

APLICAÇÕES DA ÁLGEBRA LINEAR À ANÁLISE FATORIAL DE CORRESPONDÊNCIAS

45

José Henrique de Sá Mesquita

MONOGRAFIA SUBMETIDA À COORDENAÇÃO DO
CURSO DE PÓS-GRADUAÇÃO EM MATEMÁTICA, COMO REQUISITO PARCIAL
PARA OBTENÇÃO DO GRAU DE MESTRE
UNIVERSIDADE FEDERAL DO CEARÁ

FORTALEZA - 1986

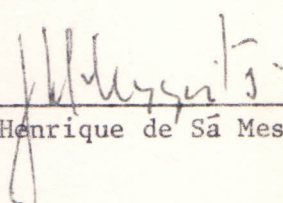
UFC/BU/BCM 01/06/1998



R818585 Aplicacoes da algebra linear a
C424412 analise
T510 N544a

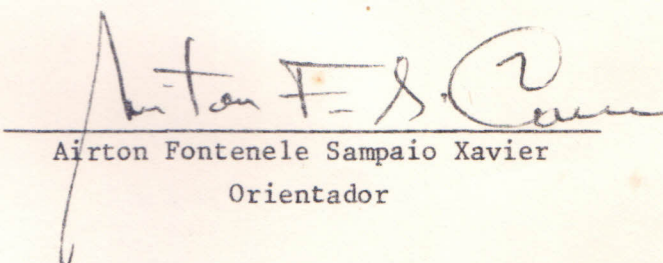
Esta Monografia foi submetida como parte dos requisitos necessá-
rios a obtenção do Grau de Mestre em Matemática, outorgado pela Universi-
dade Federal do Ceará, e encontra-se a disposição dos interessados na Bi-
blioteca Central da referida Universidade.

A citação de qualquer trecho desta Monografia é permitida, des-
de que seja feita de conformidade com as normas da ética científica.

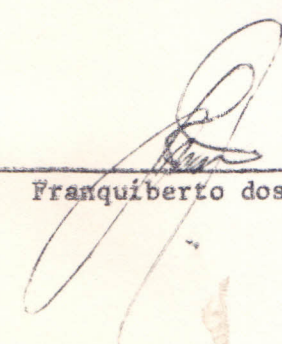


José Henrique de Sá Mesquita

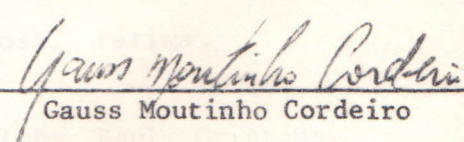
MONOGRAFIA APROVADA EM 03/09/86



Airton Fontenele Sampaio Xavier
Orientador



Franquiberto dos Santos Pessoa



Gauss Moutinho Cordeiro

DECLARAÇÃO

Eu, abaixo assinado, declaro que sou casado com Fátima, e que sou pai de Raul, Cristina, Debora, José e Mariana.

Declaro também que sou proprietário de um imóvel situado na Rua da Liberdade, nº 123, na cidade de São Paulo, Estado de São Paulo.

Declaro ainda que sou proprietário de um veículo de motorização, modelo 1980, marca Fiat, cor preta, placa ABC-1234, registro no Estado de São Paulo.

Declaro, por fim, que sou proprietário de uma conta corrente em nome de Fátima, no Banco do Brasil, Agência Central, nº 1234, sob o nº de conta 56789.

A minha esposa, Fátima.

Aos meus filhos, Raul, Cristina,
Debora, José e Mariana.

AGRADECIMENTOS

Agradeço ao Professor Airton Fontenele Sampaio Xavier, pela orientação, incentivo e encorajamento na preparação desta Monografia.

Registro também meus agradecimentos à Universidade do Amazonas, ao Instituto de Tecnologia do Amazonas e ao PICD, pelo apoio financeiro recebido.

É com satisfação que agradeço aos colegas e professores do Curso de Pós-Graduação em Matemática da Universidade Federal do Ceará, com especial atenção aos professores Raimundo Tompson Gonçalves e José de Anchieta Delgado, coordenadores do Curso, durante a realização deste trabalho.

Finalmente, agradeço ao Sr. José Alves Ferreira pelo excelente trabalho de datilografia.

José Henrique de Sã Mesquita

Fortaleza, Setembro de 1986

Í N D I C E

	página
GENERALIDADES SOBRE A ANÁLISE DE DADOS MULTIDIMENSIONAIS	1
A ANÁLISE EM COMPONENTES PRINCIPAIS	14
A ANÁLISE DE CORRESPONDÊNCIAS	31
A ANÁLISE DE CORRESPONDÊNCIAS MÚLTIPLAS	50
APÊNDICE	63
BIBLIOGRAFIA	79

INTRODUÇÃO

A Análise de Dados Multidimensionais tem recebido notável desenvolvimento nas últimas décadas, em especial atenção a partir dos anos 70. As razões para isso encontram-se, particularmente: 1º) na disponibilidade de computadores dotados de elevada capacidade de armazenamento de dados e de alta velocidade de cálculo, o que permitiu o "*mise-au-point*" (em termos operacionais) de métodos cujo interesse era apenas de ordem teórica, até então, além do estímulo para a formulação e o teste de novos métodos, cada vez mais aperfeiçoados; 2º) na necessidade de se dispor de procedimentos eficientes para a análise simultânea das relações mútuas eventualmente existentes entre as diversas variáveis envolvidas na formulação de problemas técnico-científicos ou decorrentes da aplicação sistemática de inquéritos sócio-econômicos, etc.

Por Análise de Dados Multidimensionais entende-se, em princípio, a aplicação de tais métodos, a par de sua contrapartida em termos computacionais (correspondente ao desenvolvimento e implementação de "*software*" respectivo), para a análise de grandes tabelas de dados.

A Análise de Dados agrupa, na realidade, numerosos métodos dos quais podemos distinguir dois ramos principais: a Classificação Automática, que consiste em classificar as unidades estatísticas e as variáveis estatísticas com a ajuda de algoritmos previamente estabelecidos, e a Análise Fatorial, que utiliza propriedades de espaços vetoriais euclidianos da qual faz parte a Análise de Correspondências, assunto desta monografia.

A Análise Fatorial de Correspondências é uma técnica recente, desenvolvida, em particular, pela equipe de pesquisas do Professor J.P. BENZÉCRI na Universidade de PARIS VI. Limitada, inicialmente, ao estudo das tabelas de contingência, seu emprego foi generalizado e é, frequentemente, utilizada para analisar tabelas de dados positivos. O método é construído com o propósito de descrever as linhas e as colunas de uma tabela de contingência, ou seja, de estudar a dependência entre caracteres qualitativos; tal estudo é realizado através de representações gráficas planas e de parâmetros que permitem interpretá-las.

Existem várias maneiras de introduzir a Análise de Correspondências. Aquela abordada aqui é baseada sobre a Análise em Componentes Principais, que por sua vez consiste em descrever um conjunto de indivíduos e um conjunto de caracteres quantitativos. A Análise de Correspondências será apresentada como uma dupla Análise em Componentes Principais.

Basicamente, estudaremos aqui: a Análise em Componentes Principais, a Análise de Correspondências e a Análise de Correspondências Múltiplas.

CAPÍTULO 1

GENERALIDADES SOBRE A ANÁLISE DE DADOS MULTIDIMENSIONAIS

1.1 - Os Elementos Descritivos de uma Tabela de Dados

As grandes tabelas de dados geralmente dependem ou são o resultado de elevado número de observações ou medidas realizadas sobre um conjunto finito $I = \{1, 2, \dots, n\}$ de indivíduos ou objetos. Os dados de correntes de tais observações ou medidas podem ser de natureza qualitativa ou quantitativa.

A realização de uma observação ou de uma medida sobre o conjunto I pode pensar-se como uma **variável** ou um **caracter**; isto é, como uma aplicação

$$h: I \longrightarrow Q$$

onde Q é o conjunto de todas as modalidades associadas ao caracter ou variável h .

Em geral, tem-se $Q \subset \mathbb{R}$, sendo este o caso de uma variável quantitativa, ou de uma variável qualitativa tornada quantitativa por meio de uma codificação. De fato, se Q for um conjunto de modalidades qualitativas associadas ao caracter ou variável h , a codificação corresponde a considerar uma aplicação injetiva

$$y: Q \longrightarrow \mathbb{R} ,$$

de sorte que $x = y \circ h$ torna-se uma variável quantitativa.

Por este motivo, uma variável será doravante sempre pensada como

uma aplicação

$$\mathbf{x}: I \longrightarrow \mathbb{R}$$

Se em um estudo específico estiverem envolvidos, ao todo, m variáveis, então teremos um conjunto de aplicações

$$\mathbf{x}^j: I \longrightarrow \mathbb{R}$$

onde $j \in J = \{1, 2, \dots, m\}$.

Denotaremos por x_i^j o valor assumido pela variável $\mathbf{x}^j$ no indivíduo $i \in I$; ou seja, $\mathbf{x}^j(i) = x_i^j$, para todo $j = 1, 2, \dots, m$ e todo $i = 1, 2, \dots, n$. Todos estes valores podem ser dispostos em uma tabela $\mathbf{X}$, na forma de uma matriz com m linhas e n colunas:

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \cdot & \cdot & \cdot & x_i^1 & \cdot & \cdot & \cdot & x_n^1 \\ x_1^j & \cdot & \cdot & \cdot & x_i^j & \cdot & \cdot & \cdot & x_n^j \\ x_1^m & \cdot & \cdot & \cdot & x_i^m & \cdot & \cdot & \cdot & x_n^m \end{pmatrix}.$$

No que segue, dada uma matriz $\mathbf{A}$ com m linhas e n colunas, $\mathbf{A}^T$ representará sua transposta, com n linhas e m colunas. Uma matriz diagonal será representada pela notação **diag** (..., a_i , ...), onde a_i é o elemento colocado na i -ésima posição da diagonal. Um vetor $\mathbf{u}$ será sempre pensado como vetor coluna, de modo que $\mathbf{u}^T$ será o vetor linha correspondente.

A cada indivíduo $i \in I$ associaremos o vetor do $\mathbb{R}^m$, denotado por $\mathbf{x}_i$, cujas coordenadas na base canônica

$$e_j = (\delta_1^j \dots \delta_i^j \dots \delta_m^j)^T, \quad j = 1, \dots, m, \quad (*)$$

são os elementos da i -ésima coluna da matriz X ; isto é,

$$x_i = (x_i^1 \dots x_i^j \dots x_i^m)^T.$$

A cada $i \in I$, ou a sua representação x_i no $\mathbb{R}^m$, supomos também estar associado um escalar p_i , que é a massa (ou peso) de x_i , de tal modo que seja satisfeita a condição $\sum_{i=1}^n p_i = 1$.

A massa de i (ou de x_i) poderá representar sua importância no contexto da análise. Muitas vezes considera-se a equidistribuição de massas, ou seja, cada p_i é igual a $1/n$. Entretanto, veremos que na análise de correspondências a preocupação com a atribuição de massas aos indivíduos é desnecessária; uma das peculiaridades deste tipo de análise é que a atribuição de massas é realizada de modo natural e automático.

O conjunto $\{(x_i, p_i); i = 1, \dots, n\}$ é chamado nuvem dos indivíduos, sendo denotado por $N(I)$.

O centro de gravidade de $N(I)$, denotado por g , é definido mediante

$$g = \sum_{i=1}^n p_i x_i;$$

sua j -ésima coordenada sendo então dada por

$$g^j = \sum_{i=1}^n p_i x_i^j.$$

De modo semelhante, estando o $\mathbb{R}^n$ munido da base canônica

(*) δ_i^j representa o delta de Kronecker

$$\mathbf{f}_i = (\delta_i^1 \dots \delta_i^j \dots \delta_i^n)^T, \quad i = 1, 2, \dots, n,$$

a cada variável $j \in J$ associamos o vetor

$$\mathbf{x}^j = (x_1^j \dots x_i^j \dots x_n^j)^T$$

que é a representação da j -ésima variável no $\mathbb{R}^n$. As representações das m variáveis em $\mathbb{R}^n$ constituirão a nuvem das variáveis, que será denotada por $N(J)$; isto é,

$$N(J) = \{\mathbf{x}^j, j = 1, \dots, m\}$$

Para cada variável $\mathbf{x}^j$, então, $\bar{\mathbf{x}}^j$, $\text{Var}(\mathbf{x}^j)$ e $\sigma(\mathbf{x}^j)$ denotam, respectivamente, a média, a variância e o desvio padrão, definidos por

$$\bar{\mathbf{x}}^j = \sum_{i=1}^n p_i x_i^j, \quad \text{Var}(\mathbf{x}^j) = \sum_{i=1}^n p_i (x_i^j - \bar{\mathbf{x}}^j)^2 \quad \text{e} \quad \sigma(\mathbf{x}^j) = (\text{Var}(\mathbf{x}^j))^{1/2}.$$

A média de $\mathbf{x}^j$, isto é, $\bar{\mathbf{x}}^j$, é exatamente a j -ésima coordenada g^j do centro de gravidade g de $N(J)$.

Para cada par de variáveis $\mathbf{x}^j$ e $\mathbf{x}^k$, então $\text{Cov}(\mathbf{x}^j, \mathbf{x}^k)$ e $\text{Corr}(\mathbf{x}^j, \mathbf{x}^k)$ denotam, respectivamente, a covariância e o coeficiente de correlação entre $\mathbf{x}^j$ e $\mathbf{x}^k$, definidos por

$$\text{Cov}(\mathbf{x}^j, \mathbf{x}^k) = \sum_{i=1}^n p_i (x_i^j - \bar{\mathbf{x}}^j)(x_i^k - \bar{\mathbf{x}}^k);$$

$$\text{Corr}(\mathbf{x}^j, \mathbf{x}^k) = \frac{\text{Cov}(\mathbf{x}^j, \mathbf{x}^k)}{\sigma(\mathbf{x}^j)\sigma(\mathbf{x}^k)} \quad (*)$$

A partir da matriz $\mathbf{X}$ dos dados, podemos obter a matriz $\tilde{\mathbf{X}} = (\tilde{x}_i^j)$, onde $\tilde{x}_i^j = x_i^j - \bar{\mathbf{x}}^j$.

(*) Evidentemente, supõe-se $\mathbf{x}^j$ e $\mathbf{x}^k$ não constantes, no caso de $\text{Corr}(\mathbf{x}^j, \mathbf{x}^k)$.

Como $\sum_{i=1}^n p_i \tilde{x}_i^j = \sum_{i=1}^n p_i (x_i^j - \bar{x}^j) = 0$, dizemos que os dados da

matriz $\tilde{X}$ estão centrados (ou que as variáveis estão centradas). Quando $X = \tilde{X}$, o centro de gravidade g de $N(I)$ coincide com a origem do $\mathbb{R}^m$. Observemos ainda que a matriz $\tilde{X}$ pode ser escrita $\tilde{X} = X - G$, onde G é a matriz cujas n colunas são todas iguais a g .

A matriz denotada por V e definida por

$$V = (v_k^j), \text{ onde } v_k^j = \text{Cov}(\mathbf{x}^j, \mathbf{x}^k), \quad j, k=1, \dots, m,$$

será denominada **matriz das variâncias-covariâncias associada a X** .

Se denotarmos por D_n a matriz diagonal dos pesos, $D_n = \text{diag}(p_1, \dots, p_i, \dots, p_n)$, podemos escrever a matriz V na seguinte

forma :

$$V = (X - G) D_n (X - G)^T.$$

Frequentemente, a análise realizada sobre uma tabela é iniciada após a centragem dos dados. Quando a tabela X tem os dados centrados, a matriz V assume a forma mais simples, $V = X D_n X^T$.

Uma outra matriz relacionada com a tabela X será denotada por R e definida como

$$R = (r_k^j), \text{ onde } r_k^j = \text{Corr}(\mathbf{x}^j, \mathbf{x}^k), \quad j, k=1, \dots, m.$$

Se denotarmos por $D_{1/\sigma}$ a matriz diagonal cujos elementos são os inversos dos desvios padrões das variáveis $\mathbf{x}^j$ (supostas não constantes), podemos escrever a matriz R na seguinte forma

$$R = D_{1/\sigma} V D_{1/\sigma}$$

1.2 - As Aplicações Lineares Associadas à Matriz X

Se $E = \mathbb{R}^m$ munido da base canônica $\{e_j\}_1^m$ representa o espaço dos indivíduos, denotemos por E^* o seu dual algébrico e por $\{e^j\}_1^m$ a base dual.

Se $F = \mathbb{R}^n$ munido da base canônica $\{f_i\}_1^n$ representa o espaço das variáveis, denotemos por F^* o seu dual algébrico, e por $\{f^i\}_1^n$ a base dual.

1.2.a - Dualidade

Para cada x_i , $i = 1, 2, \dots, n$, em E , e cada x^j , $j = 1, 2, \dots, m$, em F , temos:

$$e^j(x_i) = f^i(x^j) = x_i^j$$

Com efeito,

$$x_i \in E \Rightarrow x_i = \sum_{k=1}^m x_i^k e_k \quad i = 1, 2, \dots, n \quad e$$

$$x^j \in F \Rightarrow x^j = \sum_{k=1}^n x_k^j f_k \quad j = 1, 2, \dots, m$$

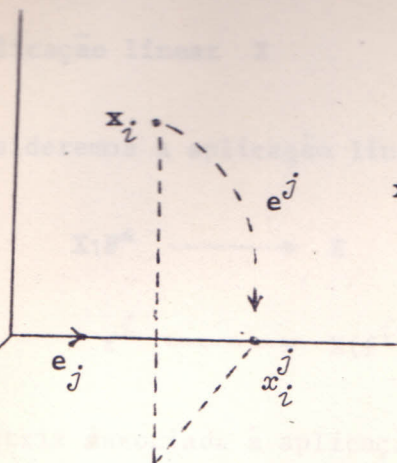
Então,

$$e^j(x_i) = \sum_{k=1}^m x_i^k e^j(e_k) = x_i^j,$$

$$f^i(x^j) = \sum_{k=1}^n x_k^j f^i(f_k) = x_i^j.$$

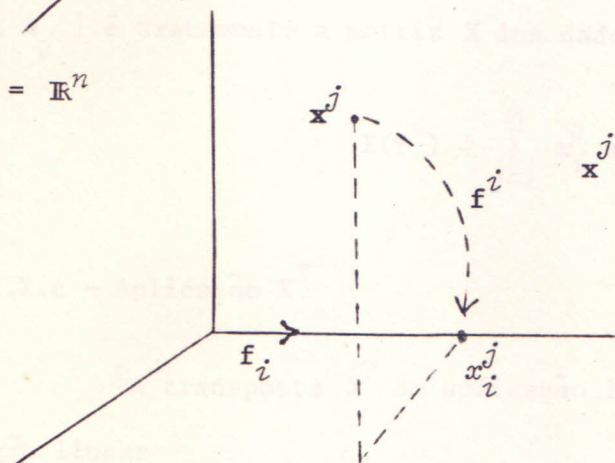
Ilustraremos tais fatos com o diagrama seguinte:

$$E = \mathbb{R}^m$$



$$x_i^j = \begin{pmatrix} f^i(x^1) \\ \vdots \\ f^i(x^j) \\ \vdots \\ f^i(x^m) \end{pmatrix} = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^j \\ \vdots \\ x_i^m \end{pmatrix}$$

$$F = \mathbb{R}^n$$



$$x_i^j = \begin{pmatrix} e^j(x_1) \\ \vdots \\ e^j(x_i) \\ \vdots \\ e^j(x_n) \end{pmatrix} = \begin{pmatrix} x_1^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{pmatrix}$$

Observemos que as componentes de $x_i^j \in E$ estão nos eixos de E gerados pelos e_j ; sendo as coordenadas respectivas iguais aos valores assumidos pela forma linear f^i nas m variáveis x^j . Deste modo, a cada $i \in I$ associamos, não só sua representação x_i em E , como também a i -ésima forma linear $f^i \in F^*$, que é a **representação do indivíduo i no F^*** .

De maneira semelhante, as componentes de x_i^j estão nos eixos de F gerados pelos e_j ; sendo as coordenadas respectivas iguais aos valores assumidos pela j -ésima forma linear $e^j \in E^*$ nos n indivíduos x_i . Assim, a cada $j \in J$ associamos, não só a sua representação x^j no F , como também a j -ésima forma linear $e^j \in E^*$, que é sua representação no E^* , e o eixo gerado por e_j .

