Análise de Dados Multivariados Através de Técnicas Baseadas na Decomposição em Valores Singulares

Rita Helena Antonelli Cardoso

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pela Sra: Rita Helena Antonelli Cardoso, e aprovada pela Comissão julgadora

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação da Universidade Estadual de Campinas, para obtenção do grau de Mestre em Estatística.

Orientador: Prof. Dr. José Ferreira de Carvalho

Campinas
1991

Aos meus pais e ao meu irmão

Santos e Ivanilde

José Luiz

Ao meu esposo e minha filha

Antonio Galvão

Juliana

Agradecimentos

Ao Prof. Dr. José Ferreira de Carvalho, a quem devo grande parte de minha formação estatística. Nunca poupou esforços e incentivos para que eu obtivesse sucesso na minha vida profissional.

Ao Eng. Cândido Pinto de Melo, por possibilitar condições para que eu atingisse este nível universitário.

Aos Estatísticos Júlia Tizue Fukushima e Silvana Dupas D. Gallucci, por terem compartilhado e contribuído em todas as fases da realização deste trabalho.

Aos Doutores Álvaro Jabur e Flair José Carrilho, pelos dados fornecidos.

Ao Eng. Eduardo Appel, pela preciosa ajuda na diagramação do texto.

À Sra. Mitsuko Oshiro Mori, pelo zelo na feitura de parte dos gráficos.

ÍNDICE

1. Introdução

1. INTRODUÇÃO

Os métodos de análise de dados multivariados têm como objetivo fornecer representações suscintas das grandes tabelas de dados, ou seja, de tabelas que contêm dados provenientes, possivelmente, de várias variáveis, aferidas em muitas unidades amostrais. Essas tabelas, em geral, apresentam dimensões muito grandes, tornando difícil interpretar os resultados e informações que elas contêm.

A Decomposição em Valores Singulares (normalmente referida por SVD), é uma das ferramentas mais úteis na álgebra matricial. Suas origens foram devidas aos trabalhos, principalmente, de matemáticos franceses e italianos nos anos de 1870 (veja Marshall e Olkin,1979). Uma das aplicações mais importantes, a aproximação de uma matriz por matriz de posto menor, deu-se por Eckart e Young(1936).

A estrutura da decomposição em valores singulares acomoda um conjunto amplo de técnicas para análise de dados multivariados, e assim pode unificar análises que parecem superficialmente muito diferentes. Dentre outras técnicas, fazem parte desse conjunto: Modelos Multiplicativos, Biplot e Análise de Correspondência, que serão abordadas neste trabalho. Estas técnicas são todas variações de um mesmo tema: a álgebra e a geometria da decomposição em valores singulares.

Modelos Multiplicativos é um método, desenvolvido por Mandel(1969), para análise de dados provenientes de

experimentos não replicados que envolvem dois fatores, preferivelmente, qualitativos. Tem-se uma matriz retangular de dados, cujas linhas representam um fator e cujas colunas representam o outro fator. O procedimento é baseado na partição do efeito da interação entre os fatores, em uma soma de termos, cada um dos quais é o produto de duas funções, uma dependendo só do fator linha e outra dependendo só do fator coluna. Os parâmetros em cada termo são estimados através da decomposição em valores singulares. O método permite não só modelar a interação, mas também estimar σ^2.

O Biplot, apresentado por Gabriel(1971), consiste de uma representação gráfica de uma matriz de dados multivariados, cujas linhas referem-se às unidades amostrais e as colunas referem-se às variáveis, ou vice-versa. Cada elemento da matriz é expresso como o produto escalar de vetores correspondendo às linhas e às colunas da matriz. No caso da matriz ser de posto superior a 2, a representação gráfica é obtida da matriz de posto 2 que melhor se aproxima da matriz original. Esta aproximação se dá através da decomposição em valores singulares.

A análise de Correspondência foi primeiramente publicada por Hirschfeld(1935), e não tendo sido muito difundida, foi redescoberta, independentemente, por vários outros autores. O tratamento dado por Hirschfeld foi puramente teórico. A forma geométrica do método originou-se na França com um grupo liderado por Jean-Paul Benzécri. É um método que se destina a encontrar a representação gráfica da associação entre linhas e colunas de uma tabela de contingência, num

espaço de dimensão menor que o original. As linhas e colunas são representadas como pontos. A determinação do espaço, bem como das coordenadas dos pontos se dá pela decomposição em valores singulares generalizada da matriz de dados, devidamente centrada e padronizada.

Neste trabalho, inicialmente, no capítulo 2, apresentamos a decomposição em valores singulares, na métrica Euclideana, e também uma generalização, onde outra métrica é utilizada. Neste caso, denominamos de decomposição em valores singulares generalizada. Em ambos os casos, a aproximação de uma matriz por matriz de posto menor foi abordada.

O capítulo 3 é dedicado à apresentação de modelos multiplicativos. Abordamos o ajuste do modelo por mínimos quadrados e por máxima verossimilhança. Decorrente de cada ajuste, temos o teste para a hipótese de aditividade do modelo, e também o estimador de máxima verossimilhança.

No capítulo 4 apresentamos o Biplot. Tratamos de matrizes de posto 2, e também de matrizes de posto maior que 2. Abordamos a representação da matriz de dados, das variâncias e covariâncias das variáveis, e das diferenças padronizadas entre unidades amostrais, através do biplot.

No capítulo 5 tratamos da Análise de Correspondência para tabelas de dupla entrada.

