Notas em Matemática Aplicada e-ISSN 2236-5915

Volume 9, 2012

Editado por

Eliana X.L. de Andrade
Universidade Estadual Paulista - UNESP
São José do Rio Preto, SP, Brasil

Rubens Sampaio
Pontifícia Universidade Católica do Rio de Janeiro -
Rio de Janeiro, RJ, Brasil

Geraldo N. Silva
Universidade Estadual Paulista - UNESP
São José do Rio Preto, SP, Brasil

Sociedade Brasileira de Matemática Aplicada e Computacional

2012

A Sociedade Brasileira de Matemática Aplicada e Computacional - SBMAC publica, desde as primeiras edições do evento, monografias dos cursos que são ministrados nos CNMAC.

Para a comemoração dos 25 anos da SBMAC, que ocorreu durante o XXVI CNMAC em 2003, foi criada a série **Notas em Matemática Aplicada** para publicar as monografias dos minicursos ministrados nos CNMAC, o que permaneceu até o XXXIII CNMAC em 2010.

A partir de 2011, a série passa a publicar, também, livros nas áreas de interesse da SBMAC. Os autores que submeterem textos à série Notas em Matemática Aplicada devem estar cientes de que poderão ser convidados a ministrarem minicursos nos eventos patrocinados pela SBMAC, em especial nos CNMAC, sobre assunto a que se refere o texto.

O livro deve ser preparado em **Latex (compatível com o Miktex versão 2.7), as figuras em eps** e deve ter entre **80 e 150 páginas**. O texto deve ser redigido de forma clara, acompanhado de uma excelente revisão bibliográfica e de **exercícios de verificação de aprendizagem** ao final de cada capítulo.

Veja todos os títulos publicados nesta série na página
http://www.sbmac.org.br/notas.php

SBMAC **Sociedade Brasileira de Matemática Aplicada e Computacional**

2012

APLICAÇÕES DE ANÁLISE FATORIAL DE CORRESPONDÊNCIAS PARA ANÁLISE DE DADOS

Dr. Homero Chaib Filho - EMBRAPA
homero@cpac.embrapa.br

Sociedade Brasileira de Matemática Aplicada e Computacional

São Carlos - SP, Brasil
2012

Coordenação Editorial: Véra Lucia da Rocha Lopes

Coordenação Editorial da Série: Geraldo Nunes Silva

Editora: SBMAC

Capa: Matheus Botossi Trindade

Patrocínio: SBMAC

Catalogação elaborada pela Biblioteca do IBILCE/UNESP
Bibliotecária: Maria Luiza Fernandes Jardim Froner

Chaib Filho, Homero.
Aplicações de Análise Fatorial de Correspondências para Análise de Dados.- São Carlos, SP: SBMAC, 2012, 60 p., 20.5 cm - (Notas em Matemática Aplicada; v. 9)

e-ISBN 978-85-86883-76-7

1. Dados Multidimensionais. 2. Análise de dados.
3. Análise Fatorial. 4. Análise de Dados Qualitativos.
1. Chaib Filho, Homero. II. Título. III. Série.

CDD - 51

Esta é uma republicação em formato de e-book do livro original do mesmo título publicado em 2004 nesta mesma série pela SBMAC.

Prefácio

Há alguns anos traduzi parte do livro *Técnicas de análisis de datos multidimensionales*, do Profº. Lucinio Judez da Universidade Politécnica de Madri, para dar um treinamento sobre essas técnicas na Embrapa Cerrados. Pensava para este mini curso fazer uma tradução integral daquele texto. Porém, pareceu mais conveniente apresentar o tratamento teórico do Prfº Judez apenas no tópico de Análise Fatorial de Correspondências que é o foco desse mini curso. Para os demais tópicos traduzi textos do livro *Analyses factorialles simples et* multiples de Brigitte Escofier e Jérôme Pagès. A Profª Escofier foi precursora nas técnicas de análise de dados multidimensionais e oferece um texto que aborda todos os aspectos dessas técnicas sem o rigor matemático apresntados pelo Profº. Judez, o que considerei mais adequado para esse mini curso. Textos de outros autores também foram traduzidos e aproveitados em diversos locais na composição do texto final. Em nossa página da Internet eles aparecerão com maior presença.

A motivação para compormos esse texto e apresentarmos esse mini curso vê do fato de que nos dias de hoje, é bastante comum afirmações do tipo: "os dados são o que há de mais importante". Porém, é bem sabido que, de um mesmo conjunto de dados podem ser extraídas as informações mais diversas, dependendo da escolha e do emprego de uma metodologia de análise.

Em 1971, ao escrever o prefácio do livro *Introduction à l'Analyse des Données*, de F. Cailliez e J.-P. Pages, G. Morlat afirmava que podia ser identificada, por aquela época, uma cisão dos estatísticos em duas categorias: uma que praticava uma estatística clássica, que pretende formalizar a indução, e segue a escola anglo-saxônica; e outra que passou a se apoiar numa visão puramente algébrica e visava a descrever, reduzir e classificar as observações multidimensionais: esta categoria segue a escola francesa da *analyse des données* (chamada análise de dados multidimensionais).

O texto aqui apresentado foi montado na intenção de tornar familiar, aos pesquisadores e estudantes, a escola francesa de analyse des données, represntada pelos chamados: métodos fatoriais de análise de dados multidimensionais. Tais métodos são utilizados para que se obtenha a síntese de vastos conjuntos numéricos, com uma visão geométrica, permitindo aos analistas observar dos próprios dados as evidencias das relações existentes em tais conjuntos. Para isso, são abordados aspectos conceituais e teóricos, extraídos de autores assumiram que a essa escola como uma alternativa ao estudo dos dados. Buscamos no texto, dar uma idéia dos diversos métodos de análise de dados multidimensionais, dissertando sobre a Análise

em Componente Principais (o método fundamental) e realizando uma abordagem teórica, um pouco mais detalhada, da Análise Fatorial de Correspondências, método que é considerado o mais útil para as mais diversas aplicações e particularmente nos assuntos relativos à produção agrícola. Também, é feita uma rápida abordagem da Análise Fatorial de correspondências Múltiplas, um método derivado da Análise Fatorial de Correspondências, para o tratamento de dados qualitativos. Por fim, são listadas algumas aplicações realizadas dentro da Embrapa Cerrados (unidade de pesquisa ecorregional da Embrapa, para os Cerrados).

Assim, este texto busca contribuir para a expansão da capacidade de análise de dados dos pesquisadores brasileiros e (é bastante pretensioso) conduzir os estudantes a uma nova seara dentro da estatística matemática. Exemplos de aplicação desses métodos encontrar-se-ão na página do mini curso, em desenvolvimento.

Agradecimentos

Gostaria de agradecer ao Prof. Lucínio Judez, que, desde Madri, enviou seu estímulo para realização deste mini curso. Ainda que o material aqui apresentado tenha sido retirado de diversos livros, dois foram os básicos, de onde traduzimos o conteúdo presente: o do Prof Judez, Técnicas de *análisis de datos multidimensionales*, de onde retiramos as linhas gerais que norteou nossa adaptação; e o *Analyses factorielles simples et* multiples, de *Brigitte Escofier e Jérome Pagès*, cuja notação foi adaptada àquela do Prof Judez.

Também agradeço à Embrapa Cerrados e aos colegas de trabalho, por permitirem a divulgação das aplicações realizadas.

Agradecemos à SBMAC a oportunidade de difundirmos nosso enfoque na análise de dados. Agradecemos especialmente à Rosana Martins de Castro Chaib, minha esposa e companheira, pela ajuda na tradução e o carinho incentivador.

Conteúdo

Capítulo 1

Sobre Dados Multidimensionais

1.1 Introdução.

Os analistas (matemáticos ou estatísticos) que executam a tarefa de interpretar os dados obtidos por outros pesquisadores, devido ao predomínio de um enfoque na condução das análises de dados, podem se deparar com situações como a seguinte: depois de realizar as análises e alcançar um resultado, alguém lhe pergunta "qual o nível de confiança dos resultados...".

Existem questionamentos feitos na condução de uma análise de dados que estão estritamente ligados ao enfoque adotado. A pergunta sobre o nível de confiança está dentro do marco das análises baseadas em inferências.

As técnicas de análise de dados multidimensionais, ligadas à escola francesa de *analyses des données* são uma alternativa para a aplicação de um enfoque diferente, por estarem baseadas em desenvolvimentos essencialmente algébricos que, além de oferecer uma interpretação geométrica dos resultados, permitem aos próprios dados evidenciarem as relações existentes dentro do conjunto analisado. Ademais, como os objetivos estão relacionados à redução, classificação e agrupamento das variáveis (ou indivíduos), não são necessários mais que elementos da estatística descritiva, para que sejam obtidas tipologias a partir de uma dada massa de dados.

Assim, para a pergunta: "mas qual é o grau de confiança dessa tipologia?", não existe uma resposta dentro do campo da análise de dados multidimensionais.

As técnicas podem ser resumidas nas seguintes: Análise em Componentes Principais, Análise Fatorial de Correspondências, Análise Fatorial de Correspondências Múltiplas, Análise Fatorial Discriminante e Técnicas de Classificação. Neste capítulo serão abordados, primeiramente, aspectos relacionados à organização dos dados, aos quais se aplicarão as técnicas. Em seguida, no capítulo 2, será dada uma noção do desenvolvimento teórico da Análise em Componentes Principais, sendo concluído o capítulo com uma rápida menção sobre a importância da Análise Discriminante e das técnicas de classificação, sem, contudo, qualquer aprofundamento teórico.

No capítulo 3 é feita uma abordagem um pouco mais detalhada dos aspectos teóricos da Análise Fatorial de Correspondências, enquanto no capítulo 4 a Análise Fatorial de Correspondências Múltiplas é levada em consideração.

1.2 Generalidades sobre dados multidimensionais.

São chamados dados multidimensionais, àqueles que compõem o conjunto de valores de um certo número de variáveis estatísticas, observadas sobre um indivíduo de uma população dada. Podemos, então, considerá-los como a realização de um vetor aleatório, definido sobre a população, com valores em um espaço a definir.

Consideremos, por exemplo, o peso, a altura, a idade e o sexo de uma pessoa que faz parte de uma população qualquer. Supomos que essa pessoa pese 60 kg, meça 1,65 m, tenha 26 anos e seja do sexo feminino. A realização do vetor aleatório (peso, altura, idade,

sexo) definido sobre um indivíduo do grupo estudado é o dado multidimensional (60 kg, 1,65 m, 26 anos, feminino).

Uma tabela multidimensional é, então, uma amostra de um vetor aleatório: as mesmas variáveis são medidas sobre um certo número de indivíduos, vindo, de outro lado, elas mesmas constituírem-se em vetores aleatórios também. A maioria das tabelas tem a seguinte forma:

	$\mathbf{x}_1$	$\mathbf{x}_2$	... $\mathbf{x}_j$...	$\mathbf{x}_p$
1			⋮	
2			⋮	
⋮			⋮	
i	...	...	... x_{ij}	
⋮				
n				

Figura 1.1. O termo da i-ésima linha e j-ésima coluna é o valor observado para a variável $\mathbf{x}_j$ sobre o indivíduo i.

A seguir são apresentados diferentes tipos de tabelas de dados encontradas na prática, e para o que se segue é definido $\boldsymbol{I}=\{1, 2,...,i,...,n\}$ e $\boldsymbol{J}=\{1, 2,...,j,...,p\}$.

1.2.1 Tabelas quantitativas (ou de *indivíduos x características*).

Esse é o tipo mais simples das tabelas, tem o formato como o da figura 1.1, acima: os componentes de vetor aleatório ($\mathbf{x}_1$, $\mathbf{x}_2$,... $\mathbf{x}_j$..., $\mathbf{x}_p$) são variáveis quantitativas, de valores reais, não necessariamente contínuos. O termo x_{ij} é, então, um número real que representa a medida da variável (característica) $\mathbf{x}_j$ sobre o indivíduo i.

1.2.2 Tabelas de Contingência.

Numa tabela de contingência fornece-se a repartição de uma população estatística por dois caracteres qualitativos expressos, cada um, por modalidades exclusivas (um indivíduo da população não pode possuir mais de uma modalidade) e exaustivas (possuirá uma e somente uma modalidade). Pelo fato de que linhas e colunas representam características, em uma tabela de contingência, os papeis que desempenham são similares. Nesse tipo de tabela, os indivíduos da população não aparecem diretamente; entretanto, certos autores se referem às linhas como indivíduos e às colunas como variáveis. No seu devido tempo estabeleceremos nossa notação.

Os dados socioeconômicos são freqüentemente apresentados na forma de tabelas de contingência; em efeito, as variáveis apresentadas são freqüentemente qualitativas: categoria sócio-profissional do chefe de família; setor de atividade de uma empresa, etc. Por outro lado, pode ocorrer, fora desse campo, situações em que será útil representar em uma tabela de contingência, dados que representam medidas físicas.

O caso dos dados socioeconômicos pode ser ilustrado nas tabelas do IBGE. Já o segundo caso é ilustrado na página seguinte, numa tabela onde são encontradas 25 variedades de manga, como indivíduos (ou observações) e, nas colunas, tipos de patologias encontradas, em diversos graus de incidência, como variáveis.

Tabela de contingência onde as variáveis (características) são: an - infestação de antracnose; **oi** - infestação de oídio; **ma** - infestação de mancha de lágrima; **mo** - ocorrência ou não de mosca do fruto; **po1** - ocorrência ou não de podridão penducular; **am** - ocorrência ou não de amolecimento do fruto; **rh** - ocorrência ou não de Rhizopus;. As linhas correspondem às variedades de manga (os indivíduos). Cada célula é a quantidade de frutos de uma dada variedade infectada por um determinado patógeno, num nível de intensidade (os números indicam este nível).

	an1	an2	an6	oi1	oi2	oi3	oi4	oi5	oi6	ma1	Ma2	ma3	ma4	ma5	ma6	mo1	mo2	po1	po2	rh1	rh2	an3	an4	an5	am1	am2
v1	2	7	0	0	2	5	12	1	0	13	5	2	0	0	0	19	1	14	6	20	0	4	7	0	15	5
v2	9	2	4	0	0	0	4	10	6	4	7	5	2	2	0	8	12	11	9	16	4	1	1	3	20	0
v3	2	2	2	1	4	6	5	4	0	13	6	1	0	0	0	15	5	7	13	20	0	7	4	3	20	0
v4	14	1	2	0	0	3	11	6	0	6	9	2	3	0	0	17	3	15	5	20	0	2	1	0	20	0
v5	5	6	2	0	2	3	6	7	2	2	8	6	4	0	0	14	6	12	8	19	1	2	3	2	20	0
v7	0	1	18	0	0	0	0	3	17	20	0	0	0	0	0	16	4	5	15	17	3	1	0	0	19	1
v8	2	6	2	0	0	5	7	7	1	14	5	1	0	0	0	12	8	9	11	16	4	7	2	1	19	1
v9	4	12	0	0	1	0	7	11	1	4	8	6	1	1	0	12	8	17	3	20	0	4	0	0	20	0
v10	5	7	0	8	8	4	0	0	0	20	0	0	0	0	0	15	5	14	6	11	9	5	1	2	13	7
v11	1	1	5	0	0	0	3	4	13	7	1	1	4	4	3	11	9	11	9	20	0	7	5	1	19	1
v12	0	8	2	0	0	1	11	7	1	9	6	3	2	0	0	12	8	9	11	18	2	9	1	0	20	0
v13	1	4	5	0	1	7	12	0	0	11	7	2	0	0	0	18	2	7	13	19	1	7	2	1	18	2
v15	4	5	1	0	0	0	1	6	13	0	0	0	1	4	15	16	4	15	5	19	1	10	0	0	20	0
v16	1	9	2	0	1	1	10	4	4	3	2	4	4	4	3	17	3	10	10	17	3	5	2	1	16	4
v17	4	6	3	5	7	3	5	0	0	17	1	1	1	0	0	11	9	15	5	15	5	4	1	2	20	0
v18	1	0	9	0	4	1	8	3	4	18	0	0	1	1	0	19	1	12	8	19	1	3	4	3	20	0
v19	3	5	4	0	0	0	5	12	2	11	4	1	3	0	0	17	2	12	7	14	5	6	1	0	18	1
v20	0	5	5	0	0	5	7	4	1	3	5	3	4	1	1	9	8	5	12	14	3	2	2	3	17	0
v21	5	4	8	0	1	4	10	5	0	8	5	4	3	0	0	11	9	12	8	15	5	2	1	0	18	2
v22	0	4	11	0	10	2	3	4	1	18	2	0	0	0	0	11	9	9	11	18	2	2	1	2	20	0
v23	0	2	11	0	0	0	1	6	13	8	0	0	4	3	5	18	2	4	16	16	4	2	4	1	15	5
v24	0	3	7	0	0	0	9	10	1	4	5	6	3	1	1	17	3	14	6	18	2	4	4	2	18	2
v25	0	5	5	0	8	4	6	0	0	12	5	1	0	0	0	17	1	10	8	18	0	3	3	2	17	1

1.2.3 Tabelas Lógicas, Tabelas Disjuntas Completas.

Nas tabelas lógicas indicam-se a pertinência de cada indivíduo de uma população estatística a um grupo particular em que, o que é evidente, a modalidade de uma variável dada possui aquela característica. O código utilizado é o código lógico: 1, quando existe a pertinência 0, caso contrário. Cada indivíduo pertence a um e somente um grupo. Os dados são apresentados na seguinte forma:

	1	2	... j ...	p
1			⋮	
2			⋮	
⋮			⋮	
i	...	...	... x_{ij}	
⋮				
n				

Tabela 1.3 **Tabela lógica.** O termo x_{ij}, da i-ésima linha e j-ésima coluna, tem valor 1 ou 0 , segundo o indivíduo pertença ao grupo j (ou se possui a modalidade j de um caráter qualitativo) ou não: assim, cada em cada linha, apenas um termo será igual a 1.

Dá-se o nome de Tabela Disjunta Completa àquela formada pela justaposição de diversas tabelas lógicas. Uma Tabela Disjunta Completa teria o seguinte aspecto, para o caso de três tabelas lógicas.

Tabela 1.4 Tabela Disjunta Completa**.**

	1	2	... j_1 ...	p_1	1	2	... j_2 ...	p_2	1	2	... j_3 ...	p_3
1			⋮				⋮				⋮	
2			⋮				⋮				⋮	
⋮			⋮				⋮				⋮	
i	...	...	... x_{ij_1}		...	...	... x_{ij_2}		...	...	... x_{ij_3}	
⋮												
n												

O termo x_{ij_k} será igual a 1 ou 0, segundo o indivíduo i pertença, ou não, ao grupo j_k, para k=1, 2, 3, neste caso. Cada linha da Tabela Disjunta Completa conterá tantos 1 quantas sejam as Tabelas Lógicas.

1.2.4 Tabelas de Dados Ordinais.

As tabelas de dados ordinais são muito utilizadas em técnicas de comercialização. Em efeito, os especialistas dessas técnicas consideram freqüentemente que uma resposta dada sob a forma de classificação aporta uma informação mais coerente que uma resposta dados, por exemplo, sob a forma de nota: é solicitado, numa enquête, que se classifique um certo número de objetos (itens, marcas), por ordem de preferência.

No caso em que os *ex-aequos*[1] não são permitidos, os dados se apresentam da seguinte forma:

	1	2	... *j* ...	*p*
1			⋮	
2			⋮	
⋮			⋮	
i	...	...	... x_{ij}	
⋮				
n				

Tabela 1.5 **Tabela de dados ordinais.** O termo x_{ij}, da *i*-ésima linha e *j*-ésima, é a classificação dada para o indivíduo *i* considerando o objeto *j*

Como os ex-aequos não são admitidos, teremos:

$$\forall j \in J \ , \forall j' \in J \ e \ \forall i \in I \quad j \neq j' \Rightarrow x_{ij} \neq x_{ij'} \quad 1 \leq x_{ij} \leq p$$

A soma de uma linha é constante e igual à soma dos p primeiros número inteiros, ou seja: p(p+1)/2.

No caso em que os ex-aequos sejam admitidos, a codificação é um pouco mais complicada; se procede como se segue: se *k* objetos tem classificação ex-aequos na posição *l*, cada um será codificado $l' = \frac{1}{k}(l + (l+1) + ... + (l+k-1))$ de maneira que a soma da linha fique igual a p(p+1)/2. Ex: dois objetos classificados na 4ª posição serão codificados como 4,5; três objetos classificados na 6ª posição serão codificados 7.