1.2.b - A aplicação linear X

Consideremos a aplicação linear X definida por

$$X: F^* \longrightarrow E$$

$$f^i \longrightarrow X(f^i) = x_i$$

A matriz associada à aplicação X , com relação às bases $\{f^i\}$ e $\{e_j\}$ é exatamente a matriz X dos dados. Com efeito,

$$X(f^i) = \sum_{j=1}^m x_{ij}^j e_j.$$

1.2.c - Aplicação X^T

A transposta X^T da aplicação linear X é, por definição, a aplicação linear

$$X^T: E^* \longrightarrow F^{**}$$

$$e^* \longrightarrow X^T(e^*) = e^* \circ X$$

para todo $e^* \in E^*$.

Em particular, $X^T(e^j) = e^j \circ X$; desde que $e^j \circ X$ atua em F^* , temos o seguinte diagrama:

$$\begin{array}{ccc} F^* & \xrightarrow{X} & E \\ & \searrow e^j \circ X & \downarrow e^j \\ & & \mathbb{R} \end{array}$$

A dimensão de F sendo finita, F^{**} e F são identificados entre si; deste modo, $e^j \circ X \in F^{**}$ se escreve como combinação linear dos elemen

tos da base $\{f_i\}$, $X^T(e^j) = a_1 f_1 + \dots + a_k f_k + \dots + a_n f_n$.

Por outro lado, para todo $f^k \in F^*$ e para $k=1, \dots, n$, temos:

$$X^T(e^j)(f^k) = (e^j \circ X)(f^k) = e^j(x_k) = x_k^j,$$

de modo que $a_k = x_k^j$ e então

$$X^T(e^j) = \sum_{k=1}^n x_k^j f_k = x^j,$$

mostrando que, com relação às bases $\{e^j\}$ e $\{f_k\}$ a matriz da aplicação X^T é a matriz transposta da matriz X dos dados. Resumindo :

$$\begin{aligned} E & \xleftarrow{X} F^* & X(f^i) &= x_i & i=1, \dots, n &; \\ E & \xrightarrow{X^T} F & X^T(e^j) &= x^j & j=1, \dots, m &. \end{aligned}$$

1.3 - Métricas Euclidianas em E e F. O Esquema de Dualidade

1.3.1 - Observação Sobre a Notação

Em $A = \mathbb{R}^D$, munido da base canônica, introduzimos uma estrutura euclidiana através de uma forma bilinear simétrica definida positiva f ; tomaremos a liberdade de usar um único símbolo para representar a forma quadrática positiva definida Q , a matriz M e o isomorfismo $\phi: A \rightarrow A^*$, associados a f ; geralmente, o símbolo M da matriz em causa. Além disso, em lugar de falar na "métrica euclidiana associada à forma bilinear simétrica definida positiva f ", referimo-nos simplesmente a "métrica dada pela matriz M ".

Assim, para todo $x, y \in A$, temos :

Produto Interno: $M(x, y) = x^T M y = y^T M x$

Norma: $\|x\|_M^2 = M(x, x) = x^T M x$;

Distância: $d^2(x, y) = \|x - y\|_M^2 = (x - y)^T M (x - y)$

1.3.2 - A Métrica Euclidiana em F

A matriz diagonal $n \times n$ cujos elementos são as massas associadas aos n elementos x_i , $D_n = \text{diag}(p_1, \dots, p_i, \dots, p_n)$, é tal que, para todo $x \in F$, $x^T D_n x \geq 0$ e, $x^T D_n x = 0 \iff x = 0_E$

Em particular, para os elementos da base canônica $\{f_1, \dots, f_n\}$ de F ,

$$(f_i)^T D_n (f_i) = p_i \quad i = 1, \dots, n,$$

$$\text{e } (f_i)^T D_n (f_k) = 0 \quad \text{se } k \neq i.$$

A métrica de F será dada pela matriz D_n .

Consideremos agora o seguinte diagrama, associado à tabela X dos dados :

$$\begin{array}{ccc} E & \xleftarrow{X} & F^* \\ \uparrow V & & \uparrow D_n \\ E^* & \xrightarrow{X^T} & F \end{array}$$

O isomorfismo D_n permite definir uma aplicação linear $V: E^* \rightarrow E$, através da composição $V = X \circ D_n \circ X^T$.

Como V pertence ao espaço das aplicações lineares de E^* em E , isomorfo ao espaço das formas bilineares em E^* , denotaremos também por V a forma bilinear associada, que é simétrica e semi-definida positiva, posto que, para todos $a, b \in E^*$, temos :

$$\begin{aligned} V(a, b) &= a^T V b = a^T X D_n X^T b = \\ &= (X^T a)^T D_n (X^T b) = D_n (X^T a, X^T b) = \\ &= D_n (X^T b, X^T a) = V(b, a) ; \end{aligned}$$

bem como,

$$V(a, a) = D_n (X_a^T, X_a^T) \geq 0 .$$

Por outro lado, se X^T é injetiva, então V é simétrica definida positiva.

Com relação à matriz associada a V , temos :

$$\begin{aligned} V(e^j, e^k) &= D_n (X^T e^j, X^T e^k) = \\ &= D_n (x^j, x^k) = (x^j)^T D_n (x^k) = \\ &= \text{Cov}(x^j, x^k) ; \end{aligned}$$

ou seja a matriz V é a matriz das variâncias-covariâncias.

1.3.2.a - Algumas Considerações Geométricas

Seja $u \in F$ o vetor que tem todas as coordenadas iguais a 1, e denotemos por $[u]$ o subespaço gerado por u . F se decompõe em uma soma direta de dois subespaços :

$$F = [u] \oplus [u]^\perp$$

onde o subespaço $[u]^\perp$ é o D_n -ortogonal de $[u]$. Uma variável $x^j \in F$ se

escreve então sob a forma

$$\mathbf{x}^j = \alpha \mathbf{u} + \beta \mathbf{v}$$

onde $\mathbf{v}$ pertence ao subespaço $[\mathbf{u}]^\perp$. Por definição $\alpha \mathbf{u}$ é a projeção D_n -ortogonal de $\mathbf{x}^j$ sobre $\mathbf{u}$ e temos:

$$\begin{aligned} D_n(\mathbf{x}^j, \mathbf{u}) &= D_n(\alpha \mathbf{u}, \mathbf{u}) + D_n(\beta \mathbf{v}, \mathbf{u}) \\ &= \alpha D_n(\mathbf{u}, \mathbf{u}) + 0 = \alpha \end{aligned}$$

Como $D_n(\mathbf{x}^j, \mathbf{u}) = \bar{\mathbf{x}}^j$, concluímos que a média da variável $\mathbf{x}^j$ é a abcissa da projeção D_n -ortogonal de $\mathbf{x}^j$ sobre $\mathbf{u}$. Além disso, o vetor $\mathbf{x}^j - \bar{\mathbf{x}}^j \mathbf{u} \in \mathbf{u}^\perp$ é a projeção D_n -ortogonal de $\mathbf{x}^j$ sobre $\mathbf{u}^\perp$, e

$$\begin{aligned} \|\mathbf{x}^j - \bar{\mathbf{x}}^j \mathbf{u}\|_{D_n}^2 &= (\mathbf{x}^j - \bar{\mathbf{x}}^j \mathbf{u})^T D_n(\mathbf{x}^j - \bar{\mathbf{x}}^j \mathbf{u}) = \\ &= \sum_{i=1}^n p_i (x_i^j - \bar{x}^j)^2 = \text{Var}(\mathbf{x}^j) . \end{aligned}$$

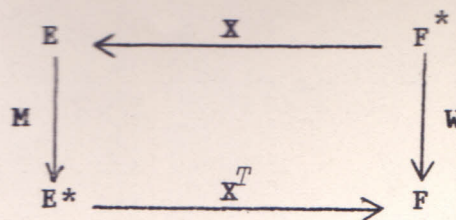
Interpretemos agora o coeficiente de correlação:

se α é o D_n -ângulo entre dois caracteres centrados j e k , temos:

$$\begin{aligned} \cos \alpha &= \frac{D_n(\mathbf{x}^j - \bar{\mathbf{x}}^j \mathbf{u}, \mathbf{x}^k - \bar{\mathbf{x}}^k \mathbf{u})}{\|\mathbf{x}^j - \bar{\mathbf{x}}^j \mathbf{u}\|_{D_n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k \mathbf{u}\|_{D_n}} = \frac{(\mathbf{x}^j - \bar{\mathbf{x}}^j \mathbf{u})^T D_n(\mathbf{x}^k - \bar{\mathbf{x}}^k \mathbf{u})}{\sigma(\mathbf{x}^j) \sigma(\mathbf{x}^k)} \\ &= \frac{\text{Cov}(\mathbf{x}^j, \mathbf{x}^k)}{\sigma(\mathbf{x}^j) \sigma(\mathbf{x}^k)} = \text{Corr}(\mathbf{x}^j, \mathbf{x}^k) . \end{aligned}$$

1.3.3 - A Métrica Euclidiana no Espaço E

Seja $\mathbf{M}$ uma forma quadrática definida positiva em $\mathbf{E}$ e considere mos o diagrama a seguir, associado à tabela $\mathbf{X}$

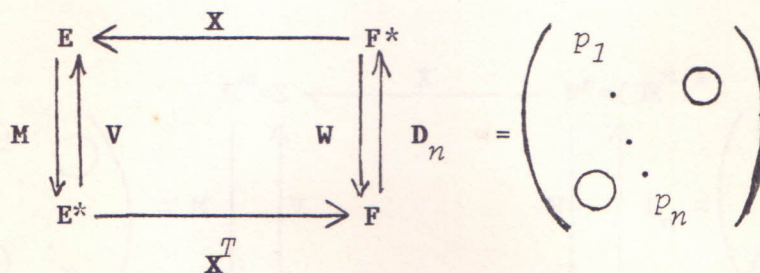


O isomorfismo M nos permite definir a aplicação linear $W: F^* \rightarrow F$ através da composição $W = X^T \circ M \circ X$. Denotemos também por W a forma bilinear associada à aplicação linear W .

De modo semelhante ao que ocorreu com a forma bilinear V , verificamos que a forma bilinear W é simétrica, semi-definida positiva, e será definida positiva se a aplicação X for injetiva.

1.3.4 - O Esquema de Dualidade

A expressão "esquema de dualidade" será empregada sempre que tivermos necessidade de fazer referência ao diagrama abaixo (ou parte dele), relacionado com a tabela X dos dados :



Observação: a matriz M geralmente é tal que

$$M = \text{diag}(r_1, \dots, r_m),$$

onde $r_1, \dots, r_m$ são números reais positivos, que serão posteriormente especificados.

A ANÁLISE EM COMPONENTES PRINCIPAIS

A técnica da análise em componentes principais constitui uma das ferramentas básicas para a análise de correspondências, sendo este o motivo pelo qual o presente capítulo é dedicado a uma breve exposição do mesmo.

2.1 - A Formulação do Problema

Os dados a analisar estão dispostos em uma tabela X tal como no capítulo 1; ou seja, dispomos de uma matriz $m \times n$ representando n pontos "indivíduos" no $\mathbb{R}^m$

Temos o já conhecido esquema de dualidade

$$\begin{array}{ccccc}
 \begin{pmatrix} x_1 & & \\ & \circ & \\ & & \ddots \\ \circ & & & x_m \end{pmatrix} & = M & \begin{array}{c} \downarrow \\ \uparrow \\ \downarrow \\ \uparrow \end{array} & \begin{array}{c} \mathbb{R}^m = E \\ (\mathbb{R}^m)^* = E^* \end{array} & \begin{array}{c} \xleftarrow{X} \\ \xrightarrow{X^T} \end{array} & \begin{array}{c} F^* = (\mathbb{R}^n)^* \\ F = \mathbb{R}^n \end{array} \\
 & & & & & \begin{array}{c} \downarrow W \\ \uparrow D_n \end{array} \\
 & & & & & \begin{pmatrix} p_1 & & \\ & \circ & \\ & & \ddots \\ \circ & & & p_n \end{pmatrix}
 \end{array}$$

Consideremos A' um subespaço afim de E e $N(I)$ a nuvem ponderada dos pontos indivíduos em E .

2.1.1 - Definição

Chamamos *momento de inércia* da nuvem $N(I)$ com relação ao subespaço afim A' e denotamos por $\text{In } \{A'\}$, o número

$$\text{In } \{A'\} = \sum_{i=1}^n p_i (\mathbf{x}_i - \mathbf{a}_i)^T \mathbf{M} (\mathbf{x}_i - \mathbf{a}_i) ,$$

onde $\mathbf{a}_i$ é a projeção $\mathbf{M}$ -ortogonal de $\mathbf{x}_i$ sobre A' . Em particular, se $\mathbf{a}$ é um ponto qualquer de E ,

$$\text{In } \{\mathbf{a}\} = \sum_{i=1}^n p_i (\mathbf{x}_i - \mathbf{a})^T \mathbf{M} (\mathbf{x}_i - \mathbf{a}) .$$

Coloca-se o seguinte problema:

Dado um inteiro positivo $q < m$, encontrar, entre todos os subespaços afins de dimensão q , de E , aquele para o qual o momento de inércia da nuvem é mínimo.

Os resultados obtidos a seguir permitirão simplificar o problema.

2.1.2 - Proposição

O centro de gravidade, $\mathbf{g}$, é o ponto relativo ao qual o momento de inércia é mínimo, qualquer que seja a métrica Euclidiana $\mathbf{M}$ de E .

Além disso, se os dados estão centrados, isto é, se $\mathbf{g} = \mathbf{0}_E$, então

$$\text{In } \{\mathbf{g}\} = \text{Traço } (\mathbf{VM}) .$$

Demonstração

Se $\mathbf{a}$ é um ponto qualquer em E , temos, para $i=1, 2, \dots, n$,

$$\mathbf{x}_i - \mathbf{a} = \mathbf{x}_i - \mathbf{g} + \mathbf{g} - \mathbf{a} ;$$

então:

$$\begin{aligned} \text{In } \{a\} &= \sum_{i=1}^n p_i ||x_i - a||^2 = \\ &= \sum_{i=1}^n p_i ||x_i - g||^2 + \sum_{i=1}^n p_i ||g - a||^2 + 2M \left(\sum_{i=1}^n p_i (x_i - g), g - a \right) \\ &= \text{In } \{g\} + ||g - a||^2, \end{aligned}$$

ou seja,

$$\text{In } \{g\} \geq \text{In } \{a\}.$$

Se os dados estão centrados, temos:

$$\text{In } \{g\} = \sum_{i=1}^n p_i x_i^T M x_i;$$

mas

$$x_i^T M x_i = \text{Traço } (x_i^T M x_i) = \text{Traço } (x_i x_i^T M),$$

e, deste modo,

$$\begin{aligned} \text{In } \{g\} &= \sum_{i=1}^n p_i \text{Traço } (x_i x_i^T M) \\ &= \text{Traço } \left[\left(\sum_{i=1}^n p_i x_i x_i^T \right) M \right] = \text{Traço } (VM) \end{aligned}$$

No que segue, os dados são supostos *centrados*.

2.1.3 - Proposição

Seja A' um subespaço afim de E com subespaço vetorial subjacente A contendo $g = 0_E$. Então $\text{In } \{0\} = \text{In } \{A\} + \text{In } \{A^\perp\}$, e $\text{In } \{A'\} \geq \text{In } \{A\}$, onde $A^\perp$ denota o complementar M -ortogonal de A em E .

Prova: Temos

$$E = A \oplus A^\perp \text{ e } x_i \in E \implies x_i = a_i + b_i ,$$

com $a_i \in A$ e $b_i \in A^\perp$.

Além disso

$$\text{In}\{A\} = \sum_{i=1}^n p_i \|b_i\|^2 \text{ e } I\left[\begin{smallmatrix} \perp \\ A \end{smallmatrix}\right] = \sum_{i=1}^n p_i \|a_i\|^2 ;$$

então:

$$\begin{aligned} \text{In}\{A\} &= \sum p_i M(a_i + b_i, a_i + b_i) = \\ &= \sum_{i=1}^n p_i M(a_i, a_i) + \sum_{i=1}^n p_i M(b_i, b_i) = \\ &= \text{In}\{A^\perp\} + \text{In}\{A\} . \end{aligned}$$

Como

$$A' \cap A^\perp = \{a\}, \quad A' = A + a ,$$

temos

$$x_i \in E \implies x_i = a_i + b_i = a_i + a + b_i - a ,$$

$$a_i + a \in A' \text{ e } b_i - a \in A^\perp ,$$

o momento de inércia com relação a A' é dado por:

$$\begin{aligned} \text{In}\{A'\} &= \sum_{i=1}^n p_i M(b_i - a, b_i - a) = \\ &= \sum_{i=1}^n p_i M(b_i, b_i) + \sum_{i=1}^n p_i M(a, a) - 2 \sum_{i=1}^n p_i M(b_i, a) . \end{aligned}$$

Mas

$$\sum_{i=1}^n p_i a_i + \sum_{i=1}^n p_i b_i = \sum_{i=1}^n p_i x_i = g = 0 ,$$

e então as duas parcelas à esquerda são iguais ao vetor nulo, pois uma está em A e a outra em $A^\perp$.

Concluimos assim que

$$\ln \{A'\} = \ln \{A\} + ||a||^2$$

ou seja

$$\ln \{A'\} \geq \ln \{A\}$$

Como consequência destes resultados, a procura de um subespaço afim de E com momento de inércia mínimo se reduz àquela de um subespaço vetorial com tal característica.

Se denotarmos por $[u_1, u_2, \dots, u_q] = A$ o subespaço de E gerado por um sistema M-ortonormal $u_1, u_2, \dots, u_q$, o problema colocado no início da seção pode ser formulado nos seguintes termos

Problema 1 - Minimizar $\ln \{A\}$, para $A = [u_1, \dots, u_q]$ onde os u_k são M-ortogonais e M-unitários.

2.2 - A Solução

2.2.1 - Proposição

$$\ln \{A\} = \text{Traço}(VM) - \sum_{k=1}^q u_k^T M V M u_k$$

Prova: Para cada $x_i \in E$, $i=1, 2, \dots, n$, temos:

$$x_i = a_i + b_i, \text{ com } a_i \in A \text{ e } b_i \in A^\perp;$$

então

$$\begin{aligned} \ln \{A\} &= \sum_{i=1}^n p_i ||x_i - a_i||^2 = \sum_{i=1}^n p_i ||b_i||^2 = \\ &= \sum_{i=1}^n p_i ||x_i||^2 - \sum_{i=1}^n p_i ||a_i||^2, \end{aligned}$$

e pelo resultado 2.1.2,

$$\text{In } \{A\} = \text{Traço (VM)} - \sum_{i=1}^n p_i \|a_i\|^2.$$

Como a_i é a projeção M-ortogonal de x_i sobre A, podemos escrever:

$$a_i = \sum_{k=1}^q M(x_i, u_k) u_k,$$

donde,

$$\|a_i\|^2 = \sum_{k=1}^q [M(x_i, u_k)]^2 = \sum_{k=1}^q (u_k^T M x_i) (x_i^T M u_k)$$

deste modo,

$$\begin{aligned} \sum_{i=1}^n p_i \|a_i\|^2 &= \\ &= \sum_{k=1}^q u_k^T M \left(\sum_{i=1}^n p_i x_i x_i^T \right) M u_k = \sum_{k=1}^q u_k^T MVM u_k, \end{aligned}$$

Concluimos então que :

$$\text{In } \{A\} = \text{Traço (VM)} - \sum_{k=1}^q u_k^T MVM u_k$$

Como $\text{Traço(VM)} = I[0]$ é constante, minimizar $\text{In}\{A\}$ é equivalente a maximizar

$$\sum_{k=1}^q u_k^T MVM u_k$$

para $u_1, \dots, u_q$ M-ortonormais, e o problema 1 é então equivalente ao

Problema 2 -

$$\text{Maximizar } \sum_{k=1}^q u_k^T MVM u_k, \text{ para } A = [u_1, \dots, u_q],$$

$\{u_1, \dots, u_q\}$ M-ortonormal

A matriz $M = \text{diag}(r_1, \dots, r_m)$ sendo simétrica, definida positiva, denotemos por $M^{1/2}$ a única matriz B tal que $B^2 = M$, especificamente:

$$M^{1/2} = \text{diag}(\sqrt{r_1}, \dots, \sqrt{r_m})$$

Ora,
$$u_k^T M V M u_k = (M^{1/2} u_k)^T (M^{1/2} V M^{1/2}) (M^{1/2} u_k).$$

Fazendo $z_k = M^{1/2} u_k$, $k = 1, 2, \dots, q$, é fácil verificar que estes vetores são ortonormais. Com efeito,

$$\begin{aligned} u_k^T M u_\ell &= u_k^T M^{1/2} M^{1/2} u_\ell = \\ &= (M^{1/2} u_k)^T I_m (M^{1/2} u_\ell) = z_k^T I_m z_\ell = \begin{cases} 1 & \text{se } k = \ell \\ 0 & \text{se } k \neq \ell \end{cases} \end{aligned}$$

O problema 1 é então equivalente ao

Problema 3 -

$$\begin{aligned} &\text{Maximizar} \quad \sum_{k=1}^q z_k^T M^{1/2} V M^{1/2} z_k, \quad \{z_1, \dots, z_q\} \\ &\quad \{z_1, \dots, z_q\} \text{ sistema ortonormal.} \end{aligned}$$

A matriz real $S = M^{1/2} V M^{1/2}$ é simétrica, semi-definida positiva, $m \times m$.