Com a finalidade de tornar clara a utilização das técnicas apresentadas neste trabalho, tivemos a preocupação de incorporar a cada capítulo uma aplicação. Os cálculos, bem como os gráficos envolvidos nas aplicações, foram feitos através do sistema SAS (Statistical Analysis System).

Consta também deste trabalho um apêndice onde revisamos alguns conceitos geométricos no espaço multidimensional, os quais são necessários para a compreensão das técnicas apresentadas.

2. Decomposição em Valores Singulares

2. DECOMPOSIÇÃO EM VALORES SINGULARES

2.1 INTRODUÇÃO

A técnica de decomposição em valores singulares (SVD) de uma matriz é uma generalização, para matrizes retangulares, da decomposição espectral de uma matriz quadrada simétrica. É de grande utilidade na análise de tabelas de dupla entrada, uma vez que essas podem ser consideradas como uma matriz.

A decomposição em valores singulares é uma ferramenta muito útil para álgebra matricial, e portanto de grande aplicabilidade em Estatística. Não são muitos os livros textos de Estatística que abordam esta técnica. Literatura relevante sobre SVD em Estatística deve-se a: Good(1969), Chambers(1977), Gabriel(1978), Rao(1980), Mandel(1982) e Greenacre e Underhill(1982), Searle (1982); em matemática vale a pena destacar Ben-Israel e Greville(1974).

Vamos considerar $A:C^n \longrightarrow C^m$ uma transformação linear em [1] $L(C^n,C^m)$, e sua representação matricial em [2] C^{mxn}. Trata-se de encontrar a representação mais simples (diagonal) desta transformação linear.

Sejam $U=\{u_1,u_2,\ldots,u_m\}$ e $V=\{v_1,v_2,\ldots,v_n\}$ duas bases

---

[1] espaço das transformações lineares que vão de C^n, espaço vetorial n-dimensional, até C^m

[2] classe das matrizes mxn complexas

de C^m e C^n, respectivamente. Denotaremos por $\mathbf{A}_{\{U,V\}}=[a_{ij}] \in C^{mxn}$ a matriz que representa a transformação linear $A:C^n \longrightarrow C^m$, a qual é determinada (unicamente) por:

$$\mathbf{A}v_j = \sum_{i=1}^{m} a_{ij}\mathbf{u}_i \qquad j = 1,2,\ldots,n \qquad (2.1.1)$$

Para quaisquer pares de base $\{\mathbf{U},\mathbf{V}\}$ a expressão (2.1.1) representa uma correspondência um a um entre a transformação linear $L(C^n,C^m)$ e matrizes de C^{mxn}.

Um dos resultados mais importantes, aquele de Eckart e Young(1936), mostra-nos que, para qualquer $\mathbf{A} \in {}^3C_r^{mxn}$, com valores singulares $\lambda(\mathbf{A})=\{\lambda_1,\lambda_2,\ldots,\lambda_r\}$, ordenados por $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$, e para quaisquer r escalares $d(\mathbf{A})=\{d_1,d_2,\ldots,d_r\}$ satisfazendo

$$|d_i| = \lambda_i \qquad i=1,2,\ldots,r$$

existem duas matrizes unitárias $\mathbf{U} \in U^{mxm}$ (conjunto das matrizes mxm unitárias), e $\mathbf{V} \in U^{nxn}$, tais que a matriz

$$\underset{m}{\mathbf{D}}_{\lambda_n} = \mathbf{U}^*\mathbf{A}\mathbf{V} = \begin{bmatrix} d_1 & & & & \vdots & \\ & d_2 & & & \vdots & \\ & & \ddots & & \vdots & \mathbf{0} \\ & & & d_r & \vdots & \\ \cdots & \cdots & \cdots & \cdots & \vdots & \cdots \\ & \mathbf{0} & & & \vdots & \mathbf{0} \end{bmatrix} \text{ é diagonal.}$$

onde $\mathbf{U}^*$ é a transposta conjugada de $\mathbf{U}$.

Assim, qualquer matriz complexa $\mathbf{A}$ mxn é unitariamente [4]equivalente à matriz diagonal $\mathbf{D}_\lambda$, ou seja

---

[3] classe das matrizes mxn, complexas de posto r

[4] matrizes equivalentes: duas matrizes A e B em C^{mxn} são chamadas equivalentes se há matrizes não singulares $S \in C^{mxn}$, $T \in C^{mxn}$ tal que $B=S^{-1}AT$. Se S e T são matrizes unitárias então A e B são chamadas unitariamente equivalentes.

$$\mathbf{A} = \mathbf{U}\, \mathbf{D}_{\lambda} \mathbf{V}^{*} \qquad (2.1.2)$$

A correspondente proposição para transformações lineares diz que para qualquer transformação linear $A:C^{n}\longrightarrow C^{m}$, com posto(**A**)=r, e para qualquer conjunto de escalares $d(\mathbf{A})=\{d_1,\ldots,d_r\}$ satisfazendo $|d_i|=\lambda_i$, existem duas bases ortonormais $\mathbf{V}=\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$ e $\mathbf{U}=\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\}$ de C^n e C^m, respectivamente, tais que a correspondente matriz representação $\mathbf{A}_{\{U,V\}}$ é diagonal,

$$\mathbf{A}_{\{U,V\}}=\begin{bmatrix} d_1 & & & & \vdots & \\ & d_2 & & & \vdots & \\ & & \ddots & & \vdots & \mathbf{0} \\ & & & d_r & \vdots & \\ \cdots & \cdots & \cdots & \cdots & \vdots & \cdots \\ & & \mathbf{0} & & \vdots & \mathbf{0} \end{bmatrix}$$

isto é :

$$\begin{cases} \mathbf{A}\,\mathbf{v}_j = d_j \mathbf{u}_j & j=1,\ldots,r \\ \mathbf{A}\,\mathbf{v}_j = \mathbf{0} & j=r+1,\ldots,n \end{cases} \qquad (2.1.3)$$

Se $d(\mathbf{A})=\lambda(\mathbf{A})$, ou seja, os escalares $\{d_1,d_2,\ldots,d_r\}$ forem escolhidos como os valores singulares de **A**, então (2.1.2) é chamado de decomposição em valores singulares de **A**.