Exemplo de dados ordinais. Uma empresa pretende lançar um novo dentifrício em caráter nacional; uma enquête aplicada a uma amostra representativa da população é efetuada para determinar as características do produto e a sensibilidade do público a certos argumentos. Solicitamos então às pessoas interrogadas que classifiquem por ordem de preferência os itens abaixo:

1. *brancura dos dentes*
2. *pureza do hálito*
3. *proteção contra cáries*
4. *limpeza da gengiva*
5. *eliminação de tártaros*
6. *possuir um gosto agradável.*

Uma pessoa interrogada responde da seguinte maneira: 4 2 1 3 6 5. O item brancura dos dentes ficou em terceiro lugar; o item pureza do hálito em segundo; o limpeza da gengiva em primeiro, etc.

[1] Pode-se entender por empates

1.2.5 Tabelas de distâncias, de proximidades.

Da mesma forma que as tabelas ordinais, as tabelas de distâncias ou de proximidades são freqüentemente utilizadas em técnicas de comercialização. Precisamos, inicialmente as noções de distâncias ou de proximidades: Seja dado uma população estatística I, chamamos:

— Índice de distância: uma função simétrica com valores reais e positivos definida entre dois indivíduos i e i'; quanto mais os indivíduos i e i'se "assemelham", mais o valor desse índice é fraco. Definimos

$$\forall i \in I, \forall i' \in I \quad tem-se \quad d(i,i') = d(i',i) \geq 0 \quad e \quad d(i,i) = 0$$

— índice de proximidade, uma função a valores reais definida entre dois indivíduos, i e i', e simétrica: quanto mais os indivíduos i e i'se "assemelham", mais o valor do índice de proximidade se eleva. Escrevemos, então:

$$\forall i \in I, \forall i' \in I \; tem-se \; p(i,i') = p(i',i)$$

Vê-se que podemos deduzir facilmente um índice de distância de um índice de proximidade e reciprocamente; escolhemos freqüentemente um índice de proximidade entre -1 e 1, por analogia com o coeficiente de correlação.

Uma tabela de distância aparece sob a seguinte forma:

	1	2	...j...	...j'...	*n*
1	0		⋮	⋮	
2		0	⋮	⋮	
⋮			⋮	⋮	
i	...	...	...0...	...d(i,i')	
⋮			⋮	⋮	
i'	...	...	...d(i',i)...	...0	
⋮		⋮	⋮	⋮	
		⋮	⋮	⋮	
n		⋮	⋮	⋮	0

Onde d(i,i')=d(i',i)≥0

1.3 Os principais métodos de análise multidimensional

Consideramos que a base para o desenvolvimento das técnicas de tipificação é a Análise em Componentes Principais, porém as Técnicas de Discriminação e Classificação ajudam muito nesse estudo. Por sua vez, a Análise Fatorial de Correspondências aborda dados de uma maneira mais geral ao ter nas tabelas de contingência o objeto de análise. A seguir, faremos uma apresentação genérica e rápida da Análise em Componentes Principais e indicamos a utilização da análise discriminante e as técnicas de classificação. Nos dois capítulos seguintes, abordamos a Análise Fatorial de Correspondências e a Análise Fatorial de Correspondências Múltiplas.

Capítulo 2
Noções de Análise em Componentes Principais

(Extraído de Escofier, B. e Pagès, j., 1998)

2.1 Dados e objetivos de estudo.

Vinte anos desde as primeiras publicações e aplicações, os métodos de análise de dados demonstraram largamente sua grande utilidade nos estudos de grandes massas complexas de informação. Esses são os métodos, ditos multidimensionais, que tratam de mais de duas variáveis simultaneamente. Eles permitem a confrontação de numerosas informações, o que é infinitamente mais rica que seu exame separadamente.

A Análise em Componentes Principais (ACP), aplica-se a tabelas, chamadas de forma concisa, de *Indivíduos*×*Variáveis quantitativas*, ou *Indivíduos*×*Características quantitativas*, como visto anteriormente.

Nessas tabelas as linhas representam os indivíduos e as colunas as variáveis. A interseção entre linha *i* e a coluna *j* exprime o valor da variável *j* observado no indivíduo *i*, como visto na tabela 2.1.

	$\mathbf{x}_1$	$\mathbf{x}_2$	... $\mathbf{x}_j$ ...	$\mathbf{x}_p$
1			⋮	
2			⋮	
⋮			⋮	
i	...	...	...x_{ij}	
⋮				
n				

Tabela 2.1. O termo da *i*-ésima linha e *j*-ésima coluna é o valor observado para a variável $\mathbf{x}_j$ sobre o indivíduo *i*.

Ressaltamos que os termos *indivíduos* e *variáveis* podem expressar noções diversas. Por exemplo, dependendo do estudo, a tabela pode ser composta de indivíduos que representam *vinhos* e as variáveis os diversos *critérios de apreciação* do vinho (acidez, adstringência, etc.). As questões que propomos aos indivíduos e para as variáveis não são da mesma natureza.

Com respeito a dois indivíduos, desejamos avaliar as semelhanças existentes entre eles: dois indivíduos se assemelharão, quanto mais próximos forem seus valores dentro do conjunto de variáveis. Em ACP, a distância d(*i*,*i'*) entre dois indivíduos *i* e *i'* é definida por:

$$\mathrm{d}^2(i,i') = \sum_{j \in J} (x_{ij} - x_{i'j})^2 \text{ , onde } J=(1, 2,..., \mathrm{p}).$$

Com respeito a duas variáveis, desejamos avaliar sua "ligação" (como se relacionam, se possuem interdependências, etc.). Na ACP, a ligação entre duas variáveis é medida pelo coeficiente de correlação linear (em raras situações utiliza-se a covariância), denotado, usualmente, como *ñ*. Ou seja:

$$r(j, j') = \frac{\text{covariânci a}(j, j')}{\sqrt{\text{variância}(j)\text{variância}(j')}} =$$

$$= \frac{1}{n}\sum_{i \in I}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)\left(\frac{x_{ij'} - \bar{x}_{j'}}{s_{j'}}\right)$$

sendo $\bar{x}_j$ e s_j a média e o desvio-padrão da variável j e $I=(1, 2,..., n)$.

Aplicada a uma tal tabela, a ACP objetiva a realização de:

- Um balanço das *semelhanças* entre indivíduos. Buscamos responder às seguintes questões: quais são os indivíduos que se assemelham? Quais são os mais distintos? Existem grupos homogêneos de indivíduos? Podemos colocar em evidência uma *tipologia dos indivíduos*?
- Um balanço das *ligações* entre variáveis. As questões podem ser as seguintes: quais variáveis se correlacionam positivamente? Quais são as variáveis, duas a duas, em que existindo a mesma característica, numa é marcante numa e noutra é débil, podendo-se dizer que essas variáveis se opõem (correlacionam-se negativamente)? Existem grupos de variáveis correlacionadas entre si? Podemos evidenciar uma *tipologia das variáveis*?

Um outro aspecto do estudo das ligações entre variáveis consiste em expressar um conjunto delas por um pequeno número de *variáveis sintéticas* chamadas aqui de **componentes principais**. Esse ponto de vista está bastante ligado ao precedente: um componente principal pode ser visto como o representante (a síntese) de um grupo de variáveis ligadas entre si (que possuem fortes relações que as ligam de uma ou outra forma).

Naturalmente, esses dois pontos de vista não são independentes da dualidade inerente ao estudo de uma tabela retangular: a estrutura da tabela pode ser analisada tanto por meio da tipologia de indivíduo, como por meio da tipologia de variáveis. Assim, buscamos, em geral, realizar as duas tipologias. Para tanto, caracterizaremos classes de indivíduos por certas variáveis (selecionaremos, então, variáveis para as quais os indivíduos de uma classe formam um conjunto que possui valores particularmente grandes ou pequenos). Da mesma forma, caracterizaremos grupos de variáveis ligadas entre si por indivíduos típicos (selecionaremos, então, os indivíduos que possuem valores particularmente grandes ou pequenos para um conjunto de variáveis correlacionadas positivamente). Enfim, numa situação ideal, as duas tipologias poderiam ser *superpostas*: cada grupo de variáveis caracteriza um grupo de indivíduos e cada grupo de indivíduos é formado pelos indivíduos típicos de um grupo de variáveis.

2.1.1 Pesos de indivíduos.

Na maior parte dos casos, os indivíduos desempenham o mesmo papel. Somos, então, levados a dar a mesma importância a cada indivíduo, atribuindo-lhes o mesmo peso. Por comodidade, tomamos o peso de maneira a que a massa total dos indivíduos seja igual a 1: a cada indivíduo atribuímos o peso $1/n$. Entretanto, em certos casos, poderemos atribuir pesos distintos aos indivíduos. Essas situações ocorrem quando os indivíduos representam,

cada um, subpopulações; atribuímos, então, a cada indivíduo os pesos proporcionais ao valor efetivo da sub-população por ele representada. Esses pesos intervêm no cálculo da média de cada variável (por assim dizer, na definição de um ***indivíduo teórico médio***), no cálculo da variância de cada variável e, assim, na medida da ligação (o coeficiente de correlação) entre variáveis. Então, chamando P_i ao peso atribuído a um indivíduo *i* (onde $\sum_i P_i = 1$):

$$\bar{x}_j = \sum_i P_i x_{ij} \qquad s_j^2 = \sum_i P(x_{ij} - \bar{x}_j)^2$$

$$r(j, j') = \sum_i P_i \left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)\left(\frac{x_{ij'} - \bar{x}_{j'}}{s_{j'}}\right)$$

Os programas de ACP permitem introduzir indivíduos ponderados.

2.1.2 Pesos das variáveis.

É muito raro, na prática, avaliarmos variáveis com importâncias distintas, ao observarmos um mesmo indivíduo. Essa assertiva é tão forte que a maioria dos programas de ACP não permite que sejam atribuídos pesos distintos às variáveis. Por isso, consideraremos, a priori, que as variáveis têm o mesmo peso. No entanto, para não desconsiderar a possibilidade de que seja dada importância distinta às variáveis, um coeficiente, chamado *peso das variáveis*, será utilizado. Denotando por m_j o peso da variável *j* a distância entre dois indivíduos *i* e *i'* é definida por:

$$\mathrm{d}^2(i, i') = \sum_{j \in J} m_j (x_{ij} - x_{i'j})^2$$

Entretanto, como veremos adiante, esses pesos não influenciam em nada os princípios gerais da análise. Assim, para não complicarmos, consideraremos, neste capítulo, os mesmos pesos aos indivíduos ($P_i = 1/n$ qualquer que seja $i \in I$) assim como para as variáveis ($m_j = 1$ qualquer que seja $j \in J$).

2.2 Transformação de dados.

Em ACP, as tabelas de dados devem ser, antes de tudo, centradas. De cada valor numérico, dever ser subtraída a média da variável em questão: $(x_{ij} - \bar{x}_j)$. Esse procedimento não provoca nenhuma alteração sobre as propriedades que possam ser encontradas no conjunto de dados original.

Embora a ACP possa ser realizada sobre o conjunto de dados centrados, seus resultados serão sensíveis às unidades de medida existentes. Geralmente, se escolhe medidas segundo unidades arbitrárias: em exemplos clássicos de medidas de animais, a variável altura pode ser expressa em metros ou centímetros. Essa escolha pode levar a uma grande influência sobre a medida de semelhança entre indivíduos.

A forma clássica para se evitar as diferenças causadas pelas diferentes unidades de medida, consiste em reduzir as variáveis centradas: divide-se o dado centrado (subtraído da

média correspondente à variável j) pelo desvio-padrão da variável $(x_{ij} - \overline{x}_j)/s_j$. Todas as variáveis apresentam então a mesma variabilidade e desse modo a mesma influência sobre o cálculo das distâncias sobre os indivíduos.

- Em estudos que não apresentam diferenças entre unidades de medida, a etapa de reduzir variáveis pode ser suprimida.

Procedendo assim, estamos atribuindo a cada variável um peso igual à sua variância, pois $d^2(i,i') = \sum_{j \in J} \frac{1}{s_j^2}(x_{ij} - x_{i'j})^2$. Segundo um outro ponto de vista, a definição de d(i,i') mostra que a variância da variável j é igual à contribuição média da variável j ao quadrado da distância entre indivíduos. Isso se deduz da definição da variância

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

Desde ponto em diante as variáveis são consideradas centradas e reduzidas.

2.3 Nuvem dos indivíduos.

É interessante considerar os indivíduos como uma justaposição de linhas. A cada indivíduo é associada uma série de p números. Segundo essa abordagem, um indivíduo pode ser representado como um ponto no espaço vetorial à p dimensões, denotado R^p, onde cada dimensão representa uma variável. O conjunto de indivíduos constitui a nuvem $\boldsymbol{I}$ com centro de gravidade G coincidindo com a origem O dos eixos, de fato centrado; G representa os *indivíduo médio* previamente citado. Essas noções são representadas na figura 2.1, abaixo.

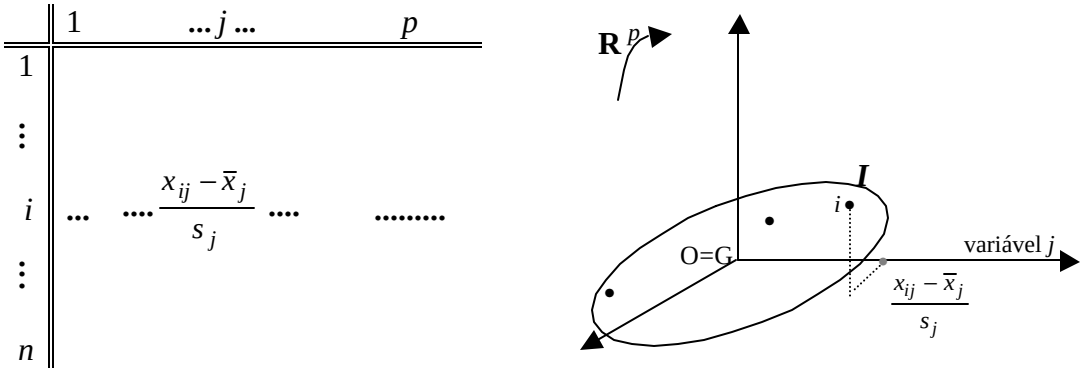


Figura 2.1 Tabela de dados centrados e reduzidos e nuvem dos indivíduos associados ao espaço $\boldsymbol{R}^p$ Devido aos dados serem centrados, a origem dos eixos coincide com o centro de gravidade da nuvem.

No espaço $\boldsymbol{R}^p$, a noção de semelhança entre dois indivíduos, não será mais que a distância euclidiana usual. Essa interpretação geométrica constitui uma justificativa, *a posteriori*, decisiva para a escolha da medida de semelhança: o fato dela ser uma distância

euclidiana confere-lhe um grande número de propriedades matemáticas indispensáveis no que segue.

O conjunto das distâncias interindividuais constitui o que chamaremos *a forma da nuvem* ***I***. Realizar um balaço dessas distâncias, equivale a estudar a forma da nuvem, ou seja, identificar partições de pontos ou das distâncias entre eles.

Sendo p superior a 3, o estudo da nuvem é impossível devido a termos uma limitação visual a três dimensões. Dai o interesse nos métodos fatoriais em geral, e da ACP em particular, que fornecem as imagens de planos que aproximam, o melhor possível, a nuvem de pontos situada num espaço de grande dimensão.

2.4 Nuvem das variáveis.

Considerando-se a tabela de dados como sendo uma justaposição de colunas, cada variável estará associada à série de n números. Define-se, dessa forma, uma variável que pode ser representada como um vetor do espaço vetorial à n dimensões, denotado por $\boldsymbol{R}^n$, onde cada dimensão representa um indivíduo: por exemplo, a variável j é representada por um vetor $\mathbf{z}_j$ onde a i-ésima componente é $(x_{ik}-\bar{x}_k)/s_k$. O conjunto das extremidades dos vetores que representam as variáveis constitui a nuvem ***J***. Observando-se a figura abaixo, tem-se uma noção dessas notações.

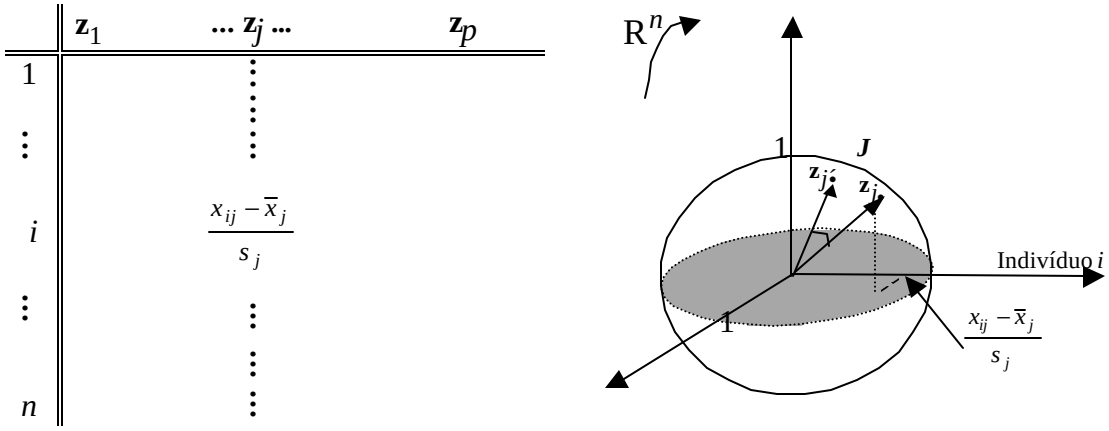


Figura 2.2. Tabela de dados e nuvem de variáveis associadas ao espaço $\boldsymbol{R}^p$.

A escolha da distancia em $\boldsymbol{R}^n$ consiste em associar a cada dimensão um coeficiente igual ao peso de cada indivíduo na nuvem ***I*** de $\boldsymbol{R}^p$ No caso geral, em que esses pesos são idênticos, a distância utilizada, com coeficientes $1/n$, é a euclidiana usual. Com essa distância, os vetores representando as variáveis centradas e reduzidas, possuem as seguintes propriedades:

a) Cada vetor representa uma variável de norma 1. Ou seja:

$$\left\| \text{variável } \mathrm{z_j} \right\|^2 = \sum_{i=1}^{n} \frac{1}{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2 = 1 \ .$$

A nuvem ***J*** está, assim, situada numa esfera de raio 1 (nos referimos a uma hiperesfera quando $\boldsymbol{R}^n$ tem dimensão maior que 3). Por essa razão, a ACP sobre dados centrados e reduzidos é dita *normalizada*[2].

b) O cosseno do ângulo formado pelos vetores representando as variáveis $\mathbf{z}_j$ e $\mathbf{z}_{j'}$, obtido ao se calcular o produto escalar, $<\mathbf{z}_j, \mathbf{z}_{j'}>$, dos dois vetores, é igual ao coeficiente de correlação entre tais variáveis:

$$\cos(\mathbf{z}_j, \mathbf{z}_{j'})=<\mathbf{z}_j, \mathbf{z}_{j'}>=\sum_i \frac{1}{n}\left(\frac{x_{ij}-\bar{x}_j}{s_j}\right)\left(\frac{x_{ij'}-\bar{x}_{j'}}{s_{j'}}\right)=\text{correlação}(\mathbf{z}_j, \mathbf{z}_{j'}).$$

Interpretar uma correlação como sendo o cosseno define uma propriedade bastante interessante porque dá um suporte geométrico, logo visual, ao coeficiente de correlação. Essa propriedade necessita que as variáveis sejam centradas, justificando assim a transformação dos dados apresentada anteriormente, como uma técnica intermediária. Essa propriedade, também justifica a escolha da métrica e implica que, na representação de variáveis, se poderá, sobretudo, conhecer as direções determinadas pelas variáveis, ou seja, conhecer os vetores porém, e sobretudo, para onde apontam suas extremidades.

O comprimento dos vetores que representam as variáveis é igual a 1, a coordenada da projeção de uma variável sobre uma outra se interpreta como um coeficiente de correlação.

Assim, buscar a coleção dos coeficientes de correlação entre as variáveis equivale a estudar os ângulos entre os vetores que definem a nuvem ***J***. A dimensão de $\boldsymbol{R}^n$ torna impossível um estudo direto desses coeficientes. O interesse na ACP é definir ***variáveis sintéticas*** que constituem *uma síntese* do conjunto das variáveis iniciais e expressam uma base para uma representação num plano aproximado, das variáveis e seus ângulos.