Sabemos que a equação

$$\det(S - \lambda I_m) = 0$$

admite m raízes reais não negativas,

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$$

Correspondendo a cada raiz λ_i , existe um vetor s_i tal que

$$S s_i = \lambda_i s_i;$$

os vetores $s_1, s_2, \dots, s_m$ podem ser tomados ortonormais, independentemente do fato dos λ_i serem distintos, ou não; além disso, a decomposição espectral nos dá:

$$S = \lambda_1 s_1 s_1^T + \dots + \lambda_m s_m s_m^T$$

$$I_m = s_1 s_1^T + \dots + s_m s_m^T$$

2.2.2 - Proposição

Sejam $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ os auto-valores de $S = M^{1/2} V M^{1/2}$ e $s_1, s_2, \dots, s_m$ os correspondentes auto-vetores.

Seja $\{z_1, z_2, \dots, z_q\}$ um conjunto qualquer de vetores ortonormais em E. Então

$$\sum_{k=1}^q z_k^T S z_k \leq \sum_{k=1}^q s_k^T S s_k = \lambda_1 + \dots + \lambda_q$$

$$q = 1, 2, \dots, m$$

e o máximo é efetivamente atingido em $z_k = s_k$.

Prova: Para $k = 1, 2, \dots, q$, cada z_k se escreve como combinação linear da base $\{s_1, s_2, \dots, s_m\}$:

$$z_k = z_{k1} s_1 + \dots + z_{kq} s_q + \dots + z_{km} s_m,$$

com a condição adicional

$$z_k^T z_k = 1$$

Como

$$S = \lambda_1 s_1 s_1^T + \dots + \lambda_q s_q s_q^T + \dots + \lambda_m s_m s_m^T,$$

obtemos

$$\mathbf{z}_k^T \mathbf{S} \mathbf{z}_k = \lambda_1 z_{k1}^2 + \dots + \lambda_q z_{kq}^2 + \dots + \lambda_m z_{km}^2$$

e, deste modo

$$\begin{aligned} \sum_{k=1}^q \mathbf{z}_k^T \mathbf{S} \mathbf{z}_k &= \lambda_1 (z_{11}^2 + \dots + z_{q1}^2) + \dots + \lambda_q (z_{1q}^2 + \dots + z_{qq}^2) + \\ &+ \dots + \lambda_m (z_{1m}^2 + \dots + z_{qm}^2) \end{aligned}$$

Observemos, agora, que os coeficientes de $\lambda_1, \lambda_2, \dots$ e λ_m são todos menores ou iguais a 1 e que a soma de todos eles é igual a q . Além disso, temos a condição $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q \geq \dots \geq \lambda_m$.

Deste modo, a escolha ótima dos coeficientes para maximizar

$$\sum_{k=1}^q \mathbf{z}_k^T \mathbf{S} \mathbf{z}_k$$

$$\bar{\epsilon} : \begin{cases} 1 & \text{para os coeficientes de } \lambda_1, \lambda_2, \dots, \lambda_q \\ 0 & \text{para os coeficientes de } \lambda_{q+1}, \dots, \lambda_m, \end{cases}$$

o que é possível, fazendo

$$\mathbf{z}_k = \mathbf{s}_k, \quad k=1, 2, \dots, q$$

A solução do problema 3 é então constituída por q auto-vetores da matriz $\mathbf{S}$, associados aos q maiores auto-valores $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$,

$$\mathbf{z}_k = \mathbf{s}_k, \quad k = 1, 2, \dots, q$$

A solução do problema 2 é deduzida de

$$\mathbf{z}_k = \mathbf{s}_k, \quad k = 1, 2, \dots, q$$

através do isomorfismo $M^{-1/2}$, por

$$u_k = M^{-1/2} z_k = M^{-1/2} s_k, \quad k = 1, 2, \dots, q$$

2.2.3 -

Os vetores

$$u_k = M^{-1/2} s_k, \quad k = 1, 2, \dots, q$$

são auto-vetores da matriz VM , correspondentes aos auto-valores $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$, os mesmos da matriz S .

De fato, para todo $k = 1, 2, \dots, q$ temos:

$$M^{1/2} V M^{1/2} s_k = \lambda_k s_k$$

$$\iff V M^{1/2} s_k = \lambda_k M^{-1/2} s_k$$

$$\iff V M M^{-1/2} s_k = \lambda_k M^{-1/2} s_k$$

$$\iff VM u_k = \lambda_k u_k$$

Como consequência dos resultados obtidos em 2.2.2 e 2.2.3, podemos enunciar a solução do problema 1 nos seguintes termos:

O subespaço A de E , de dimensão q , para o qual o momento de inércia com relação à nuvem dos indivíduos é mínimo, é aquele gerado por q auto-vetores M -ortonormais de VM , associados aos q maiores auto-valores $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$.

Observação. A matriz VM , em geral, não é simétrica; por este

motivo, na prática trabalhamos com a matriz $S = M^{1/2} V M^{1/2}$, simétrica, que possui os mesmos auto-valores de VM. Isto torna o trabalho computacional mais simples e eficiente. Os auto-vetores de VM, como vimos, são obtidos a partir dos auto-vetores de S através da transformação $M^{-1/2}$.

2.3 - Eixos e Planos Principais

Consideremos a base M-ortonormal de E, construída com a ajuda dos auto-vetores M-ortonormais de VM.

O subespaço A, dado por um sistema M-ortonormal constituído de q auto-vetores de VM correspondentes aos q maiores auto-valores,

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$$

é chamado *subespaço principal de dimensão q*.

Cada vetor M-normado u_k é chamado *vetor principal* e gera um subespaço $[u_k]$ de dimensão 1, dito *eixo principal de inércia*.

Por sua vez, os *planos principais* são os subespaços de dimensão 2 gerados por pares de vetores principais.

Para o primeiro eixo principal $[u_1]$, tal que $VM \dot{u}_1 = \lambda_1 u_1$, temos:

$$\text{In}\{u_1\} = \text{Traço VM} - \lambda_1;$$

a nuvem dos indivíduos está contida em $[u_1]$ se e somente se $\text{In}\{u_1\} = 0$; ou seja, se, e somente se $\text{Traço VM} = \lambda_1$.

Para o subespaço principal A de dimensão q, temos:

$$\text{In}\{A\} = \text{Traço VM} - \sum_{k=1}^q \lambda_k.$$

Se a nuvem não está contida em A, é evidente que

$$\frac{\sum_{k=1}^q \lambda_k}{\text{Traço VM}} < 1.$$

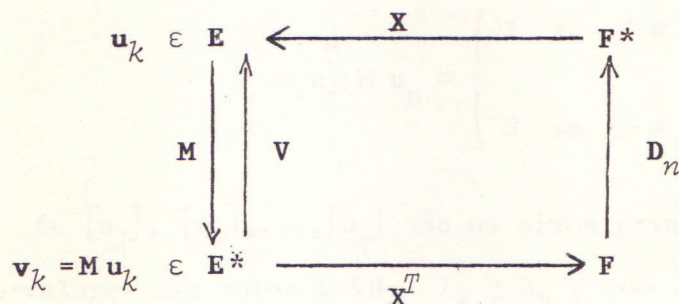
O índice

$$\frac{\sum_{k=1}^q \lambda_k}{\text{Traço VM}}$$

é chamado *parte da inércia explicada pelo subespaço principal A*.

2.4 - Fatores e Componentes Principais

Consideremos o esquema de dualidade:



Aos eixos principais $[u_1]$, $[u_2]$, ..., $[u_q]$, em E , fazemos corresponder, de acordo com o referido esquema de dualidade:

- em E^* : as formas lineares denotadas por v_k e definidas por:

$$v_k = M u_k, \quad k = 1, 2, \dots, q,$$

as quais são chamadas *fatores principais*.

- em F^* : as *componentes principais* denotadas por c^k e definidas por:

$$c^k = X^T(v_k) = v_k \circ X.$$

A i -ésima coordenada c_i^k , de c^k é dada por:

$$c_i^k = C^k(f_i) = v_k(x_i) = u_k^T M x_i = M(x_i, u_k) ;$$

ou seja $\tilde{e}$ a abcissa da projeção M-ortonormal de x_i sobre o eixo $[u_k]$.

Os fatores $v_j = M u_j$, $j = 1, 2, \dots, m$ são auto-vetores de MV correspondentes aos auto-valores λ_j , pois

$$M V v_j = M V M u_j = M(\lambda_j u_j) = \lambda_j v_j ;$$

além disso, constituem uma base M^{-1} -ortonormal de E^* ; de fato:

$$\begin{aligned} M^{-1}(v_p, v_j) &= v_j^T M^{-1} v_p \\ &= u_j^T M M^{-1} M u_p \\ &= u_j^T M u_p = \begin{cases} 1 & \text{se } j = p \\ 0 & \text{se } j \neq p \end{cases} \end{aligned}$$

Se $[u_1], [u_2], \dots, [u_q]$ são os eixos principais correspondentes aos auto-valores não nulos de VM , $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$, então:

$$\begin{aligned} \text{a) } D_n C^k &= X^T M X D_n X^T M u_k = \\ &= X^T M V M u_k = \lambda_k C^k \end{aligned}$$

$$\begin{aligned} \text{b) } \|C^k\|_{D_n}^2 &= (C^k)^T D_n C^k = u_k^T M X D_n X^T M u_k = \\ &= u_k^T M V M u_k = \lambda_k \end{aligned}$$

$$\begin{aligned} \text{c) } D_n(C^k, C^\ell) &= (C^k)^T D_n C^\ell = u_k^T M X D_n X^T M u_\ell = \\ &= u_k^T M V M u_\ell = \lambda_\ell u_k^T M u_\ell = \begin{cases} 1 & \text{se } k = \ell \\ 0 & \text{se } k \neq \ell \end{cases} \end{aligned}$$

d) As componentes principais constituem um sistema de geradores

do espaço vetorial $\mathbf{X}^T(E^*)$, como imagens de uma base de E^* . Se q é o número de auto-valores não nulos de $\mathbf{VM}$, então

$$q = \text{posto}(\mathbf{X}^T) = \dim \mathbf{X}^T(E^*);$$

neste caso, $\{\mathbf{C}^k, k = 1, 2, \dots, q\}$ é uma base D_n -ortogonal de $\mathbf{X}^T(E^*)$ e $\left\{ \frac{\mathbf{C}^k}{\sqrt{\lambda_k}}, k = 1, 2, \dots, q \right\}$ é base D_n -ortonormal de $\mathbf{X}^T(E^*)$.

Resumindo:

As componentes principais $\mathbf{C}^k$ são auto-vetores de $\mathbf{WD}_n$ com auto-valores correspondentes λ_k ; são não correlacionadas entre si e, se $q = \text{posto } \mathbf{VM}$, então $\left\{ \frac{\mathbf{C}^k}{\sqrt{\lambda_k}}, k = 1, 2, \dots, q \right\}$ é base D_n -ortonormal de $\mathbf{X}^T(E^*)$.

2.5 - Apresentação dos Resultados e Regras de Interpretação

2.5.1 - Descrição da Nuvem dos Indivíduos

Se a parte da inércia explicada por um plano principal $[\mathbf{u}_k, \mathbf{u}_\ell]$ é próxima de 1, uma boa descrição da nuvem é dada pelas projeções dos pontos $\mathbf{x}_i$ no plano (ou em um eixo).

A parte da inércia explicada por um plano principal mede a *qualidade global* da representação da nuvem no plano considerado.

Podemos medir a *qualidade isolada* da representação de um ponto $\mathbf{x}_i$ por sua projeção $\hat{\mathbf{x}}_i$, comparando as normas $||\hat{\mathbf{x}}_i||_M^2$ e $||\mathbf{x}_i||_M^2$, ou seja, por

$$\cos^2 \alpha = \frac{||\hat{\mathbf{x}}_i||_M^2}{||\mathbf{x}_i||_M^2}.$$

O exame da nuvem das projeções dos n pontos x_i em um plano (ou eixo) principal, levando em conta a qualidade das representações, permite agrupar, em classes, indivíduos "próximos" e bem representados.

Podemos também fazer figurar em um plano principal as projeções M-ortogonais dos eixos $[e_j]$, $j=1,2,\dots,m$, associados aos caracteres iniciais. Esta representação simultânea permite analisar as proximidades no interior de um mesmo conjunto, exibindo os elementos do outro responsáveis por estas proximidades.

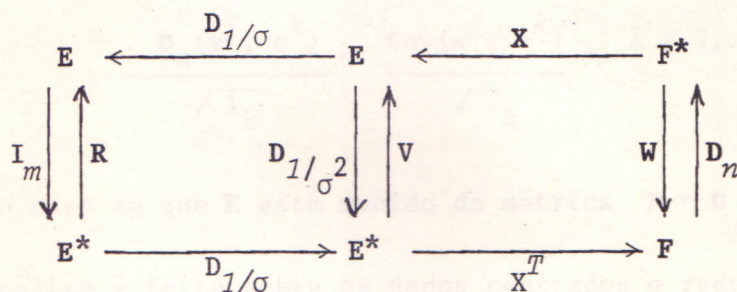
2.5.2 - Duas Métricas Particulares

Na análise em componentes principais, em geral munimos o espaço E , dos indivíduos, da métrica definida por:

$$M(e_j, e_k) = 0 \quad \text{se } j \neq k$$

$$M(e_j, e_j) = 1/\text{Var}(x^j) = 1/\sigma_j^2.$$

Consideremos o seguinte esquema de dualidade:



Fazendo $z_i = D_{1/\sigma} x_i$ e $z_k = D_{1/\sigma} x_k$, temos:

$$||z_i - z_k||_{I_m} = ||x_i - x_k||_{D_{1/\sigma^2}}$$

A nuvem dos pontos $z_i = D_{1/\sigma} x_i$ é associada à tabela $Z = D_{1/\sigma} X$.

A matriz da forma quadrática de inércia da nuvem dos pontos $\mathbf{z}_i$ é então dada por

$$\mathbf{Z} \mathbf{D}_n \mathbf{Z}^T = \mathbf{D}_{1/\sigma} (\mathbf{X} \mathbf{D}_n \mathbf{X}^T) \mathbf{D}_{1/\sigma} = \mathbf{D}_{1/\sigma} \mathbf{V} \mathbf{D}_{1/\sigma} = \mathbf{R}$$

Então, efetuar uma análise em componentes principais no espaço E , com a métrica $\mathbf{M} = \mathbf{D}_{1/\sigma^2}$, é equivalente a efetuá-la com os dados centrados e reduzidos e com a métrica clássica $\mathbf{I}_m$. Além disso, os resultados independem das unidades de medida, pois a matriz a diagonalizar $\mathbf{R} \mathbf{I}_m$, tem esta propriedade.

2.5.3 - Descrição dos Caracteres

O Conjunto

$$\left\{ \frac{\mathbf{c}^k}{\sqrt{\lambda_k}} ; k = 1, \dots, q \right\}$$

constituindo uma base $\mathbf{D}_n$ -ortonormal do subespaço vetorial gerado pelos caracteres iniciais, temos que as coordenadas de um caracter $\mathbf{x}^j$ são dadas por

$$\frac{\mathbf{D}_n(\mathbf{x}^j, \mathbf{c}^k)}{\sqrt{\lambda_k}} = \frac{\text{Cov}(\mathbf{x}^j, \mathbf{c}^k)}{\sqrt{\lambda_k}}, \quad k = 1, \dots, q$$

No caso em que E está munido da métrica $\mathbf{M} = \mathbf{D}_{1/\sigma^2}$, ou seja, quando a análise é feita sobre os dados centrados e reduzidos com a métrica dada pela matriz identidade $\mathbf{I}_m$, as coordenadas dos caracteres reduzidos $\mathbf{x}^j/\sigma(\mathbf{x}^j)$ na base $\mathbf{D}_n$ -ortonormal acima são exatamente os coeficientes de correlação dos caracteres com as componentes principais, visto que

$$\frac{1}{\sqrt{\lambda_k}} \mathbf{D}_n \left(\frac{\mathbf{x}^j}{\sigma(\mathbf{x}^j)}, \mathbf{c}^k \right) = \frac{\text{Cov}(\mathbf{x}^j, \mathbf{c}^k)}{\sigma(\mathbf{x}^j) \sqrt{\lambda_k}} = \text{Corr}(\mathbf{x}^j, \mathbf{c}^k) ;$$

bem como os caracteres estão situados sobre a esfera de centro na origem do $\mathbf{F}$ e raio unitário, pois:

$$D_n \left(\frac{\mathbf{x}^j}{\sigma(\mathbf{x}^j)} , \frac{\mathbf{x}^j}{\sigma(\mathbf{x}^j)} \right) = 1 .$$

Observação: geralmente fazemos figurar no mesmo plano principal $[\mathbf{u}_k, \mathbf{u}_l]$, onde estão representados os indivíduos, as projeções dos caracteres $\mathbf{x}^j$ no plano correspondente $[\mathbf{c}^k, \mathbf{c}^l]$. Entretanto, esta representação simultânea não tem outra conotação senão aquela de um *artifício gráfico*.

CAPÍTULO 3

A ANÁLISE DE CORRESPONDÊNCIAS

A análise de correspondências é um método de análise de dados cujo domínio ótimo de aplicação é constituído pelas *tabelas de contingência* de dupla entrada.

Duas variáveis qualitativas x e y são observadas sobre uma população. Suponhamos que as variáveis x e y apresentem, respectivamente, m e n modalidades exclusivas, e denotemos por $J = \{1, 2, \dots, m\}$ o conjunto das modalidades de x , e por $I = \{1, 2, \dots, n\}$ o das de y .

A tabela de contingência é obtida sob a forma de uma matriz $K = (k_{ij}^j)$, onde k_{ij}^j representa o número de indivíduos na população que se enquadram simultaneamente na modalidade j da variável x e na modalidade i da variável y .

A técnica utilizada é, essencialmente, a da análise em componentes principais: os conjuntos J e I serão representados por nuvens de pontos em espaços euclidianos e cada nuvem, por sua vez, será projetada em planos principais nos respectivos espaços.

Entretanto, veremos que na análise de correspondências a geometria das nuvens é muito peculiar e os espaços são munidos de métricas específicas, responsáveis por uma grande estabilidade dos resultados. Além disso, será possível obter uma representação simultânea das duas nuvens em um mesmo plano principal com relações estruturais bem definidas entre os pontos projetados.

3.1 - As Transformações Iniciais

3.1.1 - A Matriz de Correspondências. Perfis

Consideremos dois conjuntos finitos $J = \{1, 2, \dots, m\}$ e $I = \{1, 2, \dots, n\}$ suponhamos que a cada par $(j, i) \in J \times I$ está associado um número real não negativo. Denotemos por K a matriz com m linhas e n colunas cujo elemento genérico é k_{ij}^j ; suponhamos ainda que $k_{i\cdot}$ e k^j definidos por

$$k_{i\cdot} = \sum_{j=1}^m k_{ij}^j \quad k^j = \sum_{i=1}^n k_{ij}^j, \quad ,$$

sejam números reais estritamente positivos, e seja k a soma de todos os elementos de K .