2.2 A FORMA USUAL DA DECOMPOSIÇÃO EM VALORES SINGULARES

A decomposição em valores singulares que apresentaremos aqui é uma variação da decomposição provada por Beltrani, Jordan e Sylvester (veja MacDuffee,1956, pag78) para matrizes quadradas reais. A abordagem que faremos é aquela de Eckart e Young, para matrizes retangulares.

Primeiramente vamos enunciar o seguinte teorema:

TEOREMA 2.1:

Sejam $\mathbf{0 \neq A} \in C_r^{m \times n}$, $\lambda(\mathbf{A})$ os valores singulares de **A**, com

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0 \qquad (2.2.1)$$

e $d(\mathbf{A})=\{d_1, d_2, \ldots, d_r\}$ escalares complexos satisfazendo

$$|d_i| = \lambda_i \qquad i=1,\ldots,r \qquad (2.2.2)$$

Consideremos $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ sendo um conjunto ortonormal de autovetores de $\mathbf{AA}^*$ correspondendo aos seus autovalores não nulos

$$\mathbf{AA}^* \mathbf{u}_i = \lambda_i^2 \mathbf{u}_i \qquad i=1,\ldots,r \qquad (2.2.3)$$

$$(\mathbf{u}_i, \mathbf{u}_j) = \delta_{ij} \qquad i,j=1,\ldots,r \qquad (2.2.4)$$

onde δ_{ij} é o delta de Kronecker.

Seja $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ conjunto de vetores definidos por:

$$\mathbf{v}_i = \frac{1}{d_i} \mathbf{A}^* \mathbf{u}_i \qquad i=1,\ldots,r \qquad (2.2.5)$$

Então $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ é um conjunto ortonormal de autovetores de $\mathbf{A}^*\mathbf{A}$ correspondendo aos seus autovalores não nulos

$$\mathbf{A}^*\mathbf{A} \mathbf{v}_i = \lambda_i^2 \mathbf{v}_i \qquad i=1,\ldots,r \qquad (2.2.6)$$

$$(\mathbf{v}_i, \mathbf{v}_j) = \delta_{ij} \qquad i,j=1,\ldots,r \qquad (2.2.7)$$

Além disso

$$\mathbf{u}_i = \frac{1}{d_i} \mathbf{A}\, \mathbf{v}_i \qquad i=1,\ldots,r \qquad (2.2.8)$$

Reciprocamente, considere os vetores $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ satisfazendo (2.2.6) e (2.2.7), e sejam os vetores $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ definidos por (2.2.8). Então, $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$

satisfaz (2.2.3),(2.2.4) e (2.2.5).

PROVA:

Para demonstrar a equação (2.2.6), consideremos $\mathbf{v}_i$ como definido na equação (2.2.5). Então, pré-multiplicando ambos os lados da equação por $\mathbf{A}^*\mathbf{A}$ tem-se:

$$\mathbf{A}^*\mathbf{A}\mathbf{v}_i = \frac{1}{d_i}\,\mathbf{A}^*\mathbf{A}\,\mathbf{A}^*\mathbf{u}_i \qquad \text{, de (2.2.3) e (2.2.2) segue}$$

$$= d_i\mathbf{A}^*\mathbf{u}_i \qquad \text{, de (2.2.5) e (2.2.2) segue}$$

$$= \lambda_i^2\mathbf{v}_i$$

O resultado da equação (2.2.7) é dado por:

$$(\mathbf{v}_i,\mathbf{v}_j) = \frac{1}{d_i\,d_j}\,(\mathbf{A}^*\mathbf{u}_i,\,\mathbf{A}^*\mathbf{u}_j)$$

$$= \frac{1}{d_i\,d_j}\,(\mathbf{A}\mathbf{A}^*\mathbf{u}_i,\mathbf{u}_j), \qquad \text{de (2.2.3) e (2.2.2) segue}$$

$$= \frac{d_i}{d_j}\,(\mathbf{u}_i,\mathbf{u}_j)$$

$$= \delta_{ij}$$

Para demonstrar a equação (2.2.8), consideremos $\mathbf{v}_i$ como definido na equação (2.2.5), e multiplicando ambos os lados por **A**, tem-se:

$$\mathbf{A}\mathbf{v}_i = \frac{1}{d_i}\,\mathbf{A}\mathbf{A}^*\mathbf{u}_i \qquad \text{de (2.2.3) e (2.2.2) segue}$$

$$= d_i\mathbf{u}_i$$

Portanto $\mathbf{u}_i = \frac{1}{d_i}\,\mathbf{A}\,\mathbf{v}_i$

A prova da proposição recíproca é feita trocando **A** por $\mathbf{A}^*$ a cada passo. ■

Uma consequência do teorema 1 é o teorema 2, de Eckart e Young, que tem aplicação na aproximação de uma matriz por matriz de posto menor.

