right)$$

com

$$\Phi\left(t, \boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_q}\right) = \frac{\sum_{k=1}^{q} \lambda_k^t - \sum_{k=1}^{q} (\boldsymbol{\nu_k})' \mathcal{V}_t \boldsymbol{\nu_k}}{\sum_{k=1}^{q} \lambda_k^t},$$

e de

$$f_1(\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_q}) = \sum_{t=1}^{T} \sum_{k=1}^{q} \cos^2\left(\boldsymbol{\nu_k}, \boldsymbol{z_k}^t\right).$$

O **quarto critério** consiste na procura global de um sistema de eixos ortonormado $(\boldsymbol{\nu_k})_{k=1,\ldots,q}$ que maximize directamente $f_1(\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_q})$. Bouroche [2] propõe uma resolução numérica para o efeito. Contrariamente ao terceiro critério, que era sequencial e atribuía uma importância decrescente aos eixos, este critério dá a mesma importância a todos os eixos.

4.5 Compromisso e trajectórias dos indivíduos

Dazy e Le Barzic [9] sugerem formas de obter as posições compromisso e as trajectórias dos indivíduos, a partir do primeiro e segundo critérios.

As trajectórias são representadas no sistema de eixos obtido em **??**. Estes eixos são interpretados através das correlações destes com as posições compromisso das variáveis. As coordenadas destas posições correspondem às correlações médias das variáveis com o sistema de eixos encontrado.

Se se optar pelo primeiro critério, o compromisso é o objecto $\mathcal{V}_\tau$. As posições compromisso das variáveis são obtidas a partir de uma ACP sobre o quadro X_τ e as trajectórias dos indivíduos são obtidas a partir da projecção dos indivíduos de cada quadro sobre o sistema de eixos seleccionado, tomando os outros quadros como elementos suplementares na ACP sobre o quadro X_τ. As posições compromisso dos indivíduos são as posições destes no quadro X_τ.

Optando pelo segundo critério, o compromisso é o objecto $\sum_{t=1}^{T} \mathcal{V}_t$ e as posições compromisso das variáveis são obtidas através da ACP sobre esta matriz. As trajectórias são determinadas a partir das posições dos indivíduos na ACP do quadro da figura 4.2. No caso de se pretender as posições compromisso dos indivíduos far-se-á uma média das coordenadas das trajectórias de cada indivíduo sobre cada eixo.

Capítulo 5

Comparação dos métodos

Após a descrição destes métodos (STATIS/STATIS dual, AFM e DACP), interessa proceder a uma comparação entre estes de forma a recolher as vantagens e desvantagens em utilizar um em detrimento dos outros. Os aspectos a ter em conta para efectuar essa comparação são os seguintes:

- o tipo de dados;
- o tipo de quadros;
- o aspecto temporal dos dados;
- os objectos representativos;
- a inter-estrutura;
- o compromisso;
- a intra-estrutura.

5.1 Tipo de dados

A metodologia STATIS (STATIS e STATIS dual) aplica-se a dados quantitativos, bem como a DACP, enquanto que a AFM pode ser utilizada quer para dados quantitativos, quer para qualitativos ou mesmo mistos.

Oliveira [34] propõe a aplicação do método STATIS a dados qualitativos efectuando previamente uma Análise Factorial das Correspondências Múltipla.

5.2 Tipo de quadros

Em todos os métodos os quadros (matrizes) cruzam indivíduos (linhas) com variáveis (colunas). Na DACP os indivíduos, bem como as variáveis, devem ser os mesmos ao longo da sucessão de quadros. Na AFM e no método STATIS os indivíduos invariantes são cruzados com variáveis que podem diferir ao longo dos quadros. O método STATIS dual é aplicado em quadros que cruzem indivíduos, eventualmente diferentes, com as mesmas variáveis.

5.3 Aspecto temporal dos dados

Nenhum dos métodos citados toma em conta o aspecto temporal dos dados. Para esse efeito dever-se-á efectuar uma Análise de Séries Cronológicas Multidimensionais. No entanto, o aspecto temporal deve ser levado em conta na interpretação dos resultados obtidos através dos métodos citados, sobretudo na análise da inter-estrutura e das trajectórias dos indivíduos. Com a excepção da DACP, em que a interpretação da inter-estrutura se torna difícil quando os dados não são temporais, todos os outros métodos podem ser aplicados sobre dados que não apelem à noção de tempo.

5.4 Objectos representativos

O método STATIS permite a utilização de objectos não normados

$$\mathcal{W}_t = X_t Q_t X_t',$$

bem como de objectos normados

$$\frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}} = \frac{\mathcal{W}_t}{\sqrt{\sum_{i=1}^{n} \left(\lambda_i^t\right)^2}}.$$

O método STATIS dual permite igualmente a utilização de objectos não normados

$$\mathcal{V}_t = X_t' D_t X_t,$$

ou de objectos normados

$$\frac{\mathcal{V}_t}{\|\mathcal{V}_t\|_{HS}} = \frac{\mathcal{V}_t}{\sqrt{\sum_{j=1}^{p} \left(\lambda_j^t\right)^2}}.$$

A AFM apenas permite a utilização de objectos normados, da forma

$$\frac{\mathcal{W}_t}{\lambda_1^t},$$

enquanto que a DACP utiliza o objecto $\mathcal{V}_t$ segundo os primeiros dois critérios.

A normalização proposta pela metodologia STATIS é tal que a norma de Hilbert-Schmidt dos objectos representativos é igual a 1. Esta normalização tem o inconveniente de fazer desaparecer as diferenças de estrutura entre os diversos quadros. Já a normalização proposta pela AFM não altera a estrutura múltipla dos vários quadros (grupos de variáveis), uma vez que a inércia total de cada uma das nuvens não intervém nesta normalização e tem a vantagem de fazer com que a inércia da primeira direcção de cada grupo seja igual a 1, enquanto que as restantes serão majoradas por 1. É graças à ponderação $1/\lambda_1^t$ que nenhum grupo pode influenciar de forma preponderante o primeiro eixo da imagem euclidiana do compromisso. No entanto, é de salientar que esta ponderação não tem nenhuma teoria subjacente que justifique a sua optimalidade.

5.5 Inter-estrutura

5.5.1 Medida de ligação entre os objectos

Nos métodos STATIS e AFM, a ligação entre os objectos $\mathcal{W}_t$ e $\mathcal{W}_{t'}$ é feita através do produto escalar de Hilbert-Schmidt

$$\langle \mathcal{W}_t, \mathcal{W}_{t'} \rangle_{HS} = tr\left(\mathcal{W}_t D \mathcal{W}_{t'} D\right).$$

A medida de ligação entre estes objectos pelo método STATIS é o coeficiente RV:

$$RV\left(t, t'\right) = \left\langle \frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}}, \frac{\mathcal{W}_{t'}}{\|\mathcal{W}_{t'}\|_{HS}} \right\rangle_{HS},$$

que varia entre 0 e 1, e é tanto maior quanto menor for o ângulo formado pelos vectores representativos dos objectos (normados).

A medida de ligação entre os grupos de variáveis X_t e $X_{t'}$ da AFM é exprimida por:

$$\mathcal{L}_g\left(X_t, X_{t'}\right) = \left\langle \frac{\mathcal{W}_t}{\lambda_1^t}, \frac{\mathcal{W}_{t'}}{\lambda_1^{t'}} \right\rangle_{HS}.$$

Esta medida é tanto maior quanto maior for a multidimensionalidade de X_t e $X_{t'}$ e estes tenham direcções comuns numerosas e próximas das direcções de inércia máxima de cada grupo.

Tem-se que

$$RV(t,t') = \frac{\mathcal{L}_g\left(X_t, X_{t'}\right)}{\eta_g(X_t)\cdot\eta_g(X_{t'})},$$

sendo $\eta_g(X_t)$ a norma Hilbert-Schmidt do objecto $\mathcal{W}_t/\lambda_1^t$.

Tanto num método como no outro se avalia em que medida se está em presença de uma estrutura comum aos grupos de variáveis X_t e $X_{t'}$. O coeficiente RV não dá acesso à dimensão desta estrutura comum enquanto que no caso de $\mathcal{L}_g$, este toma em conta a dimensão desta, bem como da sua inércia relativamente à inércia de X_t e $X_{t'}$. As duas medidas são então complementares. Pagès [35] fornece alguns valores de referência para o coeficiente RV e $\mathcal{L}_g$ (figura 5.1).

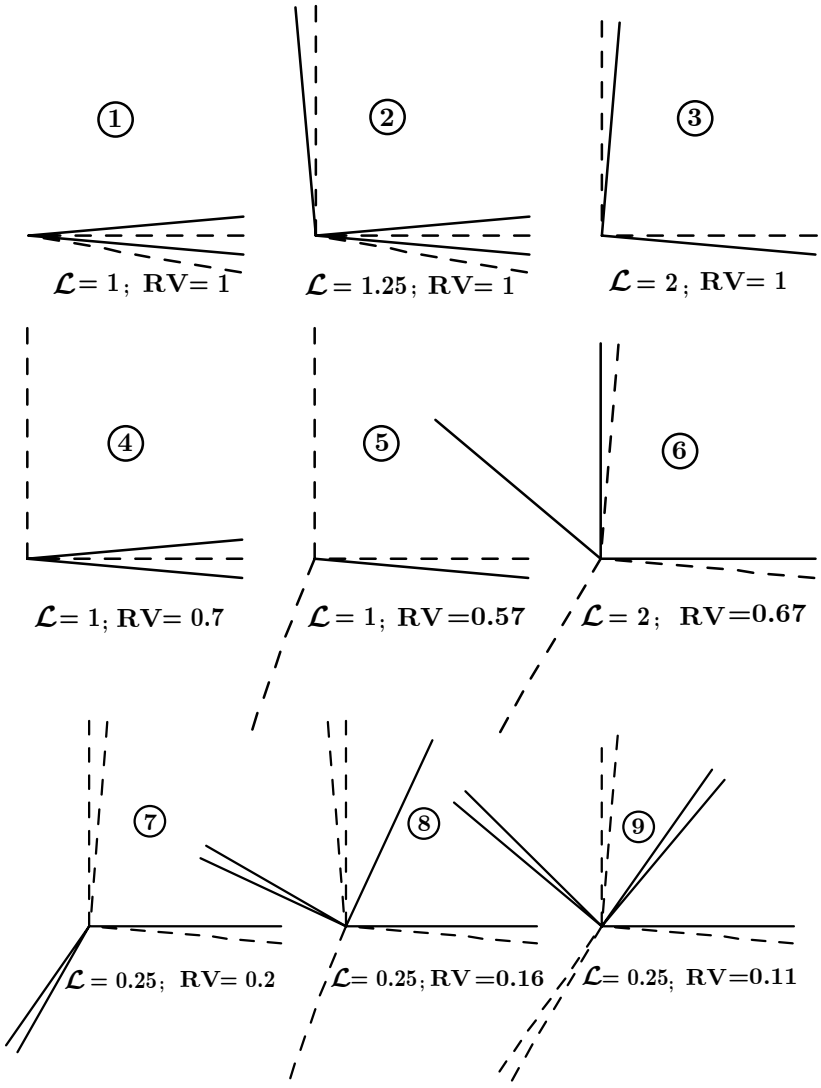


Figura 5.1 Alguns valores de $\mathcal{L}_g$ e RV.

Cada variável é representada por um segmento e pode pertencer ao grupo **I**

(segmentos a cheio) ou ao grupo **II** (segmentos a tracejado). Nos três primeiros casos, as duas estruturas são homotéticas, uma vez que o coeficiente RV é sempre igual a 1 e $\mathcal{L}_g$ aumenta com o número de dimensões comuns e a importância da componente principal de cada grupo.

Nos casos **1**, **4** e **5**, os dois grupos têm apenas a primeira componente principal em comum. É por isso que $\mathcal{L}_g$ vale sempre 1 e o coeficiente RV decresce, à medida que o número de dimensões não comuns aumenta. Do caso **1** para o caso **6**, tanto o número de dimensões comuns como não comuns aumenta, fazendo com que $\mathcal{L}_g$ aumente e o coeficiente RV diminua.

Relativamente aos casos **7**, **8** e **9** tem-se que $\mathcal{L}_g$ é constante e o coeficiente RV diminui à medida que o número de dimensões não comuns aumenta.

Tal como já foi referido em 4.2, o estudo da inter-estrutura difere ligeiramente para a DACP, uma vez que este método estuda a evolução global dos quadros através dos centros de gravidade de cada nuvem $\mathcal{N}_I^t$ $(t = 1, \ldots, T)$. Não está previsto nenhum coeficiente de associação entre objectos para este método.

5.5.2 Representação dos grupos

Na análise da inter-estrutura, tanto no método STATIS, como no método STATIS dual, há duas representações a serem feitas: a imagem euclidiana da inter-estrutura não centrada (análoga à representação das variáveis da ACP) e a imagem euclidiana da inter-estrutura centrada (análoga à representação dos indivíduos da ACP).

No que respeita à AFM, é possível interpretar os eixos de representação de $\mathcal{N}_J$, uma vez que estes coincidem com os eixos da intra-estrutura. No entanto, a qualidade de representação (medida através da razão entre a inércia projectada e a inércia total) é, em geral, má, contrariamente à metodologia STATIS (medida através da inércia projectada). Por outro lado, os eixos da inter-estrutura da metodologia STATIS são dificilmente interpretáveis, isto é, não é possível interpretar a proximidade entre os grupos num determinado eixo, em função dos dados iniciais.

Importa também salientar que, enquanto que na metodologia STATIS a qualidade de representação decresce com a ordem dos eixos, uma vez que corresponde à quantidade directamente maximizada, isso já não acontece com a AFM.

5.6 Compromisso

O único método que possui uma e uma só expressão para o compromisso é a AFM:

$$\sum_{t=1}^{T} \frac{\mathcal{W}_t}{\lambda_1^t}.$$

A vantagem deste é que tem em conta as ligações entre as variáveis de cada grupo. Por oposição, e como já foi referido, a ponderação efectuada é arbitrária, se bem que tem por objectivo majorar as inércias das várias nuvens numa direcção qualquer.

A DACP utiliza o objecto compromisso $\mathcal{V}_\tau$ no primeiro critério, onde τ corresponde ao sistema de eixos seleccionado. Este tem o inconveniente de ser bastante restritivo, isto é, a expressão do compromisso reduzir-se-á a um e um só dos objectos. No segundo critério a expressão do compromisso corresponde à soma dos objectos representativos

$$\sum_{t=1}^{T} \mathcal{V}_t,$$

com a desvantagem de ser influenciado pelos objectos com normas bastante elevadas. No que respeita ao terceiro e quarto critérios, não há, tanto quanto se saiba, nenhuma expressão que defina o objecto compromisso em função dos objectos representativos.

O compromisso relativo ao método STATIS exprime-se na forma

$$\mathcal{W} = \sum_{t=1}^{T} \alpha_t \mathcal{W}_t.$$

Se os objectos forem normados o compromisso é definido por:

$$\mathcal{W} = \sum_{t=1}^{T} \alpha'_t \frac{\mathcal{W}_t}{\|\mathcal{W}_t\|_{HS}}.$$

Quanto ao STATIS dual o objecto compromisso é da forma

$$\mathcal{V} = \sum_{t=1}^{T} \beta_t \mathcal{V}_t.$$

Se os objectos forem normados o compromisso é definido por:

$$\mathcal{V} = \sum_{t=1}^{T} \beta'_t \frac{\mathcal{V}_t}{\|\mathcal{V}_t\|_{HS}}.$$

Estando em presença de um "bom" compromisso, nenhum quadro influencia de forma preponderante a construção deste e os coeficientes α_t (ou β_t) estarão bastante próximos uns dos outros. Será, portanto, mais vantajoso, utilizar o compromisso com objectos não normados, com a desvantagem de este ser influenciado pelos objectos de normas mais elevadas. Se não for um "bom" compromisso, é conveniente utilizar os objectos normados, com a desvantagem de provocar o desaparecimento da estrutura múltipla dos quadros.

5.7 Intra-estrutura

5.7.1 Posições compromisso e trajectórias dos indivíduos

A representação das posições compromisso dos indivíduos no método STATIS obtém-se através de uma ACP sobre a justaposição dos quadros

$$\sqrt{\alpha_1}X_1, \ldots, \sqrt{\alpha_T}X_T.$$

Já na AFM, estas posições são obtidas através da justaposição dos quadros

$$\frac{X_1}{\sqrt{\lambda_1^1}}, \ldots, \frac{X_T}{\sqrt{\lambda_1^T}}.$$

As trajectórias dos indivíduos para estes dois métodos, são obtidas através da projecção dos indivíduos definidos por cada quadro sobre os eixos do compromisso, através da técnica dos elementos suplementares, em que estes são os quadros referidos.

Voisard e Lavallard [45] mostram, a nível prático, que as posições compromisso e as trajectórias dos indivíduos obtidas pelo método STATIS estão bastante próximas daquelas obtidas pela AFM, desde que os objectos utilizados no método STATIS estejam normados.

As posições compromisso para o primeiro critério da DACP são obtidas através de uma ACP sobre o quadro X_τ.

Relativamente ao método STATIS dual, a representação das posições compromisso dos indivíduos obtém-se através de uma ACP da sobreposição dos quadros

$$\sqrt{\beta_1}X_1, \ldots, \sqrt{\beta_T}X_T,$$

enquanto que no segundo critério da DACP se efectua uma ACP da sobreposição dos quadros centrados

$$Y_1, \ldots, Y_T,$$

no caso de as variáveis serem homogéneas.

Os métodos STATIS dual e DACP (segundo critério) permitem obter as trajectórias dos indivíduos relativamente às posições compromisso das variáveis. No caso de se pretender as posições compromisso dos indivíduos far-se-á uma média das coordenadas das trajectórias de cada indivíduo sobre cada eixo (Dazy e Le Barzic [9]).

No método STATIS e na AFM os eixos da intra-estrutura interpretam-se graças às correlações entre estes e as variáveis iniciais, enquanto que aqueles obtidos através do método STATIS dual e da DACP (primeiro e segundo critérios) se interpretam através das correlações entre os eixos com as variáveis compromisso.

Note-se que, estando em presença de um "bom" compromisso, as trajectórias propostas pelos três métodos não serão muito diferentes.

5.7.2 Qualidade de representação do compromisso

Relativamente à metodologia STATIS, se o coeficiente RV estiver próximo de 1 é sinal de que há uma estrutura comum aos vários quadros de dados, logo, a qualidade de representação dos indivíduos na imagem euclidiana do compromisso será razoável. Já se este coeficiente for fraco (≈ 0), pode-se depreender que não há uma estrutura de indivíduos comum aos quadros, deixando de ter interesse estudar a intra-estrutura.

A qualidade de representação do compromisso da AFM é a razão entre a inércia inter e a inércia total para cada eixo.

A DACP utiliza dois índices numéricos. O primeiro representa a perda média de inércia das nuvens $\mathcal{N}_I^1, \ldots, \mathcal{N}_I^T$ quando projectadas sobre o novo sistema de eixos:

$$\Phi(., \tau) = \frac{1}{T} \sum_{t=1}^{T} \Phi(t, \tau),$$

com

$$\Phi(t, \tau) = \frac{\sum_{k=1}^{q} \lambda_k^t - \sum_{k=1}^{q} (\boldsymbol{z_k}^\tau)' \mathcal{V}_t \boldsymbol{z_k}^\tau}{\sum_{k=1}^{q} \lambda_k^t}.$$

O segundo mede a qualidade de representação do conjunto das nuvens atrás

citadas, em termos de proximidade dos ângulos entre os sistemas de eixos:

$$f_1(\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_q}) = \sum_{t=1}^{T} \sum_{k=1}^{q} \cos^2 \left(\boldsymbol{\nu_k}, \boldsymbol{z_k}^t \right).$$

5.8 Conclusões

O método STATIS (STATIS dual) proporciona uma visualização global das semelhanças entre os grupos, enquanto que a AFM dá uma visualização parcial dessas semelhanças. Na prática, observam-se representações dos grupos bastante diferentes para estes dois métodos. Para um estudo mais pormenorizado sobre a comparação destes métodos, Pagès [35] é uma referência importante.

A DACP é um método bastante simples de ser executado, no entanto, a inter-estrutura limita este método a quadros que cruzem os mesmos indivíduos com as mesmas variáveis ao longo do tempo, fazendo com que o seu campo de aplicação se torne bastante restrito. Tal como o método STATIS dual, a DACP não possui uma representação directa das posições compromisso dos indivíduos.

Se os dados possuírem uma estrutura comum bastante forte, os resultados obtidos através dos vários métodos serão bastante próximos.

Capítulo 6

Análise Evolutiva de alguns Indicadores de Desenvolvimento em Países Europeus

6.1 Apresentação dos Dados

O World Bank Group, agência das Nações Unidas, é a fonte responsável pela recolha dos dados. Os países que fazem parte desta aplicação, bem como as respectivas abreviaturas encontram-se na tabela 6.1. Inicialmente, o objectivo era proceder a uma análise de todos os países que actualmente fazem parte da União Europeia, mas havia uma grande quantidade de dados omissos, pelo que se analisaram apenas aqueles que possuiam toda a informação solicitada.

As variáveis em estudo, respectivas abreviaturas e unidades de medida são as seguintes:

- (PU) **população urbana**, percentagem do total da população que vive em áreas classificadas como urbanas;
- (TN) **taxa de natalidade**, número de nados vivos por mil habitantes;
- (TM) **taxa de mortalidade**, número de óbitos por mil habitantes;
- (TF) **taxa de fertilidade total**, razão entre o número de nados vivos e o número de mulheres em período fértil (dos 15 aos 49 anos);
- (CO) **emissões de dióxido de carbono** em toneladas métricas per capita;

- (TR) **tractores** por cada 100 hectares de terra arável, utilizados na agricultura (exceptuando os tractores utilizados na jardinagem), no final do calendário do ano em causa ou durante os primeiros três meses do ano seguinte;
- (EA) **exportação de bens alimentares** (percentagem do total de exportações);
- (IA) **importação de bens alimentares** (percentagem do total de importações);
- (TL) **número de telemóveis** por 1000 habitantes;
- (TV) **número de aparelhos de televisão** utilizados, por 1000 habitantes;
- (CE) **consumo de energia eléctrica** em quilowatts per capita;
- (SC) **número de alunos inscritos no ensino secundário** (público e privado).

Tabela 6.1 Abreviaturas dos países em estudo.

Abreviatura	País
AU	Áustria
CH	Chipre
ES	Espanha
FI	Finlândia
FR	França
GR	Grécia
HU	Hungria
IT	Itália
MA	Malta
PB	Países Baixos
PO	Portugal
RU	Reino Unido
SU	Suécia
TU	Turquia

Os dados em causa referem-se aos anos de 1980, 1985, 1990, 1994, 1995 e 2000. Apresentam-se sob a forma de seis quadros (matrizes) correspondentes a cada ano, designados, respectivamente, por X_{80}, X_{85}, X_{90}, X_{94}, X_{95}, X_{00}. Nestas seis matrizes havia apenas três dados omissos:

- em 1985, a variável TL na Espanha;
- em 1995, a variável SC na Turquia;
- em 2000, a variável SC na Hungria.

Optou-se por preencher estes dados pelos do ano posterior correspondente, à excepção do último caso, em que se substituiu pelo valor do ano anterior (1999), uma vez que o valor do ano posterior também era omisso.

Note-se que as variáveis (em número de 12), assim como os indivíduos (em número de 14) são os mesmos para todos os quadros em análise (figura 6.1).

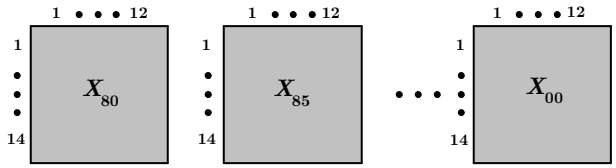


Figura 6.1 Representação esquemática dos dados analisados.

Os dados encontram-se no Anexo 1, assim como no CD-ROM [46] que consta na Bibliografia.

De início, proceder-se-á a uma análise estatística preliminar (análise univariada e bivariada). Em seguida, efectuar-se-á uma ACP normada apenas sobre as matrizes X_{80} e X_{00}, de forma a poder comparar estes anos extremos e a não tornar esta análise demasiado fastidiosa. Finalmente aplicar-se-á cada um dos métodos atrás descritos (STATIS, STATIS dual, AFM e DACP), tendo em vista os seguintes objectivos:

- analisar a eventual existência de uma estrutura comum aos quadros dos anos considerados;
- analisar as tendências evolutivas de cada um dos indivíduos (países) considerados;
- analisar as tendências evolutivas de cada uma das variáveis consideradas.

Os programas utilizados nestas análises são o Matlab (versão 6.5) e o SPAD ("Système Pour l' Analyse des Données") na versão 5.5.

6.2 Análise Preliminar

A análise das médias de cada variável (tabela 6.2) permite constatar um aumento progressivo nas variáveis PU, TR, TL, TV, CE. As médias da variável SC também aumentaram ao longo dos anos, excepto de 1995 para 2000. Já as variáveis TN, TM e TF diminuíram progressivamente, em média.

Tabela 6.2 Médias das variáveis.