2.5 Representação da nuvem de indivíduos.

Neste ponto, o objetivo é fornecer as imagens, no plano aproximado, da nuvem ***I*** situada no espaço $\boldsymbol{R}^p$. Objetivamente, buscamos uma série $\{\mathbf{u}_k : k = 1, 2,..., h\}$ de direções privilegiadas de $\boldsymbol{R}^p$ chamadas ***eixos fatoriais*** que, fixados dois a dois, definirão planos fatoriais sobre os quais projetamos a nuvem ***I***. Cada direção $\mathbf{u}_k$ retém (ou explica) a inércia máxima com respeito à origem ***O*** (que se confunde com o centro de gravidade, pelo fato dos dados terem sido centrados) da projeção de ***I*** sobre $\mathbf{u}_k$. Na busca dessa série, obrigamos que cada eixo fatorial seja ortogonal aos já encontrados. Pode ser mostrado que o plano engendrado pelos dois primeiros eixos $\mathbf{u}_1$ e $\mathbf{u}_2$ retém o máximo da inércia projetada sobre o plano por eles definido. Em se tratando de $\boldsymbol{R}^p$, para $p>2$, dar-se-á o mesmo aos três primeiros eixos e os seguintes.

[2] Quando as variáveis forem apenas centradas, seu comprimento será igual ao desvio padrão e dizemos que o ACP é *não normal*.

Isso é equivalente a reter o máximo de $\sum_i F_i^2$ ou reter o mínimo de $\sum_i e_i^2$. Essa segunda expressão, forma clássica do critério dos mínimos quadrados, mostra que os eixos

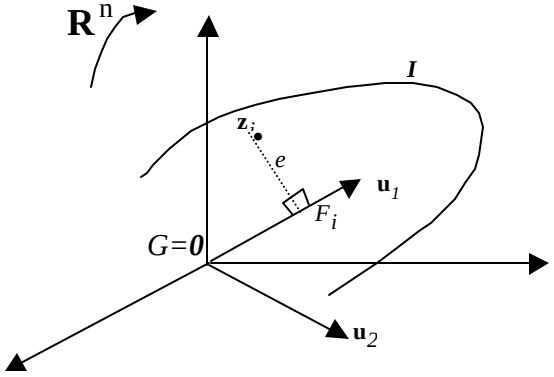


Figura 2.3 **Representação da nuvem de pontos dos indivíduos.**

O indivíduo i tem projeção F_i sobre $\mathbf{u}_1$. Buscamos encontrar u_1 que retenha o máximo da inércia $\sum_i F_i^2$. Depois, encontraremos u_2 ortogonal a u_1, que satisfaz o mesmo critério e assim por diante. No caso em que os indivíduos tenham pesos p_i diferentes, o critério consiste em reter o máximo de $\sum_i P_i F_i^2$

fatoriais retêm o menor desvio padrão entre a nuvem de indivíduos e sua projeção.

Do fato dos dados serem centrados, o critério (inércia máxima com respeito à origem ou ao centro de gravidade *G*) permite interpretar os eixos fatoriais como os de comprimento máximo da nuvem de pontos ***I***, nas direções por eles definidas. Dizemos também que esses são os principais fatores de variabilidade, pois no que é possível dão a medida da diversidade dos indivíduos.

Podemos mostrar que, ainda do fato dos dados terem sido centrados, a explicação do máximo de $\sum_i F_{1i}^2$ é equivalente a explicar o máximo de $\sum_i \sum_l \left(F_{1i} - F_{1i'}\right)^2$. Esta última expressão é de fato a distância entre os pontos projetados. A projeção ocorre de maneira a reduzir a distância entre os pontos, os eixos fatoriais aprecem numa direção tal que as distâncias entre os pontos projetados aproximam-se o mais possível das distâncias dos pontos homólogos na nuvem ***I***.

Segundo os objetivos da análise, adiante mostramos outras interpretações do critério de maximizar a inércia explicada.

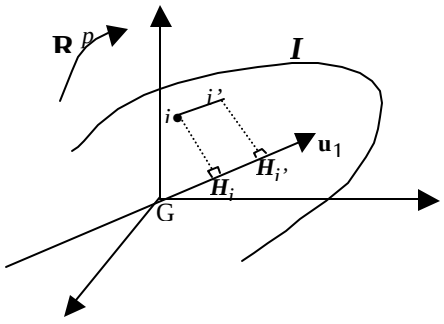


Figura 2.4 **Representação das distâncias interindividuais**. Representando por H_i e $H_{i'}$ as projeções de i e i' (os valores originais da variáveis transformadas $\mathbf{z}_i$ e $\mathbf{z}_{i'}$) sobre o eixo $\mathbf{u}_1$ (associadas a F_{1i} e $F_{1i'}$), esse eixo explicam então a inércia $\sum_i \sum_l \left(\overline{OH}_i - \overline{OH}_{i'}\right)^2$ máxima, ou seja $\sum_i \sum_{i'} d^2\left(H_1 - H_{i'}\right)$ é tal que é a mais próxima possível de $\sum_i \sum_{i'} d^2\left(i - i'\right)$.

2.5.1 Indivíduos suplementares

Com freqüência chegamos à conclusão do interessante que é de que certos indivíduos não intervenham na determinação dos eixos; porém, mesmo assim, é desejável que se conheça a posição de sua projeção sobre os eixos determinados pelo restante da

população. Isso pode ser feito, atribuindo-se pesos nulos a esses indivíduos, no critério de ajustamento.

Esses indivíduos são chamados **indivíduos suplementares**. Introduziremos um indivíduo suplementar na análise, quando é desejado que ele participe da interpretação que faremos dos dados, porém não da construção dos eixos fatoriais. Esse é o caso dos indivíduos que apresentam características excepcionais, ou que suspeitamos de erros de medida ou, enfim, consideramos que o indivíduo destaca-se dos demais por qualquer motivo.

2.6 Representação da nuvem de variáveis.

Para obtermos a série de variáveis sintéticas $\{\boldsymbol{v}_k : s = 1,..., h\}$ e uma representação aproximada das correlações entre variáveis, a ACP aplica à nuvem $\boldsymbol{J}$, das variáveis, o mesmo enfoque aplicado à nuvem de indivíduos.

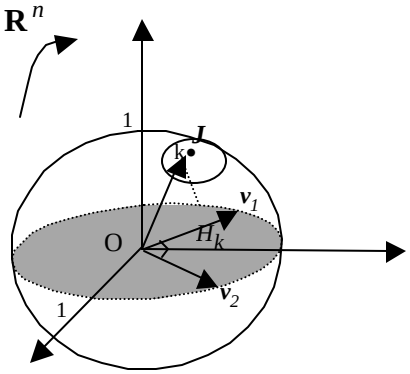


Figura.2.5 ***Representação da nuvem de variáveis****. Seja H_k projeção do ponto (G_{k1}) que representa a variável $\mathbf{y_k}$ (transformada, ou a variável* ***k*** *original) sobre o eixo $\boldsymbol{v}_1$. Procuramos $\boldsymbol{v}_1$ que explica o máximo de $\sum_k \overline{OH}_k^2$ Em seguida, procuramos $\boldsymbol{v}_2$, ortogonal à $\boldsymbol{v}_1$, que satisfaça o mesmo critério e assim por diante.*

O critério (inércia projetada máxima) que conduz à escolha dos eixos é exatamente o mesmo que para a nuvem de indivíduos. Ele, porém, terá uma significação diferente pelo fato da projeção daquela nuvem não ser mais centrada (seu centro de gravidade não é mais a origem) e de que todos seus pontos estarem situados dentro da esfera unitária: serão, então, os ângulos, entre os vetores que representam as variáveis, que serão deformados pelas projeções, e não as distâncias entre os pontos da nuvem. Em efeito, o plano $(\boldsymbol{v}_1, \boldsymbol{v}_2)$, ao maximizar a inércia à origem da nuvem projetada, explica o máximo da soma dos cossenos quadrados dos ângulos entre os vetores e sua projeção: este plano ajusta os vetores e deforma o menos possível seus ângulos.

2.6.1 Componentes principais.

O vetor $\boldsymbol{v}_1$, que caracteriza a direção da inércia máxima, define uma nova variável. As variáveis estudadas estão centradas e reduzidas, sua projeção sobre $\boldsymbol{v}_1$ é igual ao

coeficiente de correlação deste eixo com essa variável. Desse fato, procurar o vetor $\boldsymbol{v}_1$ que explica o máximo valor de $\sum_k \overline{OH}_k^2$ equivale a procurar a combinação linear o mais associado possível a estas variáveis (no sentido do critério que maximiza a soma dos quadrados das correlações). Assim, esse será o eixo que melhor sintetiza o conjunto das variáveis iniciais. Os eixos fatoriais, que são ortogonais dois a dois, põem em evidência uma série de variáveis sintéticas, a*s componentes principais*, não correlacionadas entre elas, que melhor resumem o conjunto de variáveis iniciais.

2.6.2 Variáveis suplementares.

As variáveis, como os indivíduos, podem ser tratadas como suplementares. Essas serão aquelas que simplesmente são projetadas nos eixos que foram determinados pelas outras variáveis, ditas *ativas*. Pode-se, então, conhecer as correlações entre quaisquer variáveis e os componentes principais; mesmo aquelas que não pertencem ao domínio de estudo.

2.6.3 Efeito *taille* (de valor da variável).

Se, num conjunto de dados, as variáveis são todas correlacionadas duas a duas, então a nuvem ***J*** está distante da origem, próxima da superfície da esfera. O primeiro eixo fatorial se refere, então, principalmente à posição de ***J*** com respeito à origem: paralelamente, a forma da nuvem ***J*** é mal representada, no sentido de que as projeções das variáveis são todas próximas umas das outras.

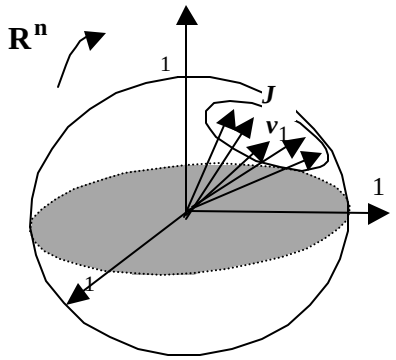


Figura 2.6. **O efeito *taille* em R^n.**
As variáveis, correlacionadas duas a duas, formam entre si ângulos agudos. A nuvem de pontos ***J*** fica concentrada sobre um pequeno setor da esfera. A projeção das variáveis sobre o primeiro eixo fatorial, definido por $\boldsymbol{v}_1$, se refere principalmente à posição de ***J*** em relação a ***O***.

O caso dessa figura é, na França, costumeiramente chamado *efeito taille*: ele corresponde à situação na qual, lembrando o fato de que as variáveis se correlacionam duas a duas, certos indivíduos possuem pequenos valores com respeito a um conjunto de variáveis, outros possuem grandes valores com respeito a outras variáveis e outros, enfim, valores intermediários entre esses extremos. Existe, neste caso, uma estrutura comum, *característica*, associada ao conjunto das variáveis: essa característica típica é o que traduz a primeira componente principal.

2.7 Dualidade e fórmulas de transição na ACP.

A nuvem ***I***, dos indivíduos, e ***J***, das variáveis, são duas representações de uma mesma tabela: uma pelas linhas e outra por suas colunas. Algumas relações bastante expressivas, ligando essas duas nuvens, são chamadas *relações de dualidade.*

2.7.1 Inércias.

Deve-se saber, antes de tudo, que a inércia total dessas duas nuvens é a mesma; ela é igual ao número de variáveis (desde que as variáveis sejam centradas e reduzidas)

$$\text{Inércia total de } \boldsymbol{I} \text{ (ou de } \boldsymbol{J}\text{)} = \frac{1}{n}\sum_j \sum_i \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2 = P$$

A projeção de cada uma dessas duas nuvens sobre uma série de eixos ortogonais corresponde a uma decomposição da inércia total. Pode-se mostrar que as decomposições são idênticas: as inércias das nuvens projetadas sobre os eixos fatoriais de mesma dimensão, são iguais. Será, para o *k-ésimo eixo encontrado*:

$$\text{Inércia}(\boldsymbol{I}/\boldsymbol{u}_k) = \text{Inércia}(\boldsymbol{J}/\boldsymbol{v}_k) = \ddot{e}_k,$$

onde $\ddot{e}_k$ é o valor próprio associado ao *k-ésimo* eixo.

2.7.2 Fatores.

O conjunto das projeções de todos os *n* indivíduos, pontos da nuvem de indivíduos ***I***, sobre os *k-ésimo* eixo fatorial $\boldsymbol{u}_k$, chamado *k-ésimo* fator sobre os indivíduos, constitui uma nova variável, denotada $\boldsymbol{F}_k$. Demonstra-se que essa variável se confunde com a *k-ésima* componente principal obtida na análise da nuvem de variáveis. Mais precisamente, o quadrado da norma do fator $\boldsymbol{F}_k$ (um vetor de $\boldsymbol{R}^{n)}$, é a soma dos quadrados de suas coordenadas, por $\ddot{e}_s$; a relação entre o *k-ésimo* fator sobre ***I*** e o *k-ésimo* eixo fatorial de $\boldsymbol{R}^n$ se escreve:

$$\mathbf{v}_k = \frac{1}{\sqrt{l_k}} \mathbf{F}_k$$

Assim, as projeções planas em $\boldsymbol{R}^p$ são representações gráficas de pares de variáveis sintéticas obtidas em $\boldsymbol{R}^{n\cdot}$ Os resultados obtidos do estudo de cada uma dessas duas nuvens possuem fundamentalmente a mesma significação, mesmos se são expressos em termos de indivíduos em uns casos ou de variáveis em outros.

Na figura 2.7, a seguir, ilustra-se esses resultados.

Os atributos da nuvem de indivíduos e da nuvem de variáveis são, em certa medida,

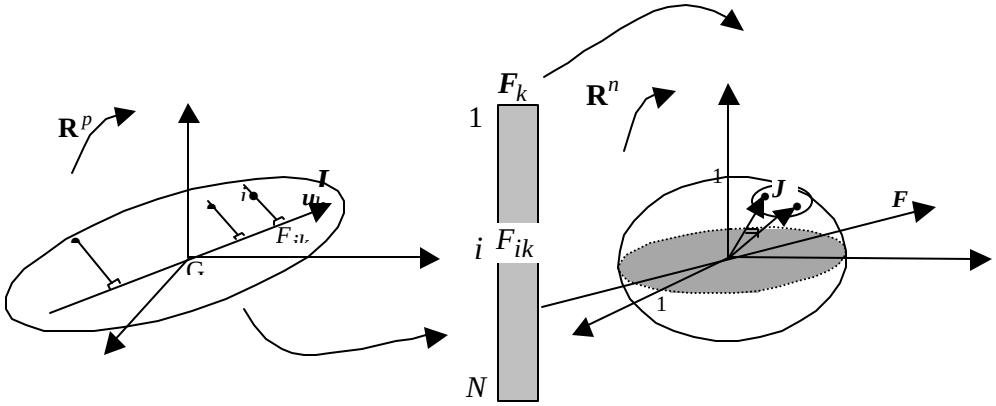


Figura 2.7. **Uma das duas formas da dualidade**. As coordenadas de I sobres u_k (k-ésimo eixo fatorial de **I**) constituem o k-ésimo fator sobre os indivíduos (denotado por $\mathbf{F}_k$). O vetor $\mathbf{F}_k$ em $\mathbf{R}^p$ é colinear à v_k (k-ésimo eixo fatorial de **J**).

simétricos, e a dualidade se formula de maneira análoga ao traçar o papel das duas nuvens: a projeção das *p* variáveis sobre o *k-ésimo* eixo fatorial $\boldsymbol{v}_k$, da nuvem $\boldsymbol{J}$, define um valor para cada uma das *p* variáveis: esses valores constituem o *k-ésimo* fator sobre as variáveis (denotado $\boldsymbol{G}_k$) que é, de alguma maneira, um *indivíduo* novo. Essa noção de indivíduo *típico* (ou *característico*) é menos clássica que a de componente principal (estamos, praticamente, considerando antes de tudo, indivíduos reais como indivíduos *típicos*). Entretanto, em casos particulares, como aqueles em que os indivíduos estão em uma curva e as variáveis são seus valores em *p* pontos de discretização, esses indivíduos são representáveis e assim utilizados.

Mostra-se que o ponto em $\boldsymbol{R}^p$ representando esse indivíduo típico está situado sobre o *k-ésimo* eixo da nuvem dos indivíduos. Mais precisamente:

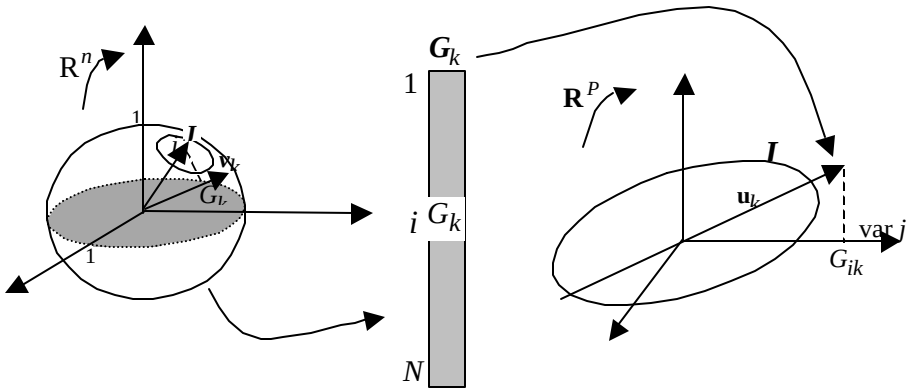


Figura 2.8. **A segunda forma da dualidade**. As coordenadas de J sobre v_k (k-ésimo eixo fatorial de **J**) constituem o k-ésimo fator sobre os indivíduos (denotado por $\mathbf{G}_k$). O vetor $\mathbf{G}_k$ em $\mathbf{R}^p$ é colinear ao k-ésimo eixo fatorial de **I,** $\mathbf{u}_k$.

$$\mathbf{u}_k = \frac{1}{\sqrt{l_k}} G_k$$

Essa relação mostra que, referindo-se ao coeficiente $\sqrt{l_k}$, valor próprio do fator $\boldsymbol{G}_k$, as coordenadas das variáveis sobre $\mathbf{v}_k$ são os coeficientes da combinação linear das variáveis que constituem o eixo $\mathbf{u}_k$ de $\boldsymbol{R}^p$. Essa propriedade é uma característica dos principais eixos (componentes principais) e essencial para a interpretação (inversamente à dificuldade de interpretação dos coeficientes da regressão múltipla quando eles não são do mesmo sinal dos coeficientes de correlação associados).

2.7.3 Relações de transição.

Chamam-se relações de transição entre os *k-ésimos* fatores, $\boldsymbol{F}_k$ e $\boldsymbol{G}_k$, à expressão algébrica das propriedades ilustradas nas duas últimas figuras. Considerando $\sqrt{l_s}$, e a inércia projetada de $\boldsymbol{I}$ (ou de $\boldsymbol{J}$) sobre o *k-ésimo* eixo, essas relações se escrevem como a seguir:

$$F_{ks}(i) = \frac{1}{\sqrt{l_k}} \sum_j \frac{x_{ij} - \bar{x}_j}{s_j} G_k(j)$$

$$G_k(j) = \frac{1}{n\sqrt{l_k}} \sum_i \frac{x_{ij} - \bar{x}_j}{s_j} F_k(i)$$

A primeira relação exprime o fato de que a projeção $F_k(i)$ de um indivíduo *i*, é uma combinação linear das projeções $G_k(j)$ de todas as variáveis. Nessa combinação linear, o coeficiente de uma variável *j* será positivo se o valor x_{ij} dessa variável, com respeito ao indivíduo *i*, ultrapassa a média $\bar{x}_k$. Caso contrário, o coeficiente será negativo. Assim se olhamos simultaneamente os dois gráficos, um indivíduo estará ao lado das variáveis que possuírem valores fortes (altos) **e** estará em oposição àquelas que possuírem valores fracos (pequenos).

O gráfico de indivíduos é uma representação aproximada da distância existente entre eles. O das variáveis pode ser considerado como o elemento explicativo dessa representação: dois indivíduos situados na mesma extremidade de um eixo estarão próximos porque os dois terão, geralmente, valores altos (significativos) com respeito às variáveis situadas no mesmo lado **e** valores insignificantes nas variáveis situadas em lado oposto.