A matriz K com estas características é a *tabela dos dados a analisar* (em particular, podendo ser uma tabela de contingência).

3.1.1.a - Definição

A matriz denotada por P e definida por

$$P = (p_{ij}^j), \text{ onde } p_{ij}^j = \frac{k_{ij}^j}{k},$$

será chamada *matriz de correspondências associada à tabela* K .

Observação. Quando K é uma tabela de contingência, P pode ser considerada como amostra de uma lei de probabilidade nas células da matriz $J \times I$. Quando K é outro tipo de tabela, temos somente uma analogia.

A partir da matriz de correspondências P definimos duas outras

matrizes que lhe são associadas. Para isto, utilizaremos as seguintes notações:

- 1) p_i : a soma dos elementos da i -ésima coluna de P ;

$$p_i = \sum_{j=1}^m p_i^j$$

- 2) p^j : a soma dos elementos da j -ésima linha de P ;

$$p^j = \sum_{i=1}^n p_i^j$$

- 3) D_n : a matriz **diag** ($p_1, \dots, p_n$) ; assim,

$$(D_n)^{-1} = \text{diag} \left(\frac{1}{p_1}, \dots, \frac{1}{p_n} \right)$$

- 4) D^m : a matriz **diag** ($p^1, \dots, p^m$) ; assim,

$$(D^m)^{-1} = \text{diag} \left(\frac{1}{p^1}, \dots, \frac{1}{p^m} \right)$$

3.1.1.b - Definição

- i) A matriz denotada por X e definida por

$$X = P (D_n)^{-1},$$

será chamada *matriz dos perfis das colunas* de K . A i -ésima coluna de X será representada pelo vetor x_i , que por sua vez é denominado *perfil da i -ésima coluna* de K .

- ii) A matriz denotada por Y e definida por

$$Y = P^T (D^m)^{-1}$$

será chamada *matriz dos perfis das linhas* de K . A j -ésima coluna de Y será representada pelo vetor y_j , que por sua vez é denominado *perfil da j -ésima linha* de K .

De acordo com esta definição, temos:

i) As coordenadas de $\mathbf{x}_i$ são dadas por

$$x_i^j = \frac{p_i^j}{p_i}, \quad j = 1, 2, \dots, m,$$

e satisfazem à relação

$$\sum_{j=1}^m x_i^j = 1;$$

isto nos permite concluir que os n pontos $\mathbf{x}_i$ pertencem a um hiperplano $(m-1)$ -dimensional de $E = \mathbb{R}^m$, que denotaremos por Γ .

ii) De modo semelhante, as componentes de $\mathbf{y}_j$, dadas por

$$y_j^i = \frac{p_j^i}{p_j}, \text{ satisfazem à relação}$$

$$\sum_{i=1}^n y_j^i = 1$$

o que nos indica que os m pontos $\mathbf{y}_j$ pertencem a um hiperplano $(n-1)$ -dimensional do $F = \mathbb{R}^n$, que denotaremos por Ω .

3.1.1.c - A Nuvem dos Perfis de Colunas. Centro de Gravidade

Agora, a cada perfil de coluna de K , $\mathbf{x}_i$, associaremos uma massa ou peso igual a p_i , refletindo a importância relativa da i -ésima coluna de K no contexto da tabela. A *nuvem dos perfis de colunas* que será denotada por $N(C)$ e definida por

$$N(C) = \{(\mathbf{x}_i, p_i); i=1, 2, \dots, n\}.$$

O centro de gravidade desta nuvem, denotado por $\mathbf{g}_c$, é então definido por

$$g_c = \sum_{i=1}^n p_i x_i ,$$

e tem como j -ésima componente a soma dos elementos da j -ésima linha de P ; com efeito,

$$g_c^j = \sum_{i=1}^n p_i x_i^j = \sum_{i=1}^n p_i^j = p^j .$$

Além disso, g_c também pertence ao hiperplano x que contém a nuvem, pois

$$\sum_{j=1}^m p^j = 1 .$$

3.1.1.d - A Nuvem dos Perfis de Linhas

De modo análogo, a cada perfil de linha de K , ou y_j , associamos a massa p^j igual à soma dos elementos da j -ésima linha de P . Denotamos por $N(L)$ a nuvem dos perfis de linha dada por

$$N(L) = \{(y_j, p^j); j=1, 2, \dots, m\} .$$

O centro de gravidade desta nuvem, denotado por g_L e definido por

$$g_L = \sum_{j=1}^m p^j y_j ,$$

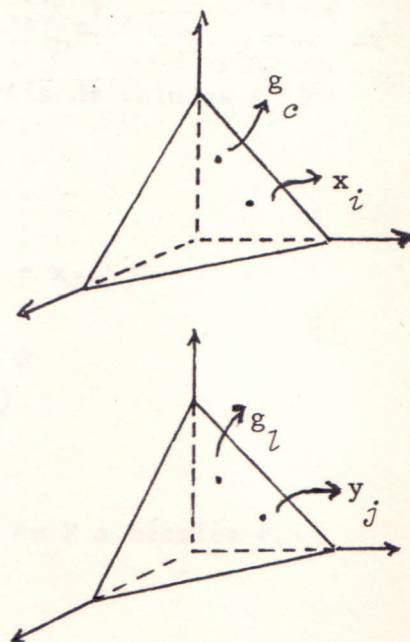
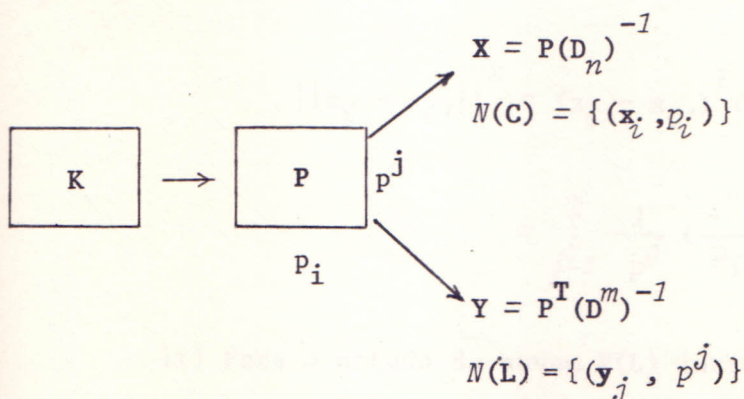
tem por i -ésima componente

$$g_L^i = \sum_{j=1}^m p^j y_j^i = \sum_{j=1}^m p_i^j = p_i .$$

Analogamente, g_L pertence ao hiperplano Ω , desde que

$$\sum_{i=1}^n p_i = 1$$

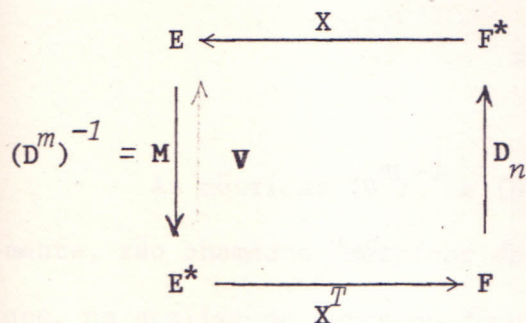
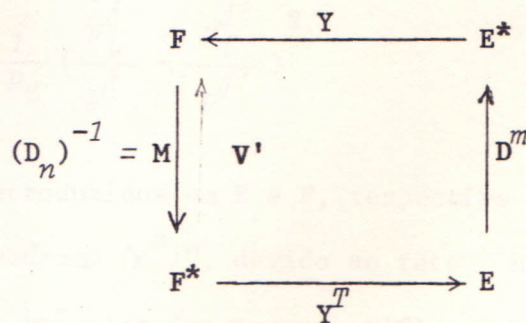
RESUMO



3.2 - As Métricas

3.2.1 - Os Esquemas de Dualidade para $N(C)$ e $N(L)$

Com as notações introduzidas no parágrafo anterior, os esquemas de dualidade serão os seguintes:

i) $N(C)$ ii) $N(L)$ 

i) Para o estudo da nuvem $N(C)$ introduzimos em E a métrica definida por

$$M(e_j, e_k) = \frac{\delta_j^k}{p_j^j}, \quad j, k = 1, 2, \dots, m,$$

ou seja, a métrica dada pela matriz $(D^m)^{-1} = \text{diag}(\frac{1}{p_1}, \dots, \frac{1}{p_m})$.

Com esta métrica, a distância entre dois perfis de colunas x_i e $x_{i'}$ é dada por

$$\begin{aligned} ||x_i - x_{i'}||^2 &= (x_i - x_{i'})^T (D^m)^{-1} (x_i - x_{i'}) = \\ &= \sum_{j=1}^m \frac{1}{p_j^j} \left(\frac{p_i^j}{p_i} - \frac{p_{i'}^j}{p_{i'}} \right)^2 \end{aligned}$$

ii) Para o estudo da nuvem $N(L)$ introduzimos em F a métrica definida por

$$M(f_i, f_k) = \frac{\delta_i^k}{p_i}, \quad i, k = 1, 2, \dots, n$$

ou seja, a métrica dada pela matriz $(D_n)^{-1} = \text{diag}(\frac{1}{p_1}, \dots, \frac{1}{p_n})$.

Com esta métrica, a distância entre dois perfis de colunas y_j e $y_{j'}$ é dada por

$$\begin{aligned} ||y_j - y_{j'}||^2 &= (y_j - y_{j'})^T (D_n)^{-1} (y_j - y_{j'}) = \\ &= \sum_{i=1}^n \frac{1}{p_i} \left(\frac{p_i^j}{p_j} - \frac{p_i^{j'}}{p_{j'}} \right)^2 \end{aligned}$$

As métricas $(D^m)^{-1}$ e $(D_n)^{-1}$ introduzidos em E e F , respectivamente, são chamados "métricas do Chi-quadrado (χ^2)", devido ao fato de que, na análise de correspondências, as inércias das nuvens $N(C)$ e $N(L)$ em relação aos respectivos centros de gravidade g_c e g_ℓ , têm uma interpretação estatística peculiar; especificamente relacionada com a estatística de lei χ^2 , como veremos a seguir.

Para a nuvem $N(C)$, temos:

$$\begin{aligned} \ln \{g_c\} &= \sum_{i=1}^n p_i (x_i - g_c)^T (D^n)^{-1} (x_i - g_c) = \\ &= \sum_{i=1}^n p_i \sum_{j=1}^n \frac{1}{p^j} \left(\frac{p_i^j}{p_i} - p^j \right)^2 = \\ &= \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i^j - p_i p^j)^2}{p_i p^j} . \end{aligned}$$

Por outro lado, para a nuvem $N(L)$, temos:

$$\begin{aligned} \ln \{g_l\} &= \sum_{j=1}^m p^j (y_j - g_l)^T (D_n)^{-1} (y_j - g_l) = \\ &= \sum_{j=1}^m p^j \sum_{i=1}^n \frac{1}{p_i} \left(\frac{p_i^j}{p^j} - p_i \right)^2 = \\ &= \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i^j - p_i p^j)^2}{p_i p^j} \end{aligned}$$

Então:

$$\begin{aligned} \ln \{g_c\} = \ln \{g_l\} &= \sum_{i=1}^n \sum_{j=1}^m \left(\frac{k_i^j}{k} - \frac{k_i k^j}{k^2} \right)^2 \frac{1}{\frac{k_i k^j}{k^2}} = \\ &= \sum_{i=1}^n \sum_{j=1}^m \frac{1}{k} \left(k_i^j - \frac{k_i k^j}{k} \right)^2 \frac{1}{\frac{k_i k^j}{k}} = \\ &= \frac{1}{k} \sum_{i=1}^n \sum_{j=1}^m (k_i^j - e_i^j) / e_i^j \end{aligned}$$

onde $e_i^j = \frac{k_i k^j}{k}$; ou seja, fazendo $\chi^2 = \sum_{i=1}^n \sum_{j=1}^m (k_i^j - e_i^j) / e_i^j$ (estatística

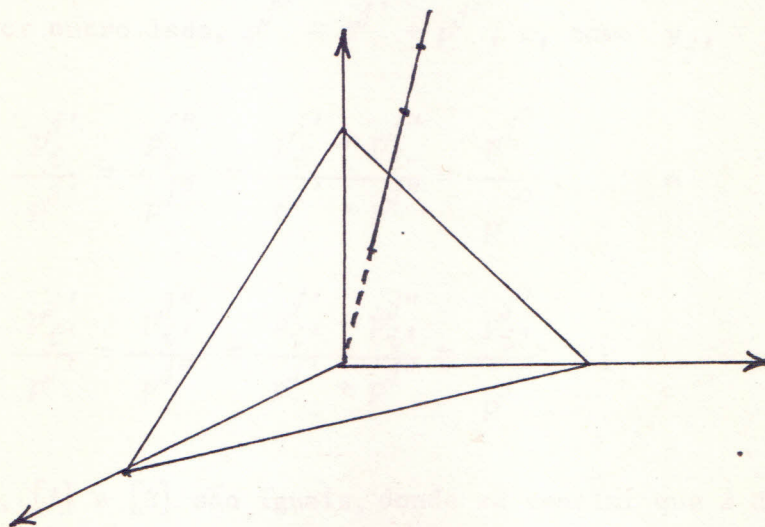
do chi-quadrado), tem-se:

$$\ln\{g_c\} = \ln\{g_l\} = \frac{\chi^2}{k}$$

Entretanto, a maior vantagem da "métrica do χ^2 " reside no fato de que ela verifica uma propriedade essencial para a estabilidade dos resultados da análise, como veremos a seguir.

3.2.2 - O Princípio da Equivalência Distribucional

Substituindo os dados brutos da tabela K por porcentagens, tornamos idênticas duas linhas proporcionais. Ou seja, duas linhas proporcionais em K terão o mesmo perfil no hiperplano Ω , como ilustra a figura:



Suponhamos, pois, que duas linhas de K, j' e j'' tenham o mesmo perfil. Como vimos $y_{j'}$ e $y_{j''}$ coincidem em Ω . Agreguemos j' e j'' em uma única j^0 e de tal modo que $p^{j^0} = p^{j'} + p^{j''}$. É claro que o peso associado a um perfil de coluna não muda. Porém, é necessário verificar se a junção de linhas acarreta mudança na distância entre dois perfis de colunas i e i' .

Note-se que na expressão da distância entre x_i e $x_{i'}$, os índices j' e j'' aparecem somente em duas parcelas:

$$||\mathbf{x}_i - \mathbf{x}_{i'}||^2 = \dots + \frac{1}{p_i^{j'}} \left(\frac{p_i^{j'}}{p_i} - \frac{p_i^{j'}}{p_{i'}} \right)^2 + \dots + \frac{1}{p_i^{j''}} \left(\frac{p_i^{j''}}{p_i} - \frac{p_i^{j''}}{p_{i'}} \right)^2 + \dots ;$$

após a junção das linhas, tais parcelas são substituídas por

$$\frac{1}{p_i^{j^0}} \left(\frac{p_i^{j^0}}{p_i} - \frac{p_i^{j^0}}{p_{i'}} \right)^2 = p_i^{j^0} \left(\frac{p_i^{j^0}}{p_i p_i^{j^0}} - \frac{p_i^{j^0}}{p_{i'} p_i^{j^0}} \right)^2 \quad [1] .$$

A soma das duas parcelas originais pode ser escrita da seguinte forma:

$$p_i^{j'} \left(\frac{p_i^{j'}}{p_i p_i^{j'}} - \frac{p_i^{j'}}{p_{i'} p_i^{j'}} \right)^2 + p_i^{j''} \left(\frac{p_i^{j''}}{p_i p_i^{j''}} - \frac{p_i^{j''}}{p_{i'} p_i^{j''}} \right)^2 \quad [2] .$$

Por outro lado, $p_i^{j^0} = p_i^{j'} + p_i^{j''}$, e, como $y_{j'} = y_{j''}$, temos:

$$\frac{p_i^{j'}}{p_i^{j'}} = \frac{p_i^{j''}}{p_i^{j''}} = \frac{p_i^{j'} + p_i^{j''}}{p_i^{j'} + p_i^{j''}} = \frac{p_i^{j^0}}{p_i^{j^0}} , \quad e$$

$$\frac{p_i^{j'}}{p_i^{j'}} = \frac{p_i^{j''}}{p_i^{j''}} = \frac{p_i^{j'} + p_i^{j''}}{p_i^{j'} + p_i^{j''}} = \frac{p_i^{j^0}}{p_i^{j^0}} ; .$$

deste modo, [1] e [2] são iguais, donde se conclui que a distância entre os perfis de colunas $\mathbf{x}_i$ e $\mathbf{x}_{i'}$ não é alterada.

Esta propriedade é conhecida como o *princípio da equivalência distribucional*, a saber:

"Se duas linhas de $\mathbf{K}$, j' e j'' , têm o mesmo perfil $y_{j'} = y_{j''}$, podemos considerar um único perfil y_0 , do qual a i -ésima componente é $p_i^{j'} + p_i^{j''}$, com massa igual a $p_i^{j^0} = p_i^{j'} + p_i^{j''}$, sem alterar as distâncias entre os perfis das colunas".

Evidentemente, temos também o *princípio dual*, que permite agre-

gar colunas com mesmo perfil, sem alterar as distâncias entre perfis de linhas.

3.3 - Nota Sobre Centragem de Dados

Sabemos do capítulo 2 que a análise em componentes principais em uma nuvem deve ser realizada com os dados centrados.

Assim, a análise em componentes principais da nuvem $N(C)$ deve ser efetuada sobre a nuvem

$$\tilde{N}(C) = \{(\mathbf{x}_i - \mathbf{g}_c), p_i\} ; i=1,2,\dots,n\} ,$$

O esquema de dualidade sendo dado por

$$\begin{array}{ccc} E & \xleftarrow{X} & F^* \\ \downarrow (D^m)^{-1} = M & \uparrow V & \uparrow D_n \\ E^* & \xrightarrow{X^T} & F \end{array}$$

a matriz a diagonalizar para a obtenção das direções dos eixos principais é então

$$\tilde{S} = \tilde{V}M = \tilde{X} D_n \tilde{X}^T (D^m)^{-1} = (X-G)D_n(X-G)^T (D^m)^{-1} ,$$

onde G é a matriz cujas n colunas são iguais a $\mathbf{g}_c$.

3.3.1 - Proposição

$$\left| \begin{array}{l} \tilde{S} \mathbf{g}_c = 0 \mathbf{g}_c \quad \text{e} \quad S \mathbf{g}_c = 1 \mathbf{g}_c \\ \text{onde} \quad S = X D_n X^T \end{array} \right.$$

Demonstração

Denotemos por $\mathbf{l}_m$ e $\mathbf{l}_n$, respectivamente, os vetores colunas $m \times 1$ e $n \times 1$, com todos seus elementos iguais a 1. É fácil verificar que $(\mathbf{D}^m)^{-1} \mathbf{g}_c = \mathbf{l}_m$, $\mathbf{X}^T \mathbf{l}_m = \mathbf{l}_n$ e $\mathbf{G}^T \mathbf{l}_m = \mathbf{l}_n$; bem como, $\mathbf{X} \mathbf{D}_n \mathbf{l}_n = \mathbf{g}_c$. Onde:

$$\tilde{\mathbf{S}} \mathbf{g}_c = (\mathbf{X}-\mathbf{G}) \mathbf{D}_n (\mathbf{X}-\mathbf{G})^T (\mathbf{D}^m)^{-1} \mathbf{g}_c = (\mathbf{X}-\mathbf{G}) \mathbf{D}_n (\mathbf{X}-\mathbf{G})^T \mathbf{l}_m =$$

$$= (\mathbf{X}-\mathbf{G}) \mathbf{D}_n \mathbf{X}^T \mathbf{l}_m - \mathbf{D}_n \mathbf{G}^T \mathbf{l}_m = (\mathbf{X}-\mathbf{G}) (\mathbf{D}_n \mathbf{l}_n - \mathbf{D}_n \mathbf{l}_n)$$

$$= \mathbf{0}_E = \mathbf{0} \mathbf{g}_c ;$$

$$\mathbf{S} \mathbf{g}_c = \mathbf{X} \mathbf{D}_n \mathbf{X}^T (\mathbf{D}^m)^{-1} \mathbf{g}_c = \mathbf{X} \mathbf{D}_n \mathbf{l}_n = \mathbf{l} \mathbf{g}_c .$$

3.3.2 - Sabemos do capítulo 2 que podemos construir em E uma base $(\mathbf{D}^m)^{-1}$ -ortonormal constituída pelos auto-vetores de $\tilde{\mathbf{S}}$. O vetor $\mathbf{g}_c$ sendo um destes, temos, para $\mathbf{M} = (\mathbf{D}^m)^{-1}$:

$$\mathbf{g}_c^T \mathbf{M} \tilde{\mathbf{u}} = \mathbf{l}_m^T \tilde{\mathbf{u}} = 0 ,$$

para todo auto-vetor $\tilde{\mathbf{u}}$ de $\tilde{\mathbf{S}}$, distinto de $\mathbf{g}_c$.