	PU	TN	TM	TF	CO	TR
1980	65.93	15.47	9.98	2.09	157887.39	8.14
1985	68.14	14.17	9.87	1.89	147294.32	9.01
1990	70.48	13.74	9.71	1.81	154971.18	9.42
1994	71.84	12.71	9.40	1.70	151226.37	9.96
1995	72.20	12.32	9.55	1.65	157156.81	10.14
2000	73.15	11.70	9.43	1.61	169703.92	10.29
	EA	IA	TL	TV	CE	SC
1980	16.23	10.93	0.35	328.14	3493.90	1915149.71
1985	14.14	10.92	1.80	374.42	4178.47	2035682.71
1990	14.12	9.97	11.24	383.52	4835.37	2058199.57
1994	13.89	10.88	41.67	448.43	5153.39	2394940.50
1995	14.01	10.80	65.46	465.84	5224.46	2716012.93
2000	10.49	8.37	558.74	536.32	5968.96	2386473.00

A tabela 6.3 evidencia as diferenças entre os desvios-padrões das respectivas variáveis. Isto deve-se ao facto de estas serem medidas em unidades diferentes.

Interessa agora estudar as relações estatísticas entre as variáveis de cada ano. Para tal determinaram-se as matrizes de correlações desde os anos 1980 até 2000: R_{80}, R_{85}, R_{90}, R_{94}, R_{95} e R_{00} (tabelas 6.4 a 6.9). Durante o período 1980-2000 há uma forte correlação positiva entre os seguintes pares de variáveis:

- TN e TF, tal como seria de esperar, uma vez que a taxa de fertilidade total se determina a partir da taxa de natalidade;
- CO e SC, embora estas duas variáveis não estejam directamente relacionadas.

A única correlação negativa razoável encontra-se em R_{80}, entre as variáveis EA e TV (-0.78). Nos anos posteriores esta correlação vai-se dissipando ligeiramente, atingindo -0.43 no ano de 2000.

Tabela 6.3 Desvios-padrões das variáveis.

	PU	TN	TM	TF	CO	TR
1980	17.38	5.11	1.66	0.68	185982.32	6.55
1985	15.14	5.16	1.76	0.60	165793.74	6.74
1990	12.90	3.89	1.71	0.48	171365.10	6.78
1994	11.77	3.43	1.83	0.40	162149.22	6.88
1995	11.50	3.32	1.87	0.39	167458.79	6.95
2000	11.22	3.02	1.69	0.32	176660.34	6.34
	EA	IA	TL	TV	CE	SC
1980	13.81	4.28	1.31	146.26	2661.70	2065347.71
1985	9.76	4.02	4.15	145.83	3510.68	2108337.90
1990	12.08	3.24	18.29	114.84	3702.69	2108245.25
1994	11.35	4.24	46.70	86.22	3730.49	2403218.53
1995	12.52	4.45	67.47	90.48	3737.72	2776358.29
2000	9.21	3.90	189.53	144.57	3898.33	2612264.52

Tabela 6.4 Matriz de correlações relativa a 1980.

1980	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
PU	1.00											
TN	-0.52	1.00										
TM	0.01	-0.21	1.00									
TF	-0.51	**0.99**	-0.18	1.00								
CO	0.45	-0.24	0.17	-0.20	1.00							
TR	0.37	-0.39	-0.07	-0.46	0.08	1.00						
EA	-0.47	**0.86**	-0.23	**0.85**	-0.21	-0.25	1.00					
IA	0.31	-0.29	-0.42	-0.32	0.12	0.09	-0.26	1.00				
TL	-0.10	-0.13	-0.13	-0.20	-0.16	0.04	-0.27	-0.26	1.00			
TV	**0.76**	-0.65	0.17	-0.67	0.19	0.27	**−0.78**	0.37	0.17	1.00		
CE	0.45	-0.54	0.21	-0.55	0.08	0.26	-0.61	-0.33	0.46	0.49	1.00	
SC	0.34	-0.09	0.00	-0.03	**0.93**	0.03	-0.08	0.06	-0.20	0.04	-0.06	1.00

Tabela 6.5 Matriz de correlações relativa a 1985.

1985	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
PU	1.00											
TN	-0.31	1.00										
TM	-0.03	-0.27	1.00									
TF	-0.31	**0.99**	-0.22	1.00								
CO	0.41	-0.14	0.21	-0.15	1.00							
TR	0.25	-0.26	-0.04	-0.40	0.02	1.00						
EA	-0.32	0.49	-0.23	0.48	-0.19	-0.13	1.00					
IA	0.15	-0.24	-0.54	-0.26	0.09	0.13	0.19	1.00				
TL	0.05	-0.16	0.16	-0.17	-0.21	0.05	-0.50	-0.52	1.00			
TV	**0.72**	-0.42	0.09	-0.42	0.11	0.22	-0.60	0.08	0.29	1.00		
CE	0.34	-0.36	0.33	-0.35	-0.04	0.16	-0.58	-0.47	**0.86**	0.41	1.00	
SC	0.28	-0.01	-0.04	-0.03	**0.89**	-0.03	-0.09	0.05	-0.29	-0.06	-0.16	1.00

Tabela 6.6 Matriz de correlações relativa a 1990.

1990	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
PU	1.00											
TN	0.02	1.00										
TM	-0.15	-0.42	1.00									
TF	0.07	**0.97**	-0.27	1.00								
CO	0.36	-0.15	0.04	-0.21	1.00							
TR	0.12	-0.24	-0.12	-0.34	0.05	1.00						
EA	-0.25	0.31	-0.19	0.32	-0.18	-0.05	1.00					
IA	-0.04	-0.11	-0.42	-0.17	0.13	0.04	0.61	1.00				
TL	0.19	-0.01	0.27	0.11	-0.06	0.05	-0.46	-0.60	1.00			
TV	0.53	-0.26	0.29	-0.19	0.34	0.41	-0.36	-0.50	0.47	1.00		
CE	0.25	-0.18	0.31	-0.05	-0.08	0.15	-0.50	-0.60	**0.96**	0.60	1.00	
SC	0.19	-0.02	-0.21	-0.12	**0.85**	-0.07	-0.08	0.11	-0.22	0.23	-0.21	1.00

Tabela 6.7 Matriz de correlações relativa a 1994.

1994	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
PU	1.00											
TN	-0.02	1.00										
TM	-0.17	-0.50	1.00									
TF	0.04	**0.96**	-0.41	1.00								
CO	0.31	-0.13	0.07	-0.22	1.00							
TR	0.03	-0.18	-0.11	-0.27	0.04	1.00						
EA	-0.26	0.32	-0.17	0.26	-0.16	0.04	1.00					
IA	-0.01	-0.29	-0.23	-0.32	0.04	0.25	0.56	1.00				
TL	0.21	-0.06	0.17	0.12	-0.05	0.03	-0.46	-0.29	1.00			
TV	0.45	-0.48	0.35	-0.41	0.55	0.17	-0.35	-0.03	0.34	1.00		
CE	0.23	-0.17	0.16	-0.01	-0.11	0.10	-0.49	-0.29	**0.94**	0.47	1.00	
SC	0.21	0.07	-0.07	-0.06	**0.92**	-0.15	-0.06	-0.06	-0.21	0.35	-0.23	1.00

Tabela 6.8 Matriz de correlações relativa a 1995.

1995	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
PU	1.00											
TN	-0.04	1.00										
TM	-0.14	-0.52	1.00									
TF	0.05	**0.95**	-0.46	1.00								
CO	0.29	-0.08	0.07	-0.17	1.00							
TR	0.04	-0.20	-0.13	-0.29	0.03	1.00						
EA	-0.25	0.29	-0.21	0.25	-0.18	0.06	1.00					
IA	-0.02	-0.11	-0.39	-0.14	-0.01	0.22	0.69	1.00				
TL	0.17	-0.12	0.24	0.06	-0.09	0.01	-0.38	-0.32	1.00			
TV	0.57	-0.46	0.30	-0.38	0.56	0.11	-0.37	-0.14	0.25	1.00		
CE	0.23	-0.22	0.23	-0.06	-0.11	0.06	-0.48	-0.41	**0.92**	0.41	1.00	
SC	0.19	0.01	-0.02	-0.11	**0.82**	-0.24	-0.07	-0.03	-0.27	0.36	-0.25	1.00

Tabela 6.9 Matriz de correlações relativa a 2000.

2000	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
PU	1.00											
TN	-0.08	1.00										
TM	-0.19	-0.56	1.00									
TF	0.17	**0.81**	**-0.71**	1.00								
CO	0.25	0.11	0.00	-0.02	1.00							
TR	-0.07	-0.32	-0.17	-0.25	0.01	1.00						
EA	-0.17	0.18	-0.35	0.19	-0.14	0.21	1.00					
IA	0.07	-0.18	-0.34	0.02	-0.12	0.37	**0.76**	1.00				
TL	0.05	-0.46	0.35	-0.52	0.29	0.46	-0.39	-0.06	1.00			
TV	0.46	-0.39	0.25	-0.17	0.49	-0.02	-0.43	-0.16	0.55	1.00		
CE	0.06	-0.31	0.18	-0.07	-0.16	0.02	-0.41	-0.22	0.57	0.49	1.00	
SC	0.24	0.36	-0.10	0.23	**0.95**	-0.16	-0.14	-0.23	0.10	0.40	-0.20	1.00

A comparação entre as matrizes R_{80} e R_{00} indica algumas diferenças no comportamento das correlações entre as variáveis. No conjunto das matrizes de correlações, aquelas que parecem ser mais semelhantes são R_{90}, R_{94} e R_{95}. Estas diferenças e semelhanças serão posteriormente analisadas.

6.3 Análise em Componentes Principais

Todos os resultados foram obtidos com o auxílio do SPAD (versão 5.5). Procedeu-se a uma ACP normada, uma vez que as variáveis são medidas em unidades diferentes. Os pesos atribuídos aos países são os mesmos, logo $D = \frac{1}{14} I_{14}$.

6.3.1 Ano de 1980

Os valores próprios da matriz de correlações R_{80} (tabela 6.4) encontram-se na tabela 6.10 por ordem decrescente de magnitude, bem como a percentagem de inércia explicada e a percentagem de inércia explicada acumulada, descritas no capítulo 1 pelas expressões (1.28) e (1.29), respectivamente.

Segundo o critério empírico habitualmente utilizado, que consiste na retenção dos factores que expliquem cerca de 80% da variabilidade total, os quatro primeiros factores são suficientes, uma vez que explicam 80.46% da variância total.

As correlações entre as variáveis e as duas primeiras componentes principais encontram-se representadas no círculo de correlações (figura 6.2) e as respectivas coordenadas nos quatro primeiros eixos principais, na tabela 6.11. As variáveis

Tabela 6.10 Valores Próprios de R_{80}.

Componente principal	Valor próprio	Inércia explicada	Inércia acumulada
1	**4.7468**	**39.56**	**39.56**
2	**2.1752**	**18.13**	**57.68**
3	**1.6361**	**13.63**	**71.32**
4	**1.0982**	**9.15**	**80.47**
5	0.8981	7.48	87.95
6	0.7147	5.96	93.91
7	0.3777	3.15	97.06
8	0.1773	1.48	98.53
9	0.1233	1.03	99.56
10	0.0359	0.30	99.86
11	0.0154	0.13	99.99
12	0.0013	0.01	100.00

PU, TN, TF, EA e TV são as que mais têm maior destaque na formação da primeira componente. Relativamente à segunda componente, as variáveis que estão mais correlacionadas de forma negativa com esta são CO e SC. A terceira componente está bastante correlacionada negativamente apenas com a variável IA. Não houve correlações significativas a registar em relação à quarta componente, de modo que se optou por restringir a análise aos três primeiros eixos.

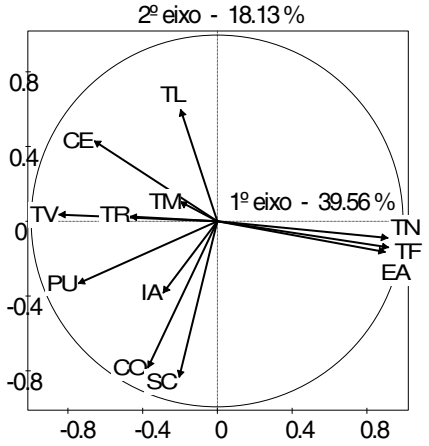


Figura 6.2 Círculo de correlações de 1980 no plano principal $(\boldsymbol{v_1}, \boldsymbol{v_2})$.

As coordenadas dos indivíduos nos três primeiros eixos, bem como as respectivas contribuições absolutas (em %) e relativas (entre 0 e 1) encontram-se

Tabela 6.11 Correlações entre as variáveis de 1980 e as componentes principais.

Variável	1	2	3	4
PU	-0.74	-0.33	-0.06	0.09
TN	0.91	-0.09	0.06	0.15
TM	-0.20	0.10	0.65	-0.67
TF	0.92	-0.14	0.12	0.11
CO	-0.37	-0.78	0.41	0.17
TR	-0.46	0.02	-0.17	0.20
EA	0.90	-0.16	0.00	0.09
IA	-0.29	-0.38	-0.81	-0.10
TL	-0.20	0.60	0.16	0.61
TV	-0.85	0.03	-0.14	-0.09
CE	-0.66	0.43	0.39	0.24
SC	-0.20	-0.83	0.37	0.27

na tabela 6.12.

Tabela 6.12 Coordenadas e contribuições dos indivíduos de 1980.

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
AU	-1.51	1.20	0.75	3.4	4.7	2.5	0.23	0.14	0.06
CH	2.25	0.26	-1.68	7.6	0.2	12.3	0.56	0.01	0.31
ES	0.43	-1.29	-0.61	0.3	5.5	1.6	0.04	0.36	0.08
FI	-1.55	3.17	0.75	3.6	33.1	2.5	0.12	0.50	0.03
FR	-0.89	-1.92	1.13	1.2	12.1	5.5	0.13	0.60	0.20
GR	1.56	0.45	-0.28	3.6	0.7	0.3	0.61	0.05	0.02
HU	0.59	0.85	1.18	0.5	2.4	6.1	0.04	0.08	0.16
IT	-1.54	-1.73	0.15	3.6	9.8	0.1	0.35	0.43	0.00
MA	-1.09	-0.02	-2.60	1.8	0.0	29.6	0.09	0.00	0.50
PB	-1.60	-0.40	-1.78	3.9	0.5	13.9	0.24	0.01	0.30
PO	1.72	0.76	-0.89	4.5	1.9	3.5	0.33	0.06	0.09
RU	-2.03	-2.49	1.35	6.2	20.4	7.9	0.32	0.48	0.14
SU	-2.23	1.58	1.12	7.5	8.2	5.5	0.38	0.19	0.10
TU	5.90	-0.41	1.42	52.3	0.5	8.8	0.89	0.00	0.05

Observando as contribuições absolutas, constata-se que os países que mais contribuem para a formação do primeiro eixo principal são a Turquia (52.3%), seguida do Chipre (7.6%) e da Suécia (7.5%), uma vez que estas são superiores ao seu peso na amostra (1/14). A Finlândia (33.1%), o Reino Unido (20.4%), a

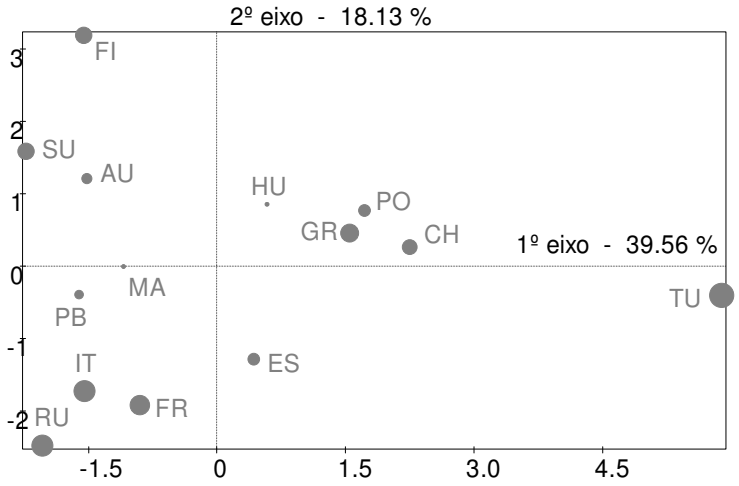


Figura 6.3 Representação dos indivíduos de 1980 no plano principal $(\boldsymbol{u_1}, \boldsymbol{u_2})$.

França (12.1%), a Itália (9.8%) e a Suécia (8.2%) são aqueles que mais contribuem para a formação do segundo eixo principal. Os que mais contribuem para a formação do terceiro eixo principal são Malta (29.6%), os Países Baixos (13.9%) e o Chipre (12.3%).

A representação dos indivíduos a partir do terceiro eixo não foi elaborada, uma vez que desta apenas surgiriam situações pontuais, que não seriam profícuas para a análise em causa. Além disso, a qualidade de representação neste eixo é insatisfatória, excepto para Malta.

Os países que têm melhor qualidade de representação no plano formado pelos dois primeiros eixos principais, são a Turquia (0.89), o Reino Unido (0.80), a Itália (0.78), a França (0.73), a Grécia (0.66), a Finlândia (0.62) e a Suécia e o Chipre (0.57).

A representação dos indivíduos no plano formado pelos dois primeiros eixos principais de inércia encontra-se na figura 6.3. O tamanho de cada marca do indivíduo é directamente proporcional à contribuição relativa do plano principal a esse mesmo indivíduo, ou seja, quanto maior for a marca deste, maior é a sua qualidade de representação no plano principal $(\boldsymbol{u_1}, \boldsymbol{u_2})$.

A Turquia é um país algo isolado no primeiro eixo, com valores superiores à média das variáveis TN, TF e EA, que estão positivamente correlacionadas com este eixo. Já para as variáveis PU e TV (correlacionadas negativamente com o primeiro eixo), a Turquia apresenta valores bastante abaixo das suas médias. O país mais próximo e com o comportamento mais semelhante ao da Turquia é o Chipre, embora com valores menos preponderantes para as variáveis citadas. A

estes dois países opõe-se a Suécia, com valores acima da média das variáveis PU e TV.

O segundo eixo opõe a Finlândia e a Suécia por um lado, e o Reino Unido, a Itália e a França, por outro. Isto significa que valores elevados das variáveis CO e SC para o Reino Unido, a Itália e a França, opõem-se a valores claramente inferiores para a Finlândia e a Suécia.

6.3.2 Ano de 2000

Através da análise da tabela 6.13, verifica-se que os quatro primeiros factores retêm 79.97% da variabilidade total.

A observação do círculo de correlações (figura 6.4) e das respectivas coordenadas nos quatro primeiros eixos principais (tabela 6.14), permite constatar que as variáveis TL e TV estão correlacionadas de forma negativa com o primeiro eixo e a variável TN de forma positiva. Tal como no ano de 1980, as variáveis CO e SC encontram-se correlacionadas negativamente com o segundo eixo e IA está correlacionada de forma positiva com o terceiro eixo. Não há correlações significativas a registar para o quarto eixo principal.

Tabela 6.13 Valores Próprios de R_{00}.

Componente principal	Valor próprio	Inércia explicada	Inércia acumulada
1	**3.5730**	**29.78**	**29.78**
2	**2.7167**	**22.64**	**52.41**
3	**1.9156**	**15.96**	**68.38**
4	**1.3910**	**11.59**	**79.97**
5	1.0150	8.46	88.43
6	0.6369	5.31	93.73
7	0.2654	2.21	95.95
8	0.2123	1.77	97.72
9	0.1559	1.30	99.01
10	0.1047	0.87	99.89
11	0.0102	0.08	99.97
12	0.0034	0.03	100.00

As coordenadas dos países nos três primeiros eixos, bem como as respectivas contribuições absolutas (em %) e relativas (entre 0 e 1) encontram-se na tabela 6.15.

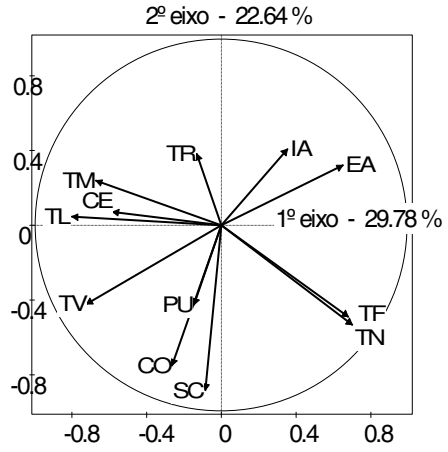


Figura 6.4 Círculo de correlações de 2000 no plano principal $(\boldsymbol{v_1}, \boldsymbol{v_2})$.

Tabela 6.14 Correlações entre as variáveis de 2000 e as componentes principais.

Variável	1	2	3	4
PU	-0.15	-0.43	0.30	-0.43
TN	0.70	-0.53	-0.21	-0.05
TM	-0.67	0.24	-0.34	0.45
TF	0.68	-0.49	-0.08	-0.46
CO	-0.27	-0.75	0.43	0.39
TR	-0.13	0.38	0.66	-0.06
EA	0.65	0.32	0.50	0.15
IA	0.35	0.41	0.73	-0.08
TL	-0.80	0.05	0.32	-0.11
TV	-0.72	-0.42	0.24	-0.24
CE	-0.58	0.07	-0.16	-0.65
SC	-0.09	-0.89	0.27	0.33

As contribuições absolutas mais elevadas para o primeiro eixo são da Turquia (39.1%), Chipre (24.9%), Reino Unido (9.8%) e Suécia (7.9%). Para o segundo eixo a contribuição mais elevada é a do Reino Unido (32.9%), seguido da Turquia (20.0%), Chipre (13.7%), França (11.5%) e Grécia (7.7%). Aqueles que mais contribuem para a formação do terceiro eixo principal são a Hungria (28.1%), o Chipre (20.5%), o Reino Unido (10.3%), a Turquia (9.9%), a Itália (8.2%) e a Finlândia (7.5%).

Uma vez que, apenas a variável IA está correlacionada com o terceiro eixo e nenhuma das variáveis está correlacionada com o quarto eixo, limitou-se a representação dos indivíduos aos dois primeiros eixos.

Os países com melhor qualidade de representação no plano formado pelos dois

Tabela 6.15 Coordenadas e contribuições dos indivíduos de 2000.

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
AU	-1.42	1.51	0.40	4.0	6.0	0.6	0.22	0.25	0.02
CH	3.53	2.28	2.35	24.9	13.7	20.5	0.50	0.21	0.22
ES	-0.57	-0.13	0.67	0.7	0.0	1.7	0.10	0.01	0.13
FI	-1.73	0.41	-1.42	6.0	0.4	7.5	0.25	0.01	0.17
FR	0.22	-2.09	0.27	0.1	11.5	0.3	0.01	0.84	0.01
GR	0.48	1.71	0.14	0.5	7.7	0.1	0.04	0.49	0.00
HU	-0.53	1.06	-2.75	0.6	3.0	28.1	0.02	0.07	0.49
IT	-1.49	0.05	1.48	4.4	0.0	8.2	0.23	0.00	0.23
MA	1.01	-0.17	-0.74	2.0	0.1	2.0	0.10	0.00	0.05
PB	0.18	0.07	1.17	0.1	0.0	5.1	0.01	0.00	0.30
PO	0.09	1.30	-0.45	0.0	4.4	0.8	0.00	0.40	0.05
RU	-2.21	-3.54	1.66	9.8	32.9	10.3	0.23	0.58	0.13
SU	-1.98	0.30	-1.16	7.9	0.2	5.0	0.39	0.01	0.13
TU	4.42	-2.76	-1.63	39.1	20.0	9.9	0.61	0.24	0.08

primeiros eixos principais são a Turquia e a França (0.85), o Reino Unido (0.81), o Chipre (0.71) e a Grécia (0.53).

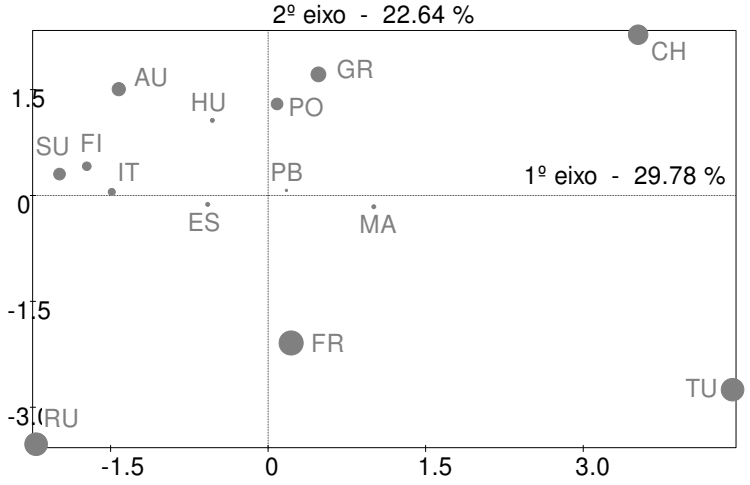


Figura 6.5 Representação dos indivíduos de 2000 no plano principal $(\boldsymbol{u_1}, \boldsymbol{u_2})$.