Reciprocamente, o gráfico dos indivíduos pode intervir de maneira a ajudar na interpretação do gráfico das variáveis: se duas variáveis são fortemente correlacionadas positivamente, elas estarão situadas do mesmo lado sobre um eixo. Sobre o eixo correspondente da nuvem de indivíduos, os indivíduos que possuírem valores expressivos para aquelas duas variáveis, ficarão situados ao lado delas e os que tiverem valores sem expressão se situarão em lado oposto. Os indivíduos mais expressivos com respeito àquelas

variáveis se encontrão mais afastados da origem. São facilmente reparados, aqueles indivíduos particulares que induzem a essas correlações fortes.

2.7.4 Representação simultânea.

A necessidade de uma interpretação conjunta das representações dos indivíduos e das variáveis conduz à sua superposição. É importante ressaltar que a justificativa para uma tal representação simultânea, de indivíduos e variáveis, é essencialmente pragmática: a representação das variáveis ajuda a fazer uma interpretação dos indivíduos e reciprocamente. Ela aponta, não obstante, o problema da representação sobre um mesmo gráfico de pontos de natureza distinta, avaliados em espaços diferentes. Essa dificuldade não é somente de principio: a presença simultânea, de indivíduos e variáveis sobre um mesmo plano, engendra as proximidades entre variáveis e indivíduos que, por sua vez, pode sugerir idéias que não se verificam dentro dos dados. As observações seguintes garantem que se pode utilizar sem perigo a representação simultânea na ACP:

- As fórmulas de transição relêem a coordenada sobre um eixo de um indivíduo com o conjunto de coordenadas de todas as variáveis sobre um eixo de mesma ordem. Não se pode interpretar a posição de um indivíduo pela ocorrência de uma só variável e reciprocamente.
- Fundamentalmente, as variáveis são vetores e não pontos. Não é a posição entre um indivíduo e um conjunto de pontos, que representam as variáveis, que é importante, mas o desvio que tem o indivíduo com respeito à direção definida por daquele conjunto de variáveis.

2.7.5 Projeção de vetores unitários da representação de indivíduos.

Uma outra idéia, tendo em vista a representação simultânea de indivíduos e variáveis, consiste em projetar os vetores unitários de $\boldsymbol{R}^p$ sobre os eixos $\mathbf{u}_k$. Obtém-se, então, uma representação sobreposta mais natural que a precedente, no sentido de que os objetos representados provêem do mesmo espaço.

Pela relação entre $\mathbf{u}_k$ e $\boldsymbol{G}_k$, e observando que a j-ésima coordenada de $\mathbf{u}_k$ é igual à projeção sobre $\mathbf{u}_k$ do vetor unitário do j-ésimo eixo de $\boldsymbol{R}^p$, essa nova representação das variáveis é homotética à precedente eixo a eixo numa razão constante de $\sqrt{l_s}$.

2.8 Ajuda à interpretação.

Os eixos fatoriais fornecem uma imagem aproximada de uma nuvem de pontos. Se faz necessário poder medir a qualidade de aproximação, tanto para cada ponto, como para a nuvem como todo. Ou seja, os planos fatoriais representam as coordenadas dos pontos e não as inércias que levam à sua determinação. É então bastante útil consultar essas inércias. Resulta, assim, num estudo do plano realizado com a consulta de um conjunto de indicadores reagrupados sob o termo de ajuda à interpretação.

2.8.1 Definições.

2.8.1.1 Qualidade de representação de um elemento num eixo.

A qualidade de representação de um elemento *i* (indivíduo ou variável) num eixo *s* é medida pela relação

$QLT_k(i)$=[*inércia da projeção do elemento i sobre o eixo k*]/[*inércia total de i*] que é também o cosseno quadrado do ângulo è entre o segmento *Oi* e o eixo $\mathbf{u}_k$ (figura 2.9).

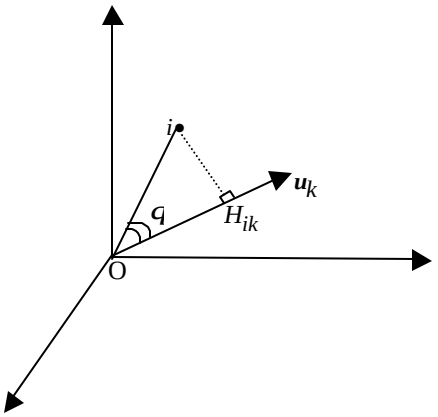


Figura 2.9 **Qualidade de representação de um elemento em um eixo.** H_{ik} é a projeção de i sobre o k-ésimo eixo.

$$QLT_k(i)= H_{ik}^2 / \overline{Oi}^2 = \cos^2 \boldsymbol{q}$$

Essa definição se generaliza no caso de um plano. Além disso, pelo fato da ortogonalidade dos eixos fatoriais, a qualidade de representação de um elemento *i* num plano (eixo *k*, eixo *h*) é a soma das qualidades de representação de *i* sobre o eixo *k* e sobre o eixo *h*. É também o cosseno quadrado do ângulo entre o vetor *Oi* e o plano de projeção. Se a qualidade de representação da projeção de um ponto sobre um eixo, ou sobre o plano, é bastante próxima de 1, então este ponto estará bem próximo do eixo ou do plano. Tratando-se de um indivíduo, sua distância ao centro de gravidade (que é o ponto médio) é então visível sobre a projeção. Ela não será visível (distinguível) no caso contrário (quando a qualidade de representação é próxima de zero). Da mesma forma, a distância entre dois pontos sobre um plano, traduz tão bem sua distância na nuvem na mesma medida em que esses pontos tiverem uma qualidade de representação. Tratando-se de uma variável centrada e reduzida o vetor tem norma igual a 1 e sua qualidade de representação é o quadrado do tamanho de sua projeção. Sobre um plano, ela se visualizada diretamente por sua proximidade ao círculo de raio 1, traçado de uma hiperesfera de raio 1 sobre o plano fatorial. Esse círculo é chamado, usualmente, *círculo das correlações*.

2.8.1.2 Qualidade de representação de uma nuvem sobre um eixo.

A definição precedente se generaliza ao conjunto de uma nuvem pela relação:

inércia da projeção da nuvem sobre o eixo/*inércia total da nuvem*

Este indicador, chamado *porcentagem de inércia associada a um eixo* mede, por sua vez, a "importância" relativa de um eixo fatorial na variabilidade dos dados.

Como no caso de um único elemento, essas percentagens podem ser acumuladas sobre vários eixos; falamos então da porcentagem de inércia extraída por um plano ou pelos K primeiros fatores. Devido à dualidade é equivalente calcular essas porcentagens de inércia a partir da nuvem de indivíduos ou daquela das variáveis.

2.8.1.3 Contribuição de um elemento à inércia de um eixo.

Um eixo fatorial contribui com o máximo (sob a condição de ortogonalidade com os eixos precedentes) da inércia projetada de uma nuvem. Essa inércia projetada da nuvem pode ser decomposta ponto por ponto. O cociente da inércia projetada do elemento *i* sobre o eixo *k* pela inércia da projeção do conjunto da nuvem sobre esse mesmo eixo, representa a contribuição do elemento *i* à inércia do eixo *k*.

Esse indicador se generaliza aos subconjuntos dos elementos. Sua contribuição à inércia de um eixo é a soma das contribuições dos elementos que o compõem. Essa razão é preciosa para por em evidência o subconjunto de elementos que contribuem principalmente à construção do eixo e sobre o qual se apóia em primeiro lugar a interpretação.

2.9 Técnicas de agrupamento

2.9.1 Análise Fatorial Discriminante

As origens da análise fatorial discriminante se inscrevem nos anos 30, do século passado, sendo os primeiros desenvolvimentos se devem a R.A. Fisher, sob a hipótese de normalidade das variáveis.

Dentro do escopo da análise de dados multidimensional a análise discriminante pode ser tomada quando, ao caracterizarmos grupos de variáveis (por exemplo, aquelas que compõem um fator qualificando uma *variável síntese*) se tem objetivos como os abaixo:

1. De caráter descritivo, quando:
   - Desejamos obter uma representação do conjunto de observações que nos permita verificar se estamos realmente na presença de grupos bem diferenciados;
   - Pretendemos encontrar a variável, ou conjunto de variáveis, que melhor discriminam a grupos preestabelecidos de observações.
2. Quando forem para uma tomada de decisão:
   - e se tratar de reclassificar certas observações do conjunto inicial ***I***,
   - tem-se por finalidade, classificar novas observações (observações que não estavam presentes no conjunto ***I*** inicial) em um dos grupos existentes.

O quadro de dados, objeto de uma análise discriminante, é um do tipo de dupla entrada, de variáveis e observações, no qual existe uma partição prévia destas últimas. Ou seja, se $\boldsymbol{I}=\{1, 2, ..., i,..., n\}$ é o conjunto de observações, deverá existir uma partição em q grupos, $I_1, I_2, ..., I_h, ..., I_q$, com $n_1, n_2, ..., n_h, ..., n_q$ elementos, respectivamente, de tal maneira que se verifica:

$$I = \bigcup_{h=1}^{q} I_h \quad \text{e} \quad n = \sum_{h=1}^{q} n_h$$

Em nosso caso cada subconjunto de ***I*** pode ser caracterizado a partir dos fatores obtidos pela aplicação de uma ACP ou de uma análise fatorial de correspondências.

Grupos	Colunas / Linhas	1, 2, ⋯ *j* ⋯ *p*
I_1	1 2 ⋮ n_1	⋮
⋮	⋮	⋮
I_h	⋮ *i* ⋮	..... x_{ij} .....
⋮	⋮	⋮
I_q	$n-n_q+1$ $n-n_q+2$ ⋮ *n*	⋮
x_{ij} = valor da variável j na observação *i*		

Figura 2.10 Quadro para definição dos dados de entrada para uma Análise Fatorial Discriminante

2.9.2 Técnicas de Classificação

Segundo Judez (p. 147) a história da classificação pode ser remontada à idade antiga, encontrando-se em trabalhos de Galen e Aristóteles, sendo posteriormente desenvolvidas no domínio biológico (séculos XVII e IXI) e da zoologia.

Os dados sobre os quais se aplicam as técnicas de classificação são do tipo de dupla entrada no qual um conjunto de observações ***I***={1, 2,..., *i*,..., *n*} está caracterizado por um conjunto de variáveis ***J***={1, 2,..., *j*,..., *p*}. As variáveis podem ser do tipo quantitativo ou qualitativo e os conjuntos *I* e *J* não apresentam nenhuma partição prévia.

As técnicas de classificação compõem o que se costuma chamar taxonomia numérica e seu objetivo é obter um conjunto de classes, disjuntas ou não, de elementos de I ou de J.

Divide-se em dois grandes grupos as técnicas que se ocupam da obtenção de classes disjuntas: as técnicas hierárquicas e não hierárquicas, ou de agrupamento.

As técnicas hierárquicas conduzem a um conjunto de partições dos elementos de I ou de J que podem ser representadas segundo uma árvore de classificação.

Entre as técnicas não hierárquicas mais utilizadas, encontram-se as chamadas técnicas otimizantes.

Essa técnica, assim como a análise discriminante é de grande ajuda para tipificação de variáveis e indivíduos na análise de dados multidimensionais

Capítulo 3
Análise Fatorial de Correspondências
Fundamentos Teóricos

(Extraído de Judez, L., 1988)

3.1 Introdução.

Embora certos trabalhos, dos anos quarenta, de R. A. Fisher e Guttman, sejam citados como precursores da Análise Fatorial de Correspondências (AFC), é reconhecido que foi nos anos sessenta que J.P. Benzecri e B. Escofier que desenvolveram suas propriedades algébricas.

Em sua origem, a AFC estava associada ao estudo de tabelas de contingência (ou tabelas cruzadas). Tais tabelas, como já foi visto, constituem-se de dados relativos ao número de elementos existentes nas modalidades combinadas, de duas características medidas. Por exemplo: considere o número de variedades de manga que contraem algum tipo de patologia. Pode-se elaborar uma tabela em que se confrontam as características *variedade*x*patologia*: cada linha da tabela corresponderá a uma modalidade devido à *variedade* e cada coluna a uma modalidade devido à *patologia*. Cada valor dentro da tabela será o número de frutos de uma variedade no qual se observou a existência de uma certa patologia.

De uma maneira geral, se cruza o caráter $I=\{1,2,...,i,...,n\}$, que expressa n modalidades, com o caráter $J=\{1,2,...,j,...,p\}$, expressando p modalidades. Os valores constantes na tabela serão N_{ij} que representam o número de unidades estatísticas que possuem, simultaneamente, a modalidade i do caráter (característica) I e a modalidade j do caráter (característica) J.

Embora, como se pode notar, a distinção entre variáveis e observações seja artificial, para esse tipo de tabela, adotaremos a prática de denominar as modalidades do caráter disposto nas colunas de variáveis e àquelas do caráter que define as linhas de observações. Isso nos permitirá ter como base os resultados devidos ao desenvolvimento da ACP.

Embora os resultados obtidos no desenvolvimento teórico da AFC se refiram, particularmente, ao estudo dos quadros de contingência, veremos que existem outros tipos de tabelas às quais a aplicação da AFC pode obter excelentes resultados.

3.1.1 Objetivos da Análise Fatorial de Correspondências.

De uma maneira geral pode-se considerar que os objetivos que se perseguem quando se aplica a AFC são similares aos buscados com a ACP . De maneira resumida são os seguintes:

- Estudo das relações existentes no interior do conjunto I e no interior do conjunto J. Ou seja, o estudo das relações entre as modalidades, dentro do caráter I ou J.

- Estudo das relações existentes entre os elementos do conjunto *I* e os elementos do conjunto *J*. Quer dizer, o estudo das relações existentes entre as modalidades das características *I* e *J*.

Desenvolvimento Teórico.

Por poder ser considerado como um caso particular da ACP, o desenvolvimento teórico da AFC é análogo ao daquela técnica. Assim, serão analisadas as nuvens de observações (indivíduos) e variáveis, como na ACP, porém, neste caso, serão construídas as nuvens de perfis dos pontos-observações e de perfis dos pontos-variáveis.

No entanto, antes de efetuarmos essas análises, examinaremos a construção de ditas nuvens, mostrando uma de suas propriedades (da equivalência distributiva), que constitui, em certos casos, a vantagem de se fazer utilização da AFC, em relação a ACP.

3.2 Nuvem de pontos-observações e pontos-variáveis.

3.2.1 Perfis de observações e variáveis.

Numa tabela de contingência sejam:

$N_{i.} = \sum_{j=1}^{p} N_{ij}$ =total de unidades estatísticas da modalidade *i.*

$N_{.j} = \sum_{i=1}^{n} N_{ij}$ =total de unidades estatísticas da modalidade *j*.

$N = \sum_{i}\sum_{j} N_{ij} = \sum_{i} N_{i.} = \sum_{j} N_{.j}$ =total de unidades estatísticas do conjunto de modalidades.

Uma primeira caracterização das modalidades *i* do caráter *I* (observações *i*) pode ser feita a partir do peso relativo de cada modalidade de *J* na modalidade *i*:

$$\frac{N_{i1}}{N_{i.}}, \frac{N_{i2}}{N_{i.}}, \ldots, \frac{N_{ij}}{N_{i.}}, \ldots, \frac{N_{ip}}{N_{i.}}$$

que chamaremos *perfil da observação i* e é a distribuição de freqüências condicionais do caráter *J* para *I=i.*

De um modo análogo podemos caracterizar as modalidades *j* da característica *J*, pelo que denominaremos *perfis da variável j*:

$$\frac{N_{1j}}{N_{.j}}, \frac{N_{2j}}{N_{.j}}, \ldots, \frac{N_{ij}}{N_{.j}}, \ldots, \frac{N_{nj}}{N_{.j}}$$

que é a distribuição de freqüências condicionais do caráter *I* para *J=j*.

Considerando, então:

$$P_{ij} = \frac{N_{ij}}{N}; \qquad P_{i.} = \sum_{j} P_{ij}; \qquad P_{.j} = \sum_{i} P_{ij}$$

como as freqüências conjuntas e das características *I* e *J*, respectivamente, pode-se reescrever a tabela de contingência original da seguinte maneira:

I \ J	$1\ 2 \cdots j \cdots p$	Total colunas
1 2 $\vdots$ i $\vdots$ n	$\vdots$ $\cdots \cdots P_{ij} \cdots \cdots$ $\vdots$	$\vdots$ $P_{i.}$ $\vdots$
Total linhas	$\cdots \cdots P_{.j} \cdots \cdots$	1

Sendo que: $\sum_i \sum_j P_{ij} = \sum_i P_{i.} = \sum_j P_{.j} = 1$

Em função dessas freqüências, o perfil da observação i será:

$$\frac{P_{i1}}{P_{i.}}, \frac{P_{i2}}{P_{i.}}, \ldots, \frac{P_{ij}}{P_{i.}}, \ldots, \frac{P_{ip}}{P_{i.}}$$

e da variável *j*, será:

$$\frac{P_{1j}}{P_{.j}}, \frac{P_{2j}}{P_{.j}}, \ldots, \frac{P_{ij}}{P_{.j}}, \ldots, \frac{P_{nj}}{P_{.j}}.$$

3.2.2 Construção da nuvem de pontos-observações.

Os perfis das observações, definidos acima, podem ser representados mediante vetores $\mathbf{x}_i$ num espaço vetorial de *p* dimensões:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{bmatrix} \quad \text{sendo} \quad x_{ij} = \frac{P_{ij}}{P_{i.}}$$

Associa-se a cada perfil, que corresponde a um ponto em $\boldsymbol{R}^p$, um peso $P_{i.}$ (que pode ser expresso como a importância relativa de *i* no conjunto *I*).

Na AFC a distância que se define entre pontos não é a euclidiana clássica, senão uma distância chamada c^2, cuja expressão é:

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

$$= (\mathbf{x}_i - \mathbf{x}_{i'})' \mathbf{M}_P (\mathbf{x}_i - \mathbf{x}_{i'})$$

onde:

$$\mathbf{M}_p = \begin{bmatrix} \frac{1}{P_{.1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{P_{.2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{P_{.p}} \end{bmatrix}$$

é a métrica associada à mencionada distância.

Transformação da caracterização dos pontos-observações para trabalhar com a métrica euclidiana clássica.

Fazendo as devidas considerações, qualquer métrica pode ser transformada na euclidiana clássica. No caso presente, bastará caracterizar as observações, segundo o seguinte vetor:

$$\mathbf{x}'_i = \begin{bmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{ij} \\ \vdots \\ x'_{ip} \end{bmatrix} \quad \text{sendo} \quad x'_{ij} = \frac{P_{ij}}{\sqrt{P_{.j}}\, P_{i.}}$$

Deste modo pode ser verificado como a distância euclidiana entre os pontos i e i' nos conduz a $(\mathbf{x}'_i - \mathbf{x}'_{i'})' \mathbf{M}_p (\mathbf{x}'_i - \mathbf{x}'_{i'})$

Vemos, pois, que a eleição da métrica c^2, que consiste em multiplicar por $1/\sqrt{P_{.j}}$ os vetores que caracterizam o perfil das observações, tem como conseqüência o fato de que a distância entre os pontos não está dominada, forçosamente, pelas variáveis de maior importância (maior peso). Essa ponderação não tem uma interpretação tão evidente, porém se justificará adiante ao ser abordada a equivalência distributiva e a condição de ótimo da representação simultânea de observações e variáveis.

Ressaltamos, enfim, que a AFC consiste basicamente em uma ACP da nuvem de pontos-observações e pontos-variáveis. Essas análises se efetuam, como visto, em valores centrados e reduzidos. Assim, tendo em conta que a média m_j da variável $\mathbf{x}'_j$ é:

$$m_j = \sum P_{i.} \frac{Pij}{\sqrt{P_{.j}} P_{i.}} = \sqrt{P_{.j}}$$

as observações *i* serão caracterizadas por vetores $\mathbf{z}_i$:

$$\mathbf{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{ij} \\ \vdots \\ z_{ip} \end{bmatrix} \quad onde \;\; z_{ij} = \frac{P_{ij}}{\sqrt{P_{.j}} P_{i.}} - \sqrt{P_{.j}} \; .$$

3.2.3 Construção da nuvem de pontos-variáveis.

Como foi visto no capítulo anterior, a caracterização da nuvem de pontos-variáveis se dá de maneira análoga à de pontos-observações. Partindo de seus perfis, associados a vetores *N*-dimensionais, tem-se:

$$\mathbf{y}_j = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{nj} \end{bmatrix} \quad onde \;\; y_{ij} = \frac{P_{ij}}{P_{.j}}.$$

Ponderados (afetados) por um peso $P_{.j}$, se define em $\boldsymbol{R}^n$, uma distância c^2, como a seguir:

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

$$= (\mathbf{y}_j - \mathbf{y}_{j'})' \mathbf{M}_n (\mathbf{y}_j - \mathbf{y}_{j'})$$

sendo:

$$\mathbf{M}_n = \begin{bmatrix} \frac{1}{P_{1.}} & 0 & \cdots & 0 \\ 0 & \frac{1}{P_{2.}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{P_{n.}} \end{bmatrix}.$$

Transformação da caracterização dos pontos-variáveis para trabalhar com a métrica euclidiana clássica.