Seja então $\tilde{\mathbf{u}}$ um auto-vetor de $\tilde{\mathbf{S}}$ distinto de $\mathbf{g}_c$ associado a um auto-valor λ não nulo. Mostraremos que $\tilde{\mathbf{u}}$ também é auto-vetor de $\mathbf{S}$, associado ao mesmo auto-valor λ .

Com efeito:

$$\tilde{\mathbf{S}} \tilde{\mathbf{u}} = (\mathbf{X}-\mathbf{G}) \mathbf{D}_n (\mathbf{X}-\mathbf{G})^T \mathbf{M} \tilde{\mathbf{u}} = \lambda \tilde{\mathbf{u}}$$

$$\iff (\mathbf{X}-\mathbf{G}) \mathbf{D}_n [\mathbf{X}^T \mathbf{M} \tilde{\mathbf{u}} - \mathbf{G}^T \mathbf{M} \tilde{\mathbf{u}}] = \lambda \tilde{\mathbf{u}}$$

$$\iff (\mathbf{X} - \mathbf{G}) \mathbf{D}_n \mathbf{X}^T \mathbf{M} \tilde{\mathbf{u}} = \lambda \tilde{\mathbf{u}}$$

$$\iff \mathbf{X} \mathbf{D}_n \mathbf{X}^T \mathbf{M} \tilde{\mathbf{u}} - \begin{pmatrix} p^1 \left[\sum_{i=1}^n p_i \mathbf{M}(\mathbf{x}_i, \tilde{\mathbf{u}}) \right] \\ \vdots \\ p^m \left[\sum_{i=1}^n p_i \mathbf{M}(\mathbf{x}_i, \tilde{\mathbf{u}}) \right] \end{pmatrix} = \lambda \tilde{\mathbf{u}}$$

$$\iff \mathbf{X} \mathbf{D}_n \mathbf{X}^T \mathbf{M} \tilde{\mathbf{u}} - \begin{pmatrix} p^1 \mathbf{M}(\mathbf{g}_c, \tilde{\mathbf{u}}) \\ \vdots \\ p^m \mathbf{M}(\mathbf{g}_c, \tilde{\mathbf{u}}) \end{pmatrix} =$$

$$\iff \mathbf{X} \mathbf{D}_n \mathbf{X}^T \mathbf{M} \tilde{\mathbf{u}} = \lambda \tilde{\mathbf{u}}$$

Deste modo, $\tilde{\mathbf{S}}$ e $\mathbf{S}$ têm os mesmos auto-vetores com auto-valores correspondentes também iguais; exceto quanto a $\mathbf{g}_c$, que é associado a 0 no caso de $\tilde{\mathbf{S}}$ e a 1, no caso de $\mathbf{S}$. A base $\mathbf{M}$ -ortonormal $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\}$ de E , fornecendo os eixos principais (de origem $\mathbf{g}_c$) da análise em componentes principais da nuvem $N(C)$, pode então ser construída com a ajuda dos auto-vetores de $\mathbf{S}$.

O vetor $\mathbf{g}_c$ pertence ao auto-espço de $\tilde{\mathbf{S}}$ associado ao auto-valor $\lambda = 0$. Se a dimensão deste auto-espço é, por exemplo, $m - q \geq 1$, a nuvem $\tilde{N}(C)$ está situada no subespço A de dimensão $q \leq m - 1$ gerado pelos q primeiros eixos principais. Notemos que A está contido no sub-espço $\mathbf{M}$ -ortogonal à direção de $\mathbf{g}_c$ e é gerado pelos auto-vetores de $\mathbf{S}$, sendo porém excluídos o vetor $\mathbf{g}_c$ associado ao auto-vetor 1 e os auto-vetores associados a um auto-valor nulo.

Evidentemente, resultado análogo é válido para o estudo da nuvem $N(C)$.

3.4 - Esquemas para a Análise em Componentes Principais das Nuvens $N(C)$ e $N(L)$.

Os esquemas de dualidade são os seguintes:

Para a nuvem $N(C)$:

$$\begin{array}{ccc}
 E & \xleftarrow{X} & F^* \\
 \downarrow (D^m)^{-1} & \uparrow V & \uparrow (D_n) \\
 E^* & \xrightarrow{X^T} & F
 \end{array}$$

$$X = P(D_n)^{-1}$$

Matriz a diagonalizar:

$$\begin{aligned}
 S &= V(D^m)^{-1} = XD_n X^T (D^m)^{-1} = \\
 &= P(D_n)^{-1} D_n (D_n)^{-1} P^T (D^m)^{-1} = \\
 &= P(D_n)^{-1} P^T (D^m)^{-1}
 \end{aligned}$$

Para a nuvem $N(L)$:

$$\begin{array}{ccc}
 F & \xleftarrow{Y} & E^* \\
 \downarrow (D_n)^{-1} & \uparrow V' & \uparrow D^m \\
 F^* & \xrightarrow{Y^T} & F
 \end{array}$$

$$Y = P^T (D^m)^{-1}$$

Matriz a diagonalizar:

$$\begin{aligned}
 S' &= V' (D_n)^{-1} = Y (D^m) Y^T (D_n)^{-1} = \\
 &= P^T (D^m)^{-1} D^m (D^m)^{-1} P (D_n)^{-1} \\
 &= P^T (D^m)^{-1} P (D_n)^{-1}
 \end{aligned}$$

3.4.1 Temos as seguintes relações entre S e S' :

$$a) \quad P(D_n)^{-1} S' = S P(D_n)^{-1}$$

$$b) \quad S' P^T(D^m)^{-1} = P^T(D^m)^{-1} S.$$

Seja $u \in E$ tal que $Su = \lambda u$, sendo u um vetor $(D^m)^{-1}$ -normado e $\lambda \neq 0$, então :

$$P^T(D^m)^{-1} Su = \lambda P^T(D^m)^{-1} u,$$

ora, a relação b) acima nos permite obter

$$S' (P^T(D^m)^{-1} u) = \lambda (P^T(D^m)^{-1} u)$$

Esta última igualdade nos mostra que $P^T(D^m)^{-1} u$ é auto-vetor de S' associado ao mesmo auto-valor λ . Além disso, verificamos facilmente que

$$\frac{1}{\sqrt{\lambda}} P^T(D^m)^{-1} u \text{ é } (D_n)^{-1}\text{-normado.}$$

De modo semelhante, seja u' e F tal que $S'u' = \lambda' u'$, com $\lambda' \neq 0$ e u' $(D_n)^{-1}$ -normado. Como

$$P(D_n)^{-1} S'u' = \lambda' P(D_n)^{-1} u',$$

a relação a) dada acima nos permite obter

$$S(P(D_n)^{-1} u') = \lambda' (P(D_n)^{-1} u').$$

Esta última igualdade mostra que $P(D_n)^{-1} u'$ é auto-vetor de S com o mesmo auto-valor λ' . Além disso,

$$\frac{1}{\sqrt{\lambda'}} P(D_n)^{-1} u' \text{ é } (D_m)^{-1}\text{-normado.}$$

S e S' possuem então os mesmos valores próprios e, uma vez determinados os eixos principais de uma das nuvens $N(C)$ ou $N(L)$, podemos facilmente deduzir os da outra nuvem.

3.4.1 - A Análise em E para a Nuvem $N(C)$

Suponhamos que o resultado de uma análise em componentes principais em E para a nuvem $N(C)$ tenham sido fornecido:

u_α : vetor próprio $(D^m)^{-1}$ -normado de $S = V(D^m)^{-1}$ associado a $\lambda_\alpha \neq 0$.

v_α : o fator associado em E^* ,

$$v_\alpha = (D^m)^{-1} u_\alpha.$$

ψ^α : a componente principal associada em F,

$$\psi^\alpha = X^T v_\alpha = X^T (D^m)^{-1} u_\alpha$$

Obtemos, imediatamente, os seguintes elementos para a análise da nuvem $N(L)$ em F:

$$1) \quad u'_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} P^T (D^m)^{-1} u_\alpha$$

é auto-vetor $(D_n)^{-1}$ -normado de S' , associado ao auto-valor λ_α ;

$$\begin{aligned} 2) \quad v_\alpha^1 &= (D_n)^{-1} u'_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (D_n)^{-1} P^T (D^m)^{-1} u_\alpha = \\ &= \frac{1}{\sqrt{\lambda_\alpha}} (D_n)^{-1} P^T v_\alpha \end{aligned}$$

é o fator principal correspondente, em F^* , e,

$$\begin{aligned} 3) \quad \psi^\alpha &= Y^T v_\alpha^1 = \frac{1}{\sqrt{\lambda_\alpha}} (D^m)^{-1} P (D_n)^{-1} P^T v_\alpha = \\ &= \frac{1}{\sqrt{\lambda_\alpha}} (D^m)^{-1} P X^T v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (D^m)^{-1} P \psi^\alpha \end{aligned}$$

a componente principal correspondente, em E.

3.4.2 - A Análise em F para $N(L)$

Como S e S' têm os mesmos auto-valores, denotando por u'_α o auto-vetor $(D_n)^{-1}$ -normado de S' associado a λ_α , temos :

$$v' = (D_n)^{-1} u'_\alpha$$

e

$$\psi^\alpha = Y^T v'_\alpha = Y^T (D_n)^{-1} u'_\alpha$$

são, respectivamente o fator principal e a componente principal associados a λ_α .

Obtemos imediatamente os seguintes resultados para a análise da nuvem $N(C)$ em E :

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} P (D_n)^{-1} u'_\alpha ;$$

1') é auto-vetor $(D^m)^{-1}$ -normado de $S = V(D^m)^{-1}$ associado ao auto-valor λ_α ;

$$\begin{aligned} v_\alpha &= (D^m)^{-1} u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (D^m)^{-1} P (D_n)^{-1} u'_\alpha = \\ &= \frac{1}{\sqrt{\lambda_\alpha}} (D^m)^{-1} P v'_\alpha ; \end{aligned}$$

2') é o fator principal correspondente, em E^* ;

$$\begin{aligned} \varphi^\alpha &= X^T v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (D_n)^{-1} P^T (D^m)^{-1} P v'_\alpha = \\ &= \frac{1}{\sqrt{\lambda_\alpha}} (D_n)^{-1} P^T Y^T v'_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (D_n)^{-1} P^T \psi_\alpha \end{aligned}$$

3') é a componente principal correspondente, em F .

Assim, as relações entre φ^α e ψ^α são dadas pelas fórmulas obtidas em 3) e 3') acima:

$$\psi^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (D^m)^{-1} P \varphi^\alpha$$

e

$$\varphi^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (D_n)^{-1} P^T \psi^\alpha$$

A j -ésima componente de ψ^α , ou seja, a abscissa da projeção $(D_n)^{-1}$ -ortogonal de y_j sobre o eixo $[u'_\alpha]$, é então dada por:

$$\psi_j^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^n \frac{p_i^j}{p^j} \varphi_i^\alpha, \quad j=1,2,\dots,m \quad [4]$$

De modo semelhante, a i -ésima componente de φ^α , ou seja, a abscissa da projeção $(D^m)^{-1}$ -ortogonal de x_i sobre o eixo $[u_\alpha]$, é dada por:

$$\varphi_i^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^m \frac{p_i^j}{p_i} \psi_j^\alpha, \quad i=1,2,\dots,n \quad [4']$$

Portanto, a menos do coeficiente $1/\sqrt{\lambda_\alpha}$, as abscissas das projeções dos pontos de uma nuvem sobre o eixo principal associado ao autovalor $\lambda_\alpha \neq 0$ são "*baricentros*" das projeções dos pontos da outra nuvem sobre o eixo principal associado ao mesmo autovalor.

As fórmulas [4] e [4'] são denominadas "*relações baricêntricas*".

Adotando uma representação cartesiana dos sistemas de eixos principais, de tal modo que os centros de gravidade das duas nuvens coincidam, assim como os eixos $[\mu_k]$ e $[\mu'_k]$ correspondentes a $\lambda_k \neq 0$, obtemos uma imagem simultânea das duas nuvens.

É necessário não esquecer que as duas nuvens projetadas em um mesmo plano pertencem, em geral, a espaços de dimensões diferentes e que somente as fórmulas baricêntricas permitem interpretar as proximidades entre as projeções dos pontos das diferentes nuvens.

A ANÁLISE DE CORRESPONDÊNCIAS MÚLTIPLAS

A análise de correspondências introduzida no capítulo 3, dita binária, pode generalizar-se para o caso em que mais de dois conjuntos são colocados em correspondência. A análise de correspondências múltiplas aqui apresentada é, pois, uma extensão do domínio de aplicação da análise binária.

4.1 - Notações

Seja I um conjunto com N indivíduos sobre os quais são observados q caracteres qualitativos, $C_1, C_2 \dots C_q$.

Supõe-se que cada caracter C_r possui m_r modalidades mutuamente exclusivas e que, para cada caracter, um indivíduo j escolhe uma e somente uma das modalidades possíveis; ou seja, que os caracteres tem a *forma disjuntiva completa*. As modalidades não escolhidas por nenhum dos N indivíduos devem ser eliminadas.

Os resultados podem ser dispostos em uma tabela com N linhas e $m_1 + m_2 + \dots + m_q = m$ colunas, através de um código binário, da seguinte maneira:

i) Consideremos um dos caracteres, C_r , com modalidades $1, 2, \dots, i, \dots, m_r$; denotemos por Z_r a matriz com N linhas e m_r colunas tal que o elemento na interseção da j -ésima linha com a i -ésima coluna vale 1, se o indivíduo j escolhe a modalidade i de C_r , e zero, caso contrário;

ii) A justaposição das q matrizes assim construídas será denotada por Z :

$$Z = \begin{bmatrix} Z_1 & \vdots & Z_2 & \vdots & \dots & \vdots & Z_r & \vdots & \dots & \vdots & Z_q \end{bmatrix}$$

4.2 - A Análise em $Z = [z_1 \mid z_2 \mid \cdots \mid z_q]$

A tabela $K = Z$ tem N linhas e $m_1 + \cdots + m_q = m$ colunas.

Conforme as notações gerais da análise fatorial de correspondências, temos:

$$1) \quad K = Z, \text{ com } k_{\cdot}^j \in \{0, 1\}, \quad k^j = q, \quad k_{\cdot}^i = \Delta^i \quad \text{e} \quad k = qN$$

$$2) \quad P = \frac{1}{qN} \cdot K = \frac{1}{qN} Z$$

$$3) \quad D_N = \text{diag}(p^1, \dots, p^N) = \text{diag}\left(\frac{1}{N}, \dots, \frac{1}{N}\right) = \frac{1}{N} I_N$$

$$D_m = \text{diag}(p_1, \dots, p_m) = \text{diag}(\Delta^1/qN, \dots, \Delta^m/qN) = \frac{1}{qN} \Delta$$

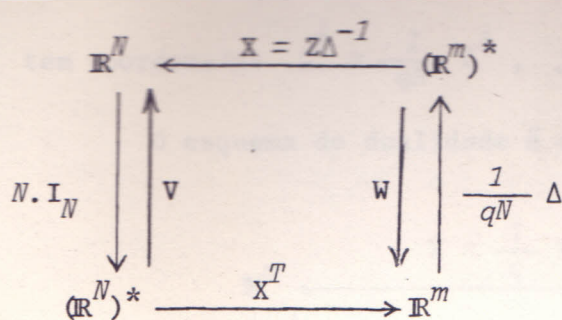
$$4) \quad X = P D_m^{-1} = \frac{1}{qN} Z \cdot qN \Delta^{-1} = Z \Delta^{-1}$$

$$5) \quad Y = P^T D_N^{-1} = \frac{1}{qN} Z^T \cdot N I_N = \frac{1}{q} Z^T$$

$$6) \quad \text{O centro de gravidade da nuvem } N(C) = \{(x_{\cdot}^i, p_{\cdot}^i), i=1, 2, \dots, m\} \quad \text{no} \\ \mathbb{R}^N, \text{ tem por coordenadas } G^j = \frac{1}{N}, \quad j=1, 2, \dots, N.$$

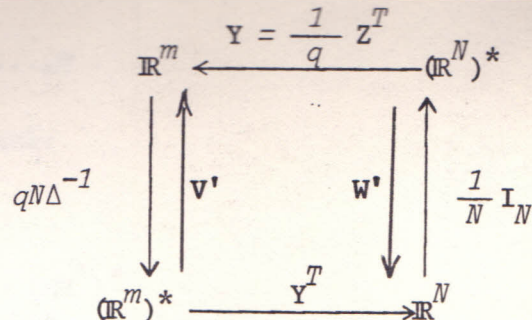
$$7) \quad \text{O centro de gravidade da nuvem } N(L) = \{(y, p^j), j=1, 2, \dots, m\} \quad \text{no} \\ \mathbb{R}^m, \text{ tem por coordenadas } H^i = \frac{1}{qN} \Delta^i, \quad i=1, 2, \dots, m.$$

Os esquemas de dualidade são:



Matriz a diagonalizar

$$\begin{aligned}
 S &= Z \cdot \Delta^{-1} \cdot \frac{1}{qN} \Delta \Delta^{-1} Z^T N \cdot T_N \\
 &= \frac{1}{q} Z \Delta^{-1} Z^T
 \end{aligned}$$



Matriz a diagonalizar

$$\begin{aligned}
 S' &= \frac{1}{q} Z^T \cdot \frac{1}{N} \cdot I_N \cdot \frac{1}{q} Z qN\Delta^{-1} \\
 &= \frac{1}{q} Z^T Z \Delta^{-1} = \frac{1}{q} B \Delta^{-1}
 \end{aligned}$$

4.3 - A Análise em $B = Z^T Z$

A matriz B é simétrica $m \times m$.

Com as notações gerais da análise fatorial de correspondências, temos:

- 1) $B = K$, onde k_{ij}^j é o número de indivíduos (entre os N) tendo escolhido as modalidades j e i ; temos

$$k_{ii} = q \Delta^i, \quad k^{jj} = q \Delta^j \quad \text{e} \quad k = q^2 N$$

$$2) \quad P = \frac{1}{q^2 N} K = \frac{1}{q^2 N} Z^T Z$$

$$\begin{aligned}
 3) \quad D_m &= \text{diag}(p_1, \dots, p_m) = \text{diag}(p^1, \dots, p^m) = \\
 &= \text{diag}(\Delta^1 / qN, \dots, \Delta^m / qN) = \frac{1}{qN} \Delta
 \end{aligned}$$

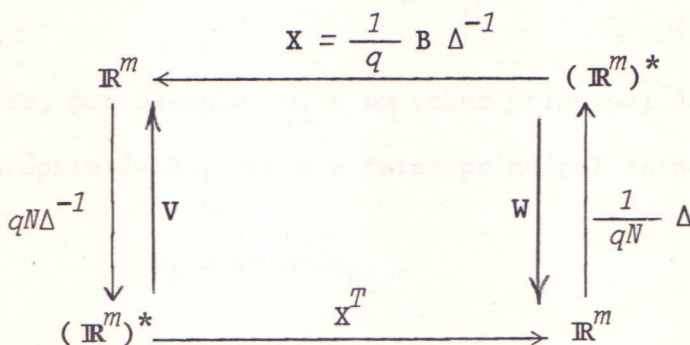
$$4) \quad X = Y = \frac{1}{q^2 N} Z^T Z qN \Delta^{-1} = \frac{1}{q} B \Delta^{-1}$$

- 5) O centro de gravidade da nuvem

$$N(C) = N(L) = \{(\mathbf{x}_i, \mathbf{p}_i) ; i = 1, 2, \dots, m\}$$

tem coordenadas $G^j = \frac{1}{qN} \Delta^j$, $j = 1, 2, \dots, m$.