A representação dos indivíduos no plano formado pelos dois primeiros eixos (figura 6.5) permite distinguir uma oposição entre o Reino Unido e o Chipre e a Turquia no primeiro eixo, o que quer dizer que os valores elevados de TL e TV para o Reino Unido opõem-se a valores bastante inferiores por parte do Chipre

e da Turquia. Como a variável TN está correlacionada de forma positiva com o primeiro eixo, a valores elevados desta para o Chipre e para a Turquia opõem-se valores inferiores para o Reino Unido. Relativamente ao segundo eixo há uma oposição entre o Reino Unido, a França e a Turquia por um lado, e a Grécia e Chipre por outro. Ou seja, a valores abaixo da médias de CO e SC, por parte destes dois últimos países, opõem-se valores mais elevados por parte do Reino Unido, França e Turquia.

6.4 Método STATIS

Este método permite a análise de uma possível estrutura comum às matrizes dos vários anos, assim como a análise das tendências evolutivas de cada um dos indivíduos. As variáveis são heterogéneas por serem medidas em unidades diferentes, de modo que todos os dados foram centrados e reduzidos.

6.4.1 Inter-estrutura

Os pesos atribuídos a cada um dos estudos (X_t, Q_t, D), $(t = 80, 85, 90, 94, 95, 00)$ são os mesmos, uma vez que se pretende atribuir a mesma importância a todos os anos, logo a matriz dos pesos dos estudos é $\Delta = \frac{1}{6} I_6$.

Aplicando uma ACP sobre a matriz dos coeficientes RV (tabela 6.16) obtém-se a imagem euclidiana da inter-estrutura não centrada (figura 6.6) dos objectos normados $\mathcal{W}_t / \|\mathcal{W}_t\|_{HS}$. As coordenadas dos objectos nesta imagem foram calculadas através do Matlab e posteriormente representadas no SPAD.

Tabela 6.16 Matriz dos coeficientes RV.

RV	1980	1985	1990	1994	1995	2000
1980	1					
1985	0.951	1				
1990	0.822	0.934	1			
1994	0.828	0.921	0.952	1		
1995	0.827	0.921	0.947	0.989	1	
2000	0.774	0.838	0.845	0.915	0.927	1

As imagens euclidianas da inter-estrutura não centrada (figura 6.6) e da inter-estrutura centrada (figura 6.7), bem como a matriz dos coeficientes RV, mostram que, durante o período 1980-2000, há uma evolução cronológica mais ou menos regular. É visível a forte proximidade entre anos consecutivos,

Tabela 6.17 Normas Hilbert-Schmidt dos objectos $\mathcal{W}_{80}, \ldots, \mathcal{W}_{00}$.

$$\|\mathcal{W}_{80}\|_{HS} = 5.714$$
$$\|\mathcal{W}_{85}\|_{HS} = 5.296$$
$$\|\mathcal{W}_{90}\|_{HS} = 5.232$$
$$\|\mathcal{W}_{94}\|_{HS} = 5.088$$
$$\|\mathcal{W}_{95}\|_{HS} = 5.101$$
$$\|\mathcal{W}_{00}\|_{HS} = 5.229$$

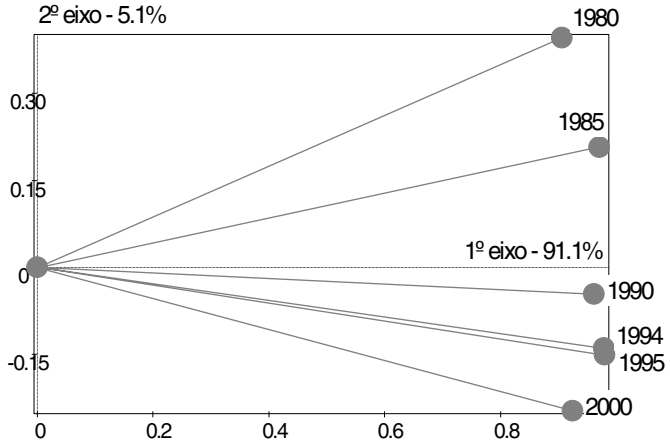


Figura 6.6 Imagem euclidiana da inter-estrutura não centrada.

nomeadamente entre 1994 e 1995, já que estes são os anos cronologicamente mais próximos. Os anos consecutivos mais afastados são os de 1995 e 2000, com um coeficiente $RV(95, 00) = 0.927$. Os anos mais afastados são os extremos, 1980 e 2000, com um coeficiente $RV(80, 00) = 0.774$.

As coordenadas das imagens euclidianas da inter-estrutura não centrada e centrada, assim como as respectivas contribuições absolutas e relativas (entre 0 e 1) encontram-se no Anexo 2.

As normas Hilbert-Schmidt dos objectos $\mathcal{W}_t$ são bastante próximas, variando entre 5.088 e 5.714 (tabela 6.17), o que poderia levar a uma implementação deste método a objectos não normados. No entanto, estes foram normados, uma vez que o SPAD efectua esta análise exclusivamente a objectos deste tipo.

6.4.2 Intra-estrutura

Os coeficientes α_t que permitem determinar o compromisso $\mathcal{W}$ são bastante próximos (tabela 6.18), reflectindo a estrutura comum existente nos vários anos, traduzida pelos elevados coeficientes RV.

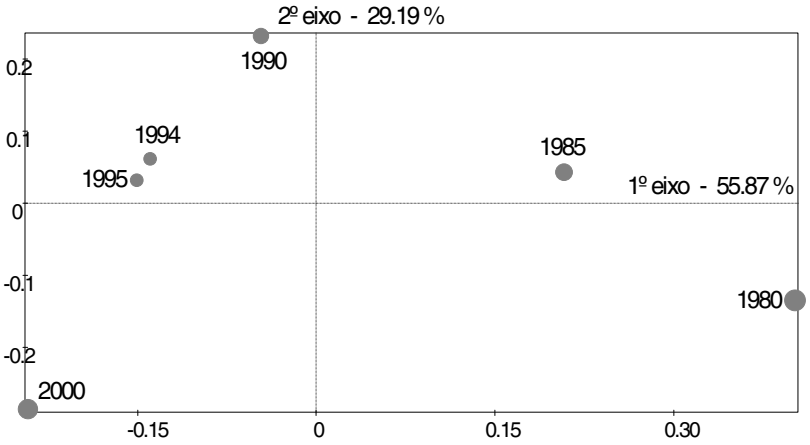


Figura 6.7 Imagem euclidiana da inter-estrutura centrada.

O objecto normado que se encontra mais próximo do compromisso é o de 1995 e aquele que se encontra mais afastado é o de 1980, segundo a distância de Hilbert-Schmidt (tabela 6.18).

Tabela 6.18 Coeficientes α_t do compromisso $\mathcal{W}$ e distâncias HS.

$\alpha_{80} = 0.166$	$d_{HS}(1980, \mathcal{W}) = 0.431$
$\alpha_{85} = 0.178$	$d_{HS}(1985, \mathcal{W}) = 0.238$
$\alpha_{90} = 0.176$	$d_{HS}(1990, \mathcal{W}) = 0.277$
$\alpha_{94} = 0.179$	$d_{HS}(1994, \mathcal{W}) = 0.201$
$\alpha_{95} = 0.179$	$d_{HS}(1995, \mathcal{W}) = 0.198$
$\alpha_{00} = 0.169$	$d_{HS}(2000, \mathcal{W}) = 0.387$

Os valores próprios de $\mathcal{W}D$ encontram-se na tabela 6.19 por ordem decrescente de magnitude, bem como a percentagem de inércia explicada e a percentagem de inércia explicada acumulada. Com os quatro primeiros eixos retém-se cerca de 76.29% da variabilidade total.

As coordenadas dos indivíduos no objecto compromisso são, como já foi exposto em 2.2, do tipo $\sqrt{\mu_k}\boldsymbol{\varepsilon_k}$, em que os vectores de $\mathcal{W}D$, $\boldsymbol{\varepsilon_k}$, são D-ortonormados. Ora o programa SPAD considera estes vectores normados pela identidade, o que quer dizer que as representações efectuadas pelos dois processos são homotéticas.

Para interpretar as posições compromisso dos países determinam-se as correlações das variáveis iniciais com os eixos do compromisso, $\boldsymbol{\varepsilon_k}$, D-ortonormados. Estas foram calculadas com o auxílio do Matlab, pelo facto de o SPAD não as fornecer para este método.

Tabela 6.19 Valores Próprios de $\mathcal{W}D$.

Componente principal	Valor próprio	Inércia explicada	Inércia acumulada
1	**0.7161**	**30.00**	**30.00**
2	**0.4539**	**19.02**	**49.02**
3	**0.3736**	**15.65**	**64.67**
4	**0.2773**	**11.62**	**76.29**
5	0.1717	7.19	83.48
6	0.1290	5.40	88.89
7	0.1174	4.92	93.81
8	0.0467	1.96	95.76
9	0.0382	1.60	97.36
10	0.0243	1.02	98.38
11	0.0145	0.61	98.99
12	0.0131	0.0131	99.54
13	0.0111	0.0111	100.00

Determinaram-se as correlações das variáveis nos quatro primeiros eixos. Há correlações significativas a registar para todos eles, excepto para o quarto eixo, de modo que se optou por exibir apenas as correlações das variáveis com os três primeiros eixos (tabela 6.20).

A variável TF é a única que está correlacionada com o terceiro eixo, ainda que de forma pouco significativa. Por conseguinte, a análise limitar-se-á aos dois primeiros eixos.

O primeiro eixo (30% de inércia explicada) marca durante todo o período uma oposição entre as variáveis EA e TN (correlacionadas negativamente com este eixo) e as variáveis CE e TV (correlacionadas positivamente com este eixo). A variável TF apresenta correlações significativas apenas para os dois primeiros anos. Nenhuma destas variáveis está correlacionada de forma preponderante com o segundo eixo.

As variáveis mais correlacionadas de forma negativa com o segundo eixo (19.02% de inércia explicada) são CO e SC, apresentando um comportamento bastante estável durante o período referido. Estas não se encontram correlacionadas de forma significativa com o primeiro eixo.

As restantes variáveis não se encontram correlacionadas significativamente com nenhum dos eixos. A representação destas correlações encontra-se na figura 6.8.

As coordenadas dos indivíduos no objecto compromisso nos três primeiros

Tabela 6.20 Correlações das variáveis com os eixos do compromisso.

	PU				CO				TL		
	1	2	3		1	2	3		1	2	3
1980	0.623	-0.343	-0.166	1980	0.386	**-0.828**	-0.271	1980	0.302	0.472	-0.208
1985	0.540	-0.360	-0.211	1985	0.347	**-0.841**	-0.290	1985	0.516	0.627	-0.339
1990	0.435	-0.360	-0.274	1990	0.312	**-0.857**	-0.279	1990	0.614	0.502	-0.405
1994	0.407	-0.345	-0.271	1994	0.296	**-0.854**	-0.290	1994	0.629	0.485	-0.332
1995	0.398	-0.339	-0.268	1995	0.277	**-0.863**	-0.287	1995	0.603	0.490	-0.298
2000	0.297	-0.384	-0.205	2000	0.223	**-0.871**	-0.311	2000	0.759	-0.084	0.206
	TN				TR				TV		
	1	2	3		1	2	3		1	2	3
1980	**-0.877**	0.018	-0.455	1980	0.337	-0.070	0.320	1980	**0.737**	0.027	-0.102
1985	**-0.784**	0.055	-0.574	1985	0.301	-0.040	0.367	1985	**0.646**	0.048	-0.142
1990	**-0.637**	0.147	-0.660	1990	0.280	-0.054	0.409	1990	**0.736**	-0.101	-0.209
1994	**-0.643**	0.160	-0.645	1994	0.226	-0.070	0.463	1994	**0.780**	-0.345	-0.075
1995	**-0.659**	0.109	-0.656	1995	0.213	-0.080	0.479	1995	**0.764**	-0.430	-0.090
2000	**-0.659**	-0.076	-0.599	2000	0.170	-0.099	0.520	2000	**0.775**	-0.309	-0.208
	TM				EA				CE		
	1	2	3		1	2	3		1	2	3
1980	0.323	0.136	-0.206	1980	**-0.905**	-0.085	-0.160	1980	**0.775**	0.395	-0.271
1985	0.428	0.152	-0.122	1985	**-0.799**	-0.116	0.202	1985	**0.724**	0.470	-0.290
1990	0.490	0.228	0.078	1990	**-0.715**	-0.090	0.261	1990	**0.727**	0.478	-0.276
1994	0.472	0.126	0.184	1994	**-0.708**	-0.118	0.271	1994	**0.712**	0.481	-0.265
1995	0.528	0.136	0.156	1995	**-0.681**	-0.085	0.317	1995	**0.724**	0.473	-0.269
2000	0.519	0.116	0.283	2000	**-0.649**	-0.131	0.378	2000	**0.710**	0.482	-0.537
	TF				IA				SC		
	1	2	3		1	2	3		1	2	3
1980	**-0.855**	-0.030	-0.461	1980	0.028	-0.375	0.439	1980	0.237	**-0.873**	-0.259
1985	**-0.783**	0.074	-0.606	1985	-0.157	-0.430	0.576	1985	0.156	**-0.865**	-0.278
1990	-0.553	0.253	**-0.682**	1990	-0.455	-0.478	0.555	1990	0.051	**-0.828**	-0.336
1994	-0.536	0.301	**-0.672**	1994	-0.239	-0.340	0.669	1994	0.057	**-0.836**	-0.447
1995	-0.548	0.235	**-0.696**	1995	-0.425	-0.324	0.592	1995	0.045	**-0.795**	-0.353
2000	-0.505	0.052	**-0.651**	2000	-0.335	-0.164	0.577	2000	0.068	**-0.788**	-0.537

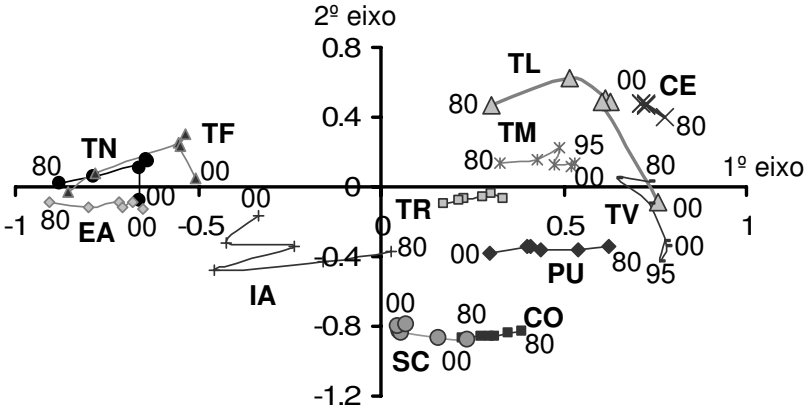


Figura 6.8 Representação das correlações entre as variáveis e o 1º e 2º eixos do compromisso.

eixos, assim como as suas contribuições (entre 0 e 1), encontram-se na tabela 6.21 e a respectiva imagem euclidiana no primeiro e segundo eixos na figura 6.9, tendo sido obtidas a partir do SPAD.

Tabela 6.21 Coordenadas e contribuições dos indivíduos no compromisso $\mathcal{W}$.

	Coordenadas			C. Absolutas			C. Relativas		
	1	2	3	1	2	3	1	2	3
AU	0.17	0.11	0.11	0.04	0.03	0.03	0.22	0.09	0.09
CH	-0.36	0.07	0.15	0.18	0.01	0.06	0.55	0.02	0.09
ES	0.00	-0.19	0.06	0.00	0.08	0.01	0.00	0.44	0.04
FI	0.25	0.31	-0.12	0.09	0.21	0.04	0.29	0.45	0.07
FR	0.07	-0.23	-0.12	0.01	0.11	0.04	0.05	0.58	0.15
GR	-0.16	0.02	0.22	0.03	0.00	0.13	0.23	0.00	0.48
HU	-0.01	0.12	0.04	0.00	0.03	0.00	0.00	0.09	0.01
IT	0.14	-0.22	0.11	0.03	0.11	0.03	0.15	0.42	0.10
MA	-0.04	0.05	-0.02	0.00	0.01	0.00	0.01	0.02	0.00
PB	0.06	-0.08	0.12	0.00	0.02	0.04	0.03	0.06	0.13
PO	-0.12	0.09	0.17	0.02	0.02	0.08	0.13	0.06	0.27
RU	0.23	-0.32	-0.19	0.08	0.23	0.09	0.25	0.48	0.16
SU	0.31	0.26	-0.16	0.13	0.15	0.07	0.42	0.31	0.12
TU	-0.53	0.02	-0.37	0.39	0.00	0.37	0.64	0.00	0.32

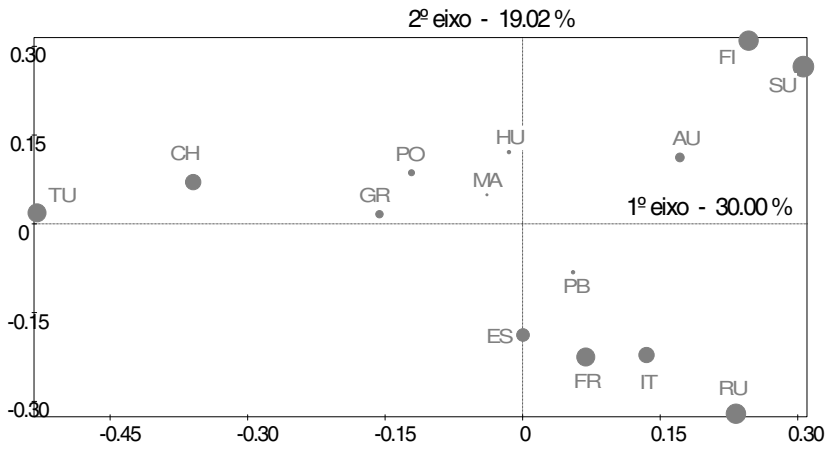


Figura 6.9 Imagem euclidiana do compromisso dos indivíduos no 1º e 2º eixos.

Os países com maior relevância no primeiro eixo são a Turquia, o Chipre, a Suécia, a Finlândia e o Reino Unido. As suas contribuições absolutas são superiores ao respectivo peso (1/14), logo a contribuição destes países para a formação deste eixo é significativa. As contribuições relativas da Turquia e do

Chipre para este eixo são superiores a 0.50, sendo a qualidade de representação destes no primeiro eixo bastante razoável. Estando as variáveis EA e TN correlacionadas negativamente com este eixo, a Turquia e o Chipre apresentam valores superiores à média destas variáveis, uma vez que se encontram na parte negativa deste eixo. Da mesma forma, se as variáveis TV e CE estão correlacionadas de forma positiva com este eixo, os países em causa apresentam valores inferiores à média destas variáveis.

Os países que mais contribuem para a explicação do segundo eixo são o Reino Unido, seguido da Finlândia, Suécia, França, Itália e Espanha. Os países melhor explicados pelo segundo eixo (com contribuições relativas não inferiores a 0.5) são a França e o Reino Unido.

O segundo eixo opõe a Finlândia e a Suécia à França, à Itália e ao Reino Unido. Isto deve-se ao facto de os dois primeiros países apresentarem valores inferiores às médias de CO e SC, por oposição aos valores mais elevados destas variáveis na França, Itália e Reino Unido.

Os indivíduos melhor representados no plano compromisso formado pelos dois primeiros eixos principais são aqueles cuja soma das contribuições relativas para estes eixos é superior a 0.50, como sendo a Finlândia (0.74), o Reino Unido (0.73), a Suécia (0.73), a Turquia (0.64), a França (0.63), o Chipre (0.57) e a Itália (0.57).

6.4.3 Trajectórias

Neste momento, é importante destacar os países responsáveis pelas diferenças entre os vários anos. A decomposição da soma das distâncias ao quadrado entre pares de objectos normados (tabela 6.22) permite distinguir aqueles países que mais contribuem para a diferença de estrutura no período 1980-2000: Turquia (19.51%), Chipre (11.22%), Suécia (9.05%), Malta (8.63%) e Grécia (8.02%).

A decomposição da distância ao quadrado entre pares de objectos normados em contribuições de indivíduos (tabela 6.23), permite destacar os países responsáveis pelas diferenças de estrutura entre pares de anos. A Turquia é responsável por essas diferenças para todos os pares de anos considerados, sendo a contribuição mais elevada entre 1980 e 1990 (27.42%) e a menos significativa entre 1994 e 1995 (7.45%). Outro país que também contribui globalmente para as diferenças estruturais é o Chipre, nomeadamente, entre 1980 e 2000 (16.71%). Malta tem contribuições significativas, sendo a mais elevada entre 1985 e 1990, com 19.49%. A Grécia tem a contribuição mais elevada entre 1990 e 1994 com 14.46%. A Suécia, entre 1980 e 1985, contribui com 20.60% e a Finlândia

Tabela 6.22 Decomposição de $\sum_t \sum_{t'} d^2_{HS}\left(\mathcal{W}_t/\|\mathcal{W}_t\|_{HS}, \mathcal{W}_{t'}/\|\mathcal{W}_{t'}\|_{HS}\right)$ em %.

AU	3.61
CH	**11.22**
ES	4.82
FI	5.97
FR	4.23
GR	**8.02**
HU	4.56
IT	4.77
MA	**8.63**
PB	4.02
PO	5.48
RU	6.10
SU	**9.05**
TU	**19.51**

com 15.92% entre 1980 e 1985. De entre os países com contribuições reduzidas destacam-se a Áustria, a Espanha, a Hungria, a Itália, os Países Baixos e Portugal.

As trajectórias dos indivíduos foram calculadas com o auxílio do Matlab. Para haver coerência entre as coordenadas do compromisso (determinadas no SPAD) e as trajectórias, os vectores próprios de $\mathcal{W}D$, $\boldsymbol{\varepsilon_k}$, são normados pela identidade e não pela métrica D.

As trajectórias foram representadas separadamente com o intuito de melhorar a legibilidade. A representação das trajectórias nos dois primeiros eixos do compromisso para a França, Itália e Reino Unido encontra-se na figura 6.10. O Chipre e a Turquia têm as suas trajectórias representadas na figura 6.11 e a Suécia e a Finlândia na figura 6.12. Cada trajectória é representada por uma linha. Os símbolos que atravessam essa linha representam os anos correspondentes e o ponto compromisso é representado por um círculo. As coordenadas das trajectórias de todos os países encontram-se no Anexo 3.

A interpretação das trajectórias deve ser feita apenas a países para os quais o co-seno ao quadrado do ângulo entre o vector representativo da posição compromisso e o plano compromisso for elevado, ou seja, com contribuições relativas superiores a 50%. Além disso há que ter cuidado com a sua interpretação, pois a trajectória de um indivíduo é apenas uma representação aproximada deste no período considerado.

Há dois tipos de trajectórias a destacar: aquelas que rodeiam a posição compromisso, como é o caso da Itália, e as mais irregulares, como o Reino Unido,

Tabela 6.23 Decomposição de $d^2_{HS}(\mathcal{W}_t/\|\mathcal{W}_t\|_{HS}, \mathcal{W}_{t'}/\|\mathcal{W}_{t'}\|_{HS})$ em %.

	80 e 85	80 e 90	80 e 94	80 e 95	80 e 00	85 e 90	85 e 94	85 e 95
AU	2.03	2.20	2.65	2.81	4.06	2.23	3.89	4.74
CH	5.06	5.40	**8.01**	**12.59**	**16.71**	5.48	**9.43**	**15.25**
ES	2.46	3.35	4.36	6.96	3.40	3.06	4.04	8.05
FI	**15.92**	7.49	6.03	6.34	2.96	3.07	3.81	3.94
FR	2.10	0.98	1.31	1.80	5.08	2.28	3.06	2.58
GR	**10.45**	**12.76**	**10.01**	8.27	5.25	**11.43**	6.25	4.34
HU	3.04	2.62	3.80	3.43	6.54	3.17	5.13	4.51
IT	3.62	4.61	4.41	4.78	5.01	4.85	4.93	6.93
MA	4.98	**10.94**	**11.44**	**9.61**	8.10	**19.49**	**17.35**	**13.56**
PB	3.58	2.63	4.12	3.72	5.19	3.60	4.57	4.03
PO	4.01	4.54	5.10	4.84	5.54	5.15	5.86	6.18
RU	3.30	2.84	4.66	4.30	7.48	2.65	6.47	6.15
SU	**20.60**	**12.21**	**8.61**	7.66	1.98	6.40	4.72	3.05
TU	**18.85**	**27.42**	**25.50**	**22.89**	**22.70**	**27.13**	**20.50**	**16.69**

	85 e 00	90 e 94	90 e 95	90 e 00	94 e 95	94 e 00	95 e 00
AU	4.33	6.70	7.24	4.01	1.98	3.81	3.61
CH	**15.67**	7.13	**11.02**	**11.97**	**14.35**	**12.82**	8.64
ES	3.16	1.61	5.13	3.65	**18.34**	6.74	**12.87**
FI	5.15	7.67	7.39	6.73	1.98	7.10	8.41
FR	5.45	6.60	5.34	6.43	**9.67**	**11.69**	**9.13**
GR	3.68	**14.46**	**11.85**	**9.92**	5.34	4.65	3.92
HU	7.00	3.79	3.76	4.83	3.97	4.08	5.16
IT	5.95	2.82	6.20	4.68	5.31	3.46	2.56
MA	7.60	4.27	3.20	2.04	8.52	2.89	3.42
PB	5.70	2.72	2.33	3.29	4.39	4.59	3.51
PO	5.99	**8.49**	**8.52**	7.47	3.17	3.81	2.94
RU	**8.00**	**12.53**	**9.90**	**9.23**	**9.31**	3.92	6.33
SU	7.58	6.11	6.78	**15.31**	6.23	**17.26**	**17.72**
TU	**14.74**	**15.10**	**11.34**	**10.45**	7.45	**13.15**	**11.78**

a França, a Turquia e a Finlândia.