TEOREMA 2.2:

Sejam $\mathbf{0 \neq A} \in C_r^{m \times n}$, $d(\mathbf{A})=\{d_1,\ldots,d_r\}$ escalares complexos satisfazendo

$$|d_i| = \lambda_i \qquad i=1,\ldots,r$$

onde

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ são valores singulares de $\mathbf{A}$.

Então existem matrizes unitárias $\mathbf{U} \in U^{mxm}$ e $\mathbf{V} \in U^{nxn}$ tais que a matriz

$$\mathbf{D}_\lambda = \mathbf{U}^*\mathbf{A}\,\mathbf{V} = \begin{bmatrix} d_1 & & & & \vdots & \\ & d_2 & & & \vdots & \\ & & \ddots & & \vdots & \mathbf{0} \\ & & & d_r & \vdots & \\ \cdots & \cdots & \cdots & \cdots & \vdots & \cdots \\ & \mathbf{0} & & & \vdots & \mathbf{0} \end{bmatrix} \text{ é diagonal.} \quad (2.2.9)$$

PROVA:

Para $\mathbf{A} \in C_r^{mxn}$, matriz dada, vamos construir 2 matrizes $\mathbf{U}$ e $\mathbf{V}$ como segue. Consideremos os vetores $\{u_1,u_2,\ldots,u_r\}$ em C^m satisfazendo (2.2.3) e (2.2.4), e assim formam uma base ortonormal de $R(\mathbf{AA}^*)=^5R(\mathbf{A})$. Seja $\{u_{r+1},\ldots,u_m\}$ uma base ortonormal de $^6R(\mathbf{A})^\perp=^7N(\mathbf{A}^*)$. Então o conjunto

---

[5] $R(A)=\{y \in C^m : y=Ax \text{ para algum } x \in C^n\}$

[6] $R(A)^\perp$: ortocomplemento de R(A)

[7] $N(A)=\{x \in C : Ax=0\}$, espaço nulo de A

$\{u_1,\ldots,u_r,u_{r+1},\ldots,u_m\}$ é uma base ortonormal de C^m satisfazendo (2.2.3) e

$$A^*u_i = 0 \qquad i=r+1,\ldots,m \tag{2.2.10}$$

A matriz **U** definida por:

$$U=[u_1,\ldots,u_r,u_{r+1},\ldots,u_m] \tag{2.2.11}$$

é assim matriz unitária mxm.

Consideremos agora os vetores $\{v_1,v_2,\ldots,v_r\}$ em C^n como definido em (2.2.5). Então esses vetores satisfazem (2.2.6) e (2.2.7), e consequentemente formam uma base ortonormal de $R(A^*A)=R(A^*)$. Seja $\{v_{r+1},\ldots,v_n\}$ uma base ortonormal de $R(A^*)=N(A)$. Então $\{v_1,\ldots,v_r,v_{r+1},\ldots,v_n\}$ é uma base ortonormal de C^n satisfazendo (2.2.6) e

$$Av_j=0 \qquad j=r+1,\ldots,n \tag{2.2.12}$$

A matriz **V** definida por:

$V=[v_1,\ldots,v_r,v_{r+1},\ldots,v_n]$ é assim uma matriz unitária nxn.

Considerando **U** e **V** da forma expressa acima, a matriz

$$D_\lambda = U^*A\,V = [d_{ij}] \qquad \begin{matrix} i=1,\ldots,m \\ j=1,\ldots,n \end{matrix} \tag{2.2.13}$$

satisfaz:

$d_{ij} = u_i^*\,A\,v_j = 0$, se i>r ou j>r, por (2.2.10) e (2.2.12) e para i,j=1,...,r

$$d_{ij} = u_i^*\,A\,v_j$$

$$= \frac{1}{d_j}\,u_i^*\,A\,A^*\,u_j \qquad \text{por (2.2.5)}$$

$$= d_j\,u_i^*\,u_j \qquad \text{por (2.2.3) e (2.2.2)}$$

$$= d_j \, \delta_{ij} \qquad \text{por (2.2.4)}$$
$$= d_j$$

Assim **D** é da forma:

$$\mathbf{D}_\lambda = \begin{bmatrix} d_1 & & & & \\ & d_2 & & & \\ & & \ddots & & 0 \\ & & & d_r & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & 0 & & 0 \end{bmatrix} \qquad \text{é diagonal.}$$

Como **U** e **V** são matrizes unitárias, de (2.2.13) tem-se:

$$\mathbf{A} = \mathbf{U} \, \mathbf{D}_\lambda \mathbf{V}^* \qquad \blacksquare$$

Uma forma equivalente de expressar a decomposição em valores singulares é:

$$\mathbf{A} = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^*$$

onde os r vetores m-dimensionais $\mathbf{u}_1,\ldots,\mathbf{u}_r$ são denominados de vetores singulares esquerdos[8], constituem uma base ortonormal para as colunas de **A** e são os autovetores de $\mathbf{AA}^*$, associados aos autovalores $\lambda_1^2,\ldots,\lambda_r^2$. Similarmente, os r vetores n-dimensionais $\mathbf{v}_1,\ldots,\mathbf{v}_r$ são denominados de vetores singulares direitos, constituem uma base ortonormal para as colunas de $\mathbf{A}^*$ e são os autovetores de $\mathbf{A}^*\mathbf{A}$, associados aos mesmos autovalores $\lambda_1^2,\ldots,\lambda_r^2$. As matrizes $\mathbf{F}=\mathbf{U}\mathbf{D}_\lambda$ e $\mathbf{G}=\mathbf{V}\mathbf{D}_\lambda$ contêm as coordenadas das linhas e colunas de **A**, relativas aos vetores bases em **V** e **U**,