O esquema de dualidade é o seguinte:



Matriz a diagonalizar

$$S = X \frac{1}{qN} \Delta X^T qN \Delta^{-1}$$

$$= X \Delta X^T \Delta^{-1} = \frac{1}{q} B \Delta^{-1} B \Delta^{-1}$$

4.4 - Comparação entre as duas análises

Consideremos os esquemas de dualidade para a análise da tabela Z, apresentados na seção 4.1.

Sabemos do Capítulo 3 que as componentes principais reduzidas da análise na tabela X são vetores próprios da matriz

$$A = W \frac{1}{qN} \Delta = \Delta^{-1} Z^T N I Z \Delta^{-1} \frac{1}{qN} \Delta$$

$$= \frac{1}{q} \Delta^{-1} Z^T Z = \frac{1}{q} \Delta^{-1} B,$$

e são $\frac{1}{qN}$ Δ -normados.

Por sua vez, os fatores principais da análise na tabela Y são os vetores próprios da matriz

$$\begin{aligned}
 qN \Delta^{-1} \mathbf{v}' &= qN \Delta^{-1} \frac{1}{q} \mathbf{Z} \frac{1}{N} \mathbf{I}_N \frac{1}{q} \mathbf{Z}^T = \\
 &= \frac{1}{q} \Delta^{-1} \mathbf{Z}^T \mathbf{Z} = \frac{1}{q} \mathbf{A} .
 \end{aligned}$$

Se, por exemplo, $\mathbf{u}_k$ é um vetor principal da análise na tabela $\mathbf{Y}$, (vetor próprio de $\mathbf{A}$), então o fator principal associado é dado por:

$$\mathbf{v}_k = qN \Delta^{-1} \mathbf{u}_k ;$$

além disso, a condição

$$\mathbf{u}_k^T qN \Delta^{-1} \mathbf{u}_k = 1 ,$$

nos permite afirmar que

$$\mathbf{v}_k^T \frac{1}{qN} \Delta \mathbf{v}_k = 1 .$$

Resumindo: os fatores principais $\mathbf{v}_k$ da análise dos perfis de linhas da tabela $\mathbf{Z}$ são vetores próprios da matriz

$$\mathbf{A} = \frac{1}{q} \Delta^{-1} \mathbf{B} ,$$

e são $\frac{1}{qN} \Delta$ -normados. Coincidem então com as componentes principais , $\frac{1}{qN} \Delta$ -normadas, da análise dos perfis de colunas da tabela $\mathbf{Z}$.

Considerando agora o esquema de dualidade obtido para a análise da tabela $\mathbf{B}$ na seção 4.3, obtemos os fatores principais como vetores próprios $\frac{1}{qN} \Delta$ -normados da matriz

$$\begin{aligned}
 qN \Delta^{-1} \mathbf{v} &= qN \Delta^{-1} \frac{1}{q} \mathbf{B} \Delta^{-1} \frac{1}{qN} \Delta \frac{1}{q} \Delta^{-1} \mathbf{B} = \\
 &= \frac{1}{q^2} \Delta^{-1} \mathbf{B} \Delta^{-1} \mathbf{B} = \mathbf{A}^2
 \end{aligned}$$

$$\text{Ora, } A v = \lambda v \implies A^2 v = \lambda^2 v .$$

Podemos então concluir que os fatores principais da análise na tabela **B** coincidem com os fatores principais da análise nos perfis das linhas da tabela **Z**. Como vimos, estes fatores são também componentes principais da análise nos perfis das linhas de **Z**. Os valores principais da análise na tabela **B**, entretanto, são os quadrados dos valores principais correspondentes obtidos na análise da tabela **Z**.

Mostramos assim que, do ponto de vista da descrição das relações entre as diferentes modalidades dos diversos caracteres, podemos efetuar **Z** ou na tabela **B**, indiferentemente.

4.5 - Propriedades da Análise de Correspondências Múltiplas

4.5.1 - Inércia de um Caracter e de uma Modalidade

Consideremos a tabela

$$Z = \begin{bmatrix} z_1 & | & z_2 & | & \dots & | & z_r & | & \dots & | & z_q \end{bmatrix} .$$

A cada modalidade i associamos o vetor coluna $x_i \in \mathbb{R}^N$, i -ésima coluna da matriz $X = Z \Delta^{-1}$, obtida em 4.2 .

O centro de gravidade da nuvem

$$N(C) = \left\{ (x_i, \frac{\Delta^i}{qN}); i=1,2,\dots,m \right\}$$

dos perfis de coluna de $K = Z$, tem por j -ésima coordenada

$$G^j = \sum_{i=1}^m x_i^j p_i = p^j = \frac{1}{N}.$$

Aqui, $E = \mathbb{R}^N$ está munido da métrica dada pela matriz

$$D_N^{-1} = N I_N,$$

de modo que

$$\begin{aligned} ||x_i - G||^2 &= (x_i - G)^T N I_N (x_i - G) \\ &= N \sum_{j=1}^N (x_i^j - \frac{1}{N})^2 \end{aligned}$$

$$\text{Mas } x_i^j = \frac{k_i^j}{k_i} = \frac{k_i^j}{\Delta^i}, \text{ onde } k_i^j \in \{0,1\} \text{ e } \sum_{j=1}^N k_i^j = k_i = \Delta^i;$$

então

$$\begin{aligned} ||x_i - G||^2 &= N \sum_{j=1}^N \left(\frac{k_i^j}{\Delta^i} - \frac{1}{N} \right)^2 \\ &= N \sum_{j=1}^N \left(\frac{k_i^j}{(\Delta^i)^2} - \frac{2}{N} \frac{k_i^j}{\Delta^i} + \frac{1}{N^2} \right) \\ &= N \left[\frac{1}{\Delta^i} - \frac{2}{N} + \frac{1}{N} \right] = N \left[\frac{1}{\Delta^i} - \frac{1}{N} \right] \end{aligned}$$

A inércia da modalidade i , ou seja, a contribuição da modalidade i para a inércia total da nuvem $N(C)$, se escreve

$$p_i ||x_i - G||^2 = \frac{\Delta^i}{qN} \left[\frac{1}{\Delta^i} - \frac{1}{N} \right]$$

$$= \frac{1}{q} \left[1 - \frac{\Delta^i}{N} \right]$$

A inércia de um caracter r , ou seja, a contribuição do caracter r para a inércia total da nuvem, se escreve

$$\sum_{i=1}^{m_r} p_i ||x_i - G||^2 = \sum_{i=1}^{m_r} \frac{1}{q} \left[1 - \frac{\Delta^i}{N} \right]$$

$$= \frac{1}{q} [m_r - 1] ,$$

lembrando que

$$\sum_{i=1}^{m_r} \Delta^i$$

é a soma dos elementos da matriz diagonal Δ_r , que vale N .

A inércia total da nuvem é então igual a:

$$\sum_{r=1}^q \frac{1}{q} [m_r - 1] = \frac{m}{q} - 1$$

Observação 1

A inércia de uma modalidade i será tanto maior quanto menor for Δ^i ; ou seja, se o efetivo dos indivíduos que escolhem a modalidade i é pequeno, a inércia da modalidade i é grande.

A inércia de um caracter C_r será tanto maior quanto maior for o número de suas modalidades, m_r .

Observação 2

A inércia da nuvem $N(C)$ é igual à soma dos auto-valores de S' diminuída de 1, pois na procura dos eixos principais suprimimos o vetor $\vec{OG}$ associado ao auto-valor 1. Os auto-valores de S' são os de $\frac{1}{q} B \Delta^{-1}$, e sua soma vale

$$\frac{1}{q} \text{Traço} (B \Delta^{-1}) .$$

Como os elementos da diagonal de B são os da diagonal de Δ , resulta que a inércia total da nuvem é igual a

$$\frac{1}{q} m - 1$$

No caso particular em que todos os caracteres tem duas modalidades, a inércia total da nuvem vale 1.

4.5.2 - Suporte da Nuvem das m Modalidades em $\mathbb{R}^N$

Na análise fatorial de correspondências efetuada a partir de Z , a cada modalidade está associada uma coluna de

$$X = Z \Delta^{-1} ;$$

estas colunas geram um subespaço do qual a dimensão é o posto de

$$Z \Delta^{-1} ,$$

ou seja, o posto de $Z = [z_1 \mid \dots \mid z_r \mid \dots \mid z_q]$.

Todos os subespaços V_r gerados pelas colunas de $Z_r, r=1, 2, \dots, q$, têm em comum o vetor 1_N , em virtude do caracter disjuntivo completo das modalidades.

A dimensão do subespaço V gerado pelas colunas de Z é inferior ou igual à soma das dimensões dos subespaços V_r gerados pelas colunas das matrizes Z_r . Deste modo,

$$\begin{aligned} \text{posto}(X) = \text{posto}(Z) = \dim V &\leq m_1 + (m_2 - 1) + \dots + (m_q - 1) = \\ &= m - q + 1 \end{aligned}$$

O posto da matriz a diagonalizar,

$$X D_m X^T (D_N)^{-1},$$

é majorado por $(m - q + 1)$, que é o número de auto-valores não nulos desta matriz. Sabemos que no estudo da nuvem em torno do centro de gravidade de G , a direção $\vec{OG}$ e o auto-valor correspondente 1 são eliminados; resulta que a nuvem das modalidades está incluída em um subespaço vetorial de dimensão máxima igual a $m - q$, que é $(D_N)^{-1}$ -ortonormal (ou ainda I_N -ortonormal) ao vetor $\vec{OG}$, que tem no presente caso a mesma direção de I_N .

4.5.3 - Nota Sobre a Representação Simultânea

Os esquemas de dualidade da análise efetuada sobre a tabela lógica Z sendo dados por

$$\begin{array}{ccc} \mathbb{R}^N & \xleftarrow{X = Z \Delta^{-1}} & (\mathbb{R}^m)^* \\ \downarrow N I_N & & \uparrow \frac{1}{2N} \\ (\mathbb{R}^N)^* & \xrightarrow{X^T} & \mathbb{R}^m \end{array} \quad , \quad \begin{array}{ccc} \mathbb{R}^m & \xleftarrow{Y = \frac{1}{q} Z} & (\mathbb{R}^N)^* \\ \downarrow 2N \Delta^{-1} & & \uparrow \frac{1}{N} I_N \\ (\mathbb{R}^m)^* & \xrightarrow{Y^T} & \mathbb{R}^N \end{array}$$

$$S = \frac{1}{q} Z \Delta^{-1} Z^T \quad , \quad S' = \frac{1}{q} B \Delta^{-1}$$

sejam λ um auto-valor de S (e de S'), u (resp. u') um auto-valor de S (resp. S') correspondente a λ , e φ (resp. ψ) a componente principal associada; as "relações baricêntricas" são dadas por:

$$\begin{cases} \varphi_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^N \frac{p_i^j}{p_i} \psi_j \\ \psi_j = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^m \frac{p_i^j}{p_j} \varphi_i \end{cases} .$$

No caso particular aqui abordado, temos

$$\frac{p_i^j}{p_j} = \frac{k_i^j}{q} \quad \text{e} \quad \frac{p_i^j}{p_i} = \frac{k_i^j}{\Delta^i} ;$$

as relações baricêntricas se escrevem então na forma

$$\begin{cases} \varphi_i = \frac{1}{\sqrt{\lambda}} \frac{1}{\Delta^i} \sum_{j=1}^N k_i^j \psi_j \\ \psi_j = \frac{1}{\sqrt{\lambda}} \frac{1}{q} \sum_{i=1}^m k_i^j \varphi_i \end{cases} .$$

Como $k_i^j \in \{0,1\}$ e Δ^i representa o número de indivíduos associados à modalidade i , concluímos que, a menos do coeficiente $1/\sqrt{\lambda}$, a abscissa da modalidade i sobre o eixo $[u]$ é a média aritmética das abscissas sobre o eixo $[u']$ dos indivíduos associados à modalidade i .

Por outro lado a forma disjuntiva completa das modalidades mostra que, para todo indivíduo j fixado, existe uma e somente uma modalidade i por caracter para a qual $k_i^j = 1$; assim, em

$$\sum_{i=1}^m k_i^j \varphi_i$$

existem exatamente q termos não nulos; deste modo, Ψ_j , abcissa do indivíduo j sobre $[u']$, é, a menos do coeficiente $1/\sqrt{\lambda}$, a média aritmética das abcissas das modalidades associadas ao indivíduo i no eixo $[u]$.

Observação: as componentes principais $\varphi \in \mathbb{R}^m$ e $\psi \in \mathbb{R}^N$, escrevem-se matricialmente como

$$\varphi = \frac{\Delta^{-1} Z^T \psi}{\sqrt{\lambda}} \quad \text{e} \quad \psi = \frac{1}{q \sqrt{\lambda}} Z \varphi.$$

APÊNDICE

APLICAÇÃO DA A.C.P. (Análise em Componentes Principais) e DA A.F.C. (Análise Fatorial de Correspondências) NO ESTUDO TAXONÔMICO DE ABELHAS SILVESTRES BRASILEIRAS.

Neste apêndice, consideramos a aplicação dos métodos de A.C.P. e de A.F.C., à análise de dados morfométricos de abelhas silvestres brasileiras (gêneros *Melipona*, *Trigona* e *Meliponula*) com base no artigo de PISANO, J.F., KERR, W.E., e CUNHA, R.A., intitulado: "Estrutura Multidimensional dos Meliponídeos" e publicado na Acta Amazônica, nº 2, (revista do Conselho Nacional de Desenvolvimento Científico e Tecnológico).

No referido artigo, os autores consideram 12 variáveis para um conjunto de 21 espécies de meliponídeos (sendo 12 pertencentes ao gênero *Melipona*, 8 pertencentes ao gênero *Trigona* e uma espécie do gênero *Meliponula*).

Na tabela 1, dispomos dos valores numéricos observados, com indicação das variáveis consideradas na pesquisa e das espécies envolvidas. Os autores empregaram a distância generalizada de Mahalanobis como índice de dissimilaridade morfológica, bem como a partir da matriz das distâncias generalizadas entre todos os pares de espécies, procuraram representar no plano a "nuvem dos indivíduos", de tal forma que se dispusesse de configurações que tanto quanto possível aproximassem as relações de distâncias existentes.

Embora tal maneira de proceder tenha sido utilizada extensivamente na literatura científica, não há dúvida que as construções de tais configurações ainda encerram certo subjetivismo. Assim, resolvemos aplicar as técnicas de A.C.P. e de A.F.C. à reanálise daqueles dados e à

obtenção de informações sobre os agrupamentos ("clusters") porventura existentes.

Note-se que os autores, igualmente, submeteram seus dados a uma análise de classificação hierárquica (método de SOKAL e SNEATH). Embora tivéssemos novamente submetido os dados à classificação hierárquica, utilizando procedimentos computacionais implantados no sistema de computação DEC-10 (Digital) da Universidade Federal do Ceará, os resultados nessa direção não serão considerados aqui.

A.1 - Os resultados obtidos pelos autores

A configuração de pontos para as 12 espécies de *Melipona*, e o dendograma representativo das distâncias generalizadas entre as espécies são exibidas na figura 1; na figura 2 encontra-se a configuração de pontos para os 9 trigonínios (a espécie *Melipona bocandei* sendo examinada conjuntamente no gênero *Trigona*), com o respectivo dendograma; bem como na figura 3 temos a configuração obtida pelos autores para as 21 espécies envolvidas na pesquisa, e o dendograma associado.

De acordo com PISANO, J.F., KERR, W.E., e CUNHA, R.A., as figuras permitem concluir o seguinte:

"A Figura 1 que resume as relações de similaridade entre as espécies de *Melipona* aqui analisadas (distância generalizada, de Mahalanobis), indica um aglomerado das espécies *rufiventris*, *seminigra* e *scutellaris* que, na classificação de Schwarz (1932), eram postas (erroneamente — pois são sexualmente isoladas) em uma única espécie (*M. fasciata*). Próxima deste grupo localiza-se *M. quinquefasciata*. Com trabalho anterior de Kerr et al. (1957) coloca-se *M. quinquefasciata* próxima de *M. favosa*. No entanto este estudo coloca-a próxima do antigo grupo *fasciata*. O estudo do grupo das *Trigonini* (veja figura 2) apresentou as espécies *Partamona*, *Trigona* e *Scaptotrigona* agrupadas, as abelhas do grupo africano *Hypotrigona* próximas a *Plebeia*; *Frieseomelitta* e *Meliponula* es

tão bem isoladas, tudo concordando bem com a citologia e outros aspectos biológicos.

A figura 3 faz também revelações importantes: revela que as *trigonini* mais próximas das *Meliponas* são as *Meliponula* e o grupo *Parta mona*, e as *Meliponas* mais próximas dos *Trigonini* seriam as *M. marginata*, *M. schwarzi*, *M. favosa* e *M. mandaçaia*".

A.2 - Nossos resultados através da A.C.P.

No que segue, utilizamos aimplesmente as siglas que aparecem na tabela 1 para representar cada espécie, tanto na A.C.P. como na A.F.C.

A.2.1 - 12 espécies de *Meliponula*.

Os dois primeiros eixos principais acumulando 95% da inércia total, nos limitamos à apresentação, na figura 4, da projeção da nuvem no plano gerado por estes dois eixos. O aglomerado das espécies MRU, MSC e MSE detectado pelos autores na figura 1 permanece, bem como MQI fica próxima de MQA.

A.2.2 - O grupo das *Trigonini*, (9 espécies).

Apresentamos na figura 5 a projeção da nuvem no plano gerado pelos dois primeiros eixos principais, que acumulam 97% da inércia total. O agrupamento das espécies TCU, TSP, TXA e TPO aparece bem nítido, principalmente quando da projeção da nuvem sobre o primeiro eixo, o horizontal, com 92% da inércia total; por outro lado, o segundo eixo principal separa TCU e TSP de TXA e TPO. As espécies MEB e TFR permanecem isoladas, bem como TDR e TBR aparecem próximas.

A.2.3 - Todas as espécies consideradas.

Na figura 6 apresentamos a projeção da nuvem no plano gerado pelos dois primeiros eixos principais, que acumulam 97% da inércia total. Observamos que MEB é o trigonínio mais próximo das *Meliponas*, aparecendo em segundo lugar TSP, em lugar da espécie TCU detectada pelos

autores; por sua vez, as *Meliponas* mais próximas das *Trigonini* são MMM, MCW, MFF e MMA como constatado pelos autores.

A.3 - Nossos resultados através da A.F.C.

Evidentemente, os dados disponíveis não constituem uma tabela de contingência, o domínio ótimo de aplicação da análise fatorial de correspondências. Há aqui o interesse didático de mostrar que este tipo de análise vai além das ideais tabelas de contingência com resultados satisfatórios. Entretanto, a motivação principal para aplicar a análise fatorial de correspondências à tabela de dados em questão surgiu do fato de que, à página 244 da revista, os autores fazem a seguinte observação: "O método empregado tem alguns defeitos ou dificuldades: a) tamanho é muito mais importante que forma, b) foram usadas poucas espécies, e c) a introdução de novas espécies afeta a distância entre as espécies".

Estas dificuldades, principalmente a primeira, sendo teoricamente amenizadas pela análise de correspondências, onde se trabalha com um "perfil", justifica-se a aplicação desta técnica à tabela em causa.

A.3.1 - 12 espécies de *Meliponula*.

A projeção da nuvem no plano gerado pelos dois primeiros eixos principais, que acumulam 90% da inércia total, é apresentada na figura 7. Como nas análises anteriores, o aglomerado das espécies MRU, MSC e MSE permanece; entretanto, MQI é agora colocada próxima a MFF, e não de MQA.

A.3.2 - O grupo das *Trigonini*, (9 espécies).

Os dois primeiros eixos acumulam 90% da inércia total, e a projeção das nuvens no plano por eles gerado é apresentada na figura 8. A espécie TFR continua bem isolada, o mesmo não ocorrendo com MEB, que fica agora próxima a TCU; por sua vez o antigo aglomerado TCU, TSP, TCA

e TXO aparece separado pelo primeiro eixo; as espécies TBR e TDR continuam próximas.

A.3.3 - Todas as espécies consideradas.