A evolução da Itália segue a evolução média, no sentido em que, para cada variável (correlacionada com o primeiro e segundo eixos) o desvio entre o valor desta e a média é regular de um ano para o outro. Isto pode ser reforçado pelas baixas contribuições deste país para a decomposição da distância ao quadrado entre objectos normados (tabela 6.23), que se situam entre os 2.56% e os 6.93%.

Observando a trajectória do Reino Unido, verifica-se uma proximidade entre os anos 1994 e 1995. Esta proximidade traduz uma estabilidade por parte das

variáveis CO e SC, durante este período. Já no período de 1990 a 1994 há um grande afastamento, traduzido pelo aumento destas variáveis durante este período. Este facto é corroborado pela contribuição elevada deste país para a decomposição da distância ao quadrado entre objectos normados neste período (12.53%).

O afastamento da trajectória da Turquia para a direita é traduzido por um aumento das variáveis TV e CE e uma diminuição das variáveis TN e EA.

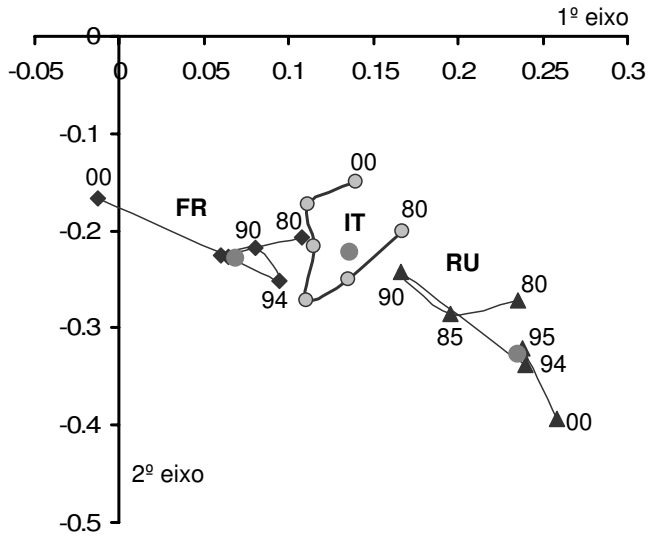


Figura 6.10 Trajectórias da França, Itália, Reino Unido.

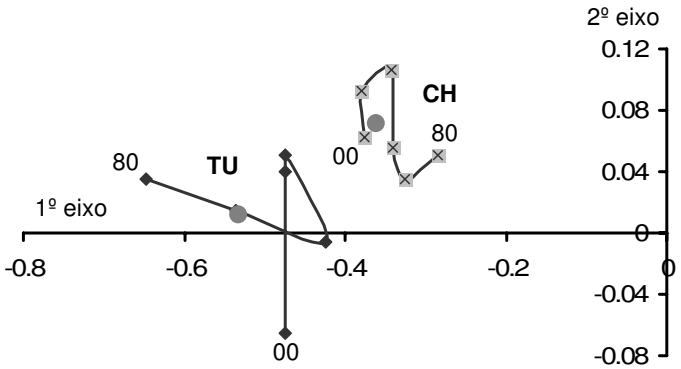


Figura 6.11 Trajectórias do Chipre e Turquia.

As trajectórias individuais num determinado eixo permitem avaliar as proximidades e oposições entre países, relativamente às variáveis correlacionadas com esse eixo, bem como a estabilidade destes ao longo dos diversos anos. A análise da figura 6.13 indica que as trajectórias que revelam maior grau de

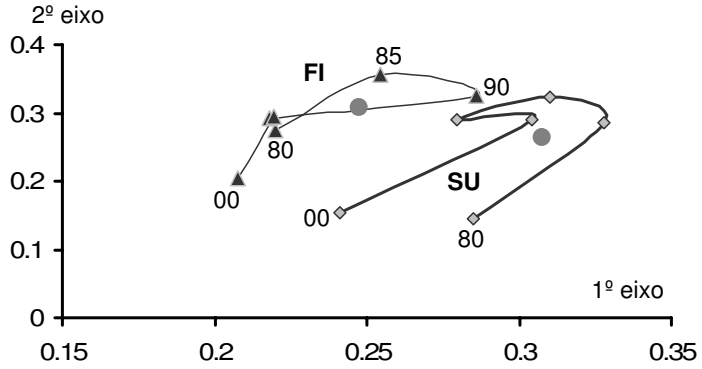


Figura 6.12 Trajectórias da Finlândia e Suécia.

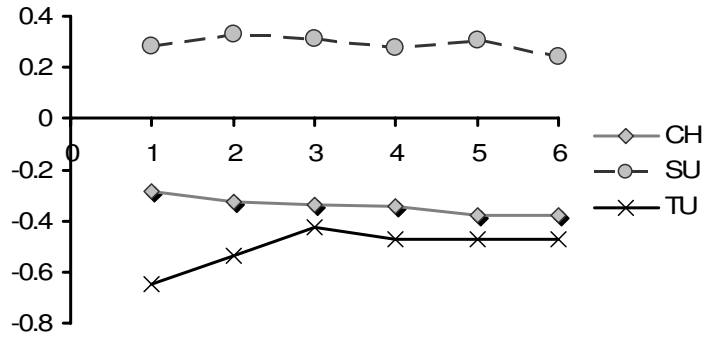


Figura 6.13 Trajectórias individuais em relação ao 1º eixo.

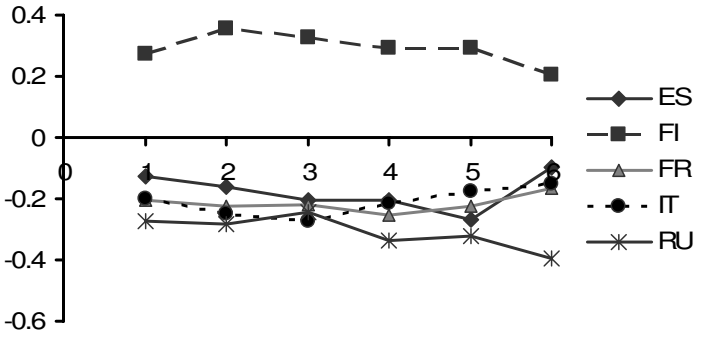


Figura 6.14 Trajectórias individuais em relação ao 2º eixo.

oposição relativamente ao primeiro eixo são o Chipre e a Turquia em relação à Suécia.

Relativamente ao segundo eixo há uma oposição entre a Itália, o Reino Unido e a França por um lado e a Finlândia por outro. Entre os anos de 1995 e 2000 há um afastamento entre os países da Itália e da França em relação ao Reino Unido.

6.5 Método STATIS dual

Este método permite a análise de uma possível estrutura comum às matrizes dos vários anos, assim como a análise das tendências evolutivas de cada uma das variáveis. Estas são heterogéneas por serem medidas em unidades diferentes, logo, tal como no método anterior, todas as observações foram centradas e reduzidas.

6.5.1 Inter-estrutura

Os pesos atribuídos a cada um dos estudos (X_t, Q, D_t), $(t = 80, 85, 90, 94, 95, 00)$ são os mesmos, sendo a matriz dos pesos dos estudos $\Delta = \frac{1}{6} I_6$.

A matriz dos coeficientes RV (tabela 6.24), obtida a partir do Matlab, é semelhante à matriz dos coeficientes RV do método STATIS (tabela 6.16) e traduz uma estrutura comum de variáveis nos vários quadros de dados.

Tabela 6.24 Matriz dos coeficientes RV.

RV	1980	1985	1990	1994	1995	2000
1980	1					
1985	0.952	1				
1990	0.821	0.933	1			
1994	0.846	0.929	0.963	1		
1995	0.836	0.931	0.969	0.990	1	
2000	0.753	0.843	0.878	0.914	0.922	1

As normas Hilbert-Schmidt dos objectos $\mathcal{V}_{80}, \ldots, \mathcal{V}_{00}$, coincidem com as normas dos objectos $\mathcal{W}_{80}, \ldots, \mathcal{W}_{00}$ (tabela 6.17), por definição desta norma e uma vez que os valores próprios de $\mathcal{W}_t D$ coincidem com os valores próprios de $\mathcal{V}_t Q$. Nesta análise considerar-se-ão os objectos não normados, em conformidade com o SPAD. Sendo os dados centrados e reduzidos, os objectos $\mathcal{V}_t$ correspondem a matrizes de correlações.

Aplicando uma ACP sobre a matriz dos produtos escalares $\mathcal{Z}_{tt'} = \langle \mathcal{V}_t, \mathcal{V}_{t'} \rangle_{HS}$, obtém-se a imagem euclidiana da inter-estrutura não centrada, efectuada com o auxílio do Matlab.

As coordenadas e contribuições das inter-estruturas não centrada e centrada encontram-se no Anexo 2.

A análise das imagens euclidianas da inter-estrutura não centrada (figura 6.15) e centrada (figura 6.16) permite visualizar a oposição entre os anos 1980 e 2000 com um coeficiente $RV(80, 00) = 0.753$, que é o mais fraco de todos. Por

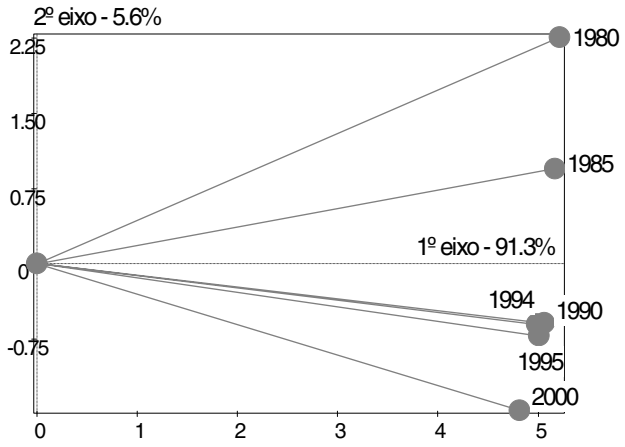


Figura 6.15 Imagem euclidiana da inter-estrutura não centrada.

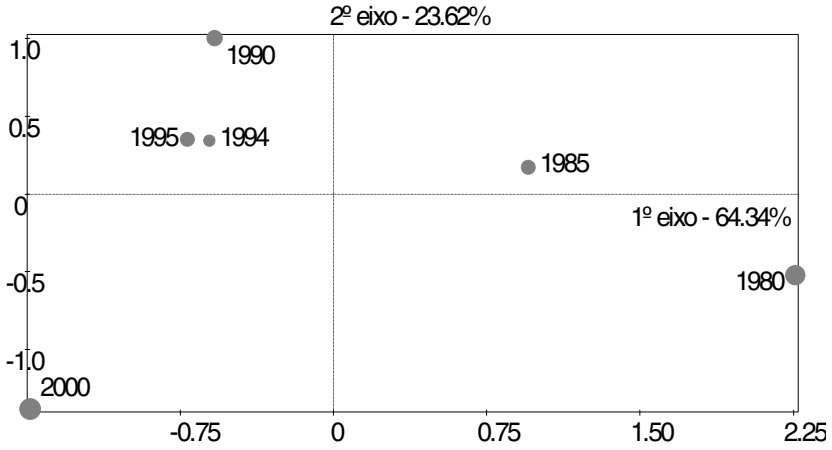


Figura 6.16 Imagem euclidiana da inter-estrutura centrada.

outro lado, os anos de 1994 e 1995 estão extraordinariamente próximos com um coeficiente $RV(94, 95) = 0.990$. De facto, observando as matrizes de correlações dos anos 1994 e 1995, verifica-se que estas são muito semelhantes, enquanto que as matrizes de correlações dos anos 1980 e 2000 apresentam algumas diferenças.

6.5.2 Intra-estrutura

Os coeficientes β_t que permitem definir o compromisso $\mathcal{V}$ encontram-se na tabela 6.25, assim como as distâncias HS entre os objectos $\mathcal{V}_t$ e o compromisso $\mathcal{V}$. Estas apresentam valores razoavelmente elevados, uma vez que os objectos não foram normados. O objecto mais próximo do compromisso é $\mathcal{V}_{95}$ e o mais afastado é $\mathcal{V}_{80}$, tal como no método STATIS.

Os valores próprios do compromisso $\mathcal{V}Q$, a percentagem de inércia explicada e

Tabela 6.25 Coeficientes β_t do compromisso $\mathcal{V}$ e distâncias HS.

$\beta_{80} = 0.172$	$d_{HS}(\mathcal{V}_{80}, \mathcal{V}) = 2.342$
$\beta_{85} = 0.171$	$d_{HS}(\mathcal{V}_{85}, \mathcal{V}) = 1.163$
$\beta_{90} = 0.167$	$d_{HS}(\mathcal{V}_{90}, \mathcal{V}) = 1.319$
$\beta_{94} = 0.165$	$d_{HS}(\mathcal{V}_{94}, \mathcal{V}) = 0.985$
$\beta_{95} = 0.165$	$d_{HS}(\mathcal{V}_{95}, \mathcal{V}) = 0.970$
$\beta_{00} = 0.159$	$d_{HS}(\mathcal{V}_{00}, \mathcal{V}) = 2.048$

a percentagem de inércia explicada acumulada encontram-se na tabela 6.26. Com os quatro primeiros eixos retém-se cerca de 76.63% da variância total.

Se os objectos $\mathcal{V}_t$ tivessem sido normados, estes valores próprios coincidiriam com os de $\mathcal{W}D$ (tabela 6.19), no entanto, a semelhança entre uns e outros é significativa, o que leva a crer que a não normalização dos objectos $\mathcal{V}_t$ não trouxe diferenças significativas na análise efectuada.

Tabela 6.26 Valores Próprios de $\mathcal{V}Q$.

Componente principal	Valor próprio	Inércia explicada	Inércia acumulada
1	**3.6263**	**30.22**	**30.22**
2	**2.2798**	**19.00**	**49.22**
3	**1.9026**	**15.86**	**65.07**
4	**1.3865**	**11.55**	**76.63**
5	0.8020	6.68	83.31
6	0.6304	5.25	88.56
7	0.6141	5.12	93.68
8	0.2960	2.47	96.15
9	0.2204	1.84	97.98
10	0.1574	1.31	99.30
11	0.0489	0.41	99.70
12	0.0355	0.30	100.00

As coordenadas das variáveis no objecto compromisso para os três primeiros eixos, bem como as suas contribuições (entre 0 e 1), encontram-se na tabela 6.27. As coordenadas no quarto eixo não foram explicitadas uma vez que as suas contribuições relativas eram extremamente fracas.

As variáveis que mais contribuem para a formação do primeiro eixo (30.22% da inércia explicada), isto é, com contribuição absoluta superior a 1/12 neste eixo são TN, TF, EA, TL, TV e CE. Destas, apenas TN, EA, TV e CE estão bem

representadas neste eixo, pois têm contribuições relativas iguais ou superiores a 0.50.

As variáveis que melhor explicam o segundo eixo (19% da inércia explicada) são CO, IA, TL, CE e SC, enquanto que as que são melhor explicadas por este são apenas a CO e a SC com contribuições relativas de 0.62 e 0.63, respectivamente.

As variáveis que melhor explicam o terceiro eixo são TN, TF, IA e SC. As contribuições relativas destas variáveis são pouco significativas neste eixo, portanto optou-se por restringir a representação dos indivíduos aos dois primeiros eixos.

As variáveis melhor representadas no plano compromisso formado pelos dois primeiros eixos principais, ou seja, com a soma das contribuições relativas para estes dois eixos superior a 0.50, são: CO (0.72), CE (0.72), TV (0.66), SC (0.64), TL (0.58), EA (0.56) e TN (0.53).

Tabela 6.27 Coordenadas e contribuições das variáveis no compromisso $\mathcal{V}$.

	Coordenadas			C. Absolutas			C. Relativas		
	1	2	3	1	2	3	1	2	3
PU	-0.45	-0.39	0.23	0.06	0.07	0.03	0.21	0.15	0.05
TN	0.72	0.11	0.60	0.14	0.01	0.19	0.52	0.01	0.36
TM	-0.46	0.22	-0.04	0.06	0.02	0.00	0.21	0.05	0.00
TF	0.67	0.18	0.62	0.12	0.01	0.20	0.45	0.03	0.39
CO	-0.32	-0.79	0.37	0.03	0.27	0.07	0.10	0.62	0.14
TR	-0.27	-0.10	-0.44	0.02	0.00	0.10	0.07	0.01	0.19
EA	0.74	-0.12	-0.22	0.15	0.01	0.02	0.55	0.01	0.05
IA	0.24	-0.48	-0.62	0.02	0.10	0.20	0.06	0.23	0.38
TL	-0.57	0.51	0.26	0.09	0.11	0.03	0.32	0.26	0.07
TV	-0.78	-0.22	0.13	0.17	0.02	0.01	0.61	0.05	0.02
CE	-0.71	0.47	0.22	0.14	0.10	0.03	0.50	0.22	0.05
SC	-0.12	-0.79	0.46	0.00	0.28	0.11	0.01	0.63	0.21

A imagem euclidiana do compromisso (figura 6.17) permite distinguir a oposição entre as variáveis TN e EA em relação às variáveis TV e CE no primeiro eixo. O segundo eixo opõe as variáveis CO e SC em relação às variáveis CE e TL. Note-se que estas semelhanças e oposições já tinham sido verificadas na implementação do método STATIS, através das correlações das variáveis com os eixos do compromisso dos indivíduos.

Há quatro grupos importantes a distinguir: o grupo formado apenas por TV, o grupo CE e TL, o grupo CO e SC e o grupo TF, TN e EA.

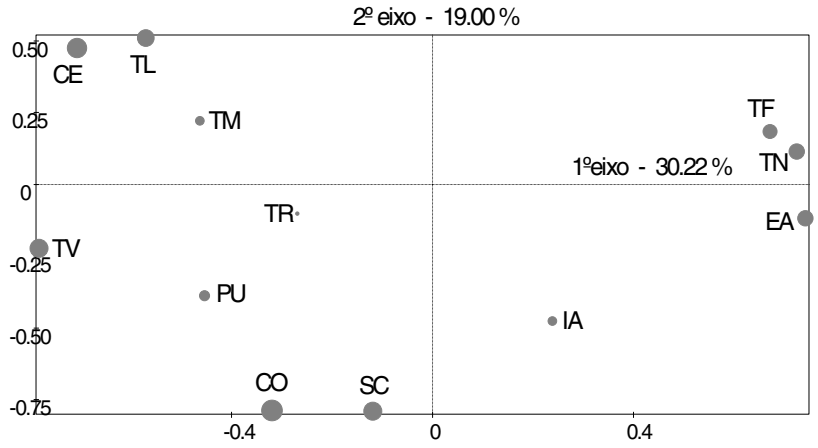


Figura 6.17 Imagem euclidiana do compromisso das variáveis no 1º e 2º eixos.

6.5.3 Trajectórias

A interpretação das trajectórias deve ser feita apenas a variáveis para as quais o co-seno ao quadrado do ângulo entre o vector representativo da posição compromisso e o plano compromisso for elevado, ou seja, com contribuições relativas superiores a 50%. Além disso há que ter cuidado com a sua interpretação, pois a trajectória de uma variável é apenas uma representação aproximada desta no período considerado, logo esta interpretação deve ser complementada com a decomposição da soma das distâncias ao quadrado entre pares de objectos (tabela 6.28) e com a decomposição das distâncias entre pares de objectos (tabela 6.29) em contribuições de variáveis.

As trajectórias das variáveis nos dois primeiros eixos do compromisso estão representadas nas figuras 6.18 a 6.21, que estão separadas apenas para melhorar a sua legibilidade. Cada trajectória é representada por uma linha. Os símbolos que atravessam essa linha representam os anos correspondentes e o ponto compromisso é representado por um círculo.

A decomposição da soma das distâncias ao quadrado entre pares de objectos (tabela 6.28) permite distinguir as variáveis cujas correlações com as restantes são instáveis: IA (13.05%), TF (12.34%), EA (12.21%), TL (10.90%) e TV (10.56%). Por outro lado, as correlações das variáveis CO e SC com as restantes são estáveis e as suas trajectórias situam-se em torno do objecto compromisso. Há apenas uma situação de menor estabilidade para SC entre 1995 e 2000, cuja contribuição na decomposição de $d^2_{HS}(\mathcal{V}_{95}, \mathcal{V}_{00})$ é de 10.83% (tabela 6.29).

As trajectórias de TF e TN quase que se sobrepõem (figura 6.20), pois tal como já se tinha analisado, estas estão fortemente correlacionadas. No entanto,

Tabela 6.28 Decomposição de $\sum_t \sum_{t'} d^2_{HS}(\mathcal{V}_t, \mathcal{V}_{t'})$ em %.

PU	7.88
TN	8.28
TM	4.65
TF	**12.34**
CO	4.06
TR	5.13
EA	**12.21**
IA	**13.05**
TL	**10.90**
TV	**10.56**
CE	6.20
SC	4.76

as contribuições de TF na decomposição das distâncias entre pares de objectos são mais significativas que as contribuições de TN.

A variável EA apresenta uma trajectória aberta (figura 6.18), em que o ano mais afastado em relação aos outros é o de 1980. A contribuição mais elevada na decomposição das distâncias entre pares de objectos é, de facto, entre 1980 e 1995 (19.47%).

No que diz respeito à variável TL, o ano que se destaca em relação aos outros é o de 2000. Este facto é reforçado pelas elevadas contribuições desta variável na decomposição entre pares de objectos normados que envolvam este ano, excepto entre 1980 e 2000.

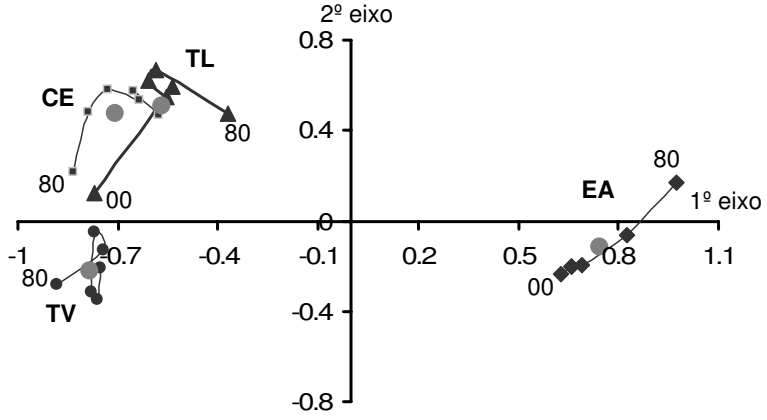


Figura 6.18 Trajectórias de CE, EA, TL e TV.

A contribuição mais significativa da variável CE dá-se no período 1980-1985 com cerca de 10.21%. Nos restantes pares de anos, as correlações desta variável

Tabela 6.29 Decomposição de $d^2_{HS}(\mathcal{V}_t, \mathcal{V}_{t'})$ em %.

	80 e 85	80 e 90	80 e 94	80 e 95	80 e 00	85 e 90	85 e 94	85 e 95
PU	6.18	**9.88**	**11.95**	**10.33**	8.95	**10.30**	**10.87**	9.46
TN	**9.94**	9.36	9.51	9.08	9.18	7.09	6.00	5.96
TM	4.31	3.02	3.79	3.68	4.99	4.61	8.60	6.34
TF	9.33	**11.37**	**12.88**	**12.13**	**13.66**	**11.07**	**11.44**	**11.03**
CO	1.44	0.90	2.40	2.76	4.26	3.06	6.73	7.40
TR	2.25	1.71	4.00	4.36	6.06	2.10	3.66	4.89
EA	**19.30**	**15.35**	**18.29**	**19.47**	**16.36**	8.63	8.76	**12.47**
IA	**13.43**	**17.19**	**10.91**	**13.94**	**11.93**	**16.77**	**10.43**	**12.06**
TL	**13.14**	7.78	6.39	5.24	7.64	5.65	5.61	4.09
TV	8.98	**14.36**	8.97	8.80	8.50	**21.26**	**15.37**	**14.99**
CE	**10.21**	7.63	8.61	7.49	4.15	5.57	6.76	4.48
SC	1.51	1.45	2.31	2.71	4.31	3.90	5.79	6.84

	85 e 00	90 e 94	90 e 95	90 e 00	94 e 95	94 e 00	95 e 00
PU	6.94	1.07	1.18	2.29	3.38	2.25	2.48
TN	6.44	5.37	4.60	7.86	9.42	8.88	8.11
TM	6.33	6.03	5.45	4.33	**9.67**	4.20	2.92
TF	**11.88**	5.42	4.98	**12.44**	9.05	**18.49**	**15.48**
CO	7.07	3.10	5.15	5.21	3.92	5.72	5.84
TR	6.73	6.71	**10.68**	9.16	3.14	6.66	7.17
EA	8.56	0.76	1.82	2.71	5.35	3.46	2.06
IA	**11.28**	**29.64**	**18.88**	**13.75**	**28.26**	7.39	5.44
TL	**13.77**	6.36	9.06	**21.97**	6.74	**27.39**	**29.80**
TV	8.59	**24.08**	**26.41**	6.02	9.00	4.19	4.65
CE	4.54	7.22	5.25	6.07	6.20	4.57	5.23
SC	7.86	4.24	6.54	8.20	5.87	6.80	**10.83**

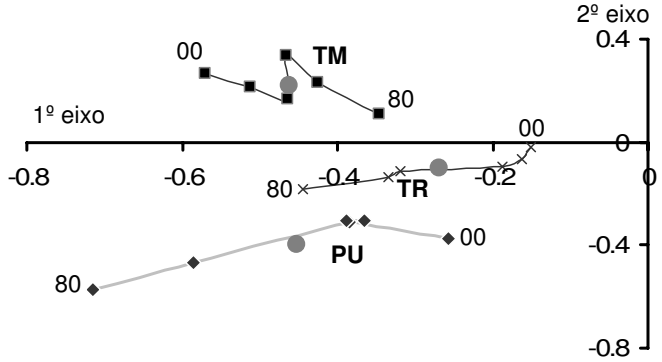


Figura 6.19 Trajectórias de PU, TM e TR.

com as outras mantêm-se bastante estáveis.