---

[8] Um par de vetores **U** e **V** são chamados de vetores singulares de uma matriz retangular **A** se as condições $\mathbf{Av}=\lambda\mathbf{u}$ e $\mathbf{A}^t\mathbf{u}=\lambda\mathbf{v}$ são válidas simultaneamente, para qualquer valor singular de **A**, $\lambda>0$.

respectivamente. Se $\mathbf{A} = \mathbf{U}\,\mathbf{D}_\lambda\mathbf{V}^*$ for expresso por $\mathbf{A}=\mathbf{F}\mathbf{V}^*$, podemos escrever a i-ésima linha de **A**, $\mathbf{a}_i^*$, como:

$$\mathbf{a}_i^* = \sum_j^r f_{ij}\mathbf{v}_j$$

assim, a i-ésima linha de **F** contém as coordenadas de $\mathbf{a}_i$. Similarmente, a j-ésima linha de **G** contêm as coordenadas da j-ésima coluna de **A**, relativas aos vetores bases de **U**.

2.3 APROXIMAÇÃO DE UMA MATRIZ POR MATRIZ DE POSTO MENOR

Para uma dada matriz $\mathbf{A} \in C_r^{mxn}$ e um inteiro k, $1 \leq k \leq r$, a melhor aproximação de **A**, de posto k, é a matriz $\mathbf{A}_{(k)} \in C_k^{mxn}$, satisfazendo

$$\| \mathbf{A} - \mathbf{A}_{(k)} \| = \inf_{X \in C_k^{mxn}} \| \mathbf{A} - X \|$$

onde $\| \;\; \|$ é a norma da matriz, ou seja

$$\| \mathbf{A} \| = (\text{traço } \mathbf{A}^*\mathbf{A})^{1/2} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij} |^2 \right)^{1/2}$$

Para as matrizes $\mathbf{D}_\lambda$, **U** e **V** do teorema 2, consideremos $\mathbf{D}_{\lambda(k)}$, $\mathbf{U}_{(k)}$ e $\mathbf{V}_{(k)}$ submatrizes definidas por:

$$\mathbf{D}_{\lambda(k)} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_k \end{bmatrix} \in C^{kxk}$$

$$\mathbf{U}_{(k)} = [\mathbf{u}_1, \ldots, \mathbf{u}_k] \in C^{mxk}$$

$$\mathbf{V}_{(k)} = [\mathbf{v}_1, \ldots, \mathbf{v}_k] \in C^{nxk}$$

Então a melhor aproximação de **A** de posto k é:

$$\mathbf{A}_{(k)} = \mathbf{U}_{(k)}\mathbf{D}_{\lambda(k)}\mathbf{V}_{(k)}^* \qquad (2.3.1)$$

a qual é única se e somente se o k-ésimo e (k+1)-ésimo valores singulares de **A** forem distintos ($\lambda_k \neq \lambda_{k+1}$).

O erro da aproximação de **A** por $\mathbf{A}_k$ é dado por:

$$\| \mathbf{A} - \mathbf{A}_{(k)} \| = \left(\sum_{i=k+1}^{r} \lambda_i^2 \right)^{1/2} \qquad (2.3.2)$$

PROVA:

A norma da matriz é unitariamente invariante. Assim, para qualquer $\mathbf{X} \in C_k^{mxn}$

$$\| \mathbf{A} - \mathbf{X} \| = \| \mathbf{U}^* \mathbf{A} \mathbf{V} - \mathbf{U}^* \mathbf{X} \mathbf{V} \| \qquad \text{por (2.2.9)}$$

$$= \| \mathbf{D}_\lambda - \mathbf{U}^* \mathbf{X} \mathbf{V} \|$$

Denotando $\mathbf{U}^* \mathbf{X} \mathbf{V} = \mathbf{Z} = [\, z_{ij} \,]$ (2.3.3)

$$\| \mathbf{A} - \mathbf{X} \| = \| \mathbf{D}_\lambda - \mathbf{Z} \|$$

Assim, vamos minimizar a norma ao quadrado:

$$\| \mathbf{A} - \mathbf{X} \|^2 = \| \mathbf{D}_\lambda - \mathbf{Z} \|^2$$

$$= \sum_{i=1}^{r} | d_i - z_{ii} |^2 + \sum_{\substack{i \neq j \\ \text{ou} \\ i > r}} | z_{ij} |^2 \qquad (2.3.4)$$

O único $Z \in C_k^{mxn}$ que minimiza (2.3.4) é portanto:

$$z_{ij} = \begin{cases} d_i & , \text{ se } 1 \leq i=j \leq k \\ 0 & , \text{ c.c.} \end{cases} \qquad (2.3.5)$$

e (2.3.1) segue de (2.3.5) e (2.3.3).

Pode-se mostrar que $\mathbf{A}_{(k)}$ pode ser reescrita como:

$$\mathbf{A}_{(k)} = \sum_{i=1}^{k} d_i \mathbf{u}_i \mathbf{v}_i^*$$

$$= \left(\sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^* \right) \mathbf{A} \qquad \text{por (2.2.5)}$$

e esta expressão pode ser usada para provar a unicidade de $\mathbf{A}_{(k)}$.