Para este caso os dois primeiros eixos acumulam 83% da inércia total, e a projeção da nuvem no primeiro plano principal é apresentada na figura 9. As espécies TCU e MEB, como anteriormente, revelam-se as trigonini reais próximas das Melíponas; por sua vez, as Melíponas mais próximas das Trigonini continuam sendo MMM, MCW, MMA e MQA, antes de MFF. A espécie MEB, próxima ao centro de gravidade, revela-se como espécie "média".

	MM	MCW	MMA	MFF	MPU	MOA	MOI	MRU	MSE	MSC	MIN	MFL
VARIÁVEIS	Melipona marginata marginata	Melipona schwarzi	Melipona monodonta	Melipona faveosa faveosa	Melipona puncticollis	Melipona quadricolor anthidioides	Melipona quinquefasciata	Melipona rufiventris	Melipona seminigra merrillae	Melipona scutellaris	Melipona interrupta grandis	Melipona flavipennis
1. Largura máxima	2,703	2,890	3,542	3,524	3,462	4,081	3,832	4,027	4,097	4,279	4,662	5,342
2. Comprimento da cabeça	2,253	2,479	2,553	2,551	2,752	2,800	2,841	3,090	3,253	3,280	3,496	3,616
3. Largura da cabeça	2,890	3,234	3,475	3,578	3,676	3,776	3,884	4,071	4,253	4,262	4,639	4,933
4. Comprimento do olho	1,942	2,176	2,071	2,172	2,464	2,254	2,407	2,526	2,633	2,631	2,928	2,900
5. Distância interorbital inferior	1,518	1,661	1,947	1,939	1,830	2,258	2,113	2,375	2,486	2,474	2,759	3,144
6. Comprimento do olho	0,864	0,939	1,061	1,056	1,036	1,273	1,134	1,267	1,382	1,331	1,363	1,522
7. Largura do clipeo	1,321	1,518	1,511	1,558	1,638	1,737	1,706	1,970	2,044	2,050	2,186	2,261
8. Largura do tórax	2,251	2,360	2,643	2,736	2,582	2,945	2,790	3,065	3,043	2,900	3,401	3,525
9. Comprimento da asa posterior	4,250	4,179	5,304	5,084	5,000	5,794	5,792	6,129	6,237	6,342	7,504	7,154
10. Largura da asa posterior	1,077	1,048	1,331	1,355	1,230	1,504	1,496	1,554	1,628	1,592	1,653	1,789
11. Comprimento do fêmur da pata posterior	1,767	2,020	2,142	2,217	2,419	2,379	2,301	2,613	2,793	2,795	2,921	3,167
12. Área da tíbia da pata posterior	1,775	2,297	2,546	2,756	2,911	2,823	3,117	4,108	4,422	4,535	5,426	6,173

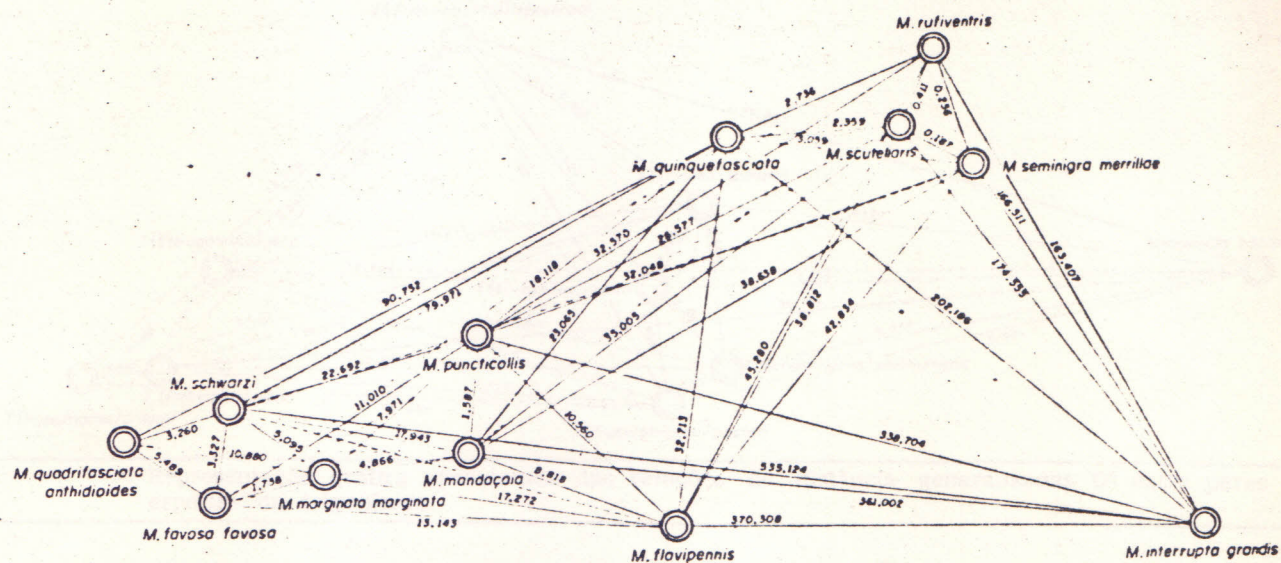
— Valores médios amostrais para as espécies de *Melipona* nos caracteres incluídos no estudo.

TABELA I

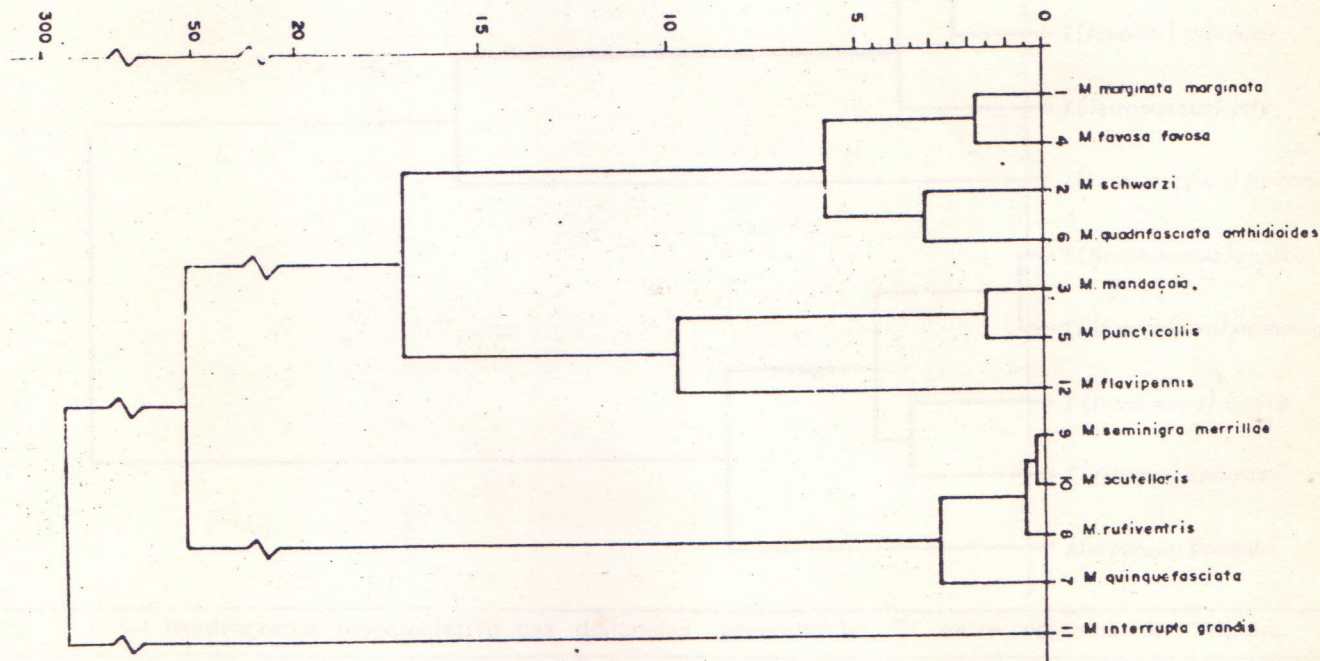
	TBR	TJA	TDR	TFR	TPO	TCU	TXA	TSP	MEB
VARIÁVEIS	Trigona (Hypotrigona) braunsi	Trigona (Tetragonisca) jaty	Trigona (Plebeia) droryana	Trigona (Frieseomelitta) freiremoiazi	Trigona (Scaptotrigona) postica	Trigona (Perfomana) cupira	Trigona (Scaptotrigona) santhofskyae	Trigona (Trigona) spinipes	Meliponula bocandei
1. Largura máxima	1,064	0,947	1,566	1,380	2,465	2,461	2,545	2,534	3,396
2. Comprimento da cabeça	0,890	1,223	1,224	1,514	1,878	1,923	1,973	1,823	2,339
3. Largura da cabeça	1,147	1,712	1,644	2,163	2,615	2,486	2,704	2,748	3,341
4. Comprimento do olho	0,800	1,133	1,110	1,377	1,492	1,591	1,529	1,616	1,947
5. Distância interorbital inferior	0,636	0,778	0,808	1,072	1,542	1,398	1,638	1,552	1,882
6. Comprimento do clipeo	0,233	0,371	0,379	0,526	0,676	0,690	0,710	0,585	0,793
7. Largura do clipeo	0,572	0,663	0,731	0,929	1,169	1,174	1,192	1,441	1,581
8. Largura do tórax	0,869	1,002	1,121	1,187	1,804	1,764	1,868	1,677	2,473
9. Comprimento da asa posterior	1,898	2,597	2,544	4,149	4,235	4,402	4,409	4,925	5,169
10. Largura da asa posterior	0,546	0,558	0,705	0,853	1,099	1,238	1,156	1,211	1,378
11. Comprimento do fêmur da pata posterior	0,703	1,125	0,955	2,002	1,433	1,643	1,509	1,924	2,077
12. Área da tíbia da pata posterior	0,319	0,809	0,571	2,322	1,418	2,305	1,596	2,242	2,851

— Valores médios amostrais para as espécies de Trigona e Meliponula nos caracteres incluídos no estudo.

CONTINUAÇÃO DA TABELA I

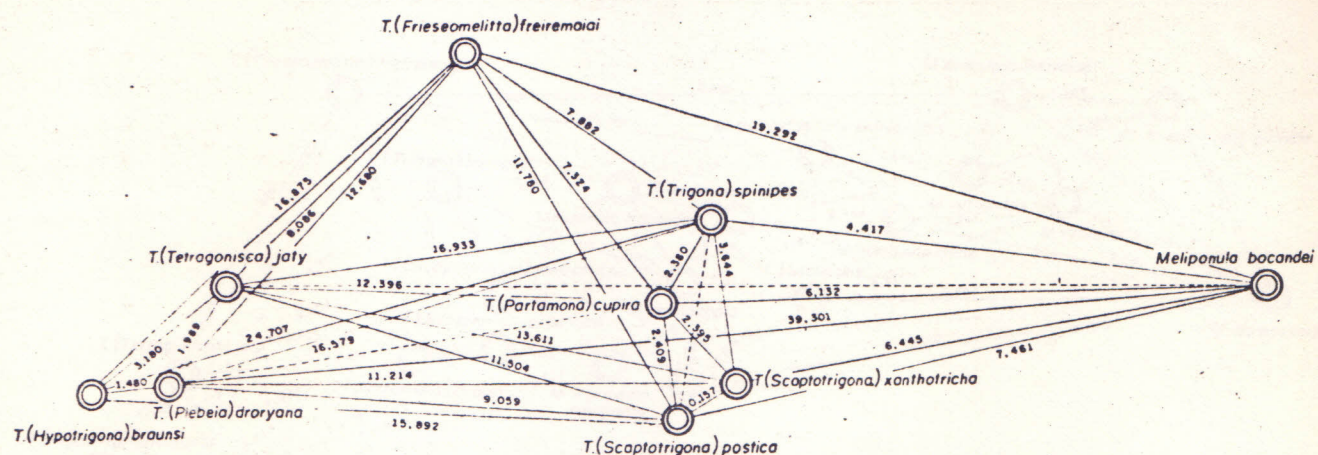


— Representação gráfica aproximada das relações de distância generalizadas D^2 entre pares de espécies de *Melipona*.

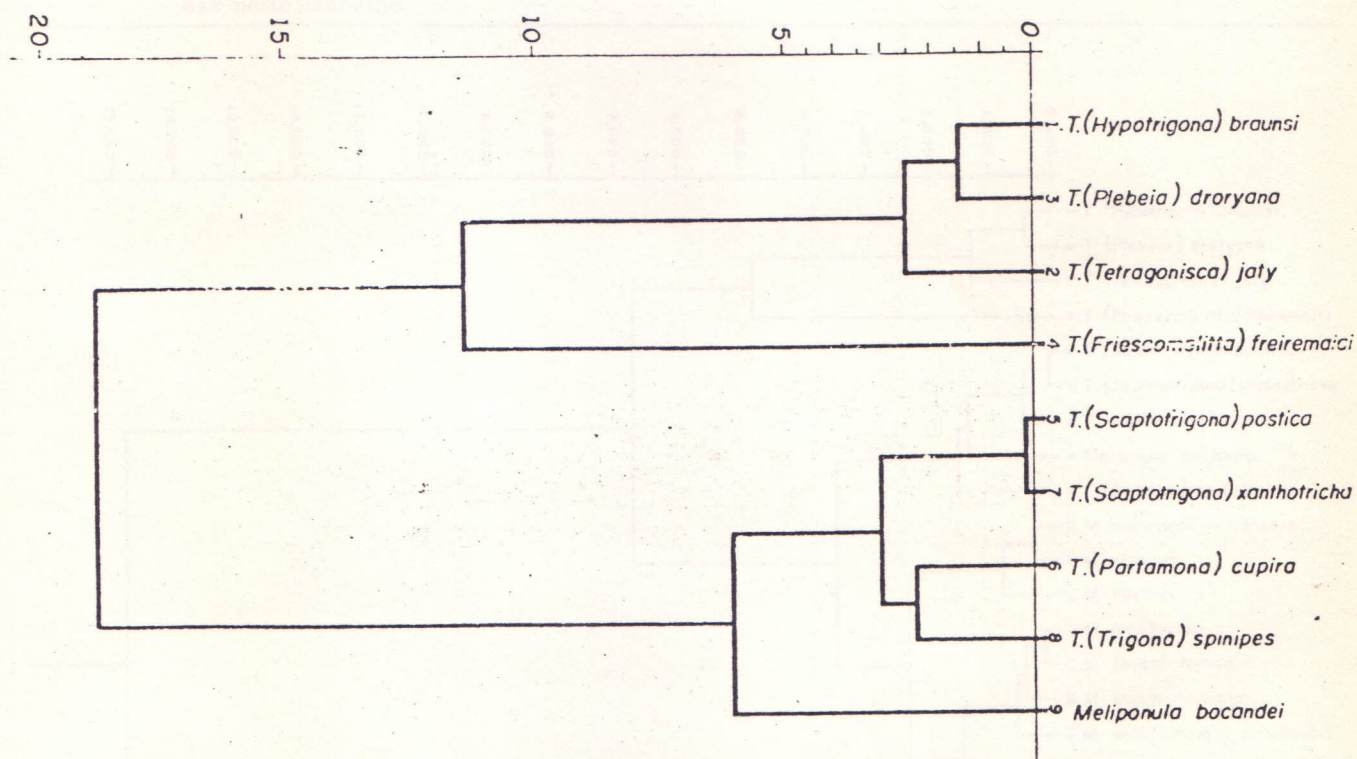


— Dendrograma representativo das distâncias generalizadas, D^2 , entre espécies de *Melipona*.

figura 1

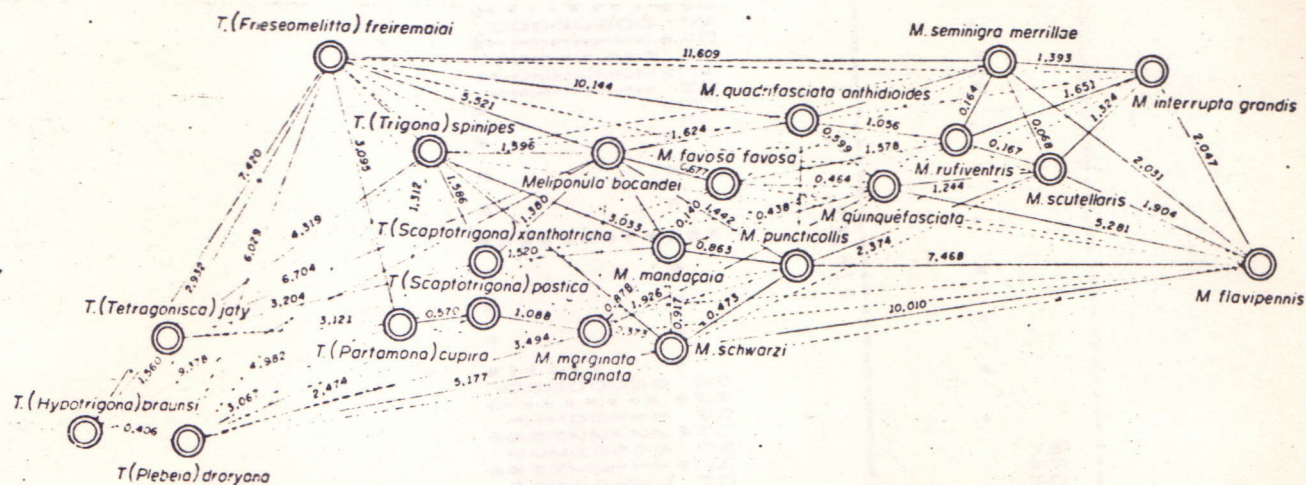


— Representação gráfica aproximada das relações de distância generalizadas D^2 entre pares de espécies de trigonínios.

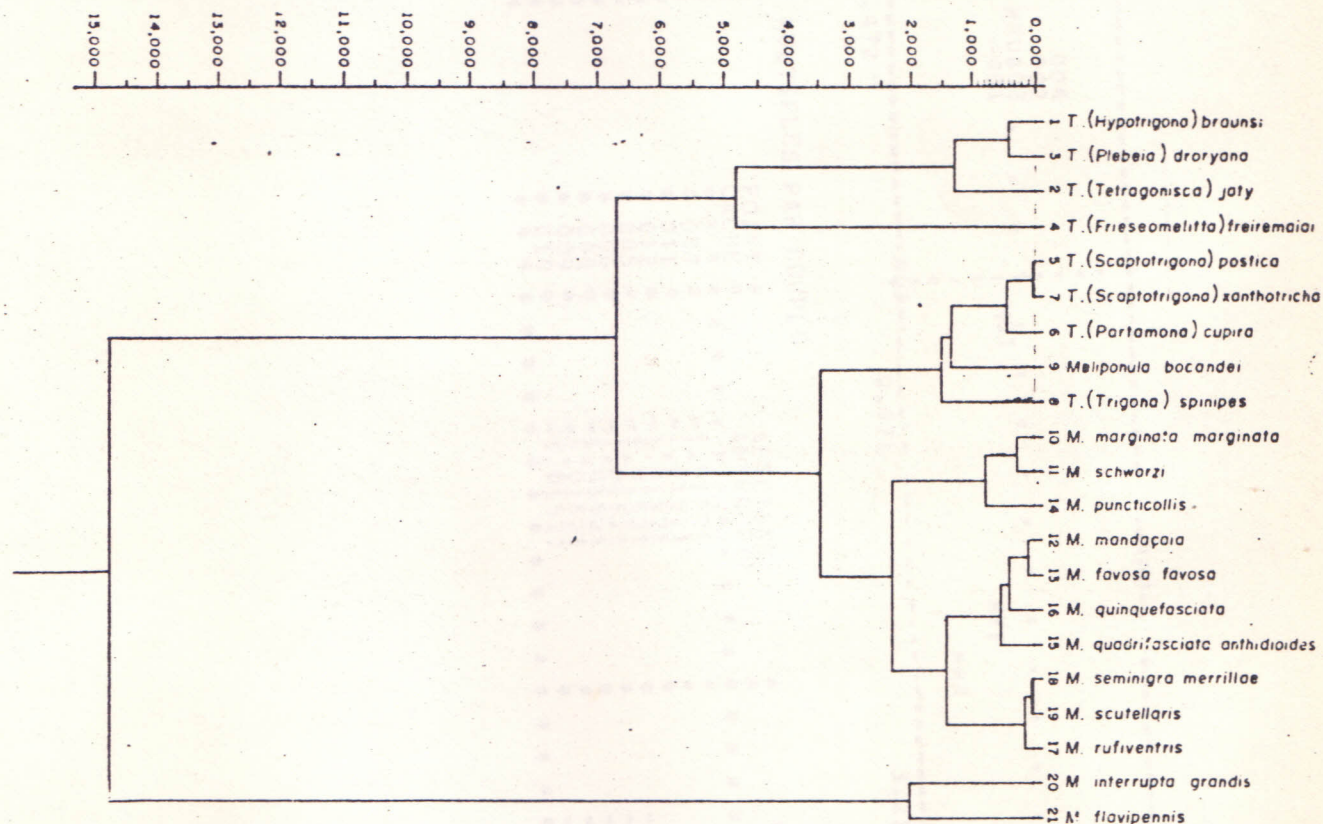


— Dendrograma representativo das distâncias generalizadas, D^2 , entre espécies de Trigona.