As contribuições mais elevadas de PU, dão-se entre 1980 e 1994 (11.95%) e

entre 1985 e 1994 (10.87%). A partir de 1995, as contribuições deixam de ser significativas.

Embora a qualidade de representação da variável IA na imagem euclidiana do compromisso (primeiro e segundo eixos) seja fraca (0.29), a sua trajectória alongada e irregular (figura 6.20) reflecte bem a decomposição de $d^2_{HS}(\mathcal{V}_t, \mathcal{V}_{t'})$. As contribuições mais elevadas de IA para a decomposição dão-se entre os anos de 1990 e 1994 (29.64%), 1994 e 1995 (28.26%) e 1980 e 1990 (17.19%).

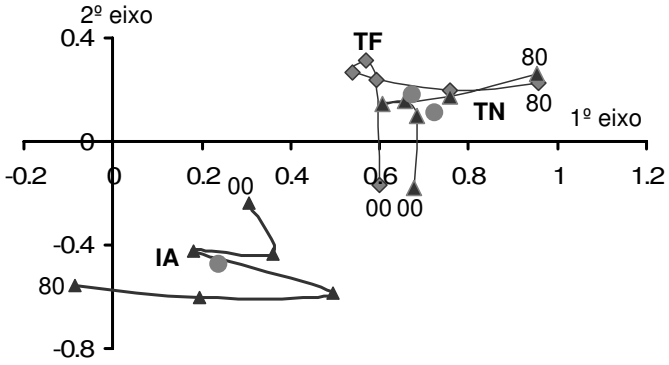


Figura 6.20 Trajectórias de IA, TF e TN.

Quanto às variáveis CO e SC, pode-se observar na figura 6.21, que as suas trajectórias são muito semelhantes para o segundo eixo, uma vez que estas estão fortemente correlacionadas (facto este comprovado através das matrizes de correlações) e a sua qualidade de representação no segundo eixo do compromisso é elevada.

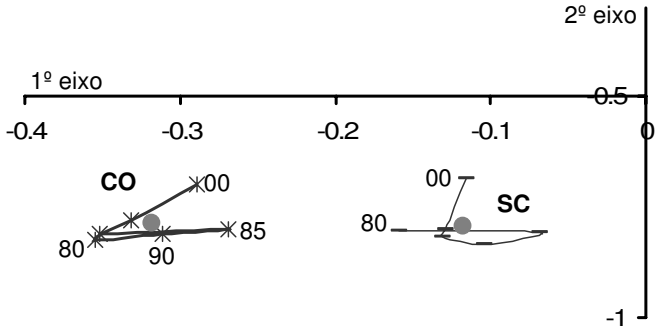


Figura 6.21 Trajectórias de CO e SC.

6.6 Análise Factorial Múltipla

A AFM permite analisar o conjunto de países caracterizados pelos seis grupos de variáveis. Nesta aplicação as variáveis são as mesmas para todos os grupos, no entanto, este método também é aplicável quando as variáveis diferem de grupo para grupo. Os indivíduos devem permanecer inalteráveis ao longo dos grupos.

A cada grupo de variáveis corresponde um quadro X_t, $(t = 80, 85, 90, 94, 95, 00)$. Tal como na metodologia STATIS, todos os dados foram centrados e reduzidos.

6.6.1 Determinação dos valores próprios de cada grupo

A primeira fase da AFM consiste em determinar os valores próprios de cada matriz $\mathcal{W}_t D$ $(t = 80, 85, 90, 94, 95, 00)$. O coeficiente de ponderação de cada objecto $\mathcal{W}_t$ será o inverso do primeiro valor próprio de $\mathcal{W}_t D$, visando equilibrar a influência de cada um destes objectos. Desta forma, nenhum dos grupos influenciará de forma preponderante a primeira direcção de inércia da análise global (intra-estrutura).

Os quatro maiores valores próprios associados a cada um dos grupos de variáveis encontram-se nas tabelas 6.30 a 6.35.

Tabela 6.30 Primeiros valores próprios de $\mathcal{W}_{80} D$.

1980			
Componente principal	Valor Próprio	Inércia explicada	Inércia acumulada
1	4.7468	39.56	39.56
2	2.1752	18.13	57.68
3	1.6361	13.63	71.32
4	1.0982	9.15	80.47

Tabela 6.31 Primeiros valores próprios de $\mathcal{W}_{85}D$.

1985			
Componente principal	Valor Próprio	Inércia explicada	Inércia acumulada
1	3.8946	32.45	32.45
2	2.5371	21.14	53.60
3	1.8189	15.16	68.76
4	1.2494	10.41	79.17

Tabela 6.32 Primeiros valores próprios de $\mathcal{W}_{90}D$.

1990			
Componente principal	Valor Próprio	Inércia explicada	Inércia acumulada
1	3.7179	30.98	30.98
2	2.5060	20.88	51.87
3	2.0179	16.82	68.68
4	1.2784	10.65	79.33

Tabela 6.33 Primeiros valores próprios de $\mathcal{W}_{94}D$.

1994			
Componente principal	Valor Próprio	Inércia explicada	Inércia acumulada
1	3.3603	28.00	28.00
2	2.5196	21.00	49.00
3	2.1464	17.89	66.89
4	1.4148	11.79	78.68

Tabela 6.34 Primeiros valores próprios de $\mathcal{W}_{95}D$.

1995			
Componente principal	Valor Próprio	Inércia explicada	Inércia acumulada
1	3.4767	28.97	28.97
2	2.4419	20.35	49.32
3	2.0633	17.19	66.52
4	1.4603	12.17	78.69

Tabela 6.35 Primeiros valores próprios de $\mathcal{W}_{00}D$.

2000			
Componente principal	Valor Próprio	Inércia explicada	Inércia acumulada
1	3.5730	29.78	29.78
2	2.7167	22.64	52.41
3	1.9156	15.96	68.38
4	1.3910	11.59	79.97

6.6.2 Intra-estrutura

Os eixos do compromisso são então obtidos através de uma ACP normada sobre o quadro completo $\mathcal{X}$ formado pela justaposição dos quadros

$$\frac{X_{80}}{\sqrt{4.75}}, \frac{X_{85}}{\sqrt{3.89}}, \frac{X_{90}}{\sqrt{3.72}}, \frac{X_{94}}{\sqrt{3.36}}, \frac{X_{95}}{\sqrt{3.48}}, \frac{X_{00}}{\sqrt{3.57}},$$

a partir da métrica

$$\mathcal{Q} = \begin{pmatrix} \boxed{Q_{80}} & 0 & \dots & 0 \\ 0 & \boxed{Q_{85}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boxed{Q_{00}} \end{pmatrix},$$

de dimensão 72×72, em que cada Q_t é a métrica associada ao espaço dos indivíduos do quadro X_t ($t = 80, 85, 90, 94, 95, 00$).

A expressão geral do compromisso é

$$\mathcal{W} = \mathcal{X}\mathcal{Q}\mathcal{X}' = \sum_{t=80}^{00} \frac{\mathcal{W}_t}{\lambda_1^t},$$

e os valores próprios de $\mathcal{W}D$ encontram-se na tabela 6.36 por ordem decrescente de magnitude, assim como a inércia explicada e a inércia explicada acumulada (em %). Os valores da percentagem de inércia explicada são muito semelhantes aos do método STATIS (tabela 6.19).

No método AFM, o SPAD determina os vectores próprios de $\mathcal{W}D$, $\boldsymbol{\varepsilon_k}$, ortonormados pela métrica D. A interpretação dos eixos do compromisso através das variáveis é feita com base na determinação das correlações destas (não ponderadas) com os eixos do compromisso, de forma análoga à do método STATIS. A sua representação num círculo de correlações pode-se obter no SPAD.

Tabela 6.36 Valores Próprios de $\mathcal{W}D$.

Componente principal	Valor Próprio	Inércia explicada	Inércia acumulada
1	**5.7356**	**29.85**	**29.85**
2	**3.6515**	**19.00**	**48.85**
3	**3.0411**	**15.82**	**64.67**
4	**2.2413**	**11.66**	**76.33**
5	1.3783	7.17	83.50
6	1.0447	5.44	88.94
7	0.9404	4.89	93.83
8	0.3728	1.94	95.77
9	0.3035	1.58	97.35
10	0.1982	1.03	98.39
11	0.1170	0.61	98.99
12	0.1061	0.55	99.55
13	0.0873	0.45	100

No entanto, optou-se por não a expor, uma vez que a sua legibilidade se torna difícil dado o grande número de variáveis de todos os grupos e a sua sobreposição por parte daquelas que tinham correlações mais estáveis.

Estas correlações coincidem com as do método STATIS (tabela 6.20), a menos do sinal para o primeiro e terceiro eixos. Tal facto é devido à diferença de sinal dos vectores próprios associados a estes eixos nos dois métodos. Desta forma, correlações positivas com estes eixos no método STATIS, são negativas no método AFM e vice-versa.

O primeiro eixo explica 29.85% da inércia e marca uma oposição entre as variáveis CE e TV (correlações negativas) e as variáveis EA, TN e TF (correlações positivas).

O segundo eixo (19% da inércia explicada) está correlacionado de forma positiva com as variáveis CO e SC.

A variável TF é a única que está correlacionada de forma positiva com o terceiro eixo, ainda que de forma pouco significativa. Não há correlações relevantes a registar para o quarto eixo, por conseguinte, a análise limitar-se-á aos dois primeiros eixos, como para o método STATIS.

A imagem euclidiana do compromisso nos dois primeiros eixos (figura 6.22) tem um aspecto bastante semelhante àquela obtida pelo método STATIS (figura 6.9), a menos de sinal para o primeiro eixo, e as contribuições obtidas pelos dois métodos são as mesmas (tabela 6.37).

Tabela 6.37 Coordenadas e contribuições dos indivíduos no compromisso $\mathcal{W}$.

	Coordenadas			C. Absolutas			C. Relativas		
	1	2	3	1	2	3	1	2	3
AU	-1.81	1.20	-1.20	0.04	0.03	0.03	0.21	0.09	0.09
CH	3.85	0.77	-1.52	0.18	0.01	0.05	0.56	0.02	0.09
ES	-0.04	-1.99	-0.64	0.00	0.08	0.01	0.00	0.44	0.05
FI	-2.62	3.24	1.30	0.09	0.21	0.04	0.29	0.44	0.07
FR	-0.73	-2.41	1.21	0.01	0.11	0.03	0.05	0.59	0.15
GR	1.63	0.19	-2.40	0.03	0.00	0.14	0.23	0.00	0.49
HU	0.13	1.28	-0.45	0.00	0.03	0.00	0.00	0.08	0.01
IT	-1.44	-2.34	-1.22	0.03	0.11	0.03	0.15	0.41	0.11
MA	0.46	0.54	0.31	0.00	0.01	0.00	0.01	0.02	0.01
PB	-0.54	-0.85	-1.22	0.00	0.01	0.03	0.02	0.06	0.13
PO	1.24	0.92	-1.86	0.02	0.02	0.08	0.13	0.07	0.28
RU	-2.49	-3.46	1.94	0.08	0.23	0.09	0.25	0.49	0.15
SU	-3.25	2.80	1.77	0.13	0.15	0.07	0.42	0.31	0.12
TU	5.59	0.11	3.99	0.39	0.00	0.37	0.64	0.00	0.32

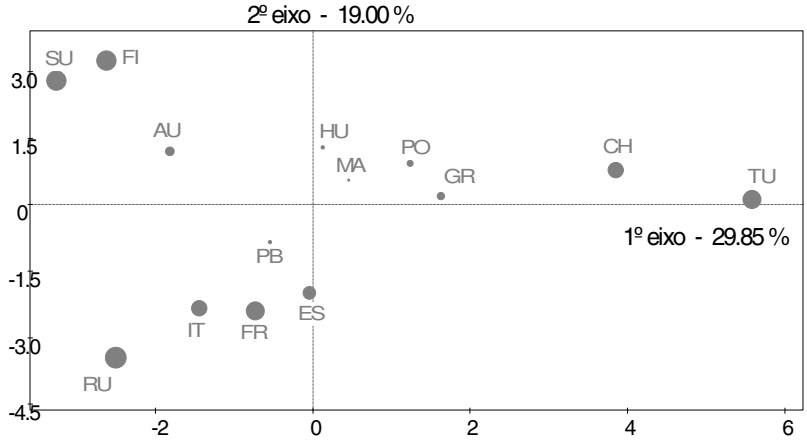


Figura 6.22 Imagem euclidiana do compromisso dos indivíduos no 1º e 2º eixos.

Neste momento é importante saber se cada um dos países se exprime de forma homogénea no conjunto das variáveis, ou se há instantes em que se destaca. Esta questão deve ser colocada não em termos de variáveis, como o foi até ao momento, mas em termos de grupos de variáveis. Neste sentido, a representação simultânea de todas as nuvens é bastante útil, uma vez que sintetiza o comportamento de cada um dos países do ponto de vista de cada um dos grupos de variáveis.

A representação simultânea das nuvens de indivíduos $\mathcal{N}_I^t$ $(t = 80, 85, 90, 94, 95, 00)$ nos dois primeiros eixos encontra-se na figura 6.23 e as coordenadas,

bem como respectivas contribuições nos três primeiros eixos encontram-se no Anexo 4. Há semelhanças a registar entre esta representação e a das trajectórias obtidas pelo método STATIS.

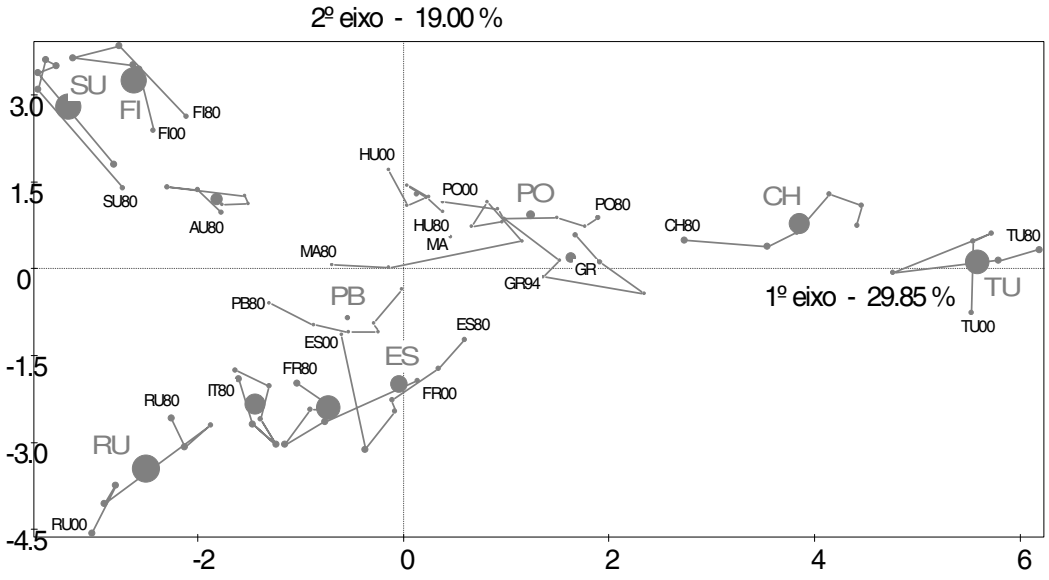


Figura 6.23 Representação simultânea das nuvens de indivíduos.

6.6.3 Inter-estrutura

A ligação entre os grupos de variáveis pode ser avaliada através de dois coeficientes, RV e $\mathcal{L}_g$, que fornecem informação complementar. A matriz dos coeficientes de ligação $\mathcal{L}_g$ entre os vários grupos de variáveis, incluindo o compromisso (última linha e coluna) encontra-se na tabela 6.38, enquanto que a matriz dos coeficientes RV se encontra na tabela 6.39. Os valores destes coeficientes são bastante elevados, nomeadamente entre X_{94} e X_{95}, traduzindo estruturas de indivíduos bastante semelhantes, uma vez que o coeficiente RV se encontra próximo de 1. O valor de $\mathcal{L}_g(X_{94}, X_{95})$ indica que cada uma das variáveis de X_{94} está bastante correlacionada com cada variável de X_{95} e que há duas direcções comuns aos dois grupos, cuja inércia é comparável à inércia máxima de cada grupo.

O quadrado da norma do objecto representativo de cada grupo de variáveis, η_g^2, definido pela expressão (3.3), encontra-se na tabela 6.40 e foi calculado com o auxílio do Matlab. Este indica o número de direcções de inércia comparável à inércia do primeiro eixo associado a esse mesmo objecto. O objecto X_{80} é o que possui um índice de dimensionalidade mais fraco (1.449), enquanto que X_{94}

Tabela 6.38 Matriz dos coeficientes de ligação $\mathcal{L}_g$.

$\mathcal{L}_g$	1980	1985	1990	1994	1995	2000	AFM
1980	1.449						
1985	1.556	1.849					
1990	1.393	1.787	1.980				
1994	1.509	1.897	2.029	2.292			
1995	1.460	1.837	1.955	2.196	2.152		
2000	1.364	1.668	1.739	2.027	1.990	2.141	
AFM	1.522	1.847	1.898	2.084	2.021	1.906	1.966

Tabela 6.39 Matriz dos coeficientes RV.

RV	1980	1985	1990	1994	1995	2000	AFM
1980	1						
1985	0.951	1					
1990	0.822	0.934	1				
1994	0.828	0.921	0.952	1			
1995	0.827	0.921	0.947	0.989	1		
2000	0.774	0.838	0.845	0.915	0.927	1	
AFM	0.902	0.969	0.962	0.981	0.982	0.929	1

possui o índice η_g^2 mais elevado (2.292), com duas direcções de inércia comparáveis à direcção de inércia máxima deste grupo.

Tabela 6.40 Índice de multidimensionalidade η_g^2.

$$\eta_g^2(X_{80}) = 1.449$$
$$\eta_g^2(X_{85}) = 1.849$$
$$\eta_g^2(X_{90}) = 1.980$$
$$\eta_g^2(X_{94}) = 2.292$$
$$\eta_g^2(X_{95}) = 2.152$$
$$\eta_g^2(X_{00}) = 2.141$$

As coordenadas dos grupos de variáveis nos três primeiros eixos encontram-se na tabela 6.41, bem como as contribuições relativas (entre 0 e 1) e absolutas (em %). A sua imagem euclidiana encontra-se na figura 6.24. Geralmente a qualidade de representação dos grupos de variáveis para este método é pouco satisfatória, o que também acontece neste caso, uma vez que os únicos grupos

que estão bem representados no primeiro eixo e só no primeiro são o de 1980 e 1985. Por outro lado, é possível interpretar cada eixo da inter-estrutura em função da componente principal que define esse eixo. A coordenada de um grupo relativamente a um determinado eixo é a contribuição absoluta desse grupo para a inércia da componente principal associada a esse mesmo eixo.

As coordenadas relativas ao primeiro eixo estão próximas de 1 para todos os grupos, o que significa que este é uma direcção de inércia importante para todas as nuvens. Ora este eixo está correlacionado de forma negativa com as variáveis CE, TV e de forma positiva com as variáveis EA e TN.

Os grupos que possuem coordenadas mais elevadas relativamente ao segundo eixo são o de 1994 e o de 1995. Pode-se dizer que, globalmente, as variáveis CO e SC são as as mais características destas nuvens, por estarem correlacionadas de forma positiva com este eixo. Já a nuvem de 1980 é caracterizada por uma posição menos dominante em relação a estas variáveis.

Tabela 6.41 Coordenadas e contribuições dos grupos de variáveis.

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
1980	0.92	0.45	0.22	16.0	12.2	7.3	0.58	0.14	0.03
1985	0.98	0.62	0.42	17.0	17.1	13.8	0.51	0.21	0.09
1990	0.94	0.65	0.54	16.4	17.8	17.8	0.44	0.21	0.15
1994	0.98	0.71	0.65	17.2	19.6	21.4	0.42	0.22	0.19
1995	0.98	0.68	0.60	17.1	18.5	19.6	0.45	0.21	0.17
2000	0.94	0.54	0.61	16.4	14.9	20.1	0.41	0.14	0.17

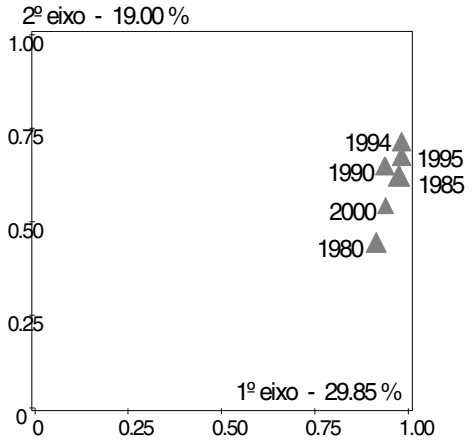


Figura 6.24 Representação dos grupos de variáveis.

As direcções de inércia existentes nos outros eixos são mais fracas, o que não significa forçosamente que não sejam importantes. Como tal calcularam-se os coeficientes de correlação entre as componentes principais associadas a cada nuvem $\mathcal{N}_I^t$ e as cinco primeiras componentes principais associadas ao compromisso (tabela 6.42), que são bastante satisfatórios em todos os casos.

Tabela 6.42 Coeficientes de correlação entre ${\boldsymbol{F_{u_k}}}^t$ e $\boldsymbol{F_{u_k}}$.

r	1	2	3	4	5
1980	0.97	0.98	0.83	0.78	0.86
1985	0.99	0.99	0.96	0.91	0.91
1990	0.98	0.98	0.96	0.96	0.87
1994	1.00	0.99	0.99	0.97	0.97
1995	0.99	0.99	0.99	0.98	0.95
2000	0.98	0.94	0.93	0.93	0.94

6.6.4 Interpretação das posições compromisso e das trajectórias

A qualidade da representação simultânea das nuvens é avaliada através da razão entre a inércia inter e a inércia total. Quanto maior for este quociente para cada eixo da intra-estrutura, maior é a qualidade de representação das nuvens. Como se pode observar na tabela 6.43, esta é bastante satisfatória para os cinco primeiros eixos.

Tabela 6.43 Razão [inércia inter/inércia total].

eixo	1	2	3	4	5
razão	0.97	0.94	0.85	0.85	0.83

A inércia intra pode ser decomposta em contribuições de indivíduos. As contribuições mais significativas (em %) encontram-se na tabela 6.44 e traduzem uma maior variabilidade dos pontos representativos de cada país.

Os indivíduos cuja contribuição para a inércia intra é menos significativa são aqueles que apresentam uma maior estabilidade nas suas trajectórias (tabela 6.45). Esta informação pode ser cruzada com a decomposição da soma do quadrado das distâncias entre pares de objectos normados em contribuições de

indivíduos (tabela 6.23). As contribuições mais reduzidas são também aquelas cuja contribuição para a inércia intra é mais fraca.

Tal como no método STATIS, a interpretação das trajectórias só deverá ser feita para aqueles países que possuírem elevadas contribuições relativas.

O primeiro explica 29.85% da inércia total e traduz essencialmente o comportamento das variáveis CE, TV, EA e TN. Os países com maior importância neste eixo são o Chipre e a Turquia, opondo-se aos restantes. As trajectórias destes são abertas e irregulares. Este facto é confirmado pelas suas contribuições elevadas para a inércia intra.

O segundo eixo marca uma oposição entre o grupo da Finlândia e Suécia em relação ao grupo da Itália e do Reino Unido. Estes últimos apresentam valores superiores à média das variáveis CO e SC, enquanto que a Finlândia e a Suécia apresentam valores inferiores.

O país que evidencia o comportamento mais regular durante o período analisado é a Itália. Isto pode ser comprovado através da representação simultânea e da sua fraca contribuição para a inércia intra.