Caracterizando as variáveis segundo vetores $\mathbf{y}'_j$:

$$\mathbf{y}'_j = \begin{bmatrix} y'_{1j} \\ y'_{2j} \\ \vdots \\ y'_{ij} \\ \vdots \\ y'_{nj} \end{bmatrix} \quad \text{sendo} \quad y'_{ij} = \frac{P_{ij}}{\sqrt{P_{i.}}P_{,j}}$$

se obtém a distância $d(j,j')$ expressada em $(\mathbf{y}_j - \mathbf{y}_{j'})'\mathbf{M}_n(\mathbf{y}_j - \mathbf{y}_{j'})$ utilizando a fórmula para definir a distância euclidiana clássica.

O translação da origem de $\boldsymbol{R}^n$ ao centro de gravidade da nuvem de pontos-variáveis faz com que, para proceder a análise, seus pontos devam ser caracterizados por vetores $\mathbf{w}^j$, tais que:

$$\mathbf{w}_j = \begin{bmatrix} w_{1j} \\ w_{2j} \\ \vdots \\ w_{ij} \\ \vdots \\ w_{nj} \end{bmatrix} \quad \text{onde} \quad w_{ij} = \frac{P_{ij}}{\sqrt{P_{i.}}P_{,j}} - \sqrt{P_{i.}}$$

Devemos notar finalmente que a caracterização que fizemos dos pontos-variáveis corresponde à que seria feita às observações se transpuséssemos a tabela de contingência. Esse aspecto de simetria entre observações e variáveis, da análise de correspondência, não se encontra na ACP.

3.2.4 Propriedade de equivalência distributiva.

1. *Equivalência distributiva em $\boldsymbol{R}^p$.*

Se dois pontos $\mathbf{z}_{i_1}$ e $\mathbf{z}_{i_2}$ se confundem em $\boldsymbol{R}^p$ então:

$$\frac{P_{i_1 j}}{\sqrt{P_{j.}}P_{i_1.}} - \sqrt{P_{j.}} = \frac{P_{i_2 j}}{\sqrt{P_{j.}}P_{i_2.}} - \sqrt{P_{j.}} \text{ para todo } j,$$

ou seja, se temos:

$$\frac{P_{i_1 j}}{P_{i_1.}} = \frac{P_{i_2 j}}{P_{i_2.}} \text{ para todo } j,$$

a substituição dos pontos i_1 e i_2 por um outro, i_0, cujos valores dos efetivos sejam $N_{i_0 j} = N_{i_1 j} + N_{i_2 j}$ não variará a distância entre os pares de pontos nem em $\boldsymbol{R}^p$ ou em $\boldsymbol{R}^n$.

Demonstração da invariância das distâncias em R^p.

A distância entre pares de pontos, excluídos i_1 e i_2, não variará. Deve-se demonstrar, então, que a distância entre qualquer ponto i e i_1 (ou i_2) é a mesma que aquela entre i e i_0. Como para todo j, tem-se:

$$\frac{P_{i_1 j}}{P_{i_1.}} = \frac{P_{i_2 j}}{P_{i_2.}} = \frac{P_{i_1 j} + P_{i_2 j}}{P_{i_1.} + P_{i_2.}} = \frac{N_{i_1 j} + N_{i_2 j}}{N_{i_1.} + N_{i_2.}} = \frac{N_{i_0 j}}{N_{i_0.}} = \frac{P_{i_0 j}}{P_{i_0.}} \quad \textbf{(E1)}$$

e assim, a expressão da distância entre o ponto i_1 e outro qualquer em $\boldsymbol{R}^p$ é escrita da seguinte maneira:

$$d^2(i, i_1) = \sum_{j=1}^{p} \frac{1}{P_{.j}} \left(\frac{P_{ij}}{P_{i.}} - \frac{P_{i_1 j}}{P_{i_1.}} \right)^2$$

e, por **(E1)**, pode-se obter $d(i, i_1) = d(i, i_2) = d(i, i_0)$

Demonstração da invariância das distâncias em R^n.

Como:

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

$$= \sum_{i \neq i_1, i_2} P_{i.} \left(\frac{P_{ij}}{P_{i.}P_{.j}} - \frac{P_{ij'}}{P_{i.}P_{.j'}} \right)^2 + P_{i_1.} \left(\frac{P_{i_1 j}}{P_{i_1.}P_{.j}} - \frac{P_{i_1 j'}}{P_{i_1.}P_{.j'}} \right)^2 + P_{i_2.} \left(\frac{P_{i_2 j}}{P_{i_2.}P_{.j}} - \frac{P_{i_2 j'}}{P_{i_2.}P_{.j'}} \right)^2$$

para demonstrar a invariância da distância em $\boldsymbol{R}^n$ basta mostrar que:

$$P_{i_1.} \left(\frac{P_{i_1 j}}{P_{i_1.}P_{.j}} - \frac{P_{i_1 j'}}{P_{i_1.}P_{.j'}} \right)^2 + P_{i_2.} \left(\frac{P_{i_2 j}}{P_{i_2.}P_{.j}} - \frac{P_{i_2 j'}}{P_{i_2.}P_{.j'}} \right)^2 = P_{i_0.} \left(\frac{P_{i_0 j}}{P_{i_0.}P_{.j}} - \frac{P_{i_0 j'}}{P_{i_0.}P_{.j'}} \right)^2 \quad \textbf{(E2)}$$

Considerando que

$$\frac{P_{i_1 j}}{P_{i_1.}} = \frac{P_{i_2 j}}{P_{i_2.}} = \frac{P_{i_0 j}}{P_{i_0.}}$$

o primeiro membro da expressão **(E2)** será, então:

$$(P_{i_1.} + P_{i_2.})\left(\frac{P_{i_0 j}}{P_{i_0.}P_{.j}} - \frac{P_{i_0 j'}}{P_{i_0.}P_{.j'}}\right)^2$$

o que demonstra a igualdade **(E2)**, pois $N_{i_1.} + N_{i_2.} = N_{i_0.} \Rightarrow P_{i_1.} + P_{i_2.} = P_{i_0.}$.

2. *Equivalência distributiva em* $\boldsymbol{R}^n$.

Pela simetria existente entre a nuvem de pontos-variáveis e a nuvem de pontos-observações a demonstração anterior põe também em evidência que se em $\boldsymbol{R}^n$ dois pontos $\mathbf{w}_{j_1}$ e $\mathbf{w}_{j_2}$ se confundem, ou seja, se:

$$\frac{P_{ij_1}}{P_{.j_1}} = \frac{P_{ij_2}}{P_{.j_2}} \text{ para todo } i$$

sua substituição por um único $\mathbf{w}_{j_0}$ tal que $N_{ij_1} + N_{ij_2} = N_{ij_0}$ para todo i não modifica as distâncias entre pares de pontos nem em $\boldsymbol{R}^n$ nem em $\boldsymbol{R}^p$.

Esta propriedade de equivalência distributiva em $\boldsymbol{R}^n$ ou em $\boldsymbol{R}^p$ mostra que a substituição de modalidades (tanto em I como em J) que têm um perfil aproximado por uma única, não muda os resultados da análise, o que faz com que tais resultados sejam, em certa medida, independentes das modalidades que se estabeleçam para caracteres objeto de estudo.

3.3 Análise da nuvem de pontos-observação.

Esta análise consiste em uma ACP. O ponto de partida é o conjunto de elementos de I caracterizados pelos vetores $\mathbf{z}_i$ de um espaço vetorial $\boldsymbol{R}^p$ dotado de uma métrica euclidiana clássica. Por outra parte, como já foi assinalado, a cada ponto-observação se associará um peso $P_{i.}$.

A matriz de dados para a análise será:

$$\underset{p\times n}{\mathbf{Z}} = [\mathbf{z}_1 \cdots \mathbf{z}_i \cdots \mathbf{z}_N]$$

Antes de obter os eixos fatoriais a partir desta matriz de dados, apresentaremos as expressões da inércia com respeito à origem, da inércia explicada por uma direção e da matriz de inércia, para esse caso particular.

Inércia da nuvem de pontos I com respeito à origem é definida por:

$$\mathrm{H}(I,0) = \sum_{i=1}^{n} P_{i.} d^2(i,0) = \sum_{i=1}^{n} P_{i.} \sum_{j=1}^{p} z_{ij}^2 = \sum_i \sum_j P_{i.}\left(\frac{P_{ij}}{\sqrt{P_{.j}}\,P_{i.}} - \sqrt{P_{.j}}\right)^2 = \sum_i \sum_j \frac{\left(P_{ji} - P_{i.}P_{.j}\right)}{P_{i.}P_{.j}}$$

Esta expressão é a da estatística para o teste c^2 de independência e será nula se os caracteres I e J forem independentes.

Define-se a *Inércia da nuvem de pontos I explicada pela direção* $\mathbf{u}_1$ *de* $\boldsymbol{R}^{p.}$

$$\mathrm{H}(I,\mathbf{u}_1) = \sum_{i=1}^{n} P_{i.} F_{ij}^2$$

onde, como na ACP, F_{ij} é a projeção do ponto i sobre o eixo unitário $\mathbf{u}_1$ (que chamaremos simplesmente eixo $\mathbf{u}_1$). Assim:

$$F_{i1} = (\mathbf{z}_i)'\mathbf{u}_1.$$

Denominando $\mathbf{F}_1$ ao vetor cujas componentes dão as projeções dos elementos de I sobre $\mathbf{u}_1$, como na ACP, teremos:

$$\mathbf{F}_1 = (\mathbf{Z})'\mathbf{u}_1$$

utilizando-se a expressão da inércia explicada de I com relação à direção $\mathbf{u}_1$ resulta em:

$$\mathrm{H}(I,\mathbf{u}_1) = \mathbf{F}_1'\mathbf{P}_I\mathbf{F}_1$$

tendo:

$$\mathbf{P}_I = \begin{bmatrix} P_{1.} & 0 & \cdots & 0 \\ 0 & P_{2.} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ddots & P_{n.} \end{bmatrix}$$ a matriz dos perfis.

Com isso podemos notar que $\mathrm{H}(I,\mathbf{u}_1) = \mathbf{u}'_1\,\mathbf{Z}\mathbf{P}_I\mathbf{Z}'\mathbf{u}_1$ e definir a *matriz de inércia* $\mathbf{S}_I$ como sendo: $\mathbf{S}_I = \mathbf{Z}\mathbf{P}_I(\mathbf{Z})'$.

O termo geral desta matriz, $s_{jj'}$,é a covariância das variáveis $\mathbf{z}_j$ e $\mathbf{z}_{j'}$ que é a mesma que das variáveis $\mathbf{x}'_j$ e $\mathbf{x}'_{j'}$:

$$s_{jj'} = \sum_{i=1}^{n} P_{i.} z_{ij} z_{ij'} = \sum_{i=1}^{n} P_{i.} \left(\frac{P_{ij}}{\sqrt{P_{.j}}\,P_{i.}} - \sqrt{P_{.j}} \right) \left(\frac{P_{ij'}}{\sqrt{P_{.j'}}\,P_{i.}} - \sqrt{P_{.j'}} \right) = \sum_{i=1}^{n} \frac{P_{ij}P_{ij'}}{\sqrt{P_{.j}}\sqrt{P_{.j'}}\,P_{i.}} - \sqrt{P_{.j}}\sqrt{P_{.j'}}$$

A matriz de inércia $\mathbf{S}_I$ é, então, uma matriz de variâncias-covariâncias e tem as propriedades:

a) é uma matriz simétrica;

b) pelo fato de ser o produto de uma matriz por sua transposta, se trata de uma matriz definida não negativa e não possui autovalores negativos;

d) O traço da matriz (a soma de sua diagonal principal) é igual à inércia de $\boldsymbol{J}$ com respeito à origem (é a soma das variâncias das variáveis).

3.3.1 Eixos fatoriais.

Na abordagem da ACP, na busca dos eixos fatoriais e na projeção dos indivíduos sobre eles, foram encontradas relações que, nesta abordagem para a AFC, pode ser escrita como se segue:

$$\mathbf{S}_I \mathbf{u}_k = l_k \mathbf{u}_k .$$

Tal expressão evidencia que os eixos fatoriais $\mathbf{u}_k$ são na verdade vetores próprios da matriz de inércia $\mathbf{S}_I$.

Algumas propriedades dos vetores e valores próprios

1. O vetor $\mathbf{u}_0$, definido por:

$$\mathbf{u}_0 = \begin{bmatrix} \sqrt{P_{.1}} \\ \sqrt{P_{.2}} \\ \vdots \\ \sqrt{P_{.j}} \\ \vdots \\ \sqrt{P_{.n}} \end{bmatrix}$$

é um vetor próprio de $\mathbf{S}_I$ associado ao valor próprio $l_0 = 0$.

2. Os valores próprios de $\mathbf{S}_I$ são menores que 1.

Conseqüências dessas propriedades:

a) Chamando $\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_k, \cdots, \mathbf{u}_{p-1}$ aos *p* eixos fatoriais da nuvem de pontos *I,* associados aos valores próprios $l_0, l_1, \cdots, l_k, \cdots, l_{p-1}$ a condição de ortogonalidade dos vetores $\mathbf{u}_k$ implica em:

$$\mathbf{u}'_0 \mathbf{u}_k = \sum_{j=1}^{p} \sqrt{P_{.j}}\, u_{kj} = 0 \text{ para } k=1,2,..., p\text{-}1.$$

b) A inércia explicada por um eixo fatorial é igual, como na ACP, ao valor próprio associado ao eixo. Assim, a inércia explicada por $\mathbf{u}_0$ é nula, o que implica, como veremos, que todos os elementos de *I* se projetem sobre um só ponto do eixo $\mathbf{u}_0$.

c) A inércia com respeito à origem da nuvem de pontos-observações *I* será, como na ACP, igual à soma dos *p* valores próprios de $\mathbf{S}_I$. Sendo $l_0 = 0$, essa inércia terá a seguinte expressão:

$$\mathrm{H}(I,0) = \sum_{k=1}^{p-1} l_k .$$

A parte de inércia explicada (PHE) pelos *h* primeiros eixos fatoriais, será:

$$\mathrm{PHE}_h = \frac{\sum_{k=1}^{h} l_k}{\sum_{k=1}^{p-1} l_k} .$$

Demonstração da propriedade 1.

Se $\mathbf{u}_0$ é um vetor próprio de $\mathbf{S}_I$, associado a um valor próprio $l_0 = 0$, se verificará:

$$\mathbf{S}_I \mathbf{u}_0 = l_0 \mathbf{u}_0 = \mathbf{0}.$$

ou seja:

$$\sum_{h=1}^{P} s_{ij} u_{0h} = l_0 u_{0h} = 0 \text{ para } j=1, 2,..., p.$$

Porém, como:

$$\sum_{h=1}^{p} s_{jh} u_{0h} = \sum_{h=1}^{p} \sum_{i=1}^{n} P_{i.} z_{ij} z_{ih} u_{0h} = \sum_{i} \sum_{h} P_{i.} \left(\frac{P_{ij}}{\sqrt{P_{.j}} P_{i.}} - \sqrt{P_{.j}} \right) \left(\frac{P_{ih}}{\sqrt{P_{.h}} P_{i.}} - \sqrt{P_{.h}} \right) \sqrt{P_{.h}} =$$

$$= \sum_{i} P_{i.} \left(\frac{P_{ij}}{\sqrt{P_{.j}} P_{i.}} - \sqrt{P_{.j}} \right) \sum_{h} \left(\frac{P_{ih}}{P_{i.}} - P_{.h} \right) = 0$$

desde que :

$$\sum_{h} \left(\frac{P_{ih}}{P_{i.}} - P_{.h} \right) = 0, \text{ o que demonstra a propriedade.}$$

3.3.2 Cálculo prático dos eixos fatoriais.

Lembremos:

$$s_{jj'} = \sum_{i=1}^{n} \frac{P_{ij} P_{ij'}}{\sqrt{P_{.j}} \sqrt{P_{.j'}} P_{i.}} - \sqrt{P_{.j}} \sqrt{P_{.j'}},$$

Consideremos agora a matriz $\mathbf{Q}_I=[q_{jj'}]$ onde

$$q_{jj'} = \sum_{i=1}^{n} \frac{P_{ij} P_{ij'}}{\sqrt{P_{.j}} \sqrt{P_{.j'}} P_{i.}}$$

Esta matriz tem as seguintes propriedades:

1. Para todo $k \neq 0$ o vetor próprio $\mathbf{u}_k$ associado ao valor próprio λ_k da matriz $\mathbf{S}_I$ é também vetor próprio associado a um mesmo valor próprio da matriz $\mathbf{Q}_I$.

Demonstração:

Sejam $\mathbf{u}_k$ vetor próprio de $\mathbf{Q}_I$ associado ao valor próprio λ_k, então:

$$\mathbf{Q}_I \mathbf{u}_k = l_k \mathbf{u}_k \text{ e } \sum_{h=1}^{p} q_{jh} u_{kh} = \ddot{e}_k u_{kj}, \text{ para } j=1,2,.., p$$

desenvolvendo, temos:

$$\sum_{h=1}^{p} q_{jh} u_{kh} = \sum_{h} \left(\sum_{i} \frac{P_{ij} P_{ih}}{\sqrt{P_{.j}} \sqrt{P_{.h}} P_{i.}} - \sqrt{P_{.j}} \sqrt{P_{.h}} \right) u_{kh} =$$

$$= \sum_i \sum_h \frac{P_{ij} P_{ih}}{\sqrt{P_{.j}}\sqrt{P_{.h}} P_{i.}} - \sqrt{P_{.j}} \sum_h \sqrt{P_{.h}} u_{kh} ,$$

pela condição de ortogonalidade $\sum_h \sqrt{P_{.h}} u_{kh} = 0$ para k≠0. Logo os vetores $\mathbf{u}_k$, para k≠0, são também vetores próprios da matriz $\mathbf{Q}_I$ como definida acima.

2. O vetor próprio $\mathbf{u}_0$ de $\mathbf{S}_I$, que está associado ao valor próprio λ_0=0, é também vetor próprio de $\mathbf{Q}_I$, porém associado a um valor próprio λ_0=1.

Demonstração:
Considerando que para *j*=1,2,..., *p* se verifica:

$$\sum_{h=1}^{p} q_{jh} u_{0h} = \sum_i \sum_h \frac{P_{ij} P_{ih}}{\sqrt{P_{.j}}\sqrt{P_{.h}} P_{i.}} \sqrt{P_{.h}} = \sqrt{P_{.j}}$$

tem-se que $\mathbf{Q}_I \mathbf{u}_0 = \mathbf{u}_0$.

3. Chamando **A** à uma matriz *n* x *p*, cujos elementos a_{ij} têm a seguinte definição:

$$a_{ij} = \frac{P_{ij}}{\sqrt{P_{.j}}\sqrt{P_{i.}}}$$

pode ser verificada a seguinte igualdade: $\mathbf{Q}_I = \mathbf{A}\mathbf{A}'$.

Estas propriedades implicam em:

a) A propriedade 1 tem como conseqüência que os eixos fatoriais de *I* são os vetores próprios de $\mathbf{Q}_I$.
b) A segunda propriedade evidencia que a inércia total com respeito à origem da nuvem pontos-observações *I* será a soma de todos os valores próprios de $\mathbf{Q}_I$, à exceção do valor próprio λ_0=1, já que o vetor próprio $\mathbf{u}_0$ não contribui para a explicação da inércia.
c) A terceira propriedade será útil para a definição das relações de transição.

3.3.3 Componentes principais. Características.

Uma vez obtidos os eixos fatoriais $\mathbf{u}_k$ podemos representar, como na ACP os elementos de *I* mediante suas projeções sobre os eixos $\mathbf{F}_k$. Ou seja, poderemos caracterizar as observações a partir das novas *observações*, ou *variáveis*, (que são as componentes principais) $\mathbf{F}_k$.