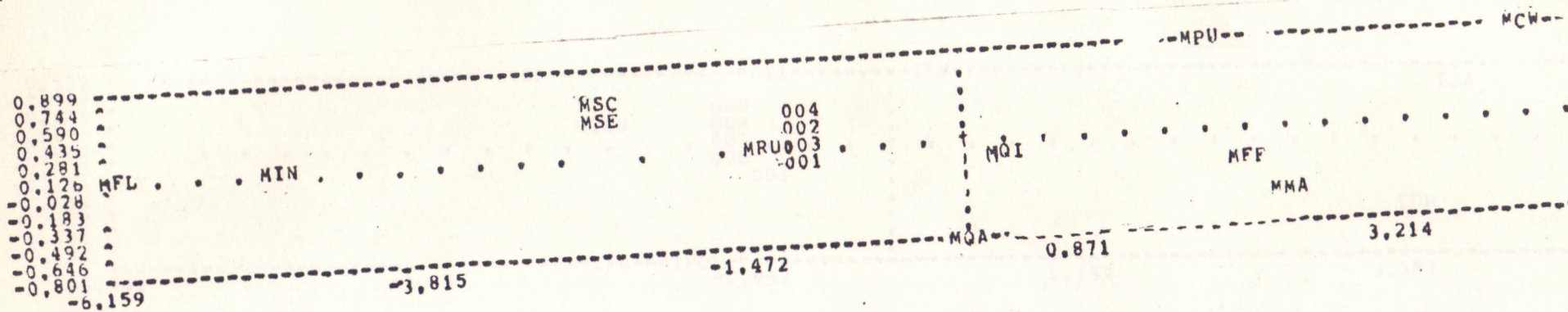
figura 2



— Representação gráfica aproximada das relações de similaridade entre todas as abelhas estudadas neste trabalho.



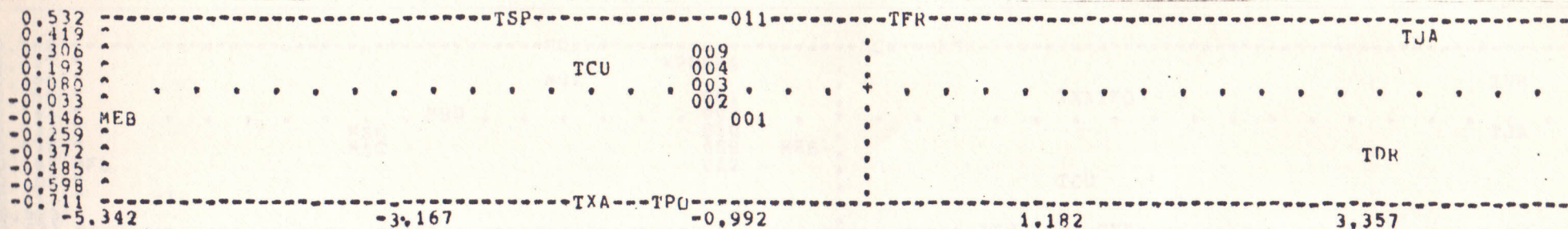
— Dendrograma representativo das distâncias generalizadas, D^2 , entre todas as espécies de abelhas incluídas no estudo.



POINTS MULTIPLES PAR COUPLE

POINT VU	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE
* 002 *	-1.01521	0.12617
* 002 *	-1.01521	0.12617
* 002 *	-1.01521	0.12617
* 002 *	-1.01521	-0.02834
* 003 *	-1.01521	-0.02834
* 003 *	-1.01521	-0.18286
* 001 *	-1.01521	-0.18286
* 001 *	-1.01521	-0.18286
* 001 *	-1.01521	-0.18286

POINT CACHE	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE
* 007 *	-1.01521	0.12617
* 011 *	-1.01521	0.12617
* 012 *	-1.01521	0.12617
* 005 *	-1.01521	-0.02834
* 006 *	-1.01521	-0.02834
* 008 *	-1.01521	-0.18286
* 009 *	-1.01521	-0.18286
* 010 *	-1.01521	-0.18286



POINTS MULTIPLES PAR COUPLE

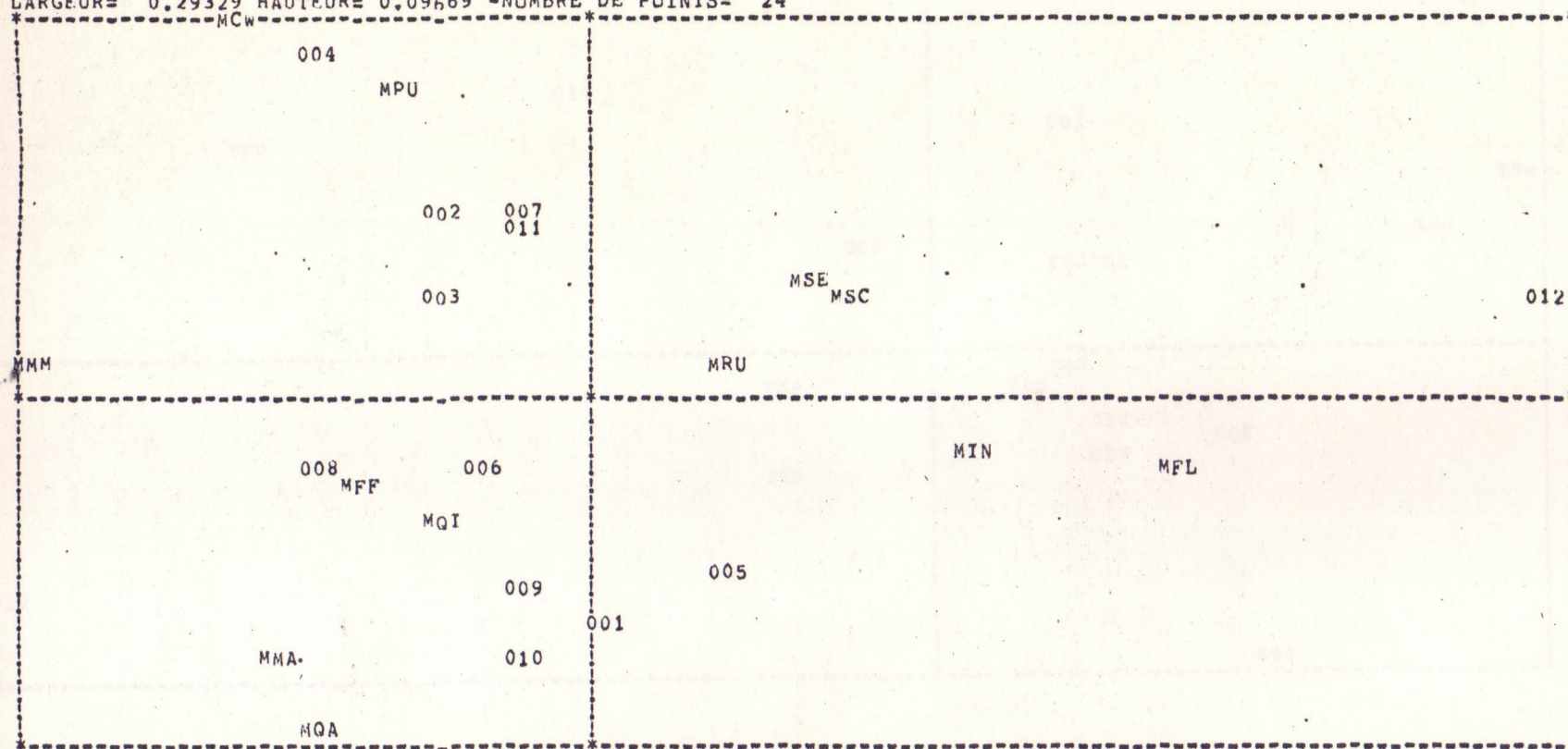
POINT VU	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE
* 011 *	-0.83324	0.41887
* 003 *	-1.09844	-0.03296
* 002 *	-1.09844	-0.14592
* 002 *	-1.09844	-0.14592
* 002 *	-1.09844	-0.14592
* 001 *	-0.83324	-0.25888

POINT CACHE	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE
* 012 *	-0.83324	0.41887
* 007 *	-1.09844	-0.03296
* 005 *	-1.09844	-0.14592
* 006 *	-1.09844	-0.14592
* 010 *	-1.09844	-0.14592
* 008 *	-0.83324	-0.25888

POINT VU	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE	POINT CACHE	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE
MMA	-0.84129	0.36588	MFF	-0.84129	0.36588
001	-0.84129	0.14381	006	-0.84129	0.14381
001	-0.84129	0.14381	008	-0.84129	0.14381
002	-0.84129	0.03277	003	-0.84129	0.03277
002	-0.84129	0.03277	004	-0.84129	0.03277
002	-0.84129	0.03277	005	-0.84129	0.03277
002	-0.84129	0.03277	007	-0.84129	0.03277
009	-0.84129	-0.18931	011	-0.84129	-0.18931

AXE HORIZONTAL(1)--AXE VERTICAL(2)--TITRE:[ANAL. FAT. CORR. / 12 ESPECIES MELIPONA]

LARGEUR= 0.29329 HAUTEUR= 0.09669 -NOMBRE DE POINTS= 24



AXE HORIZONTAL(1)--AXE VERTICAL(2)--TITRE:[ANAL. FAT. CORR. / 8 ESPECIES "TRIGONA" + 1 "MELIPONULA"]
 LARGEUR= 0.45371 HAUTEUR= 0.26146 -NOMBRE DE POINTS= 21

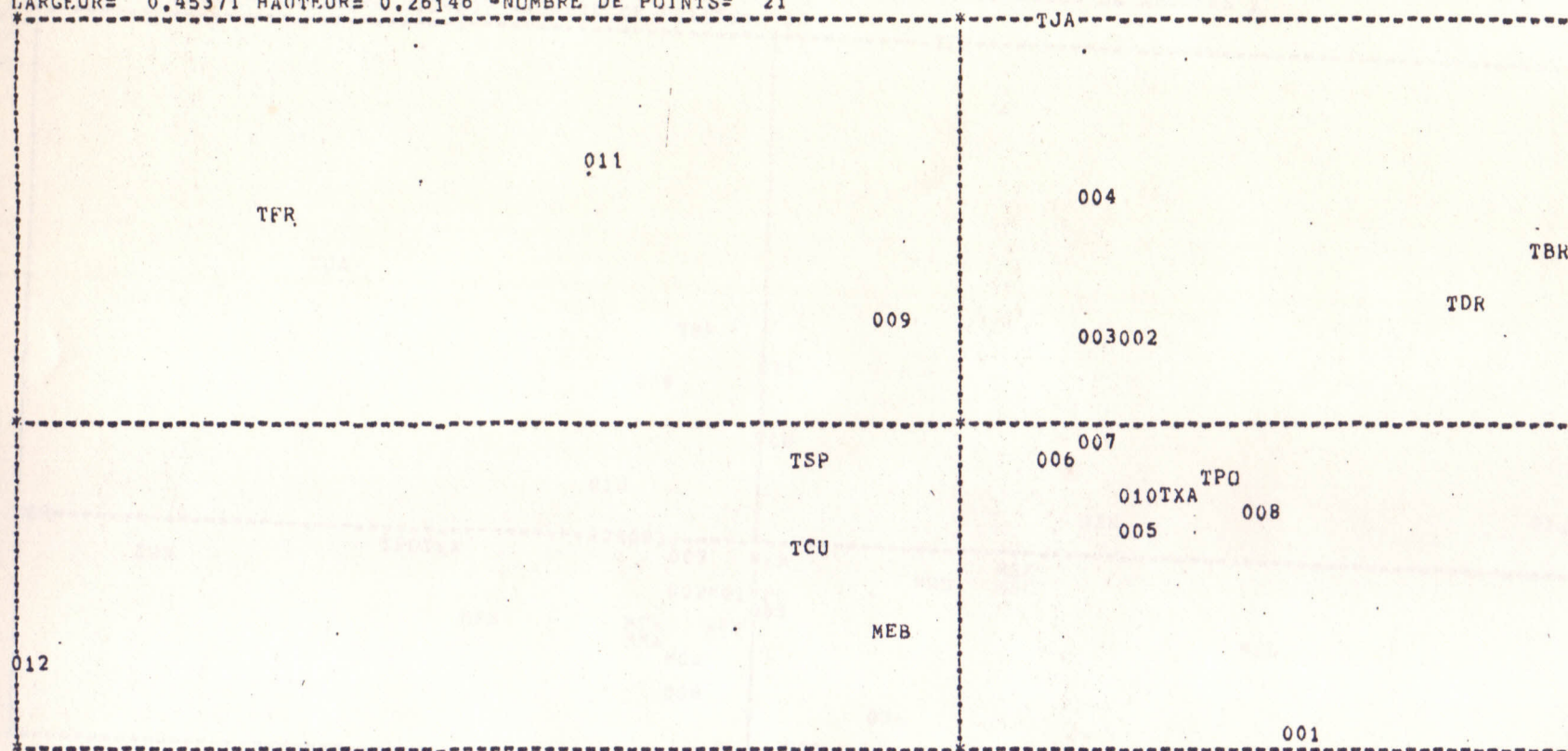


Figura 8

AXE HORIZONTAL(1)--AXE VERTICAL(2)--TITRE:[ANAL. FAT. CORR. / 21 ESPECIES DE ABELHAS]
 LARGEUR= 0.46434 HAUTEUR= 0.29497 -NOMBRE DE POINTS= 33

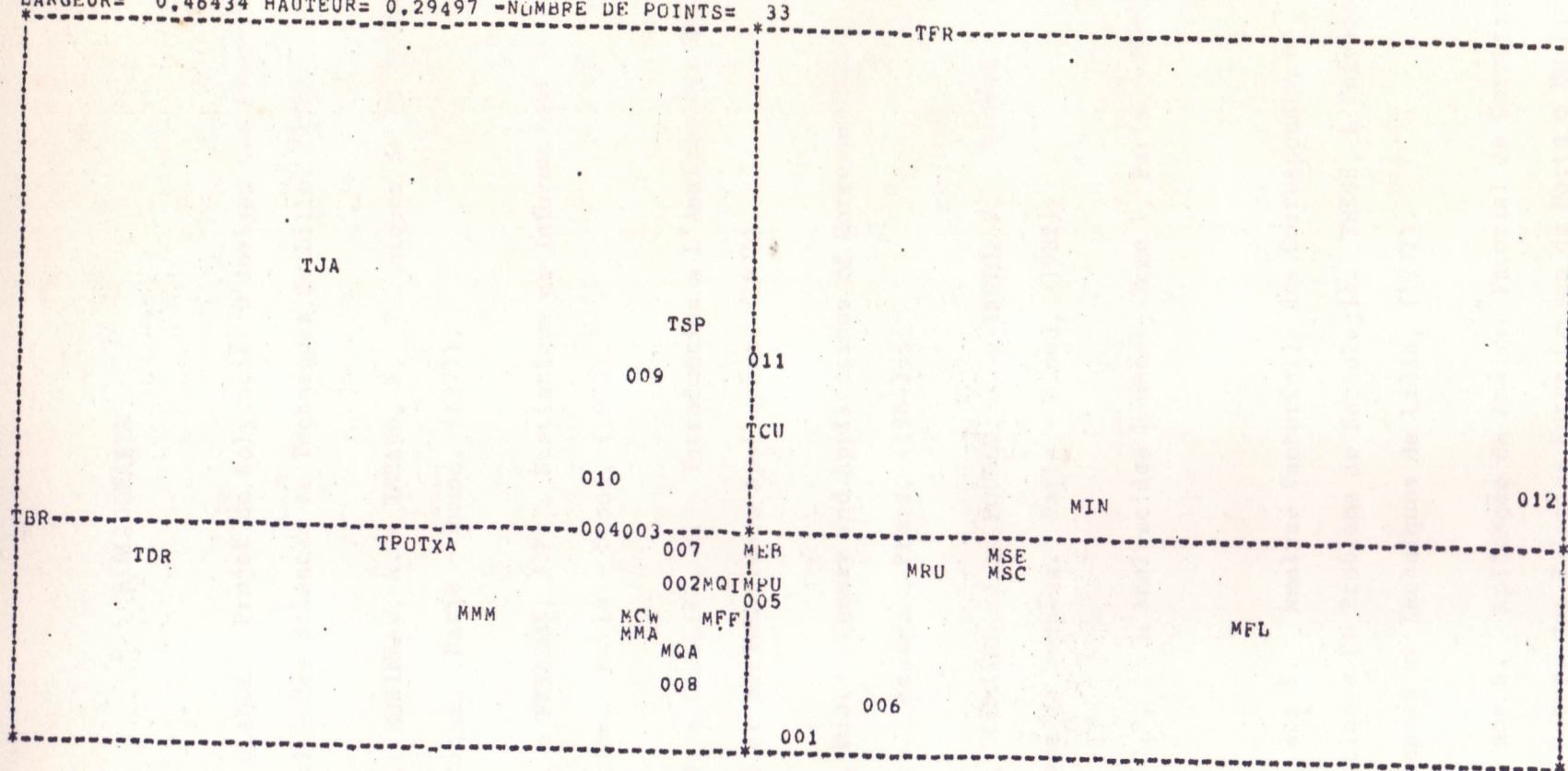


Figura 9

BIBLIOGRAFIA

- [1] LANGRAND, CLAUDE. Projet de Polycopié D'Analyse des Données.
Université des Sciences et Techniques de Lille, (1983)
- [2] LEBART, L.; MORINEAU, A. e TABARD, N. Techniques de la Description
Statistique. Paris - Dunod, (1977).
- [3] LEBART, L. e FENELON, J.P. Statistique et Informatique Appliquées
3^a édition. Paris - Dunod, (1975).
- [4] PAGES, J.P. e CAILLIEZ, F. Introduction a L'Analyse des Données.
Société de Mathématiques Appliquées, (1976).
- [5] GREENACRE, M.J. Theory and Applications of Correspondence
Analysis. Academic Press, (1984).
- [6] DIDAY, E.; LEMAIRE, J.; POUGET, J. e TESTU, F. Elements
d'Analyse de Données. Paris - Dunod, (1982).
- [7] BENZÉCRI, J.P. L'Analyse des Données, Tome 1, Paris - Dunod,
(1973).
- [8] PESSOA, F. dos S. Analyse Factorielle des Correspondences. Une
Application a Un Probleme de Sociologie. THESE, L'Université
des Sciences et Techniques de Lille, (1972).
- [9] PESSOA, F. dos S. Aplicação da Análise Fatorial de Correspondên-
cias no Estudo das Relações Entre Causas de Morte e Municípios
Brasileiros, U.F.C., (1977).

- [10] GREENACRE, M.J. "Practical Correspondence Analysis", In:
Interpreting Multivariate Data (Barnett, V, ed.), pp 119-146.
Wiley, Chichester, UK. (1981).
- [11] McCabe, G.P. "Principal Variables" Technometrics, vol 26, nº 2 ,
pp 137-144, 1984.
- [12] PISANI, J.F.; KERR, W.E.; CUNHA, R.A. "Estrutura Multidimensio-
nal dos meliponídeos". Acta Amazônica nº 2, pp 223-245. INPA,
1977.
- [13] RAO, C.R. "The Use and Interpretation of Principal Component
Analysis Applied Research", Sankyā (The Indian Journal of
Statistics); vol 26A, pp 329-356, 1964.