Tabela 6.44 Indivíduos com inércias intra mais significativas em %.

eixo 1	inércia	acum.	eixo 2	inércia	acum.	eixo 3	inércia	acum.
MA	17.30	17.30	SU	23.22	23.22	GR	18.61	18.61
CH	14.05	31.35	RU	16.19	39.42	SU	15.35	33.96
PO	11.13	42.48	ES	15.15	54.57	TU	12.03	45.99
RU	7.50	49.97	FI	8.88	63.44	CH	8.49	54.48
PB	7.50	57.47	IT	6.72	70.17	MA	8.3	62.78
TU	7.40	64.87	TU	6.25	76.42	RU	8.21	70.99
GR	7.30	72.18	GR	5.75	82.16	FR	5.65	76.64

Tabela 6.45 Indivíduos com inércias intra menos significativas em %.

eixo 1	inércia	acum.	eixo 2	inércia	acum.	eixo 3	inércia	acum.
IT	0.85	0.85	PO	0.57	0.57	PB	2.41	2.41
HU	1.19	2.04	AU	0.73	1.30	IT	2.66	5.07
AU	2.91	4.95	HU	1.68	2.98	HU	2.71	7.78
FI	4.48	9.43	PB	2.31	5.29	ES	2.75	10.53
SU	4.76	14.19	CH	3.21	8.50	PO	3.95	14.49
ES	6.55	20.74	FR	4.34	12.84	AU	4.14	18.63
FR	7.08	27.82	MA	5.00	17.84	FI	4.73	23.36

6.7 Dupla Análise em Componentes Principais

A DACP pode ser implementada neste tipo de dados, uma vez que os mesmos indivíduos e as mesmas variáveis são medidos em cada instante $t = 80, 85, 90, 94, 95, 00$. Pelas mesmas razões invocadas nos outros métodos, os dados serão centrados e reduzidos, após a análise da inter-estrutura.

O SPAD não efectua este método integralmente, como para a AFM e a metodologia STATIS, mas uma vez que consiste essencialmente na realização de Análises em Componentes Principais, estas podem ser obtidas através do SPAD.

6.7.1 Inter-estrutura

A inter-estrutura consiste numa ACP normada sobre a matriz G dos centros de gravidade das nuvens $\mathcal{N}_I^t$, que se encontram na tabela 6.2.

Os valores próprios da inter-estrutura encontram-se na tabela 6.46 por ordem decrescente de magnitude, bem como a percentagem de inércia explicada e a percentagem de inércia explicada acumulada. Os dois primeiros eixos explicam cerca de 94.17% da inércia logo, a representação gráfica da inter-estrutura nos dois primeiros eixos é mais do que suficiente.

Tabela 6.46 Valores Próprios da inter-estrutura.

Componente principal	Valor próprio	Inércia explicada	Inércia acumulada
1	**9.3776**	**78.15**	**78.15**
2	**1.9232**	**16.03**	**94.17**
3	0.4303	3.59	97.76
4	0.1611	1.34	99.10
5	0.1078	0.90	100.00

As correlações das variáveis com os dois primeiros eixos da inter-estrutura são bastante satisfatórias, uma vez que todas elas se aproximam da fronteira do círculo de correlações (figura 6.25) e possuem correlações elevadas essencialmente com o primeiro eixo (tabela 6.47).

As variáveis mais correlacionadas de forma negativa com o primeiro eixo são TN, TF, TM e EA. Através da imagem euclidiana da inter-estrutura (figura 6.26) verifica-se que o percurso dos anos estudados no primeiro eixo se desloca no sentido positivo. Quer isto dizer que há uma diminuição das médias destas variáveis. Por outro lado, as variáveis TV, CE e PU estão correlacionadas de

forma positiva com o primeiro eixo, logo há um aumento das médias destas variáveis durante o período analisado. O primeiro eixo traduz uma evolução regular por parte das variáveis correlacionadas com este.

As variáveis mais correlacionadas com o segundo eixo, ainda que de forma bastante menos significativa, são CO (-0.67) e IA (0.67). As coordenadas da inter-estrutura relativamente a este eixo são bastante menos regulares.

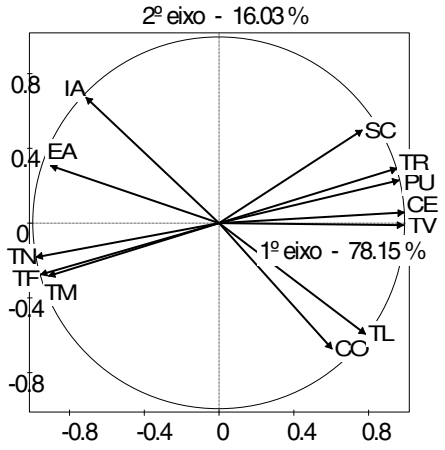


Figura 6.25 Círculo de correlações (inter-estrutura da DACP).

Tabela 6.47 Correlações entre as variáveis e as componentes principais.

Variável	1	2	3	4
PU	0.96	0.23	-0.06	-0.13
TN	-0.98	-0.18	0.01	-0.05
TM	-0.91	-0.28	0.03	0.18
TF	-0.95	-0.28	0.08	-0.01
CO	0.60	-0.67	0.41	-0.13
TR	0.95	0.29	-0.06	-0.01
EA	-0.90	0.31	0.25	-0.16
IA	-0.71	0.67	0.15	0.11
TL	0.78	-0.59	0.04	0.15
TV	0.99	-0.01	0.07	0.11
CE	0.99	0.06	-0.09	-0.08
SC	0.77	0.50	0.39	0.12

A imagem euclidiana da inter-estrutura (figura 6.26) bem como as suas coordenadas (tabela 6.48) evidenciam a proximidade entre os anos 1994 e 1995, pois são os anos cronologicamente mais próximos, e a oposição entre os anos de 1980 e 2000, pois são os anos cronologicamente mais afastados.

Tabela 6.48 Coordenadas e contribuições da inter-estrutura (DACP).

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
1980	-4.63	-1.28	0.79	38.1	14.2	24	0.90	0.07	0.03
1985	-2.44	0.24	-0.80	10.6	0.5	24.7	0.84	0.01	0.09
1990	-0.69	-0.19	-0.62	0.9	0.3	14.9	0.33	0.02	0.26
1994	1.08	1.67	-0.28	2.1	24.3	3.1	0.27	0.63	0.02
1995	1.79	1.63	0.93	5.7	23.2	33.4	0.46	0.39	0.13
2000	4.90	-2.08	-0.02	42.7	37.6	0.00	0.85	0.15	0.00

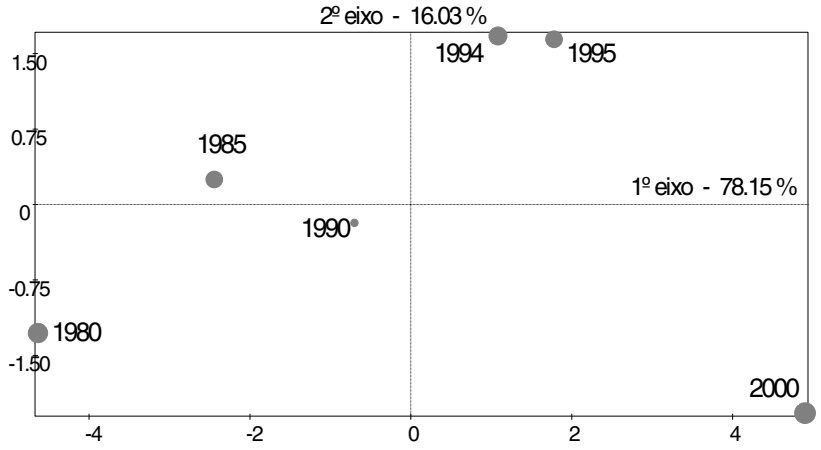


Figura 6.26 Imagem euclidiana da inter-estrutura.

6.7.2 Análise das nuvens de indivíduos

Esta etapa tem por objectivo avaliar a deformação das nuvens de indivíduos e consiste na realização de uma ACP normada (uma vez que as variáveis são heterogéneas) sobre cada um dos quadros representativos de cada ano. Embora esta análise seja vantajosa, tornar-se-ia bastante demorada, uma vez que efectuar-se-iam seis Análises em Componentes Principais. Na secção 6.3 efectuaram-se estas análises somente aos quadros de 1980 e 2000.

Os quatro maiores valores próprios das matrizes de correlações dos anos em causa, bem como a percentagem de inércia explicada e a percentagem de inércia explicada acumulada encontram-se nas tabelas 6.30 a 6.35, uma vez que foi necessário proceder à sua determinação na primeira etapa da AFM. Os quatro primeiros valores próprios de todas as matrizes de correlações encontram-se na tabela 6.49.

Tabela 6.49 Valores próprios das matrizes de correlações.

	1980	1985	1990	1994	1995	2000
1	4.7468	3.8946	3.7180	3.3603	3.4767	3.5730
2	2.1752	2.5372	2.5060	2.5196	2.4419	2.7167
3	1.6361	1.8189	2.0179	2.1464	2.0633	1.9156
4	1.0982	1.2494	1.2784	1.4148	1.4603	1.3910

6.7.3 Intra-estrutura

Esta etapa consiste na pesquisa de um espaço de representação comum às diversas nuvens. Para tal foram aplicados os dois primeiros critérios descritos na secção 4.4.

Primeiro Critério

Os índices $\Phi(t, \tau)$ encontram-se na tabela 6.50 e representam a perda de inércia da nuvem $\mathcal{N}_I^t$ quando se projectam os "seus" indivíduos sobre os q primeiros eixos factoriais da nuvem $\mathcal{N}_I^\tau$.

Uma vez que a percentagem de inércia explicada pelos valores próprios de cada uma das matrizes de correlações ronda os 80% com os quatro primeiros valores próprios (tabelas 6.30 a 6.35) tem-se que $q = 4$.

O primeiro critério consiste em seleccionar o sistema de eixos para os quais a perda média de inércia é mínima. Como se pode verificar na tabela 6.50 o sistema de eixos escolhido é o de 1994, logo a matriz compromisso das variáveis é a matriz das correlações R_{94}.

As posições compromisso das variáveis estão representadas no círculo de correlações da figura 6.27, obtido através de uma ACP sobre X_{94}. As respectivas coordenadas encontram-se na tabela 6.51.

Os dois primeiros eixos da intra-estrutura explicam 49% da inércia. As variáveis mais correlacionadas com o primeiro são TV (-0.84), TN e EA (0.67), CE (-0.64) e TF (0.59) e as variáveis mais correlacionadas com o segundo eixo são TL (0.70) e CE (0.65). As correlações das variáveis com o terceiro e quarto eixo não são significativas, de modo que se limitou a representação aos dois primeiros eixos.

As trajectórias dos indivíduos obtêm-se por projecção destes nos eixos principais de 1994 (vectores próprios de R_{94}). A representação das trajectórias nos dois primeiros eixos do compromisso encontra-se na figura 6.28 e as respectivas

Tabela 6.50 Índices Φ (1º critério).

$\Phi(80,80)$	0.0000	$\Phi(85,80)$	-0.0165	$\Phi(90,80)$	-0.0143
$\Phi(80,85)$	0.0162	$\Phi(85,85)$	0.0000	$\Phi(90,85)$	0.0021
$\Phi(80,90)$	0.0141	$\Phi(85,90)$	-0.0021	$\Phi(90,90)$	0.0000
$\Phi(80,94)$	0.0223	$\Phi(85,94)$	0.0062	$\Phi(90,94)$	0.0083
$\Phi(80,95)$	0.0222	$\Phi(85,95)$	0.0061	$\Phi(90,95)$	0.0082
$\Phi(80,00)$	0.0062	$\Phi(85,00)$	-0.0101	$\Phi(90,00)$	-0.0080
$\Phi(\cdot,80)$	0.0135	$\Phi(\cdot,85)$	-0.0027	$\Phi(\cdot,90)$	-0.0006
$\Phi(94,80)$	-0.0228	$\Phi(95,80)$	-0.0227	$\Phi(00,80)$	-0.0063
$\Phi(94,85)$	-0.0062	$\Phi(95,85)$	-0.0061	$\Phi(00,85)$	0.0100
$\Phi(94,90)$	-0.0084	$\Phi(95,90)$	-0.0083	$\Phi(00,90)$	0.0079
$\Phi(94,94)$	0.0000	$\Phi(95,94)$	0.0001	$\Phi(00,94)$	0.0162
$\Phi(94,95)$	-0.0001	$\Phi(95,95)$	0.0000	$\Phi(00,95)$	0.0161
$\Phi(94,00)$	-0.0164	$\Phi(95,00)$	-0.0163	$\Phi(00,00)$	0.0000
$\Phi(\cdot,94)$	-0.0090	$\Phi(\cdot,95)$	-0.0089	$\Phi(\cdot,00)$	0.0073

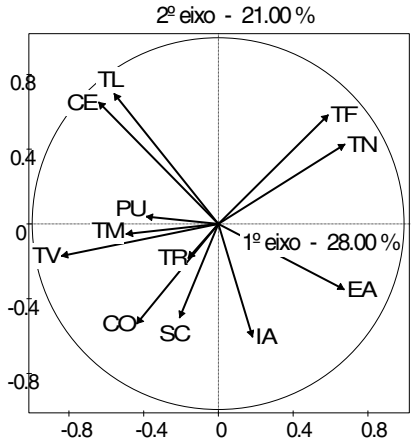


Figura 6.27 Círculo de correlações (1º critério da DACP).

coordenadas encontram-se no Anexo 5.

Segundo Critério

O sistema de eixos procurado, segundo este critério, é constituído pelos vectores próprios da matriz $\sum_t R_t$, que é a matriz compromisso das variáveis no período analisado. Os quatro primeiros eixos explicam cerca de 76.62% de inércia.

O primeiro eixo explica 30.16% da inércia e está fundamentalmente relacionado com as variáveis TV (-0.78), EA (0.74), TN (0.72), CE (-0.71)

Tabela 6.51 Correlações entre as variáveis e as componentes principais (1º critério da DACP).

Variável	1	2	3	4
PU	-0.38	0.04	0.46	-0.46
TN	0.67	0.43	0.54	-0.09
TM	-0.49	-0.05	-0.34	0.58
TF	0.59	0.58	0.49	-0.09
CO	-0.44	-0.53	0.66	0.02
TR	-0.16	-0.18	-0.29	-0.61
EA	0.67	-0.35	-0.11	-0.22
IA	0.18	-0.60	-0.33	-0.53
TL	-0.56	0.70	-0.01	-0.21
TV	-0.84	-0.17	0.19	-0.15
CE	-0.64	0.65	-0.08	-0.25
SC	-0.21	-0.50	0.78	0.15

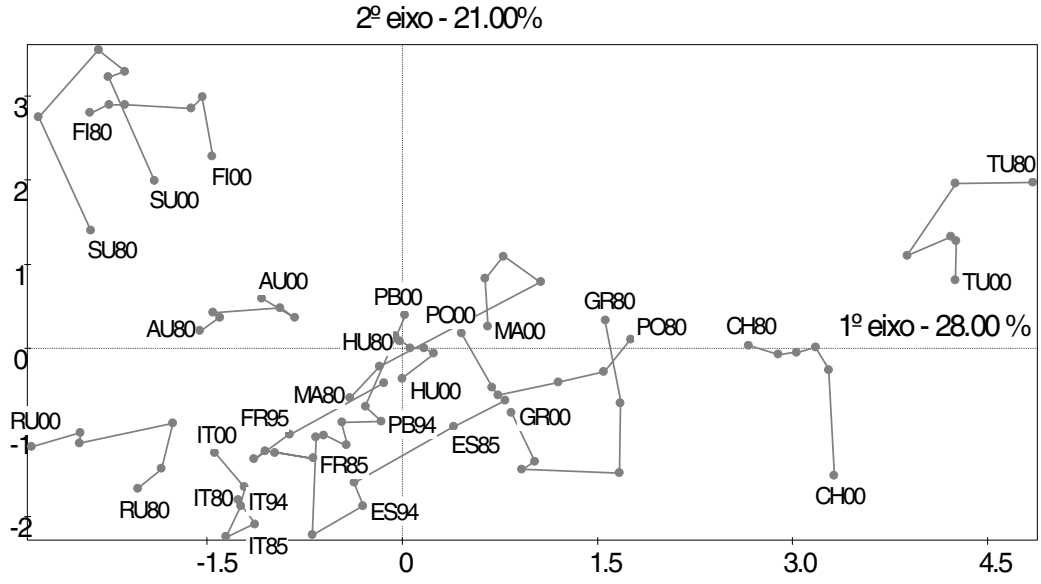


Figura 6.28 Trajectórias dos indivíduos (1º critério da DACP).

e TF (0.67). O segundo eixo explica 18.96% da inércia e as variáveis mais correlacionadas com este são CO e SC (-0.80). As correlações das variáveis com o terceiro não são significativas e para o quarto eixo há apenas que registar a correlação com a variável TM (-0.69), de modo que se limitou a representação aos dois primeiros eixos.

As trajectórias dos indivíduos são obtidas através de uma ACP normada sobre o quadro formado pela sobreposição dos quadros $X_{80}, \ldots, X_{00}$. As

Tabela 6.52 Valores Próprios de $\sum_t R_t$ (2º critério da DACP).

Componente principal	Valor próprio	Inércia explicada	Inércia acumulada
1	**3.6194**	**30.16**	**30.16**
2	**2.2752**	**18.96**	**49.12**
3	**1.9103**	**15.92**	**65.04**
4	**1.3892**	**11.58**	**76.62**
5	0.8034	6.69	83.31
6	0.6263	5.22	88.53
7	0.6182	5.15	93.68
8	0.2950	2.46	96.14
9	0.2206	1.84	97.98
10	0.1572	1.31	99.29
11	0.0490	0.41	99.70
12	0.0363	0.30	100.00

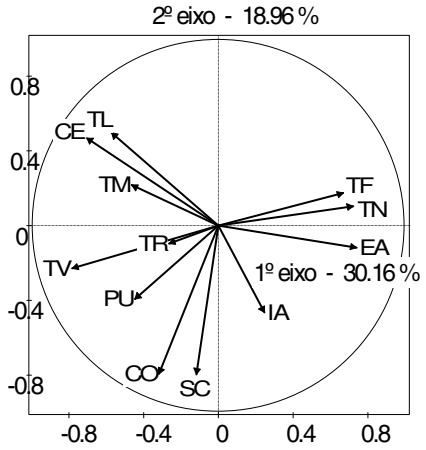


Figura 6.29 Círculo de correlações (2º critério da DACP).

coordenadas bem como as respectivas contribuições encontram-se no Anexo 5 e a sua representação na figura 6.30. Comparando esta representação com a representação obtida através do primeiro critério, verifica-se que estas são bastante semelhantes, nomeadamente para aqueles países cujas contribuições relativas são superiores a 0.50.

Comparando as trajectórias obtidas pelo método STATIS, AFM e DACP nos paises com contribuições relativas superiores a 50%, verifica-se que estas são bastante semelhantes na forma.

Tabela 6.53 Correlações entre variáveis-compromisso e as componentes principais (2º critério).

Variável	1	2	3	4
PU	-0.45	-0.39	0.23	0.50
TN	0.72	0.10	0.61	0.20
TM	-0.47	0.22	-0.05	-0.69
TF	0.67	0.17	0.63	0.24
CO	-0.32	-0.80	0.36	-0.17
TR	-0.26	-0.10	-0.44	0.45
EA	0.74	-0.12	-0.22	0.08
IA	0.25	-0.47	-0.62	0.33
TL	-0.57	0.50	0.26	0.25
TV	-0.78	-0.23	0.13	0.21
CE	-0.71	0.47	0.23	0.26
SC	-0.12	-0.80	0.45	-0.20

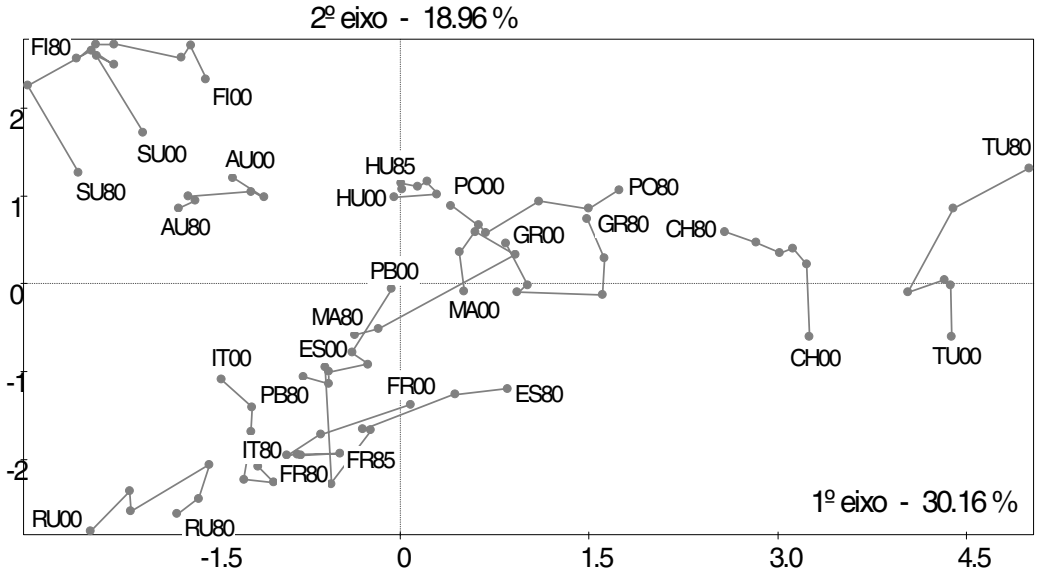


Figura 6.30 Trajectórias dos indivíduos (2º critério da DACP).

6.8 Conclusões

Os resultados obtidos através dos vários métodos são bastante próximos, o que significa que os dados possuem uma estrutura comum bastante forte.

Há três grupos de países a distinguir:

- a Turquia e o Chipre;
- a França, a Itália e o Reino Unido;

- a Finlândia e a Suécia.

O primeiro grupo apresenta valores de EA, TN e TF superiores à média destas variáveis por oposição aos restantes grupos. O segundo grupo apresenta valores elevados de CO e SC por oposição ao terceiro grupo.

Os países que revelam um comportamento mais estável ao longo do período analisado são a Itália, a Áustria, a Hungria e os Países Baixos. Por outro lado, a Turquia é o país que mais contribui para as diferenças de estrutura no período em causa.

De entre os anos analisados, aqueles que possuem uma estrutura comum mais demarcada são de 1994 e 1995. Pelo contrário, aqueles que se encontram mais afastados são os de 1980 e 2000.

Conclusão

O método STATIS, complementado com o método STATIS dual permite o tratamento da maior parte dos quadros de medidas de variáveis quantitativas sobre os indivíduos. A AFM é concebida para analisar os mesmos quadros que o método STATIS. De entre os métodos citados, somente a AFM permite o tratamento de quadros de dados qualitativos e até mesmo mistos. A DACP é mais restritiva quanto ao tipo de dados tratados. Esta exige que os quadros analisados sejam idênticos, no sentido de estes conterem as mesmas variáveis quantitativas sobre os mesmos indivíduos.

A normalização dos objectos proposta pela metodologia STATIS é tal que a norma Hilbert-Schmidt desses mesmos objectos é igual a 1. Esta normalização tem o inconveniente de fazer desaparecer a estrutura múltipla dos quadros, isto é, de atenuar as ligações entre variáveis de um mesmo quadro. A DACP não prevê nenhum tipo de normalização aos objectos em relação aos critérios abordados. Dazy e Le Barzic [9] propõem um outro critério no qual os objectos $\mathcal{V}_t$ são normalizados pelas inércias totais de cada nuvem. Esta ponderação também provoca um desaparecimento das ligações entre variáveis de um mesmo quadro. A normalização proposta pela AFM não altera a estrutura múltipla dos diferentes grupos de variáveis. Esta faz com que a inércia da primeira direcção de cada grupo seja igual a 1 e permite a majoração das inércias das restantes direcções por 1, com a vantagem de nenhum grupo influenciar de forma preponderante o primeiro eixo da imagem euclidiana compromisso.

O método STATIS e a AFM propõem uma representação das posições compromisso dos indivíduos na imagem euclidiana do compromisso, que correspondem a posições médias dos indivíduos o período analisado. A DACP não prevê este tipo de representação.

As trajectórias são obtidas a partir da projecção dos indivíduos provenientes de cada quadro sobre os eixos determinados na intra-estrutura. Estas permitem descrever a evolução do fenómeno analisado. É, no entanto, conveniente interpretar as trajectórias daqueles indivíduos cuja qualidade de representação é elevada, de forma a evitar interpretações fantasistas. Se os dados possuírem um estrutura comum bastante forte, isto é, um "bom" compromisso, as trajectórias obtidas através destes métodos serão mais ou menos semelhantes. A interpretação dos eixos da intra-estrutura do método STATIS e da AFM é feita através das correlações destes com as variáveis iniciais. Para a DACP e o STATIS dual esta interpretação é feita recorrendo às correlações dos eixos com as variáveis compromisso.