São as seguintes, as características dessa novas variáveis:

a) O valor da variável $\mathbf{F}_k$ na observação *i*, F_{ki}, em função das freqüências e das componentes do eixo fatorial $\mathbf{u}_k$, é o seguinte:

$$F_{ik} = (\mathbf{z}_i)' \mathbf{u}_k = \sum_{j=1}^{p} \left(\frac{P_{ij}}{\sqrt{P_{.j}} P_{i.}} - \sqrt{P_{.j}} \right) u_i = \sum_{j=1}^{p} \frac{P_{ij}}{\sqrt{P_{.j}} P_{i.}} u_{kj} , \text{ para k≠0.}$$

Deve ser notado que $F_{0i}=0$ para todo *i*, pois $F_{0k} = (\mathbf{z}_i)'\mathbf{u}_0 = \sum_{j=1}^{p}\left(\frac{P_{ij}}{\sqrt{P_{.j}}P_{i.}} - \sqrt{P_{.j}}\right)\sqrt{P_{.j}}$. O que reafirma o fato de que, neste caso, todos os elementos de *I* se projetam em um só ponto demonstrando que $\mathbf{u}_0$ não contribui para explicar a inércia obtida.

b) A média de $\mathbf{F}_k$ é nula:

$$m_{\mathbf{F}_k} = \sum_{i=1}^{p} P_{i.}F_{ik} = \sum_i \sum_j P_{i.}\frac{P_{ij}}{\sqrt{P_{.j}}P_{i.}}u_{kj} = 0$$

c) A variância de $\mathbf{F}_k$ é a inércia explicada pelo eixo $\mathbf{u}_k$, λ_k:

$$v(\mathbf{F}_k) = \sum_i P_{i.}F_{ik}^2 = \mathbf{u}'_k\,\mathbf{S}_I\,\mathbf{u}_k = l_k$$

d) A correlação entre $\mathbf{F}_k$ e $\mathbf{F}_{k'}$ é nula (os fatores são ortogonais)

3.3.4 Estudo da dispersão dos pontos-observações.

Como na ACP, a dispersão dos pontos observação é estuda nos planos fatoriais definidos pelos fatores, como na figura abaixo.

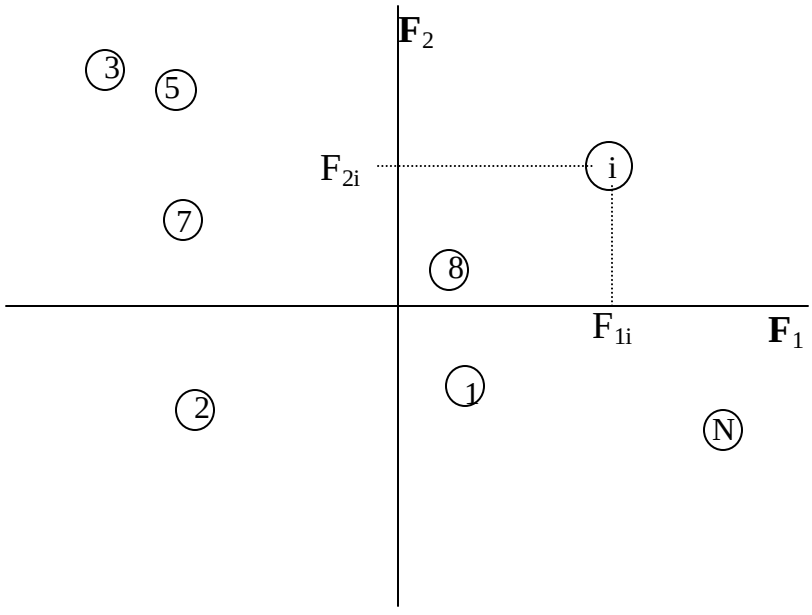


Figura 3.1 Dispersão dos pontos-observações no primeiro plano fatorial

No entanto alguns valores ajudam neste estudo.

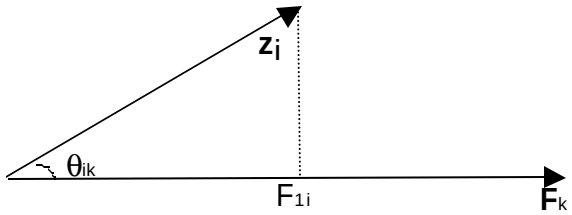
3.3.4.1 Contribuição absoluta da observação i ao eixo k.

Expressa a parte da inércia explicada pelo eixo de ordem *k* atribuída à observação *i*: se a inércia explicada pelo eixo $\mathbf{u}_k$ é $l_k = \sum_{i=1}^{n} P_{i.} F_{ik}^2$, então a contribuição absoluta de *i* ao eixo *k* será: $C_{ik}^{(a)} = \frac{P_{i.} F_{ik}^2}{l_k}$, e se vê que $\sum_{i=1}^{n} C_{ik}^{(a)} = \sum_{i=1}^{n} \frac{P_{i.} F_{ik}^2}{l_k} = 1$.

3.3.4.2 Contribuição relativa da observação i ao eixo k.

Se define como sendo $C_{ik}^{(r)} = \frac{F_{ik}^2}{d^2(i,0)}$, tendo em conta que $d^2(i,0) = \sum_{j=1}^{p} z_{ij}^2 = \sum_{k=1}^{p-1} F_{ik}^2$, então $\sum_{k=1}^{p-1} C_{ik}^{(r)} = \sum_{k=1}^{p-1} \frac{F_{ik}^2}{d^2(i,0)} = 1$

Figura 3.2 Projeção de um vetor variável sobre um fator e o ângulo, cujo cosseno determinará a



correlação entre a variável e o fator.

A contribuição relativa é o cosseno quadrado ($\cos C_{ik}^{(r)} = \cos^2 q_{ik}$) do ângulo formado pelo vetor $\mathbf{z}_i$ e o eixo $\mathbf{F}_k$. Quanto mais próximo a 1, mais próximo do eixo estará o ponto. Essa contribuição pode ser interpretada como um coeficiente de correlação entre a observação e o eixo. Os pontos que possuem uma forte contribuição relativa, sem ter uma contribuição absoluta importante, denominam-se *pontos ilustrativos do eixo.*

3.4 Representação simultânea ótima de pontos-observações e pontos-variáveis sobre um eixo.

Sejam F_{1i}, para i=1,2,..., n as projeções dos pontos-observações sobre o eixo 1, e sejam G_{1j}, para j=1,2,..., p as projeções dos pontos-variáveis sobre o mesmo eixo.

Definamos

$$F_{1i} = \sum_{j=1}^{p} \frac{P_{ij}}{P_{i.}} G_{1j} ,$$

como o centro de gravidade das projeções das variáveis, sendo estas ponderadas por seu peso na observação (à variável j será atribuído um peso $P_{ij} / P_{i.}$).

Seja, por outra parte, definido sobre o mesmo eixo, a variável j como centro de gravidade das projeções das observações F_{1i}, atribuindo a estas uma ponderação igual a seu peso na variável $P_{ij} / P_{.j}$:

$$G_{1j} = \sum_{i=1}^{n} \frac{P_{ij}}{P_{.j}} F_{1i} \ .$$

Em geral, é impossível que as expressões acima se verifiquem simultaneamente. Considera-se que a representação conjunta ótima de variáveis e observações é aquela que associa o mais possível, as coordenadas de variáveis e observações a F_{1i} e G_{1j}.

Dadas, então, as seguintes relações:

$$F_{1i} = c \sum_{j} \frac{P_{ij}}{P_{i.}} G_{1j} \ ,$$

$$G_{1j} = c \sum_{i} \frac{P_{ij}}{P_{.j}} F_{1i} \ .$$

trata-se de encontrar um valor de c>0 que mais se aproxime de 1.

Se conservamos, para a coordenada F_{i1}, o valor da projeção de i sobre o primeiro eixo fatorial, então:

$$F_{1i} = \sum_{j} \frac{P_{ij}}{\sqrt{P_{.j}}\, P_{i.}} U_{1j} \ ,$$

para que esta expressão seja igual ao valor obtido anteriormente:

$$F_{1i} = \frac{1}{\sqrt{l_1}} \sum_{j} \frac{P_{ij}}{P_{i.}} G_{1j} \ ,$$

o que define $c = 1/\sqrt{l_1}$ e a expressão para G_{1j}, que dá a representação ótima simultânea sobre o eixo de variáveis e de observações, será:

$$G_{1j} = \frac{U_{1j} \sqrt{l_1}}{\sqrt{P_{j.}}} .$$

Veremos que esta é a projeção do ponto-variável j sobre o primeiro eixo fatorial da nuvem de pontos $\boldsymbol{J}$.

Para demonstrar que $c = 1/\sqrt{l_1}$ é o maior valor próprio de $\boldsymbol{Q_I}$ distinto de 1, teremos dois pontos:

1. Considerando que, de um lado, $F_{1i} = c\sum_j \frac{P_{ij}}{P_{i.}} G_{1j}$ então F_{1i}/c é o centro de gravidade dos pontos G_{11}, G_{12},..., G_{1j},.... G_{1p}, em conseqüência, se tais pontos não coincidem:

$$\min_j G_{1j} < \frac{F_{1i}}{c} < \max_j G_{1j},$$

logo

$$\max_i \frac{F_{1i}}{c} < \max_j G_{1j}.$$

Por outro lado, de $G_{1j} = c\sum_i \frac{P_{ij}}{P_{i.}} F_{1i}$ tem-se que G_{1j}/c é o centro de gravidade dos pontos F_{11}, F_{21},..., F_{i1},.... F_{n1}, novamente se esses pontos não coincidem, teremos:

$$\min_i F_{1i} < \frac{G_{1j}}{c} < \max_i F_{1i},$$

logo

$\max_j \frac{G_{ij}}{c} < \max_i F_{1i}$, o que conduz a $\max_j G_{ij} < \max_i cF_{1i}$, resultando em:

$$\max_i \frac{F_{1i}}{c} < \max_j G_{1j} < \max_i cF_{1i},$$

o que implica que $c>1$; c será um se tanto os pontos F_{1i} como os G_{1j} se confundirem em um só ponto.

2. Nesta segunda etapa, mostraremos que $\ddot{e}_1 = 1/c^2$ é um valor próprio da matriz $\boldsymbol{Q_I}$, pelo que, para satisfazer ao ponto 1, $\ddot{e}_1$ será o maior valor próprio distinto de um, desta matriz.

Das expressões precedentes para F_{1i} e G_{1j} obtemos:

$$G_{1j} = c^2 \sum_{i=1}^{n} \frac{P_{ij}}{P_{.j}} \sum_{h=1}^{p} \frac{P_{ih}}{P_{i.}} G_{1h}.$$

Fazendo a seguinte mudança de variável:

$D_{1h} = \sqrt{P_{.h}}\, G_{1h}$, tem-se:

$$\frac{D_{1j}}{\sqrt{P_{.j}}} = c^2 \sum_h \sum_i \frac{P_{ij} P_{ih}}{P_{.j} P_{i.} \sqrt{P_{.h}}} D_{1h}$$

ou seja:

$$\frac{1}{c^2} D_{1j} = \sum_h \sum_i \frac{P_{ij} P_{ih}}{\sqrt{P_{.j}} P_{i.} \sqrt{P_{.h}}} D_{1h} \quad \textbf{(E3)}$$

Se dissermos que:

$$\mathbf{D_1} = \begin{bmatrix} D_{11} \\ D_{12} \\ \vdots \\ D_{1j} \\ \vdots \\ D_{1p} \end{bmatrix}$$

a expressão **(E3)** corresponderá à *j*-ésima equação do seguinte sistema:

$$\mathbf{Q_I D}_1 = \frac{1}{c^2} \mathbf{D}_1$$

onde $1/c^2$ é um valor próprio da matriz $\boldsymbol{Q_I}$.

3.5 Análise da nuvem de pontos-variáveis (análise dual).

Como para a nuvem ***I***, a análise da dispersão e das relações entre os elementos da nuvem ***J*** das variáveis, é feita a partir de uma ACP desta nuvem.

O ponto partida desta análise é a matriz de dados

$$\boldsymbol{W} = \left[\mathbf{w}_1 \mathbf{w}_2 \cdots \mathbf{w}_j \cdots \mathbf{w}_p\right]$$

na qual os vetores $\mathbf{w}_j$ representam os pontos-variáveis pertencentes a um espaço $\mathbf{R}^n$ dotado de uma métrica euclidiana clássica. Esses pontos estão associados a um peso $P_{.j}$.

A simetria existente entre o conjunto ***I*** e o conjunto ***J*** na AFC, faz com que as expressões já obtidas para o caso da nuvem ***I***, possam ser utilizadas neste caso simplesmente invertendo-se os índices.

Assim, nos limitaremos a recapitular conceitos e expressões já visto para aquele caso.

*Inércia da nuvem **J** com respeito a origem:* $H(\boldsymbol{J},0) = \sum_j \sum_i \frac{\left(P_{ij} - P_{i.} P_{.j}\right)^2}{P_{i.} P_{.j}}$

*Inércia da nuvem **J** com respeito a origem.*

Chamando $G_{1j} = \left(\mathbf{w}_j\right) \mathbf{v}_1$ à projeção da variável *j* sobre o eixo unitário $\mathbf{v}_1$, então a inércia explicada por esse vetor, será: $H(\boldsymbol{J}, \mathbf{v}_1) = \sum_j P_{.j} G_{1j}^2$.

Definindo adequadamente a $\mathbf{G}_1$, pode-se chegar a $H(\boldsymbol{J}, \mathbf{v}_1) = \mathbf{G}'_1 \mathbf{P}_j \mathbf{G}_1$

Se a matriz de inércia se escreve como $\mathbf{S}_J = \mathbf{w}\mathbf{P}_J\mathbf{w}'$ então a inércia explicada pode ser escrita como: $H(J, \mathbf{v}_1) = \mathbf{v}'_1 \mathbf{S}_J \mathbf{v}_1$.

Os eixos fatoriais $\mathbf{v}_k$ da nuvem J serão os vetores próprios de $\mathbf{S}_J$ expressos na equação: $\mathbf{S}_J \mathbf{v}_k = \ddot{e}_k \mathbf{v}_k$.

A simetria entre as nuvens de dados conduz a que as propriedades da matriz de inércia associada à nuvem de indivíduos sejam análogas às da nuvem de pontos-variáveis. Ou seja:

- O valor próprio $\ddot{e}_0$ associado ao vetor próprio $\mathbf{v}_0$ tem valor 0, o que implica que

$$\sum_{i=1}^{n} v_{ki}\sqrt{P_{i.}} = 0 \text{ para todo } k \quad 0.$$

- Os valores de $\mathbf{S}_J$ são menores que 1.
- A matriz $\mathbf{Q}_J$ tem os mesmos valores próprios e vetores próprios $\mathbf{v}_k$ que $\mathbf{S}_J$, para k=1, 2,..., n-1.
- O vetor $\mathbf{v}_0$ é um vetor próprio de $\mathbf{Q}_J$ associado ao valor próprio igual a 1.
- A matriz $\mathbf{Q}_J$ pode ser colocada em função da matriz $\mathbf{A}$, anteriormente definida. Essa propriedade põe em evidência que o número de vetores próprios associados a valores próprios não nulos da matriz $\mathbf{Q}_J$ será o mesmo que o da matriz $\mathbf{Q}_I$=**AA'**, já que ambas são simétricas e têm o mesmo posto.

3.5.1 Característica das variáveis G_k e estudo da dispersão dos pontos-varáveis.

Novamente, a simetria existente entre observações e variáveis na AFC leva consigo que as características das variáveis $\boldsymbol{G}_k$ (componentes principais da nuvem $\boldsymbol{J}$) sejam análogas às de $\boldsymbol{F}_k$, ou seja:

1. As variáveis $\boldsymbol{G}_k$ são combinações lineares das variáveis iniciais. Neste caso, a expressão da projeção da variável j no eixo k será, para k 0:

$$G_{kj} = \mathbf{w}'_j \mathbf{v}_k = \sum_i \left(\frac{P_{ij}}{\sqrt{P_{i.}}P_{.j}} - \sqrt{P_{i.}} \right) v_{ki} = \sum_i \frac{P_{ij}}{\sqrt{P_{i.}}P_{.j}} v_{ki}$$

pois como já foi visto $\sum_{i=1}^{n} v_{ki}\sqrt{P_{i.}} = 0$ para k 0

O valor de G_{0j} será nulo para todo j, posto que os elementos de $\mathbf{v}_0$ são $\sqrt{P_{i.}}$ e em conseqüência

$$G_{kj} = \sum_i \left(\frac{P_{ij}}{\sqrt{P_{i.}}P_{.j}} - \sqrt{P_{i.}} \right) \sqrt{P_{i.}} = 0.$$

2. A média das variáveis $\boldsymbol{G}_k$ é nula e sua variância é igual à inércia explicada pelos eixos $\mathbf{v}_k$.
3. As varáveis $\boldsymbol{G}_k$ são não correlacionadas duas a duas.

Por outra parte, a dispersão e as relações entre os pontos do conjunto ***J*** se estudam da mesma forma como as dos pontos do conjunto ***I***: mediante as projeções dos pontos em planos fatoriais, utilizando-se como ajudas para a interpretação as contribuições absolutas e relativas das variáveis aos eixos, que neste caso serão:

Contribuição absoluta da variável j ao eixo k:

$$C_{kj}^{(a)} = \frac{P_{.j} G_{kj}^2}{\ddot{e}_k}$$

com

$$\sum_{j=1}^{p} C_{kj}^{(a)} = 1$$

Contribuição relativa da variável j ao eixo k:

$$C_{kj}^{(r)} = \frac{G_{kj}^2}{d^2(j,0)}$$

com

$$\sum_{k=1}^{p-1} C_{kj}^{(r)} = 1$$

3.6 Relações de transição.

Como no caso da ACP, as relações de transição permitem obter os eixos fatoriais da nuvem de pontos ***J*** (nuvem de pontos ***I***) e coordenadas dos elementos de ***J*** (elementos de ***I***) sobre ditos eixos, em função dos eixos fatoriais do conjunto ***I*** (conjunto ***J***), o que evita ter que efetuar duas ACP, sobre cada nuvem. Estas relações de transição nos permitem, por outra parte, ressaltar as propriedades baricêntricas das projeções dos pontos-variáveis e dos pontos-observações. Estas propriedades permitem a representação simultânea de variáveis e observações, que não está plenamente justificada na ACP.

As relações de transição na AFC são as seguintes:

$$\begin{cases} \mathbf{v}_k = \dfrac{1}{\sqrt{\ddot{e}_k}} \mathbf{A}' \mathbf{u}_k \\ \mathbf{u}_k = \dfrac{1}{\sqrt{\ddot{e}_k}} \mathbf{A}' \mathbf{v}_k \end{cases}$$

Se $\mathbf{u}_k$ é um vetor próprio de $\mathbf{Q_I} = \frac{1}{n}\mathbf{AA}'$ associado ao valor próprio $\ddot{e}_k$ diferente de zero, se verifica :

$$\frac{1}{n}\mathbf{AA}'\mathbf{u}_k = \ddot{e}_k\mathbf{u}_k$$

pré-multiplicando por **A**', temos:

$$(\frac{1}{n}\mathbf{A'A})\,(\mathbf{A'u}_k) = \ddot{e}_k(\mathbf{A'u}_k)$$

logo $\mathbf{A'u}_k$ é um vetor próprio de $\mathbf{Q_I}$ associado ao valor próprio $\ddot{e}_k$.

Um vetor unitário $\mathbf{v}_k$ que tenha a mesma direção de $\mathbf{A'u}_k$ será o eixo fatorial de ***J***, que explicará uma inércia igual a $\ddot{e}_k$, logo:

$$\mathbf{v}_k = c\mathbf{A'v}_k$$

onde *c* é uma constante tal que se verifique $\mathbf{v'}_k\mathbf{v}_k=1$, ou seja:

$$c^2\,\mathbf{v'}_k\mathbf{AA'v}_k=1$$

como já vimos que:

$$c = 1/\sqrt{l_k}$$

logo:

$$\mathbf{v}_k = \frac{1}{\sqrt{l_k}}\mathbf{A}'\mathbf{u}_k$$

A equação para $\mathbf{u}_k$ é obtida de maneira análoga.

Como conseqüência dessas relações de transição, as coordenadas F_{ki} das observações podem ser obtidas em função dos eixos fatoriais $\mathbf{v}_k$ de ***J***, e as coordenadas G_{kj} das variáveis em função dos eixos fatoriais de ***I***. Em efeito:

$$F_{ki} = \sqrt{\frac{l_k}{P_{i.}}}v_{ki} \quad \textbf{(E4)}$$

$$G_{kj} = \sqrt{\frac{l_k}{P_{.j}}}u_{kj} \quad \textbf{(E5)}$$

Demonstração.