Finalmente, (2.3.2) segue de

$$\| \mathbf{A} - \mathbf{A}_{(k)} \|^2 = \sum_{i=k+1}^{r} \lambda_i^2 \text{ por (2.3.4),(2.3.5) e (2.2.2).} \quad \blacksquare$$

2.4 DECOMPOSIÇÃO EM VALORES SINGULARES GENERALIZADA

Vamos considerar Ψ(mxm) e Φ(nxn) matrizes simétricas positivas semi-definidas conhecidas; então qualquer matriz $\mathbf{A} \in C_r^{m \times n}$ pode ser expressa como:

$$\mathbf{A} = \mathbf{N}\, \mathbf{D}_\lambda \mathbf{M}^* = \sum_{i=1}^{r} \lambda_i \mathbf{n}_i \mathbf{m}_i^* \qquad (2.4.1)$$

onde $\mathbf{N}^* \Psi\, \mathbf{N} = {}^{9}\mathbf{I}_m$ e $\mathbf{M}^* \Phi\, \mathbf{M} = \mathbf{I}_n$, (2.4.2)

ou seja, as colunas de **N** e **M** são ortonormalizadas com respeito a Ψ e Φ, respectivamente. Esta decomposição é denominada de **Decomposição em Valores Singulares Generalizada nas métricas Ψ e Φ .**

As colunas de **N** e **M** são chamadas de vetores singulares generalizados esquerdos e direitos,respectivamente. Eles ainda são bases ortonormais para as colunas e linhas de **A**, mas a métrica imposta nos espaços m e n-dimensionais não é a métrica Euclideana simples, mas sim a métrica Euclideana generalizada (ou ponderada), definida por Ψ e Φ respectivamente. Similarmente, os elementos da diagonal da matriz $\mathbf{D}_\lambda$ podem ser chamados de valores singulares generalizados, ordenados do maior para o menor.

A SVD generalizada é facilmente provada quando se

---

[9] $\mathbf{I}_m$: matriz identidade mxm.

assume a decomposição em valores singulares usual de $\Psi^{1/2}\mathbf{A}\ \Phi^{1/2}$, onde usamos a raiz quadrada da matriz simétrica (isto é, se Ψ é decomposta como $\Psi=\mathbf{W}\mathbf{D}_\alpha\mathbf{W}^*$ então $\Psi^{1/2}=\mathbf{W}\mathbf{D}_\alpha^{1/2}\mathbf{W}^*$):

$$\Psi^{1/2}\mathbf{A}\ \Phi^{1/2}= \mathbf{U}\ \mathbf{D}_\lambda\mathbf{V}^* \qquad \text{onde} \qquad \mathbf{U}^*\mathbf{U} = \mathbf{I}_m \quad \text{e} \quad \mathbf{V}^*\mathbf{V} = \mathbf{I}_n \tag{2.4.3}$$

Considerando $\mathbf{N} = \Psi^{-1/2}\mathbf{U}$ e $\mathbf{M} = \Phi^{-1/2}\mathbf{V}$ (2.4.4)

tem-se (2.4.1) e (2.4.2).

A correspondente generalização para aproximação de **A** por uma matriz de posto menor, a qual é induzida pelo resultado do item (2.3) é como segue: se os últimos r-k termos de (2.4.1) são retirados, então $\mathbf{A}_{(k)}= \sum_i^k \lambda_i \mathbf{n}_i \mathbf{m}_i^*$ = $\mathbf{N}_{(k)}\mathbf{D}_{\lambda(k)}\mathbf{M}^*_{(k)}$ é a aproximação de mínimos quadrados generalizada de **A**, de posto k, a qual minimiza:

$$\text{traço}\ \{\ \Psi(\mathbf{A}-\mathbf{X})\Phi(\mathbf{A}-\mathbf{X})^*\} \tag{2.4.5}$$

entre todas as matrizes **X** de posto menor ou igual a k.

Quando Ψ é uma matriz diagonal $\mathbf{D}_w$ de números positivos $w_1,\ldots,w_m$, a expressão (2.4.5) pode ser escrita como:

$$\text{traço}\{\mathbf{D}_w(\mathbf{A}-\mathbf{X})\Phi(\mathbf{A}-\mathbf{X})^*\} = \sum_i^m w_i(\mathbf{a}_i- \mathbf{x}_i)^*\Phi(\mathbf{a}_i-\mathbf{x}_i) \tag{2.4.6}$$

onde $\mathbf{a}_i$ e $\mathbf{x}_i$ são as linhas de **A** e **X** respectivamente, escritas como vetores coluna. Os valores $w_1,\ldots,w_m$ são massas (pesos), os quais são atribuidas as linhas. As linhas de **X** são pontos desconhecidos no subespaço k-dimensional e o mínimo de (2.4.6), atingido por $\mathbf{X}=\mathbf{A}_{(k)}$, identifica o subespaço o qual está mais próximo dos pontos em termos de soma ponderada de distâncias ao quadrado. Neste caso os vetores $m_1,\ldots,m_k$

definem os eixos principais ortonormais do subespaço, e a matriz $\mathbf{N}_{(k)}\mathbf{D}_{\lambda(k)}$ define as coordenadas (com relação a esses eixos) das projeções dos pontos no subespaço. É importante ressaltar que a ortonormalidade dos eixos e das projeções é definida em termos da métrica Φ.

Usualmente as linhas de $\mathbf{F}=\mathbf{N}_{(k)}\mathbf{D}_{\lambda(k)}$ são desenhadas como pontos no subespaço Euclideano k-dimensional (k frequentemente sendo 2 ou 3) para explorar a configuração multidimensional das linhas da matriz **A**. Entretanto, é possível desenhar também as linhas de $\mathbf{G}=\mathbf{M}_{(k)}\mathbf{D}_{\lambda(k)}$ no mesmo espaço, como descrito por Gabriel(1971,1980,1981).