As diferenças entre os métodos destacam-se na inter-estrutura. É nesta etapa que se procura representar cada um dos quadros por um ponto, com vista a visualizar globalmente as semelhanças e diferenças entre estes. Os eixos da inter-estrutura para o método STATIS e STATIS dual não são interpretáveis. Em contrapartida a sua qualidade de representação (medida através dos coeficientes RV) é, geralmente bastante satisfatória. Na AFM, os eixos já podem ser interpretados, uma vez que coincidem com os eixos da intra-estrutura, no entanto, a sua qualidade de representação (medida através da razão entre a inércia inter e a inércia total para cada eixo) é menos satisfatória. A qualidade de representação das trajectórias da DACP é medida através da percentagem da perda de inércia do conjunto das nuvens quando projectadas no novo sistema de eixos, ou então a qualidade de representação do conjunto das nuvens é avaliada em termos de proximidade dos ângulos entre sistemas de eixos.

Os métodos analisados não têm em conta o aspecto temporal dos dados, uma vez que em nenhum momento da análise este aspecto intervém. Em contrapartida, este deve ser tomado em conta na interpretação dos resultados. Em particular, a interpretação da inter-estrutura da DACP torna-se mais delicada se os dados não forem cronológicos.

Não se pode concluir que um deste métodos seja mais eficaz que os outros. Os resultados obtidos pelo método STATIS e pela AFM são relativamente semelhantes. A DACP já se afasta um pouco mais destes e tem o inconveniente de o seu campo de aplicação ser bastante mais restrito. No entanto, é um método cuja implementação é bastante simples, baseado na ACP.

Nos últimos anos, têm sido desenvolvidos numerosos trabalhos nesta área, em particular, Cadima *et. al* [3] desenvolveram estudos em aspectos computacionais para a selecção de variáveis no contexto da Análise em Componentes Principais, nomeadamente com o coeficiente RV.

Também Vivien e Sabatier [44] realizaram em 2004 uma generalização do método STATIS, DO-ACT ("Double Analyse Conjointe de Tableaux"), que consiste em ter dois períodos a analisar, isto é, duas séries de quadros de dados. Um possível desenvolvimento futuro, no âmbito desta dissertação poderá ser o estudo aprofundado desta metodologia. Outro aspecto interessante a desenvolver seria o estudo de um "novo" produto escalar, alternativo ao de Hilbert-Schmidt e, por conseguinte, um novo coeficiente de correlação entre os objectos alternativo ao coeficiente RV.

A importância da Análise Conjunta de Quadros de Dados é justificada pela necessidade emergente de tratar séries de quadros de dados que, sem esta técnica seriam sub-analisadas.

Anexos

Anexo 1

Tabela I Dados relativos a 1980.

1980	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
AU	67.23	12.10	12.30	1.62	52398.50	20.84	4.16	6.44	0.00	390.78	4371.24	937484.00
CH	51.94	19.72	8.48	2.46	3197.57	10.49	33.55	14.78	0.00	164.07	1494.27	47599.00
ES	72.79	15.20	7.70	2.22	200004.57	3.37	17.98	12.51	0.00	254.77	2401.17	3976747.00
FI	59.83	13.10	9.20	1.63	56886.90	8.95	3.09	7.10	4.91	414.23	7778.87	449322.00
FR	73.31	14.90	10.20	1.95	482684.73	8.43	16.05	10.28	0.00	369.71	3881.05	5013666.00
GR	57.73	15.40	9.10	2.23	51737.15	4.83	25.76	8.76	0.00	171.13	2063.98	740058.00
HU	56.87	13.90	13.60	1.91	82492.40	1.10	22.24	8.30	0.00	310.14	2389.00	357334.00
IT	66.64	11.30	9.80	1.64	371947.30	11.31	7.02	13.13	0.00	389.96	2831.33	5307989.00
MA	83.13	15.40	9.10	2.05	989.28	3.42	5.33	19.36	0.00	623.46	1362.64	25501.00
PB	88.40	12.80	8.10	1.60	152962.84	22.53	19.87	14.92	0.00	399.29	4056.89	1391485.00
PO	29.44	16.20	9.70	2.19	27079.52	3.51	11.90	13.67	0.00	166.09	1468.77	398320.00
RU	88.79	13.40	11.70	1.89	580348.65	7.41	6.77	13.18	0.00	401.21	4159.99	5341849.00
SU	83.09	11.60	11.00	1.68	71354.57	6.08	2.40	7.10	0.00	460.84	10216.13	606833.00
TU	43.78	31.52	9.72	4.26	76339.44	1.72	51.07	3.50	0.00	78.24	439.33	2217909.00

Tabela II Dados relativos a 1985.

1985	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
AU	67.12	11.60	11.70	1.46	54088.70	22.52	4.28	6.37	1.29	431.33	4913.04	847188.00
CH	58.61	19.44	8.52	2.38	3098.64	12.93	32.12	15.13	0.00	248.27	1820.99	46159.00
ES	74.21	11.70	8.00	1.63	189724.85	4.07	14.89	10.64	0.04	270.09	2677.25	4555541.00
FI	59.81	12.80	9.80	1.64	47920.36	10.56	3.37	5.77	13.80	469.20	9897.80	424076.00
FR	73.65	13.90	10.00	1.82	377363.79	8.32	16.32	10.99	0.00	433.94	4520.43	5371593.00
GR	58.44	11.70	9.30	1.68	60312.74	6.35	28.17	12.96	0.00	190.86	2399.13	813534.00
HU	59.54	12.20	13.90	1.83	80868.88	1.10	21.01	6.86	0.00	401.74	2851.78	422323.00
IT	66.83	10.10	9.60	1.39	357313.28	13.56	7.45	14.29	0.11	413.11	3074.96	5361579.00
MA	85.53	14.20	7.40	1.99	1179.81	3.71	6.87	16.42	0.00	683.14	1921.51	27779.00
PB	88.53	12.30	8.50	1.51	135410.81	22.20	19.37	14.90	0.33	462.36	4241.17	1620011.00
PO	37.16	12.80	9.60	1.74	30603.56	4.65	9.42	14.82	0.00	183.09	1762.46	580248.00
RU	88.93	13.30	11.80	1.80	551357.25	7.52	6.59	11.90	0.88	432.72	4244.30	4877000.00
SU	83.10	11.80	11.30	1.74	60132.84	6.33	2.86	6.58	8.74	464.07	13607.78	624835.00
TU	52.45	30.48	8.80	3.79	112744.94	2.37	25.21	5.24	0.00	157.90	566.00	2927692.00

Tabela III Dados relativos a 1990.

1990	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
AU	67.02	11.60	10.70	1.45	57518.94	23.74	3.32	5.20	9.55	472.92	5595.78	746272.00
CH	64.98	18.30	8.40	2.42	4648.52	13.68	42.04	14.51	5.38	325.04	2629.96	44614.00
ES	75.35	10.30	8.50	1.33	211793.49	4.83	14.76	10.79	1.37	388.60	3239.24	4755322.00
FI	61.43	13.10	10.00	1.78	52897.53	10.75	2.40	4.88	51.60	494.20	11821.70	426864.00
FR	74.04	13.40	9.30	1.78	357497.95	8.00	15.84	9.78	4.99	539.06	5321.44	5521862.00
GR	58.84	10.10	9.30	1.40	72228.07	7.44	29.51	15.33	0.00	193.90	2801.99	851353.00
HU	62.03	12.10	14.00	1.84	58497.59	0.98	22.75	7.56	0.25	417.35	3048.05	514076.00
IT	66.73	10.00	9.40	1.26	398852.05	15.87	6.28	12.23	4.61	419.70	3784.04	5117897.00
MA	87.58	15.20	7.70	2.05	1659.79	3.73	2.14	10.38	0.00	323.12	2527.78	32544.00
PB	88.70	13.20	8.60	1.62	149983.28	20.73	19.95	12.61	5.28	481.57	4917.00	1401739.00
PO	46.66	11.80	10.40	1.43	42326.89	5.63	7.14	11.54	0.66	185.56	2379.14	670035.00
RU	89.08	13.90	11.10	1.83	569335.04	7.63	7.00	10.27	19.35	432.59	4767.67	4335600.00
SU	83.10	14.50	11.10	2.13	48538.11	6.01	2.16	6.12	53.69	465.62	14060.87	588474.00
TU	61.19	24.80	7.40	3.00	143819.33	2.80	22.41	8.32	0.56	229.99	800.51	3808142.00

Tabela IV Dados relativos a 1994.

1994	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
AU	66.94	11.50	10.00	1.44	57110.77	24.24	3.75	5.45	34.64	473.23	5751.89	786156.00
CH	67.79	16.40	7.80	2.23	5257.84	16.34	39.21	18.36	37.00	354.19	3322.31	57804.00
ES	76.27	9.33	8.58	1.20	216505.76	5.33	15.87	13.60	10.52	431.68	3499.09	4744829.00
FI	63.58	12.80	9.40	1.85	58052.42	10.02	3.36	6.98	132.49	509.90	12782.67	454707.00
FR	74.58	12.20	9.00	1.65	329485.20	7.26	14.78	11.17	15.31	591.41	5847.59	6003797.00
GR	59.16	10.00	9.40	1.36	76867.06	8.28	29.46	15.94	14.67	442.95	3136.68	842633.00
HU	63.04	11.30	14.30	1.64	54883.06	1.94	19.68	6.53	13.91	424.25	2688.24	1128911.00
IT	66.63	9.30	9.50	1.22	386544.67	17.94	6.74	12.73	39.11	427.81	4053.78	4825719.00
MA	89.01	13.10	7.30	1.89	2627.09	3.75	1.74	9.06	20.30	398.43	3248.66	35527.00
PB	88.94	12.70	8.70	1.57	134424.83	19.89	21.09	15.11	20.86	493.79	5285.87	1508772.00
PO	54.31	11.00	10.00	1.44	46708.67	6.73	7.09	14.31	17.50	379.64	2722.18	945077.00
RU	89.19	12.90	10.70	1.74	542697.02	8.46	7.57	10.26	67.46	602.73	4887.96	6677836.00
SU	83.15	12.60	10.30	1.88	50358.02	6.17	2.21	7.81	156.64	467.31	13947.04	791848.00
TU	63.18	22.80	6.60	2.69	155646.72	3.09	21.93	4.95	2.89	280.64	973.49	4725551.00

Tabela V Dados relativos a 1995.

1995	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
AU	66.92	11.00	10.10	1.40	58422.48	25.18	3.79	5.70	47.66	484.65	5880.72	791453.00
CH	68.51	15.40	7.70	2.13	5129.60	16.60	45.74	20.53	70.67	362.48	3036.89	59845.00
ES	76.50	9.10	8.80	1.18	233294.21	5.74	15.45	13.64	24.10	489.67	3593.75	8234104.00
FI	64.12	12.30	9.60	1.81	54729.17	9.11	2.35	5.95	200.72	511.88	12750.78	460878.00
FR	74.71	12.50	9.10	1.71	349593.23	7.16	14.16	10.76	22.52	600.24	5922.43	5980518.00
GR	59.24	9.90	9.40	1.32	77456.96	8.37	29.56	15.99	26.12	443.06	3259.42	835158.00
HU	63.30	11.00	14.20	1.57	57587.09	1.92	20.70	5.56	25.86	427.00	2712.19	2224298.00
IT	66.60	9.20	9.70	1.18	410335.02	18.42	6.55	11.57	68.42	436.05	4165.30	4708406.00
MA	89.37	12.40	7.30	1.83	2693.04	5.10	1.63	10.18	29.00	447.47	3330.69	35044.00
PB	89.00	12.30	8.80	1.53	138491.87	19.57	19.83	13.93	34.79	522.79	5373.80	1479682.00
PO	56.41	10.80	10.40	1.38	50500.91	6.97	7.19	13.65	34.36	385.96	2896.33	947478.00
RU	89.22	12.50	11.00	1.71	544466.74	8.42	7.46	10.01	97.86	643.21	5059.61	6696772.00
SU	83.16	11.70	11.00	1.73	46591.42	6.23	2.21	6.67	227.21	475.25	14105.54	809653.00
TU	63.69	22.40	6.60	2.65	170903.62	3.15	19.56	6.99	7.09	292.00	1055.01	4760892.00

Tabela VI Dados relativos a 2000.

2000	PU	TN	TM	TF	CO	TR	EA	IA	TL	TV	CE	SC
AU	67.28	9.60	9.50	1.36	60848.05	23.59	5.12	5.75	753.23	541.81	6574.76	749135.00
CH	69.94	11.20	7.10	1.91	6422.99	16.98	35.61	18.69	321.54	385.27	3957.73	64065.00
ES	77.61	9.91	9.00	1.24	282934.08	6.68	13.55	9.16	604.78	560.79	4653.31	3183282.00
FI	58.97	11.00	9.50	1.73	53428.45	8.89	1.64	5.22	720.37	691.65	14594.35	493187.00
FR	75.37	13.20	9.10	1.88	362431.89	6.85	10.96	7.87	493.31	628.26	6539.16	5876047.00
GR	60.10	11.70	10.50	1.32	89603.12	9.12	21.76	11.30	561.51	492.19	4086.27	743462.00
HU	64.55	9.60	13.30	1.29	54161.25	2.46	7.32	2.91	307.47	444.88	2937.05	1001855.00
IT	66.94	9.30	9.70	1.23	428171.38	19.85	6.10	8.61	737.30	493.91	4731.76	4473362.00
MA	90.92	10.80	7.60	1.81	2813.95	6.20	2.11	8.47	293.45	558.13	4017.95	36243.00
PB	89.49	13.00	9.10	1.72	138865.60	16.43	15.02	9.79	672.73	537.93	6152.27	1402928.00
PO	64.43	11.60	10.50	1.51	59833.12	8.49	6.84	11.26	664.95	419.43	3788.06	813172.00
RU	89.48	11.40	10.20	1.68	567843.06	8.51	5.46	7.94	727.04	891.81	5596.69	8374404.00
SU	83.28	10.20	10.50	1.55	46943.17	6.10	2.50	6.25	717.56	574.29	14514.04	928424.00
TU	65.76	21.35	6.45	2.36	221554.75	3.90	12.80	3.93	247.09	288.06	1422.03	5271056.00

Anexo 2

Tabela VII Coordenadas e contribuições da inter-estrutura não centrada (STATIS).

	Coordenadas			C. Absolutas			C. Relativas		
	1	2	3	1	2	3	1	2	3
1980	0.91	0.40	0.13	0.15	0.51	0.11	0.82	0.16	0.02
1985	0.97	0.21	-0.04	0.17	0.14	0.01	0.94	0.04	0.00
1990	0.96	-0.05	-0.24	0.17	0.01	0.37	0.92	0.00	0.06
1994	0.98	-0.14	-0.07	0.18	0.06	0.03	0.96	0.02	0.00
1995	0.98	-0.15	-0.04	0.18	0.08	0.01	0.96	0.02	0.00
2000	0.93	-0.25	0.27	0.16	0.20	0.48	0.86	0.06	0.08

Tabela VIII Coordenadas e contribuições da inter-estrutura centrada (STATIS).

	Coordenadas			C. Absolutas			C. Relativas		
	1	2	3	1	2	3	1	2	3
1980	0.40	-0.14	-0.02	0.53	0.11	0.01	0.87	0.1	0
1985	0.21	0.04	0.03	0.14	0.01	0.02	0.76	0.03	0.02
1990	-0.05	0.23	0.13	0.01	0.33	0.4	0.03	0.69	0.22
1994	-0.14	0.06	-0.10	0.06	0.02	0.22	0.48	0.09	0.23
1995	-0.15	0.03	-0.09	0.07	0.01	0.18	0.58	0.02	0.19
2000	-0.24	-0.29	0.09	0.19	0.51	0.18	0.39	0.55	0.05

Tabela IX Coordenadas e contribuições da inter-estrutura não centrada (STATIS dual).

	Coordenadas			C. Absolutas			C. Relativas		
	1	2	3	1	2	3	1	2	3
1980	5.22	2.26	0.51	0.18	0.54	0.08	0.83	0.16	0.01
1985	5.17	0.95	-0.17	0.17	0.10	0.01	0.95	0.03	0.00
1990	5.06	-0.59	-1.00	0.17	0.04	0.29	0.94	0.01	0.04
1994	4.99	-0.61	-0.34	0.16	0.04	0.03	0.96	0.01	0.00
1995	5.01	-0.72	-0.35	0.16	0.05	0.03	0.96	0.02	0.00
2000	4.82	-1.47	1.39	0.15	0.23	0.56	0.85	0.08	0.07

Tabela X Coordenadas e contribuições da inter-estrutura centrada (STATIS dual).

	Coordenadas			C. Absolutas			C. Relativas		
	1	2	3	1	2	3	1	2	3
1980	2.26	-0.52	-0.17	0.54	0.08	0.03	0.93	0.05	0.01
1985	0.95	0.17	0.42	0.10	0.01	0.16	0.67	0.02	0.13
1990	-0.58	1.00	0.53	0.04	0.29	0.26	0.19	0.58	0.16
1994	-0.61	0.34	-0.63	0.04	0.03	0.36	0.38	0.12	0.41
1995	-0.72	0.35	-0.4	0.05	0.04	0.14	0.54	0.13	0.17
2000	-1.49	-1.39	0.25	0.23	0.55	0.06	0.53	0.46	0.01

Anexo 3

Tabela XI Trajectórias do método STATIS em 1980, 1985, 1990.

	Coordenadas		
	1	2	3
AU80	0.19	0.10	0.06
CH80	-0.29	0.05	0.07
ES80	-0.06	-0.13	0.00
FI80	0.22	0.28	-0.03
FR80	0.11	-0.21	-0.07
GR80	-0.18	0.06	0.02
HU80	-0.04	0.10	-0.02
IT80	0.17	-0.20	0.04
MA80	0.07	0.01	0.10
PB80	0.14	-0.06	0.13
PO80	-0.20	0.09	0.09
RU80	0.23	-0.27	-0.07
SU80	0.28	0.15	-0.03
TU80	-0.65	0.04	-0.28
AU85	0.18	0.13	0.06
CH85	-0.33	0.04	0.11
ES85	-0.03	-0.16	0.02
FI85	0.25	0.36	-0.13
FR85	0.06	-0.22	-0.06
GR85	-0.18	0.01	0.16
HU85	-0.02	0.11	-0.02
IT85	0.13	-0.25	0.08
MA85	0.02	0.00	0.06
PB85	0.08	-0.09	0.16
PO85	-0.16	0.07	0.16
RU85	0.20	-0.28	-0.12
SU85	0.33	0.29	-0.17
TU85	-0.54	0.01	-0.32
AU90	0.21	0.13	0.09
CH90	-0.34	0.06	0.09
ES90	0.01	-0.20	0.07
FI90	0.29	0.32	-0.16
FR90	0.08	-0.22	-0.10
GR90	-0.21	-0.04	0.31
HU90	0.00	0.13	0.05
IT90	0.11	-0.27	0.12
MA90	-0.10	0.04	-0.04
PB90	0.05	-0.10	0.11
PO90	-0.13	0.08	0.21
RU90	0.17	-0.24	-0.16
SU90	0.31	0.32	-0.28
TU90	-0.42	-0.01	-0.30

Tabela XII Trajectórias do método STATIS em 1994, 1995, 2000.

	Coordenadas		
	1	2	3
AU94	0.13	0.10	0.10
CH94	-0.34	0.11	0.13
ES94	0.01	-0.20	0.11
FI94	0.22	0.29	-0.15
FR94	0.09	-0.25	-0.09
GR94	-0.11	-0.01	0.28
HU94	-0.02	0.10	0.07
IT94	0.11	-0.22	0.13
MA94	-0.07	0.10	-0.08
PB94	0.02	-0.09	0.14
PO94	-0.08	0.07	0.20
RU94	0.24	-0.34	-0.22
SU94	0.28	0.29	-0.22
TU94	-0.47	0.05	-0.40
AU95	0.13	0.10	0.12
CH95	-0.38	0.09	0.16
ES95	0.03	-0.27	0.08
FI95	0.22	0.29	-0.16
FR95	0.06	-0.23	-0.10
GR95	-0.13	0.01	0.27
HU95	0.00	0.09	0.05
IT95	0.11	-0.17	0.13
MA95	-0.06	0.06	-0.07
PB95	0.03	-0.08	0.11
PO95	-0.08	0.09	0.18
RU95	0.24	-0.32	-0.20
SU95	0.30	0.29	-0.19
TU95	-0.47	0.04	-0.37
AU00	0.15	0.09	0.19
CH00	-0.38	0.06	0.27
ES00	0.05	-0.10	0.05
FI00	0.21	0.21	-0.07
FR00	-0.01	-0.17	-0.22
GR00	-0.08	0.08	0.21
HU00	0.01	0.15	0.07
IT00	0.14	-0.15	0.13
MA00	-0.08	0.07	-0.08
PB00	0.00	-0.03	0.04
PO00	-0.03	0.10	0.14
RU00	0.26	-0.39	-0.25
SU00	0.24	0.15	-0.04
TU00	-0.47	-0.07	-0.45

Anexo 4

Tabela XIII Coordenadas da Áustria, Chipre, Espanha, Finlândia e França na AFM.

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
AU	-1.8	1.2	-1.2	4.1	2.8	3.4	0.21	0.09	0.09
80	-1.8	1.0	-0.6	0.0	0.3	0.8	0.04	0.01	0.01
85	-2.0	1.4	-0.7	0.2	0.1	0.6	0.04	0.02	0.01
90	-2.3	1.4	-1.0	1.6	0.2	0.1	0.05	0.02	0.01
94	-1.5	1.2	-1.2	0.5	0.0	0.0	0.02	0.02	0.02
95	-1.5	1.1	-1.4	0.6	0.0	0.1	0.02	0.01	0.02
00	-1.8	1.1	-2.2	0.0	0.1	2.5	0.03	0.01	0.05
CH	3.9	0.8	-1.5	18.5	1.2	5.5	0.56	0.02	0.09
80	2.7	0.5	-0.6	8.3	0.4	1.9	0.11	0.00	0.01
85	3.5	0.4	-1.2	0.7	0.8	0.3	0.12	0.00	0.01
90	3.8	0.6	-0.9	0.0	0.1	0.8	0.10	0.00	0.01
94	4.1	1.3	-1.5	0.6	1.4	0.0	0.09	0.01	0.01
95	4.5	1.1	-1.8	2.4	0.5	0.2	0.09	0.01	0.02
00	4.4	0.7	-3.0	2.1	0.0	5.3	0.08	0.00	0.04
ES	0.0	-2.0	-0.6	0.0	7.8	1.0	0.00	0.44	0.05
80	0.6	-1.2	0.0	2.7	3.0	1.0	0.01	0.04	0.00
85	0.3	-1.7	-0.2	0.9	0.4	0.4	0.00	0.06	0.00
90	-0.1	-2.3	-0.8	0.0	0.4	0.0	0.00	0.10	0.01
94	-0.1	-2.5	-1.3	0.0	1.1	1.1	0.00	0.09	0.03
95	-0.4	-3.1	-0.9	0.7	6.5	0.2	0.00	0.11	0.01
00	-0.6	-1.1	-0.6	2.1	3.7	0.0	0.01	0.04	0.01
FI	-2.6	3.2	1.3	8.5	20.5	4.0	0.29	0.44	0.07
80	-2.1	2.6	0.3	1.7	1.9	2.3	0.03	0.05	0.00
85	-2.8	3.8	1.4	0.1	1.8	0.0	0.05	0.09	0.01
90	-3.2	3.6	1.8	2.4	0.8	0.5	0.07	0.09	0.02
94	-2.6	3.5	1.8	0.0	0.4	0.5	0.05	0.09	0.02
95	-2.6	3.4	1.9	0.0	0.2	0.7	0.05	0.09	0.03
00	-2.4	2.4	0.7	0.2	3.8	0.7	0.05	0.05	0.00
FR	-0.7	-2.4	1.2	0.7	11.4	3.4	0.05	0.58	0.15
80	-1.0	-2.0	0.6	0.6	0.9	0.7	0.02	0.08	0.01
85	-0.7	-2.4	0.7	0.0	0.0	0.6	0.01	0.12	0.01
90	-0.9	-2.4	1.1	0.2	0.0	0.0	0.01	0.09	0.02
94	-1.1	-3.0	1.1	1.2	2.0	0.0	0.02	0.12	0.01
95	-0.8	-2.6	1.2	0.0	0.3	0.0	0.01	0.11	0.02
00	0.1	-1.9	2.6	5.0	1.1	4.2	0.00	0.07	0.12

Tabela XIV Coordenadas da Grécia, Hungria, Itália, Malta e Países Baixos na AFM.