Recordando que:

$$a_{ij} = \frac{P_{ij}}{\sqrt{P_{.j}}\sqrt{P_{i.}}}$$

e que

$$F_{ki} = \sum_{j=1}^{p} \frac{P_{ij}}{\sqrt{P_{.j}}\sqrt{P_{i.}}} u_{kj}$$

a equação (E4) pode ser obtida do valor para $\mathbf{v}_{ki}$:

$$\mathbf{v}_{ki} = \frac{1}{\sqrt{l_k}} \sum_{j=1}^{p} a_{ij} u_{kj} = \frac{1}{\sqrt{l_k}} \sum_{j=1}^{p} \frac{P_{ij}}{\sqrt{P_{.j}}\sqrt{P_{i.}}} u_{kj} = \frac{\sqrt{P_{i.}}}{\sqrt{l_k}} \sum_{j=1}^{p} \frac{P_{ij}}{\sqrt{P_{.j}}\sqrt{P_{i.}}} u_{kj} = \frac{\sqrt{P_{i.}}}{\sqrt{l_k}} F_{ki}$$

De modo similar pode ser deduzida a equação para $\mathbf{u}_{kj}$, partindo-se do valor de G_{kj} em (E5).

Pelas equações obtidas para F_{ki} e G_{kj} pode ser mostrado como as projeções dos pontos-observações ficam próximos do centro de gravidade dos pontos-variáveis e vice-versa.

Partindo-se da expressão anterior para G_{kj} e substituindo $\mathbf{v}_{ki}$ por (E4), obtém-se:

$$G_{kj} = \sum_{i} \frac{P_{ij}}{\sqrt{P_{i.}}P_{.j}} v_{ki} = \frac{1}{\sqrt{l_k}} \sum_{i} \frac{P_{ij}}{P_{.j}} F_{ki}$$

De modo análogo, obtém-se:

$$F_{ki} = \frac{1}{\sqrt{l_k}} \sum_{j} \frac{P_{ij}}{P_{i.}} G_{kj}$$

As equações acima têm formas análogas às obtidas na busca da representação ótima das coordenadas. Característica que não existe na ACP.

Esta propriedade de "ótimo" juntamente com a de equivalência distributiva, constituem as vantagens mais notáveis da AFC diante da ACP.

3.7 Representação de observações e variáveis suplementares.

Em algumas circunstâncias tais como quando existe um grande número de elementos nas nuvens ***I*** ou ***J***, ou quando se deseja ressaltar a presença de uma certas características em relação às demais etc., é conveniente, para facilitar a interpretação, realizar a AFC considerando somente uma parte dos elementos dos conjuntos (aqueles que se deseja dar relevância). Os elementos excluídos da obtenção dos eixos podem ser posteriormente projetados sobre tais eixos e, como já foi visto, se denominam observações ou variáveis suplementares.

As equações acima proporcionam as expressões para obter as mencionadas projeções:

- A projeção do ponto suplementar i_0 da nuvem ***I*** sobre o fator k, será:

$$F_{ki_0} = \frac{1}{\sqrt{l_k}} \sum_{j=1}^{p} \frac{N_{i_0 j}}{N_{i_o .}} G_{kj}$$

- e a projeção do ponto suplementar j_0 da nuvem $\boldsymbol{J}$ sobre o fator k, será:

$$G_{kj_0} = \frac{1}{\sqrt{l_k}} \sum_{i=1}^{n} \frac{N_{ij_0}}{N_{. j_0}} F_{ki}$$

Desta forma é possível num mesmo plano fatorial compor variáveis sintéticas que permitem tipificar as relações entre os dados e observar o comportamento daquelas tratadas como suplementares diante das que dão a forma da análise dos dados originais.

Capítulo 4
Análise Fatorial de Correspondências Múltiplas
Noções e conceitos Teóricos

(Extraído de Escofier, B. e Pagès, j., 1998
Extraído de Judez, L., 1988)

4.1 Introdução.

4.1.1 Os Dados.

A Análise de Fatorial Correspondências Múltiplas (AFCM) permite estudar uma população de *I* indivíduos descritos por *J* variáveis qualitativas.

Uma variável qualitativa (ou nominal) é uma aplicação de um conjunto de *I* indivíduos num conjunto finito sobre o qual não definimos uma estrutura particular: Por exemplo, um conjunto de três cores (azul, branca e vermelha). Os elementos desse conjunto são chamados modalidades da variável e dizemos, por exemplo, que um indivíduo azul possui a modalidade *azul*.

A aplicação mais comum da AFCM aparece no tratamento de dados de resposta a uma enquête. Cada questão constitui uma variável na qual as modalidades são respostas a propostas (entre as quais de que o entrevistado deve fazer apenas uma escolha).

Faremos uma revisão de diferentes formas de transcrever numericamente o conjunto desses dados.

4.1.2 Codificação condensada.

Esses dados podem ser rearranjados uma tabela do tipo *Indivíduos*x*Variáveis* em tudo análoga àquela estudada na ACP. As linhas representam indivíduos (ou observações), as colunas representam as variáveis: na interseção entre a linha *i* e a coluna *j* encontramos o valor x_{ij} (diremos também a codificação condensada) do indivíduo *i* devido à variável *j*. Geralmente, x_{ij} indica em *i* o número da modalidade (variável *j*).

Naturalmente, os valores x_{ij} são codificações que não possuem propriedades numéricas. Se a variável *j* é a cor dos indivíduos, essa cor pode ser codificada assim: azul=1; branca=2; vermelha=3. É claro que a média de azul e vermelha não faz sentido e não pode ser considerada como branca. Parece, então, ser impossível tratar tal conjunto de dados pela ACO (ou AFC). As tabelas *Indivíduos*x*Variáveis qualitativas* constituem-se de suas especificidades e devem ser tratadas com um método específico.

4.1.3 Tabela Disjunta Completa.

Uma outra forma de representar os mesmos dados é pela construção de uma Tabela Disjunta Completa (TDC). Nessa tabela, as linhas representam os indivíduos e as colunas as modalidades das variáveis: o valor x_{ij} que representa a interseção da linha *i* com a coluna

j, será igual a 1 ou 0, segundo o indivíduo *i* possua a modalidade *j*, ou não. A origem da terminologia Tabela Disjunta Completa é a seguinte: o conjunto de valores x_{ij} de um mesmo indivíduo, para as modalidades de uma mesma variável, comporta o valor 1 uma vez (completa) e somente uma vez (disjunta).

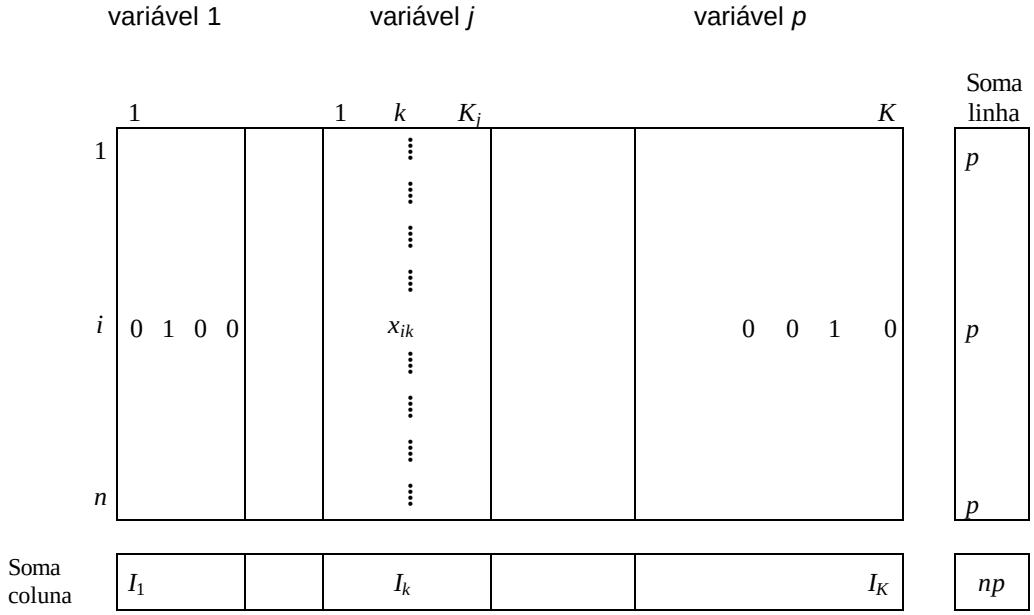


Figura 4.1 Tabela de dados sob a forma disjunta completa. Ali se tem:

K_j = número e conjunto das modalidades da variável j;

$K=\sum_{j=1}^{j=p} K_j$ é número e conjunto das modalidades de todas as variáveis indistintamente;

x_{ik}=1 se o indivíduo i possui a modalidade k, 0 senão;

$\sum_{k=1}^{k=K_j} x_{ik} = 1$ para todo (i,j) $\sum_{k=1}^{k=K} x_{ik} = p$ para todo i

$\sum_{i=1}^{i=I} x_{ik} = I_k$ para todo k $\sum_{k=1}^{k=K_j} I_k = n$ para todo j

As colunas dessa tabela têm funções numéricas definidas sobre o conjunto dos indivíduos, são chamadas indicatrizes da modalidade (obs: indicamos por uma indicatriz a variável que tem maior peso, pela qual os indivíduos podem ser caracterizados).

4.1.4 Tabela de Burt.

Pode-se fazer uma generalização da análise de correspondências no estudo de mais de duas variáveis, construindo uma tabela contendo o conjunto de tabelas de contingência entre as variáveis duas a duas. A Tabela de Burt não é exatamente uma tabela de contingência, porém uma justaposição de tais tabelas; cada indivíduo aparece J^2 vezes. As tabelas cruzam na diagonal as variáveis com elas mesmas: elas apresentam valores iguais a zero fora da diagonal que, por sua vez, contêm os efetivos das modalidades.

4.2 Objetivos.

A problemática da AFCM se parece com aquela da ACP mas pode ser também considerada uma generalização de AFC. Esses dois aspectos estão mais ou menos presentes nos objetivos de análise da AFCM apresentados aqui em três famílias de objetos que intervêm na AFCM: os indivíduos, as variáveis e as modalidades das variáveis.

4.2.1 Estudo dos indivíduos.

De maneira análoga à ACP, um dos objetivos da AFCM é realizar uma tipologia dos indivíduos. Essa tipologia se apóia numa noção de semelhança tal que dois indivíduos serão bastante próximos se possuírem um grande número de modalidades em comum.

Por outro lado, em grande parte das aplicações de AFCM, os indivíduos são muito numerosos e não são conhecidos, senão pelas suas características, presentes na tabela de dados. Por exemplo, numa pesquisa de opinião, não dispomos para cada indivíduos de qualquer outro conhecimento que não aquele presente nas respostas. Nesse caso, os indivíduos são estudados através de classes definidas pelas variáveis. Assim, numa enquête de opinião, nos interessamos, por exemplo, pelas mulheres, pelos jovens, pelos aposentados etc. Uma análise dos indivíduos por suas classes deve ser tal que duas classes se assemelharão tanto mais quanto os perfis das partições do conjunto de modalidades sejam próximos.

4.2.2 Estudo das variáveis.

Procedendo da mesma forma que na ACP, podemos adotar dois pontos de vista no estudo das variáveis.

O primeiro é aquele de realizar um inventário das ligações entre as variáveis. O estudo das relações entre duas variáveis qualitativas necessita levar em consideração a tabela cruzada de suas modalidades. Um levantamento pouco detalhado das relações implica, então, em se situar no nível das modalidades, aquém daquele das variáveis.

O segundo consiste em expressar o conjunto de variáveis originais (qualitativas) por um pequeno número de variáveis numéricas. Por exemplo, podemos buscar expressar um conjunto de variáveis sócio-profissionais por um indicador de *status social*. O interesse nessas variáveis sintéticas provém do fato delas serem ligadas às variáveis estudadas. Assim, uma variável somente poderá ser considerada como um indicador de *status social* se estiver ligada, ao mesmo tempo, à categoria sócio-profissional, a *diploma* etc.

4.2.3 Estudo das modalidades.

Estudar o conjunto das modalidades corresponde a realizar um levantamento de seus relacionamentos. Onde uma modalidade pode ser considerada segundo dois pontos de vista:

- Uma variável indicatriz, correspondente a uma coluna, que abarca os indivíduos, da Tabela Disjunta Completa;
- Uma classe de indivíduos em que a repartição sobre o conjunto de modalidades, seja uma linha ou coluna na Tabela de Burt.

A noção de semelhança que se estabelece entre as modalidades difere segundo o ponto de vista adotado. No primeiro caso, as semelhanças entre duas modalidades devem repousar sobre sua associação mútua: duas modalidades se assemelham quanto mais elas estiverem presentes, ou ausentes, simultaneamente em um grande número de indivíduos. As outras modalidades não intervirão.

No segundo caso, a semelhança entre duas modalidades é análoga àquela utilizada na tabela de freqüências. Uma linha da Tabela de Burt caracteriza a associação da modalidade com as modalidades de todas as variáveis: duas modalidades se assemelham quanto forte, ou fraca, é sua associação às mesmas modalidades.

4.2.4 Conclusão sobre os objetivos.

O estudo de uma tabela *Indivíduos*x*Variáveis qualitativas* pões em jogo três famílias de objetos: indivíduos, variáveis e modalidades. Resulta numa problemática mais rica e complexa que a tipologia clássica: tipologia de linhas, tipologia das colunas e as relações entre as duas tipologias. Essa riqueza não deve, entretanto, deixar esquecer a unicidade da tabela: ela ao pode ser um problema para o estudo separado dos diferentes aspectos da problemática por métodos sem relação entre eles. Praticamente, essa unicidade é verificada ao se articular as interpretações em torno da tipologia das modalidades. Em efeito, essa tipologia permite estudar a associação mútua entre as modalidades, ou seja, as ligações entre os pares de variáveis. Ela permite abordar os indivíduos aos examinar o comportamento médio de classes de indivíduos.

4.3 A ACF aplicada a uma Tabela Disjunta Completa.

4.3.1 AFCM e AFC.

Embora a AFC, como método de análise de tabelas de freqüência, não seja apropriada a aplicação a TDC, os cálculos que são realizados pelos programas de AFC, podem ser aplicados a essas tabelas. Porém, dependendo do caso, esses cálculos devem ser reinterpretados em função da natureza particular da tabela. Esses cálculos, munidos dessa nova interpretação, constituem um método à parte, conhecido por Análise de Correspondências Fatorial Múltiplas (em alguns casos Análise de Correspondências Múltiplas) . A AFC de uma TDC não é mais que uma forma prática de realizar os cálculos, de resto incompleto porque ignora a noção de variável e não lhes fornece algum resultado que as concerne.

Dessa forma, seguimos esse caminho histórico e cômodo para apresentar a Análise de Correspondências Múltiplas.

Uma TDC possui não somente uma natureza distinta daquela de uma tabela de contingência (se codifica os dados diferentemente) mas também das propriedades numéricas particulares. As mais importantes são:

- os valores na tabela são 0 ou 1;
- as coluna podem ser reagrupadas por pacotes (que correspondem cada um a uma variável) onde o resultado é uma coluna composta de 1;

- a soma dos números de uma mesma linha é constante e igual a *J*, número total de variáveis.

As seções seguintes mostram que as distâncias, os pesos e os fatores da AFC de uma TDC corresponde aos objetivos previamente fixados.

4.3.2 Nuvem de indivíduos

A margem sobre I é constante, a transformação em perfis linha praticamente não modifica os dados. Um indivíduo é representado pelas modalidades que possui. Dois indivíduos se assemelham se apresentam globalmente as mesmas modalidades. Mais precisamente, a distância entre dois indivíduos *i* e *i é* definida por:

$$d^2(i,i') = \sum_k \frac{np}{I_k}\left(\frac{x_{ik}}{p} - \frac{x_{i'k}}{p}\right)^2 = \frac{1}{p}\sum_k \frac{n}{I_k}\left(x_{ik} - x_{i'k}\right)^2$$

A expressão $(x_{ik} - x_{lk})^2$ vale 0 ou 1 e não é mais que 0, senão para aquelas modalidades *k* presentes em um só dos dois indivíduos considerados. A distância d(*i,l*) cresce com o número de modalidades que diferem para os indivíduos *i* e *l* (o que é lógico!). Uma modalidade *k* intervém nessa distância com pesos N/I_k, que são o inverso de sua freqüência. A presença de uma modalidade rara distancia seu ou seus possessores de todos os outros indivíduos.

A distância induzida pela AFC aplicada a uma TDC é então satisfeita. O peso correspondente a cada um indivíduo é o mesmo (pelo fato da margem ser constante).

4.3.3 Nuvem das modalidades

A modalidade *k* é representada pelo perfil da coluna *k*. Os número da TDC não são diferentes de 0 ou 1, o perfil da coluna *k* somente poderá ter, por sua vez, os valores 0 ou $1/I_k$. Em outras palavras, o centro de gravidade da nuvem de modalidades, que se confunde com o perfil da margem sobre *I,* é caracterizado por um perfil perfeitamente plano. Resulta que o perfil da coluna *k* se assemelha tanto mais ao perfil médio quanto maior é o número de indivíduos possuidores da modalidade *k*. Reciprocamente, uma modalidade rara se assemelhará o menos possível do centro de gravidade da nuvem de modalidades.

A distância entre duas modalidades *k* e *h* é definida por:

$$d^2(k,h) = \sum_i n\left(\frac{x_{ik}}{I_k} - \frac{x_{ih}}{I_k}\right)^2$$

Utilizando o fato de que $(x_{ik})^2 = x_{ik}$ e desenvolvendo o termo quadrado, obtemos:

$d^2(k,h) = \dfrac{n}{I_k I_h}$ *número de indivíduos que possuem uma e somente uma das modalidades h e k.*

Essa distância cresce com o número de indivíduos que possuem uma e somente uma das duas modalidades *h* e *k*, e decresce com o efetivo de cada uma dessas modalidades. Duas modalidades de uma mesma variável são obrigatoriamente bastante distante uma da

outra no espaço. Duas modalidades possuídas pelo mesmo indivíduo se confundem. As modalidades raras estarão afastadas de todas as outras. Essa distância traduz bem o primeiro dos dois pontos de vista sobre a semelhança entre modalidades indicados nos objetivos.

Ao aplicar esse cálculo à distância entre uma modalidade k e o centro de gravidade G_k da nuvem de modalidades (equivalente à uma modalidade possuída por todos os indivíduos), encontramos: $d^2(k, G_k) = (n / I_k) - 1$; que especifica a influência dos efetivos de uma modalidade sobre sua distância ao ponto médio.

O peso da modalidade k vale I_k / np; o que é proporcional ao efetivo I_k.

Observações

Um elemento (linha ou coluna) influencia a construção dos eixos por intermédio de sua inércia com relação ao centro de gravidade. Um cálculo simples resulta:

$$\text{Inércia de } k \text{ com relação à} G_k = \frac{1}{p}(1 - \frac{I_k}{n})$$

Esse resultado mostra que, sob influência de uma modalidade rara, o pequeno peso será suficiente para compensar seu distanciamento. Por exemplo, uma modalidade presente em apenas 1% da população possui uma inércia (ou exercerá uma influência) duas vezes maior que uma modalidade presente em 50% a população. Concretamente isso significa que é costumeiro ver os primeiros eixos fatoriais de uma AFCM determinados quase que exclusivamente por essas modalidades bastante raras repartidas pelos mesmos indivíduos. Como, freqüentemente, será bem mais interessante observar os fenômenos de uma maneira geral, buscamos, na prática, evitar essas modalidades raras.

Pela soma das inércias das modalidades, mostramos facilmente que a inércia total da nuvem estudada vale $(K / p) - 1$. Na AFCM, como na ACP, mas diferentemente da AFC, a inércia total das nuvens não intervém na interpretação.

A inércia das K_j modalidades da variável j vale $(K_j - 1) / p$. Essa inércia, estando ligada diretamente ao número de modalidades da variável j, leva a exigir que o número de modalidades seja igual para todas as variáveis ativas. De fato, essa diferença de inércia entre as variáveis que possuem número diferente de modalidades vale para todo o espaço $\mathbf{R}^n$. Desde o momento em que consideramos apenas uma direção de $\mathbf{R}^n$, que é o caso dos eixos fatoriais, a inércia da nuvem de K_j modalidades de uma mesma variável j é certamente inferior a 1/p, quantidade que não depende de K_j. Isso resulta em que não será danoso, realizar uma intervenção simultânea nos efetivos das variáveis que possuem números distintos de modalidades.