Note que a matriz aproximação $\mathbf{A}_{(k)}$, equivalentemente, o subespaço ótimo definido por $\mathbf{M}_{(k)}$, é única se λ_k é estritamente maior que λ_{k+1}.

3. Modelos Multiplicativos

3. MODELOS MULTIPLICATIVOS

3.1 INTRODUÇÃO

São frequentes situações em que o objetivo de um experimento é avaliar a dependência entre características de interesse (variável resposta) e fatores. Em geral, o que se faz é escolher um número fixo de níveis para cada um dos fatores, e obter, através do experimento, respostas para todas as combinações dos níveis dos fatores. Tal procedimento produz uma matriz de dados.

Vamos considerar situações em que a resposta é uma variável aleatória quantitativa e que a natureza dos dados depende de dois fatores que são representados pelas linhas e colunas da matriz.

Denotaremos por α e β os fatores com m e n níveis respectivamente, e y_{ij} a resposta associada ao i-ésimo nível do fator α e j-ésimo nível do fator β. Temos portanto um conjunto de mn observações. Conjuntos como esse são frequentemente analisados através do ajuste de modelos que expliquem, o melhor possível, o comportamento da variável resposta em função dos fatores considerados, com um mínimo de parâmetros. Dessa forma, as seguintes representações poderiam ser tentadas sucessivamente:

(i) $\quad y_{ij} = \mu$

(ii) $\dot{y}_{ij} = \mu + \alpha_i$ ou $\dot{y}_{ij} = \mu + \beta_j$

(iii) $\dot{y}_{ij} = \mu + \alpha_i + \beta_j$

(iv) $\dot{y}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$

onde:

μ constante, estimada pela média global

α_i "efeito principal da linha"

β_j "efeito principal da coluna"

γ_{ij} "interação entre linha i e coluna j"

Em (i) temos o modelo mais simples possível, se considerarmos adequado informa que ambos os fatores não tem efeito. Em (ii) temos a influência de um só fator e esta é aditiva; em (iii) a dependência aditiva de ambos os fatores. A expressão (iv) informa apenas que a suposição de aditividade não é válida. Notemos que a redução para os dados só é obtida nos casos (i), (ii) e (iii). Os mn números são representados por um parâmetro em (i), por m ou n parâmetros em (ii) e por m+n parâmetros em (iii). Em (iv), desde que $\mu=0$, $\alpha_i \equiv 0$, $\beta_j \equiv 0$, $\gamma_{ij}=y_{ij}$ produz zeros para os resíduos. Esta representação tem mn parâmetros e não há redução dos dados. No caso do modelo expresso por (iii), temos uma aproximação que depende de i e de j através da soma de duas funções, uma dependendo só de i e outra dependendo só de j. O mérito desse modelo é a simplicidade de interpretação.

A suposição de inexistência de interação não deveria ser feita sem justificativa de sua validade. Entretanto, encontramos muitos pesquisadores que lançam mão dessa suposição, especialmente quando esta impossibilita o emprego

de algum teste estatístico. No caso de experimentos em que se tem apenas uma observação a cada combinação dos níveis dos fatores (experimentos sem replicação), o problema de interação torna-se crítico pois:

- os métodos usuais não se prestam à verificação da validade do modelo aditivo, isto é, não determinam a presença de interação;

- a não aditividade faz com que os modelos da forma (iv) não representem redução alguma dos dados, ou seja, as mn observações são representadas por mn parâmetros;

- os dados não fornecem um estimador não tendencioso para a variância do erro aleatório;

Devemos observar que existem alguns procedimentos que se prestam a testar a não aditividade, no caso de dois fatores (veja Freitas,1986), tal como "um grau de liberdade para não aditividade de Tukey", (Tukey,1949). Entretanto, o maior interesse reside em modelar a interação. O método que iremos abordar foi desenvolvido por Mandel(1969) e permite modelar termos da interação, estimar σ^2 e, consequentemente, determinar sua presença, mesmo em experimentos sem replicação. A idéia é particionar a função de duas variáveis em duas funções multiplicativas de i e j. Isto porque uma partição aditiva é impossível, dado que as partes aditivas já foram extraídas do modelo.

Embora essa técnica possa ser estendida para um número maior de fatores (Carvalho,1977 e Giannotti,1982), estudaremos apenas o caso de 2 fatores (veja Miliken e Johnson,1989).

3.2 O MODELO DE MANDEL PARA A INTERAÇAO DE DOIS FATORES

Considere um experimento fatorial mxn sem replicação onde a variável resposta pode ser representada pelo modelo

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij} \qquad \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,n \end{matrix} \qquad (3.2.1)$$

onde e_{ij} é considerado erro aleatório não observável com média zero e variância constante.

Usualmente as seguintes restrições são impostas:

$$\sum_{i=1}^{m} \alpha_i = 0 \qquad \sum_{j=1}^{n} \beta_j = 0 \qquad \sum_{i=1}^{m} \gamma_{ij} = \sum_{j=1}^{n} \gamma_{ij} = 0$$

Pode-se notar que o modelo (3.2.1) expressa y_{ij}, função de duas variáveis, como duas funções que dependem de uma só variável, α_i e β_j, mas também envolve uma função de duas variáveis, γ_{ij}.

No caso em que se tem $\gamma_{ij}=0$, o ajuste do modelo acima é feito por mínimos quadrados, e produz um resíduo que será apenas um erro aleatório. No caso da interação estar presente, o resíduo conterá, além do erro aleatório, um componente sistemático. Baseados nisso, vamos procurar uma representação para o resíduo de tal forma que o componente sistemático possa ser isolado e que a representação obtida seja informativa à análise.