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
GR	1.6	0.2	-2.4	3.3	0.1	13.6	0.23	0.00	0.49
80	1.7	0.6	-0.2	0.0	0.8	11.4	0.09	0.01	0.00
85	1.9	0.1	-1.7	0.5	0.0	1.0	0.06	0.00	0.05
90	2.3	-0.4	-3.5	3.4	2.0	2.7	0.05	0.00	0.11
94	1.4	-0.1	-3.4	0.5	0.6	2.3	0.02	0.00	0.14
95	1.5	0.1	-3.1	0.1	0.0	1.2	0.03	0.00	0.13
00	1.0	0.9	-2.5	2.8	2.4	0.0	0.02	0.01	0.10
HU	0.1	1.3	-0.4	0.0	3.2	0.5	0.00	0.08	0.01
80	0.4	1.0	0.2	0.4	0.4	0.9	0.00	0.02	0.00
85	0.2	1.2	0.1	0.1	0.0	0.8	0.00	0.02	0.00
90	0.0	1.4	-0.6	0.1	0.1	0.1	0.00	0.02	0.00
94	0.3	1.2	-0.9	0.1	0.0	0.4	0.00	0.01	0.01
95	0.0	1.1	-0.6	0.1	0.2	0.1	0.00	0.01	0.00
00	-0.1	1.7	-0.9	0.5	0.9	0.4	0.00	0.02	0.01
IT	-1.4	-2.3	-1.2	2.6	10.7	3.5	0.15	0.41	0.11
80	-1.6	-1.9	-0.4	0.2	0.9	1.6	0.05	0.07	0.00
85	-1.5	-2.7	-0.9	0.0	0.6	0.2	0.03	0.10	0.01
90	-1.2	-3.0	-1.3	0.3	2.5	0.0	0.02	0.10	0.02
94	-1.4	-2.6	-1.6	0.0	0.4	0.3	0.02	0.08	0.03
95	-1.3	-2.0	-1.6	0.1	0.5	0.3	0.02	0.05	0.03
00	-1.6	-1.7	-1.5	0.3	1.8	0.2	0.03	0.03	0.03
MA	0.5	0.5	0.3	0.3	0.6	0.2	0.01	0.02	0.01
80	-0.7	0.1	-0.9	8.8	1.1	3.4	0.00	0.00	0.01
85	-0.1	0.0	-0.6	2.4	1.4	1.9	0.00	0.00	0.00
90	1.2	0.5	0.5	3.3	0.0	0.1	0.02	0.00	0.00
94	0.8	1.2	1.1	0.9	1.9	1.3	0.01	0.01	0.01
95	0.7	0.7	0.8	0.3	0.2	0.6	0.01	0.01	0.01
00	1.0	0.8	1.0	1.7	0.3	1.1	0.01	0.01	0.01
PB	-0.5	-0.8	-1.2	0.4	1.4	3.5	0.02	0.06	0.13
80	-1.3	-0.6	-1.2	3.9	0.3	0.0	0.02	0.00	0.02
85	-0.9	-1.0	-1.7	0.7	0.1	0.5	0.01	0.01	0.03
90	-0.5	-1.1	-1.2	0.0	0.3	0.0	0.00	0.02	0.02
94	-0.2	-1.1	-1.6	0.6	0.3	0.4	0.00	0.02	0.03
95	-0.3	-1.0	-1.2	0.4	0.1	0.0	0.00	0.01	0.02
00	0.0	-0.4	-0.4	1.8	1.2	1.5	0.00	0.00	0.00

Tabela XV Coordenadas de Portugal, Reino Unido, Suécia e Turquia na AFM.

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
PO	1.2	0.9	-1.9	1.9	1.7	8.1	0.13	0.07	0.28
80	1.9	0.9	-0.8	2.8	0.0	2.3	0.05	0.01	0.01
85	1.8	0.7	-1.8	1.8	0.2	0.0	0.03	0.01	0.03
90	1.5	0.9	-2.4	0.4	0.0	0.7	0.02	0.01	0.06
94	1.0	0.8	-2.4	0.4	0.0	0.7	0.01	0.01	0.08
95	0.9	1.0	-2.1	0.7	0.1	0.1	0.01	0.02	0.07
00	0.4	1.2	-1.7	4.9	0.3	0.1	0.00	0.03	0.06
RU	-2.5	-3.5	1.9	7.8	23.4	8.8	0.25	0.49	0.15
80	-2.3	-2.6	0.7	0.4	3.9	3.6	0.05	0.07	0.00
85	-2.1	-3.1	1.3	0.9	0.8	1.0	0.04	0.08	0.01
90	-1.9	-2.7	1.8	2.6	2.9	0.0	0.03	0.07	0.03
94	-2.9	-4.1	2.7	1.1	1.8	1.2	0.05	0.09	0.04
95	-2.8	-3.8	2.3	0.6	0.4	0.3	0.05	0.09	0.03
00	-3.0	-4.6	2.9	1.9	6.4	2.0	0.04	0.10	0.04
SU	-3.2	2.8	1.8	13.1	15.3	7.3	0.42	0.31	0.12
80	-2.7	1.4	0.3	1.8	10.0	4.8	0.08	0.02	0.00
85	-3.6	3.1	1.8	0.6	0.5	0.0	0.08	0.06	0.02
90	-3.5	3.6	3.2	0.3	3.3	4.4	0.07	0.07	0.05
94	-3.4	3.5	2.6	0.1	2.6	1.6	0.06	0.07	0.04
95	-3.6	3.4	2.3	0.6	1.7	0.5	0.07	0.06	0.03
00	-2.8	1.8	0.4	1.3	5.1	4.0	0.08	0.03	0.00
TU	5.6	0.1	4.0	38.9	0.0	37.3	0.64	0.00	0.32
80	6.2	0.3	2.6	2.4	0.2	4.4	0.13	0.00	0.02
85	5.8	0.1	3.5	0.3	0.0	0.6	0.11	0.00	0.04
90	4.8	-0.1	3.3	4.6	0.2	1.0	0.10	0.00	0.05
94	5.7	0.6	4.9	0.1	1.2	1.8	0.10	0.00	0.07
95	5.5	0.5	4.4	0.0	0.6	0.3	0.10	0.00	0.06
00	5.5	-0.8	5.3	0.0	4.0	3.9	0.09	0.00	0.09

Anexo 5

Tabela XVI Trajectórias da DACP (1º critério) em 1980, 1985, 1990.

	Coordenadas		
	1	2	3
AU80	-1.55	0.21	1.37
CH80	2.66	0.03	0.92
ES80	0.79	-0.62	-1.17
FI80	-2.40	2.80	0.85
FR80	-0.98	-1.24	-1.70
GR80	1.56	0.33	0.49
HU80	-0.02	0.09	1.17
IT80	-1.26	-1.80	-0.85
MA80	-0.40	-0.59	0.41
PB80	-0.60	-1.03	0.54
PO80	1.76	0.11	1.35
RU80	-2.03	-1.67	-1.91
SU80	-2.39	1.40	0.57
TU80	4.85	1.97	-2.05
AU85	-1.40	0.36	1.30
CH85	2.89	-0.07	0.85
ES85	0.40	-0.93	-0.92
FI85	-2.25	2.89	0.81
FR85	-0.68	-1.30	-1.62
GR85	1.68	-0.65	1.29
HU85	-0.05	0.15	1.05
IT85	-1.13	-2.09	-0.67
MA85	-0.17	-0.22	-0.06
PB85	-0.43	-1.15	0.46
PO85	1.55	-0.28	1.71
RU85	-1.85	-1.43	-2.06
SU85	-2.79	2.75	0.42
TU85	4.25	1.96	-2.56
AU90	-1.45	0.43	1.23
CH90	3.03	-0.05	0.67
ES90	-0.37	-1.59	-0.58
FI90	-2.13	2.89	0.66
FR90	-1.05	-1.22	-1.78
GR90	1.67	-1.48	2.04
HU90	0.06	0.00	1.17
IT90	-1.35	-2.24	-0.41
MA90	1.07	0.79	-0.27
PB90	-0.46	-0.88	0.21
PO90	1.20	-0.40	1.97
RU90	-1.76	-0.89	-2.17
SU90	-2.33	3.55	-0.14
TU90	3.88	1.10	-2.62

Tabela XVII Trajectórias da DACP (1º critério) em 1994, 1995 e 2000.

	Coordenadas		
	1	2	3
AU94	-0.94	0.48	1.24
CH94	3.18	0.01	0.89
ES94	-0.30	-1.87	-0.13
FI94	-1.62	2.85	0.56
FR94	-1.14	-1.31	-1.65
GR94	0.92	-1.44	1.82
HU94	0.17	0.00	1.22
IT94	-1.24	-1.87	0.01
MA94	0.78	1.09	-0.30
PB94	-0.16	-0.87	0.30
PO94	0.74	-0.56	1.72
RU94	-2.48	-1.13	-2.84
SU94	-2.13	3.29	0.10
TU94	4.22	1.33	-2.94
AU95	-0.82	0.37	1.34
CH95	3.28	-0.26	1.07
ES95	-0.69	-2.21	-0.78
FI95	-1.53	2.99	0.55
FR95	-0.86	-1.02	-1.67
GR95	1.02	-1.34	1.83
HU95	0.24	-0.06	0.99
IT95	-1.21	-1.64	0.11
MA95	0.64	0.83	-0.23
PB95	-0.28	-0.69	0.23
PO95	0.69	-0.46	1.71
RU95	-2.47	-1.00	-2.64
SU95	-2.26	3.22	0.31
TU95	4.26	1.28	-2.81
AU00	-1.08	0.59	1.59
CH00	3.32	-1.51	1.58
ES00	-0.66	-1.05	-0.11
FI00	-1.46	2.28	0.78
FR00	-0.14	-0.41	-2.05
GR00	0.84	-0.76	1.65
HU00	0.00	-0.36	1.43
IT00	-1.44	-1.24	0.13
MA00	0.66	0.26	-0.19
PB00	0.02	0.40	-0.18
PO00	0.46	0.18	1.32
RU00	-2.85	-1.17	-3.17
SU00	-1.90	1.99	0.52
TU00	4.25	0.81	-3.28

Tabela XVIII Coordenadas e contribuições das trajectórias da DACP (2º critério).

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
AU80	-1.76	0.85	-0.93	1.0	0.4	0.5	0.30	0.07	0.09
CH80	2.58	0.59	-1.04	2.2	0.2	0.7	0.74	0.04	0.12
ES80	0.86	-1.19	0.45	0.2	0.7	0.1	0.16	0.31	0.04
FI80	-2.57	2.57	0.59	2.2	3.4	0.2	0.33	0.33	0.02
FR80	-0.78	-1.95	0.94	0.2	2.0	0.6	0.10	0.61	0.14
GR80	1.48	0.74	-0.28	0.7	0.3	0.0	0.55	0.14	0.02
HU80	0.02	1.08	-0.36	0.0	0.6	0.1	0.00	0.14	0.02
IT80	-1.12	-2.08	-0.16	0.4	2.3	0.0	0.18	0.63	0.00
MA80	-0.35	-0.58	-0.93	0.0	0.2	0.5	0.01	0.02	0.06
PB80	-0.77	-1.06	-1.58	0.2	0.6	1.6	0.06	0.10	0.23
PO80	1.74	1.07	-0.96	1.0	0.6	0.6	0.34	0.13	0.10
RU80	-1.77	-2.62	0.97	1.0	3.6	0.6	0.24	0.52	0.07
SU80	-2.55	1.27	0.34	2.1	0.8	0.1	0.50	0.12	0.01
TU80	5.00	1.32	2.95	8.2	0.9	5.4	0.64	0.04	0.22
AU85	-1.63	0.94	-0.86	0.9	0.5	0.5	0.27	0.09	0.08
CH85	2.83	0.47	-1.05	2.6	0.1	0.7	0.70	0.02	0.10
ES85	0.44	-1.26	0.19	0.1	0.8	0.0	0.04	0.32	0.01
FI85	-2.42	2.72	0.79	1.9	3.9	0.4	0.34	0.43	0.04
FR85	-0.48	-1.93	0.85	0.1	2.0	0.4	0.04	0.69	0.13
GR85	1.63	0.29	-1.45	0.9	0.0	1.3	0.41	0.01	0.33
HU85	0.01	1.15	-0.15	0.0	0.7	0.0	0.00	0.13	0.00
IT85	-1.00	-2.26	-0.50	0.3	2.7	0.2	0.13	0.64	0.03
MA85	-0.17	-0.51	-0.48	0.0	0.1	0.1	0.00	0.02	0.02
PB85	-0.56	-1.14	-1.49	0.1	0.7	1.4	0.03	0.14	0.24
PO85	1.50	0.86	-1.49	0.7	0.4	1.4	0.23	0.07	0.22
RU85	-1.60	-2.44	1.24	0.8	3.1	1.0	0.20	0.47	0.12
SU85	-2.96	2.25	1.13	2.9	2.7	0.8	0.52	0.30	0.08
TU85	4.40	0.86	3.26	6.4	0.4	6.6	0.61	0.02	0.34
AU90	-1.68	1.00	-0.80	0.9	0.5	0.4	0.26	0.09	0.06
CH90	3.02	0.35	-0.92	3.0	0.1	0.5	0.61	0.01	0.06
ES90	-0.30	-1.65	-0.35	0.0	1.4	0.1	0.02	0.49	0.02
FI90	-2.26	2.73	1.03	1.7	3.9	0.7	0.33	0.48	0.07
FR90	-0.81	-1.94	1.01	0.2	2.0	0.6	0.10	0.56	0.15
GR90	1.61	-0.12	-2.56	0.8	0.0	4.1	0.22	0.00	0.56
HU90	0.14	1.10	-0.29	0.0	0.6	0.1	0.00	0.10	0.01
IT90	-1.24	-2.23	-0.78	0.5	2.6	0.4	0.16	0.54	0.07
MA90	0.92	0.34	0.22	0.3	0.1	0.0	0.10	0.01	0.01
PB90	-0.56	-1.00	-1.09	0.1	0.5	0.7	0.04	0.13	0.15
PO90	1.10	0.94	-1.66	0.4	0.5	1.7	0.11	0.08	0.26
RU90	-1.51	-2.06	1.53	0.8	2.2	1.5	0.20	0.37	0.21
SU90	-2.45	2.65	1.99	2.0	3.7	2.5	0.31	0.37	0.21
TU90	4.03	-0.09	2.67	5.3	0.0	4.5	0.67	0.00	0.30

Tabela XIX Coordenadas e contribuições das trajectórias da DACP (2º critério).

	Coordenadas			C. Absolutas (%)			C. Relativas		
	1	2	3	1	2	3	1	2	3
AU94	-1.18	1.04	-0.84	0.5	0.6	0.4	0.15	0.12	0.08
CH94	3.12	0.40	-1.19	3.2	0.1	0.9	0.57	0.01	0.08
ES94	-0.24	-1.66	-0.85	0.0	1.4	0.5	0.01	0.46	0.12
FI94	-1.73	2.58	1.03	1.0	3.5	0.7	0.24	0.52	0.08
FR94	-0.90	-1.95	0.88	0.3	2.0	0.5	0.11	0.51	0.10
GR94	0.93	-0.10	-2.20	0.3	0.0	3.0	0.11	0.00	0.63
HU94	0.22	1.17	-0.30	0.0	0.7	0.1	0.00	0.10	0.01
IT94	-1.18	-1.68	-0.93	0.5	1.5	0.5	0.17	0.33	0.10
MA94	0.60	0.59	0.43	0.1	0.2	0.1	0.04	0.04	0.02
PB94	-0.25	-0.92	-1.12	0.0	0.4	0.8	0.01	0.12	0.18
PO94	0.69	0.58	-1.59	0.2	0.2	1.6	0.07	0.05	0.37
RU94	-2.14	-2.59	2.04	1.5	3.5	2.6	0.27	0.40	0.25
SU94	-2.27	2.49	1.56	1.7	3.2	1.5	0.30	0.37	0.14
TU94	4.32	0.05	3.08	6.2	0.0	5.9	0.63	0.00	0.32
AU95	-1.08	0.99	-1.01	0.4	0.5	0.6	0.13	0.11	0.11
CH95	3.23	0.23	-1.53	3.4	0.0	1.5	0.51	0.00	0.12
ES95	-0.54	-2.28	-0.42	0.1	2.7	0.1	0.03	0.58	0.02
FI95	-1.66	2.71	1.13	0.9	3.8	0.8	0.21	0.56	0.10
FR95	-0.63	-1.71	1.03	0.1	1.5	0.7	0.06	0.48	0.18
GR95	1.01	-0.01	-2.17	0.3	0.0	2.9	0.14	0.00	0.64
HU95	0.29	1.02	-0.14	0.0	0.5	0.0	0.01	0.08	0.00
IT95	-1.17	-1.40	-0.88	0.5	1.0	0.5	0.17	0.25	0.10
MA95	0.47	0.36	0.21	0.1	0.1	0.0	0.03	0.02	0.01
PB95	-0.38	-0.78	-0.94	0.0	0.3	0.6	0.02	0.10	0.14
PO95	0.63	0.67	-1.51	0.1	0.2	1.4	0.06	0.07	0.36
RU95	-2.15	-2.36	1.94	1.5	2.9	2.4	0.29	0.35	0.24
SU95	-2.41	2.59	1.40	1.9	3.5	1.2	0.34	0.39	0.11
TU95	4.37	-0.01	2.89	6.3	0.0	5.2	0.66	0.00	0.29
AU00	-1.32	1.20	-1.20	0.6	0.8	0.9	0.19	0.16	0.16
CH00	3.25	-0.60	-2.66	3.5	0.2	4.4	0.43	0.01	0.29
ES00	-0.59	-0.95	-0.43	0.1	0.5	0.1	0.11	0.27	0.05
FI00	-1.54	2.32	0.69	0.8	2.8	0.3	0.20	0.45	0.04
FR00	0.09	-1.38	1.65	0.0	1.0	1.7	0.00	0.36	0.52
GR00	0.84	0.47	-1.59	0.2	0.1	1.6	0.12	0.04	0.42
HU00	-0.05	0.99	-0.69	0.0	0.5	0.3	0.00	0.06	0.03
IT00	-1.41	-1.08	-0.72	0.7	0.6	0.3	0.21	0.12	0.05
MA00	0.51	-0.09	-0.09	0.1	0.0	0.0	0.03	0.00	0.00
PB00	-0.06	-0.05	-0.01	0.0	0.0	0.0	0.00	0.00	0.00
PO00	0.41	0.89	-0.95	0.1	0.4	0.6	0.04	0.19	0.21
RU00	-2.46	-2.82	2.28	2.0	4.2	3.2	0.28	0.37	0.24
SU00	-2.04	1.72	0.62	1.4	1.5	0.2	0.41	0.29	0.04
TU00	4.38	-0.60	3.08	6.3	0.2	5.9	0.60	0.01	0.30

Bibliografia

[1] Benzécri, J. P. (1976); *L'Analyse des Données*, Dunod.

[2] Bouroche, J.-M. (1975); *Analyse des données ternaires: la Double Analyse en Composantes Principales*,Thèse de $3^{\text{ème}}$ cycle. Université de Paris VI.

[3] Cadima, J.; Cerdeira, J.; Minhoto, M. (2004); *Computational aspects of algorithms for variable selection in the context of principal components*, Computational Statistics & Data Analysis, **47**, 225-236.

[4] Cailliez, F.; Pagès J. (1976); *Introduction à l'analyse des données*, Smash.

[5] Carroll, J. D. (1968); *A generalization of canonical correlation analysis to three or more sets of variables*, Proceedings of the 76th Anual Convention of the American Psychological Association, 227-228.

[6] Cattell, R. B. (1966); *The scree test for the number of factors*, Multivariate Behav. Res. **1**, 245-276.

[7] Centre International de Statistique et d' Informatique Appliquées (2001); *SPAD (version 5), Analyse de Tableaux Multiples, Manuel de Reference*, Copyright CISIA.

[8] Cox, T.; Cox, M. (2001); *Multidimensional Scaling*, Chapman & Hall/CRC.

[9] Dazy, F.; Le Barzic, J. (1996); *L'Analyse des Données Évolutives, methodes et applications*, Éditions Technip.

[10] Escofier, B.; Pagès, J. (1985); *Mise en oeuvre de l'AFM pour des tableaux numériques, qualitatifs ou mixtes*, Publication interne de l'IRISA, **429**.

[11] Escofier, B.; Pagès, J. (1994), *Multiple Factor Analysis (AFMULT package)*, Computational Statistics & Data Analysis, **18**, 121-140.

[12] Escofier, B.; Pagès, J. (1998), *Analyses factorielles simples et multiples: objectifs, méthodes et interprétation*, 3ª ed., Dunod, Paris.

[13] Escoufier, Y. (1973); *Le traitement de variables vectorielles*, Biometrics, vol. **29**, nº **4**, 751-760.

[14] Escoufier, Y. (1985); *Objectifs et Procedures de l'Analyse Conjointe de Plusieurs Tableaux de Données*, Statistiques et Analyse de Données, vol. **10**, nº **1**, 1-10.

[15] Figueiredo, A. M. (2003); *Aplicação do método STATIS a dados económicos*, Actas do XI Congresso da Sociedade Portuguesa de Estatística, Faro, 235-247.

[16] Foucart, T. (1981); *Suites de Tableaux et de Sous-Tableaux, Revue de Statistique Appliquée*, $XXIX$ **2**, 31-42.

[17] Glaçon, F. (1981); *Analyse Conjointe de Plusieurs Matrices de Données. Comparaison de Différentes Méthodes*, Thèse de 3ème cycle. Grénoble.

[18] Golub, G. H.; Van Loan, C. F. (1996); *Matrix Computations*, 3ª ed., The Johns Hopkins University Press.

[19] Gomes, P. (s. d.); *Análise em Componentes Principais: uma abordagem geométrica*, Laboratório de Análise de Dados, Faculdade de Economia, Universidade do Porto.

[20] Harman, H. H. (1967) *Modern Factor Analysis*, 2ª ed., University of Chicago Press, Chicago.

[21] Hotelling, H. (1933) *Analysis of a Complex of Statistical Variables in to Principal Components*, Journal of Educational Psychology, **24**, 417-441, 498-520.

[22] Jaffrenou, P. A. (1978); *Sur l'analyse des familles finies de variables vectorielles*, Thèse de 3ème cycle. Lyon I.

[23] Johnson, R. A.; Wichern, D. W. (1992); *Applied Multivariate Statistical Analysis*, 3ª ed., Prentice Hall International Editions.

[24] Kaiser, H. F. (1958); *The varimax criterion for analytic rotation in factor analysis*, Psychometrika, **23**, 187-200.

[25] Kiers, H. (1989); *Three-Way Methods for the Analysis of Qualitative and Quantitative Two-Way Data*, DSWO Press.

[26] Kroonenberg, P. (1989); *Three-Mode Principal Component Analysis*, DSWO Press.

[27] Lavit, C. (1988); *Analyse Conjointe de Tableaux Quantitatives*, Collection Méthodes+Programmes, Masson.

[28] Lavit, C.; Escoufier, Y.; Traissac, P. (1994); *The ACT (STATIS method)*, Computational Statistics & Data Analysis, **23**, 97-119.

[29] Lebart, L.; Morineau, A.; Piron, M. (2000); *Statistique Exploratoire Multidimensionnelle*, 3ª ed., Dunod, Paris.

[30] L'Hermier des Plantes, H. (1976); *Structuration des Tableaux à Trois Indices de la Statistique*, Thèse de 3ème cycle. Université de Montpellier II.

[31] Mardia, K. V.; Kent, J. T.; Bibby, J. M. (1979); *Multivariate Analysis*, Academic Press.

[32] Morrison, D. F. (1967); *Multivariate Statistical Methods*, 2ª ed., McGraw-Hill.

[33] Oliveira, M. M. (s. d.); *Análise dos resultados eleitorais e análise da evolução do grau de desfoliação dos sobreiros, ACT-Método STATIS*, Dissertação de Mestrado, Faculdade de Ciências da Universidade de Lisboa.

[34] Oliveira, M. M., Leal, M. M.; Nicolau, F. (1995); *Análise Evolutiva das Eleições Legislativas Autárquicas de 1975 a 1993*, Actas do III Congresso Anual da Sociedade Portuguesa de Estatística, Guimarães, 629-642.

[35] Pagès, J. (1996); *Eléments de Comparaison entre l' Analyse Factorielle Multiple et la Méthode STATIS*, Revue Statistique Apliqueé, $XLIV$, **4**, 81-95.

[36] Pearson, K. (1901); *On lines and planes of closest fit to systems of points in space*, Philosophical Magazine, vol. **2**, **6**, 559-572.

[37] Reis, E. (2001); *Estatística Multivariada Aplicada*, 2ª ed., Edições Sílabo.

[38] Robert, P.; Escoufier, Y. (1976); *A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient*, Applied Statistics, **25**, 257-265.

[39] Saporta, G. (1978); *Théories et Méthodes de la Statistique*, Éditions Technip.

[40] Saporta, G. (1990); *Probabilités, Analyse des Données et Statistique*, Éditions Technip.

[41] Simier, M.; Blanc, L.; Pellegrin, F.; Nandris, D. (1999); *Approche Simultanée de k Couples de Tableaux: Application à l' étude des relations pathologie végétale-environnement*, Revue Statistique Appliquée, $XLVII$, **1**, 31-46.

[42] Sousa, F. (1992); *Estabilidade nos Resultados de uma Análise em Componentes Principais*, Provas de Aptidão Pedagógica e Aptidão Científica, Faculdade de Engenharia, Universidade do Porto.

[43] Spearman, C. (1904); *General intelligence objectively determined and measured*, American Journal of Psychology, vol. **15**, 201-293.

[44] Vivien, M.; Sabatier, R. (2004); *A generalization of STATIS-ACT strategy: DO-ACT for two multiblocks tables*, Computational Statistics & Data Analysis, **46**, 155-171.

[45] Voisard, J.; Lavallard, F. (1995); *Le reflexe institutionnel des Français à l'épreuve des présidentielles*, GÉRI-La documentation Française.

[46] World Bank (2004); *World Development Indicators 2004* CD-ROM.