4.3.4 Relações de transição e representação simultânea.

Com lãs notações já utilizadas na ACP e na AFC, as relações de transição da AFC aplicadas à uma TDC, se escrevem como a seguir:

$$F_s(i) = \frac{1}{\sqrt{l_s}} \sum_{k \in K} \frac{x_{ik}}{p} G_s(k)$$

$$G_s(k) = \frac{1}{\sqrt{l_s}} \sum_{i \in I} \frac{x_{ik}}{I_k} F_s(i)$$

Do fato de que x_{ik} não possui valores outros que 0 ou 1, essas relações de transição têm uma interpretação simples. Na projeção sobre o eixo *s*, o indivíduo *i* é colocado, com coeficiente $1/\sqrt{l_s}$ próximo do baricentro das modalidades que ele possui. Inversamente, a modalidade *k* é colocada, com coeficiente $1/\sqrt{l_s}$ próximo do baricentro dos indivíduos que a possuem. Disso resulta que, sobre um eixo, uma modalidade (coluna da TDC) representa uma dilatação próxima da média dos indivíduos que a possuem. Também, no estudo de sua projeção, podemos considerar uma modalidade como o baricentro de uma classe de indivíduos e como a indicatriz de uma variável. O coeficiente de dilatação varia com os eixos, o que não é incômodo pois a interpretação dos resultados se faz fator a fator e prima por examinar conjuntamente as preferências dos eixos inerciais comparáveis (princípio comum à todas as análises fatoriais).

Essa equivalência deve ser feita sem que se esqueça que as modalidades, de uma parte enquanto indicatrizes e de outra parte enquanto baricentros, estão situadas em espaços diferentes. Disso resulta que as qualidades de representação de uma mesma modalidade segundo cada ponto de vista não são relacionadas. Em outras palavras, a noção de proximidade entre esses dois tipos de objetos, difere.

Em efeito, a proximidade entre indicatrizes dá a medida de sua associação mútua. De outra parte, a proximidade das médias das classes dos indivíduos se depreende das distâncias definidas entre os indivíduos: duas classes de indivíduos *k* e *h* estarão próximas daquelas com características idênticas no que diz respeito ao conjunto de variáveis, ou seja, as modalidades *k* e *h* se associam da mesma maneira às modalidades de todas as variáveis. Essa noção de proximidade corresponde ao segundo ponto de vista sobre as semelhanças entre modalidades listado nos objetivos. Observamos que, exceto as dilatações nos eixos, as duas noções de proximidade, fundamentadas em princípios diferente, conduzem aos mesmos gráficos na análise de uma TDC.

Na prática as duas noções de proximidade se utilizam conjuntamente; em particular, interpretamos freqüentemente a proximidade entre modalidades de variáveis diferentes tanto quanto a associação de modalidades e a proximidade entre modalidades de uma mesma variável, como do ponto de vista da semelhança das classes de indivíduos. Por exemplo, ao descrever um plano fatorial no qual aparecem diferente aspectos sociais, interpretamos a proximidade entre as duas modalidades seguintes, ***aposentados*** e ***mais de 60 anos***, em termos de associação (serão próximos os mesmos indivíduos que possuem essas modalidades) e a proximidade entre ***60 a 65 anos*** e ***mais de 65 anos*** em termos de semelhança (as duas classes de indivíduos possuirão características idênticas no que diz respeito às outras variáveis). Assim, as relações de transição, mesmo se não são utilizadas no marco estrito de uma representação simultânea, conferem à representação das modalidades as propriedades esperadas, emanadas na exposição dos objetivos.

4.3.5 As variáveis através de suas modalidades.

As variáveis qualitativas não são introduzidas explicitamente numa AFC de uma TDC. Elas aparecem nada mais que através do conjunto de suas modalidades. As *subnuvens* de modalidades de uma mesma variável têm propriedades que são interessantes de compreender para interpretar os resultados, mas também para codificar quaisquer variáveis em vias de serem tratadas como qualitativas numa AFCM.

4.3.5.1 Baricentro das modalidades de uma variável.

Como mostra a relação abaixo, o baricentro das modalidades de uma variável se confunde com aquele do conjunto da nuvem.

$$\sum_{k \in K} \frac{I_k}{n} \frac{x_{ik}}{I_k} = \frac{1}{n}$$

A projeção conserva essa propriedade. O conjunto das modalidades de uma mesma variável é então centrada sobre a origem para todos os gráficos; os fatores permitem que se comparem, ao mesmo tempo, as modalidades de todas as variáveis com as de uma só variável.

4.3.5.2 Subespaço engendrado pelas modalidades de uma variável.

Considerando a característica disjunta de uma TDC, os vetores de $\mathbf{R}^n$,que passam pela origem, (já centrada) definida pelas modalidades de uma mesma variável são, ortogonais entre eles. O conjunto de *r* modalidades de uma variável engendra um subespaço de dimensão igual a *r*. Considerando a característica de uma TDC ser completa, todos esses subespaços possuem uma direção comum: aquela que converge apara a origem do centro de gravidade da nuvem. Por essa direção estar eliminada devido a centralização, podemos considerar que , em AFCM, uma variável apresentando *r* modalidades engendra um subespaço de dimensão igual a *r*-1. daí resulta que, por representar perfeitamente as *r* modalidades de uma mesma variável, são necessários ao menos (*r*-1) fatores.

Essa propriedade possui várias conseqüências práticas:

- qualquer que seja a estrutura da tabela, a porcentagem de inércia associada a cada fator, em particular o primeiro, é necessariamente baixa desde que as variáveis apresentam muitas modalidades;
- mesmo que um fator esteja fortemente ligado a uma variável (no sentido de que ele reagrupa os indivíduos que possuem a mesma modalidade daquela variável), é impossível que todas essas modalidades sejam bem representadas por esse fator;
- na elaboração de uma tabela de dados, mesmo quando o número de indivíduos é muito grande, não é útil multiplicar as modalidades de uma mesma variável: o ganho de definição que se pode obter obtido, corre o risco de não ser aproveitado.

A inércia de uma variável com *r* modalidades (igual a $(r-1)/p$) é então repartida num subespaço de r-1 dimensões. De outra forma, pode ser mostrado que ela é igual à 1/*P* em todas as direções desse subespaço. Daí resulta que uma variável com um grande número de modalidades, ainda que engendrando uma inércia importante em $\mathbf{R}^n$, não influencia na orientação do primeiro fator de forma privilegiada, porque essa importante inércia está, de qualquer forma, diluída num subespaço de grande dimensão.

4.3.6 Síntese das variáveis qualitativas.

Um aspecto do estudo de um conjunto de variáveis é colocar em evidência um pequeno número de variáveis sintéticas, ligadas mais possível ao conjunto de variáveis originais. Para mostrar que os fatores da AFCM constituem essas variáveis sintéticas, nos utilizamos correlação, que mede a ligação entre uma variável numérica (aqui o fator) e uma variável qualitativa.

Relembrando a definição de correlação. Uma variável qualitativa define as partições sobre o conjunto de indivíduos em tantas classes quantas sejam as modalidades. Utilizando o teorema de Huygens, a inércia total (ou variância) de uma variável numérica pode se decompor na soma da inércia inter (i.e. inércia dos centros de gravidades das classes) e das inércias intra (i.e. inércia dos indivíduos com relação ao centro de gravidade da classe à qual ele pertence, a qual ele particiona). A correlação é o quociente da inércia inter pela inércia total. Ela varia entre 0 e 1. Quando ele é próximo de 1, os indivíduos de uma mesma classe estão bem reagrupados e as classes são nitidamente separadas umas das outras: essa é onde existe uma situação de ligação bastante forte entre a variável numérica (o fator) e a variável qualitativa. Quando de é próximo de 0, as médias das classes são bastante próximas da média geral e os indivíduos de uma mesma classe são bastante dispersos: a variável qualitativa e a variável numérica não são ligadas.

Na figura abaixo, ilustra-se a situação

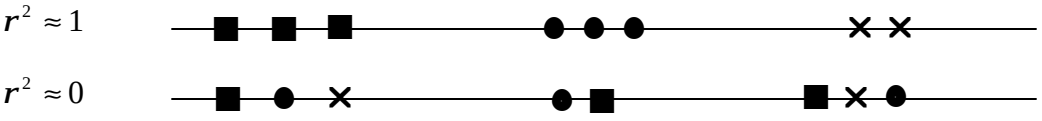


Figura 4.2 Ilustração de dois valores extremos da correlação para 8 indivíduos representados por um símbolo diferente segundo sua modalidade para uma variável qualitativa, sobre um eixo representando a variável numérica.

Denotando por G_k o baricentro dos indivíduos onde se encontra a modalidade *k*, a correlação entre uma variável *j* e o fator F_s vale:

$$r^2(F_s, j) = \frac{\text{inércia inter}}{\text{inércia total}} = \frac{\sum_{k \in K} \frac{I_k}{n} (F_s(G_k))^2}{l_k}$$

Ao utilizar o fato de que, em AFCM, a modalidade *k* tem peso I_k / np e se encontra próxima (por um coeficiente) do baricentro dos indivíduos que a possuem, seja:

$$G_S(k) = F_S(G_k) / \sqrt{l_S}$$

Encontramos:

$$r^2(F_s, j) = p \sum_{k \in K} \text{inércia das modalidade s de j, projetadas sobre o eixo de ordem s}$$

Notamos que a relação de correlação está compreendida entre 0 e 1, a inércia da sub-nuvem das modalidades de uma mesma variável sobre um eixo está compreendida entre 0 e 1/*P*: ela vale 1/*P* se F_S pertence ao subespaço engendrado pelas modalidades da variável.

A quantidade maximizada pelos eixos fatoriais no subespaço $\mathbf{R}^n$ é a inércia projetada da nuvem do conjunto das modalidades. Ao reagrupar as modalidades de uma mesma variável, esse critério não é outro que a média das correlações entre o fator e cada uma das variáveis. Isso resulta em que os fatores da AFCM são as variáveis numéricas com maior ligação ao conjunto de variáveis qualitativas estudadas e, nesse sentido, constituem então as variáveis sintéticas anunciadas.

As propriedades enunciadas nos dois últimos parágrafos permitem apreciar a influência relativa de uma variável na AFCM: para um dado eixo a importância a priori de cada variável é a mesma, mas o número dos eixos sobre os quais uma variável pode influir está diretamente ligado ao número de suas modalidades. Isto implica notadamente que se algumas variáveis muito ricas em modalidades são ligadas entre elas, os primeiros fatores não expressarão mais que essas ligações e será, então, necessário avançar na seqüência dos fatores para perceber outras ligações.

4.3.7 Representação das variáveis na AFCM.

O conceito de variável (e não mais de modalidades) aparece na AFCM e conduz a ajuda à interpretação. Esses índices completam aqueles já obtidos em uma AFC simples da TDC concernentes aos indivíduos e modalidades.

A contribuição de uma variável à inércia de um fator é a soma das contribuições de todas suas modalidades, ela permite também medir a ligação (a correlação) entra a variável e o fator. É interessante começar a análise dos resultados de uma AFCM pela consulta sistemática desses coeficientes que coloca em evidência as variáveis que estão mais ligadas a cada um dos fatores.

Pode ser útil construir para avaliar a dispersão no primeiro plano fatorial onde na abscissa e na ordenada figuram dois fatores, por exemplo: F_s e F_t. Nesse quadrante, podemos representar cada variável j por um ponto cuja coordenada sobre F_s (respectivamente F_t) é a correlação entre a variável j e F_s (respectivamente F_t).

Ademais, mostramos que esse gráfico se interpreta também como a projeção de uma nuvem na qual cada ponto representa uma variável, a proximidade entre dois pontos-variáveis expressam a semelhança entre as partições engendradas pelas duas variáveis.

4.4 Codificação das variáveis qualitativas

Na prática as variáveis qualitativas estudadas na AFCM resultam freqüentemente de uma transformação de variáveis numéricas. Além disso, mesmo quando a variável é por natureza qualitativa, existe com freqüência, para descrevê-la, uma escolha entre várias partições mais ou menos finas. Os resultados dependem da escolha das partições associadas às variáveis, esse problema é crucial.

Na análise de dados chamamos geralmente codificação a construção, a partir de dados brutos de uma tabela pronta para ser analisada: nesse sentido, o problema da escolha das classes é um problema de codificação e não existe método sistemático para realizá-la. A prática e a teoria têm, entretanto, esboçado um certo número de princípios que é prudente respeitar. Ademais, os resultados de uma análise permitem a validação ou a reconsideração

da codificação utilizada. Detalharemos aqui apenas alguns problemas relativos à codificação das variáveis numéricas como variáveis qualitativas.

4.4.1 Por que transformar as variáveis contínuas em qualitativas?

Dois objetivos principais conduzem a codificar por classes das variáveis contínuas partindo seu intervalo de variação.

Antes de tudo, podemos querer tornar homogêneos dados que se compõem inicialmente por variáveis numéricas e qualitativas. Assim, na análise de um conjunto de eventos sociais (sexo, profissão, idade, renda etc.) o fato de transformar as variáveis numéricas idade e renda em variáveis qualitativas permite tratar o conjunto dessas variáveis pela AFCM.

Podemos também ter interesse em realizar uma codificação qualitativa mesmo quando dispomos de um conjunto de variáveis numéricas sobre o qual uma ACP pode de toda maneira ser aplicada. Em efeito, na AFCM sobre essas mesmas variáveis codificadas em classes oferece uma outra aproximação dos dados. Ao representar cada variável pela mesma quantidade de pontos quantas são as classes, a AFCM pode colocar em evidência, se elas existem, as ligações não lineares entre as variáveis. Esse tipo de ligação é bastante freqüente porque muitos fenômenos apresentam efeitos de umbral: um estado patológico pode ser caracterizado por um valor muito fraco ou muito elevado; um queijo será tão mais apreciado quanto for salgado até um certo ponto (desse ponto de vista, os dois extremos do intervalo de variação do caráter *salgado* são tão próximos entre eles como de sua média). Concretamente, sobre os gráficos a proximidade de modalidades extremas demonstra a atitude da AFCM em colocar em evidência as ligações não lineares.

Tais fenômenos são naturalmente invisíveis nos resultados de uma ACP que não dão conta mais do que de suas ligações lineares. Paradoxalmente, reduzindo a informação tratada (a pertinência a uma classe ou a um intervalo é menos preciso que um valor numérico), aumentamos a riqueza do resultado! Notamos por exemplo que a média de uma classe de indivíduos compreendendo indivíduos muito grandes e indivíduos muito pequenos, corresponde a um indivíduo médio para uma variável numérica embora ela corresponda a uma partição dentro dos dois extremos por ela mesma codificada como qualitativa.

A AFCM de variáveis numérica codificadas como variáveis qualitativas é uma aproximação de uma análise não linear, no seguinte sentido: tratamos de encontrar as variáveis sintéticas que são combinações lineares de funções quaisquer das variáveis estudadas e não, como na ACP, das variáveis elas mesmas. Esse problema encontra-se tão-somente dentro dos limites de um modelo onde a população é infinita. Na prática, na AFCM sobre uma população finita, em lugar de considerar o conjunto das funções de uma variável, dividimos o intervalo dos valores da função em sub-intervalos e consideramos o conjunto das funções constantes em cada sub-intervalo. Em efeito, quando tratamos pela AFCM uma variável qualitativa j, essa variável está representada em $\mathbf{R}^n$ pelo subespaço $\mathbf{E}_j$ engendrado pelas indicatrizes de suas classes; $\mathbf{E}_j$ não é outro senão o conjunto das variáveis que possuíam o mesmo valor para todos os elementos de uma mesma classe. O primeiro fator é a combinação linear dos elementos desses P subespaços $\mathbf{E}_j$ (cada elemento é uma

função constante sobre as classes de uma variável) que melhor se aproxima desses subespaços.

Essa codificação permite também estudar variáveis em que as distribuições são bastante irregulares e para as quais o coeficiente de correlação é uma medida de ligação inadequada. Por exemplo se um elemento tem um valor muito afastado dos valores dos outros elementos influi de maneira preponderante sobre os coeficientes de correlação e uma codificação qualitativa o neutraliza.

4.4.2 Escolha do número de classes.

Para codificar por classes uma variável contínua, quer dizer dividir seu intervalo de variação em sub-intervalos que definem a mesma quantidade de modalidades, é preciso determinar de uma parte o número classes e de outra parte seus limites. Essa separação é um pouco formal na medida em que as duas escolhas são freqüentemente efetuadas simultaneamente.

Quantas classes é preciso utilizar? Nem muitas, nem poucas.

Diminuindo o excesso de número de classes, reagrupamos cada vez mais indivíduos diferentes, perdemos assim muita informação. As modalidades atingem então as situações mais variadas e seu estudo não coloca em evidência nada mais que os fenômenos já os mais destacados.

Aumentando o número de classes arriscamos obter classes com um pequeno efetivo com todos os inconvenientes que isto comporta. Se o efetivo da população é muito grande, descartamos esse risco e somos tentados a tomar um grande número de classes. Contudo, o número de classes excessivamente grande não deixa de apresentar problemas. Quanto mais criamos classes mais arriscamos fazer aparecer as ligações pontuas entre as modalidades. De outra parte, cada variável intervém na análise pelo subespaço de dimensão r-1 engendrado por suas r modalidades. Assim que aumentamos r, o número de fatores sobre os quais uma variável pode influir aumenta e o aspecto sintético da análise não melhora, muito pelo contrário!

Indicamos, para fixar as idéias, que a experiência mostra que não é útil ultrapassar o número de oito modalidades na codificação de variáveis quantitativas e que 4 ou 5 são freqüentemente bem suficientes.

4.4.3 Escolha das classes.

Para escolher as classes, examinamos primeiramente se existem limites naturais, ou clássicos, para a variável medida. Dessa maneira, em um estudo social, a idade para a aposentadoria é um limite “natural”.

Ainda que esse ponto de vista não é suficiente, estudamos as irregularidades da repartição dos valores. Para isto construímos um histograma com numerosas classes. Os vazios na repartição sugerem cortes no intervalo de variação. Ainda que os dois princípios precedentes não imponham nenhum limite, realizamos uma divisão sistemática do intervalo de variação. O princípio a respeitar nesta operação é o de obter *classes de mesmo efetivo* ao invés de intervalos com mesma amplitude. Esse procedimento de divisão está sempre previsto nos programas para análise de dados.

Existe justificações teóricas para essa prática. Um certo número de argumentos diretos militam por essa escolha.

- As modalidades representam um conjunto de indivíduos e para que a comparação entre eles tenham sentido é desejado que esses conjuntos sejam análogos do ponto de vista de seus efetivos. Isso é particularmente importante na AFCM onde a distância das modalidades ao baricentro cresce quando o efetivo decresce.
- Esse procedimento evita as modalidades de efetivo pequeno que nos apontamos onde ressaltamos o efeito perturbador. Além disso, o perfil dessas modalidades é muito sensível às pequenas variações da população; isto é particularmente embaraçoso mesmo que essa população seja proveniente de um amostra.

Bibliografia

[1] Bouroche, J.-M., Saporta, G. L'Analyse des données; . - 5ème éd. corrigée . - Paris: Presses Universitaires de France, 1992 . - 127 p.

[2] Escofier, B. e Pagès, Analices factorielles simples et multiples - Objectifs, méthodes et interprétation. Paris, Dunod, 1998284p.

[3] Escofier, B., Analyse des Correspondences - Recherche au couer de L'analyse des donées. Presses Universitaires de Rennes, Rennes, 2003, 234p.

[4] Foucart, T., Analyse factorielle - programmation sur micro-ordinateurs. Paris, Masson, 19885, 2ª ed., 234 p.

[5] Greenacre, M., J. Correspondence analysis in practice. London [etc.]: Academic Press, cop. 1993 . - 195 p.

[6] Jambu, M., Classification Automatique pour l'Analyse des Données 1- methods et algorithms. Paris, Dunod, 1978, 310p.

[7] Jobson, J. D., Applied multivariate data analysis. New York [etc.]: Springer, cop. 1991-1992 . - 2 v.

[8] Lebart, L., Morineau, A., Piron, M., Statistique exploratoire multidimensionnelle. Paris: Dunod, 1995 . - XV, 439 p.

[9] Lebart, L., Morineau, A., Fénelon, J. P., Tratamiento estadístico de datos : métodos y programas. Barcelona [etc.]: Marcombo, DL 1985 . - XI, 520 p.

[10] Lebart, L., Salem, A., Analyse statistique des donnéss textuelles.Paris, Dunod, 1988, 202p.

[11] Saporta, G., Probabilités, analyse des données et statistique. Paris: Technip, cop. 1990 . - XXVI, 493 p.

[12] Volle, M., Analyse des données.Paris: Economica, 3e éd, 1985 . - 323 p.

Índice