A suposição adotada por Mandel(1969) é expressa pela seguinte equação:

$$\gamma_{ij} = \theta\, u_i v_j + \theta' u'_i v'_j + \theta'' u''_i v''_j + \ldots + e^*_{ij} \qquad (3.2.2)$$

onde e^*_{ij} é uma variável aleatória com média zero e

desvio-padrão σ.

Sem perda de generalidade, são impostas as seguintes restrições:

$$\sum_i u_i = \sum_i u'_i = \ldots = \sum_j v_j = \sum_j v'_j = \ldots = 0 \qquad (3.2.3)$$

$$\sum_i u_i^2 = \sum_i u_i'^2 = \ldots = \sum_j v_j^2 = \sum_j v_j'^2 = \ldots = 1 \qquad (3.2.4)$$

Quando γ_{ij} é expresso apenas por poucos termos multiplicativos do tipo $\theta u_i v_j^t$, a equação (3.2.2) ainda constitui uma real simplificação, pois esta representa uma aproximação de mn valores com (m-1)(n-1) graus de liberdade, por k(m+n-1) parâmetros, sendo k o número de termos.

As equações (3.2.1) e (3.2.2) com as restrições (3.2.3) e (3.2.4) constituem o modelo de Mandel(1969), sendo μ, α_i, β_j, θ, u_i, v_j, θ', u'_i, v'_j, etc... os parâmetros estruturais.

Trata-se agora de encontrar estimativas para os novos parâmetros θ, θ', u, u',..., v, v', ... e para o desvio padrão do erro aleatório (σ).

3.3 O AJUSTE POR MÍNIMOS QUADRADOS

Vamos retomar o modelo expresso em (3.2.1)

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij} \qquad \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,n \end{matrix}$$

Um conjunto de soluções de mínimos quadrados para os parâmetros do modelo acima é dado por:

$$\hat{\mu} = y_{..}$$

$$\hat{\alpha}_i = y_{i.} - y_{..}$$

$$\hat{\beta}_j = y_{.j} - y_{..}$$

$$\hat{\gamma}_{ij} = y_{ij} - y_{i.} - y_{.j} + y_{..} = r_{ij} \qquad (3.3.1)$$

Os símbolos "$\wedge$" e "." são usualmente utilizados para representar a solução de mínimos quadrados do parâmetro e a média da variável sobre o índice omitido, respectivamente.

Em (3.3.1) temos o resíduo do modelo após o ajuste da parte aditiva de (3.2.1). Quando a interação estiver presente no modelo, este deverá conter um componente sistemático, o qual poderá ser detectado através de uma aproximação de mínimos quadrados de r_{ij} pelo modelo (3.2.2).

Seja $\mathbf{R}=(r_{ij})$ uma matriz mxn de resíduos e $\mathbf{u}$ e $\mathbf{v}$ vetores coluna com m e n elementos, respectivamente. (Tem-se $\mathbf{1}^t\mathbf{R}=\mathbf{0}^t$ e $\mathbf{R1}=\mathbf{0}$, sendo $\mathbf{1}^t=(1...1)$). Então a equação (3.2.2) pode ser reescrita em forma matricial,

$$\mathbf{R} = \sum_{i=1}^{p} \theta_i \, \mathbf{u}_i \, \mathbf{v}_i^t + \mathbf{E} \qquad (3.3.2)$$

onde:

$p \leq$ posto($\mathbf{R}$), inteiro fixado

$\mathbf{E}$ matriz de erros aleatórios

$$\theta_1 \geq \theta_2 \geq \theta_3 \geq \ldots \geq \theta_p \geq 0$$

$$\| \mathbf{u}_i \|^2 = \| \mathbf{v}_i \|^2 = 1 \qquad i=1,2,\ldots,p$$

Pode-se observar que $\mathbf{u}_i\mathbf{v}_i^t$ forma uma matriz de posto menor ou igual a 1 e, como o posto da soma de duas matrizes não pode exceder a soma dos postos dos somandos, tem-se:

$$\text{posto}\left(\sum_{i=1}^{p} \theta_i \, \mathbf{u}_i \, \mathbf{v}_i^t\right) \leq p$$

Para a obtenção dos estimadores de mínimos quadrados dos parâmetros em (3.3.2) utiliza-se o resultado do teorema 2.2 (Eckart e Young), apresentado no capitulo 2. Este teorema mostra que a melhor aproximação de uma matriz por matriz de posto menor é dada pela soma de p matrizes de posto 1, determinadas pela decomposição em valores singulares de **R**, para $p \leq \text{posto}(\mathbf{R})$.

As soluções de mínimos quadrados dos parâmetros de

$$\mathbf{R} = \sum_{i=1}^{p} \theta_i \, \mathbf{u}_i \, \mathbf{v}_i^t + \mathbf{E}$$

são dadas por:

$\hat{\theta}_1=\lambda_1 \geq \hat{\theta}_2=\lambda_2 \geq \ldots \geq \hat{\theta}_p=\lambda_p$	$\lambda_1, \lambda_2, \ldots, \lambda_p$ são os p primeiros valores singulares de **R**;
$\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_p$	autovetores ortonormalizados de $\mathbf{RR}^t$, correspondendo aos autovalores $\hat{\theta}_1^2, \hat{\theta}_2^2, \ldots, \hat{\theta}_p^2$;
$\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_p$	autovetores ortonormalizados de $\mathbf{R}^t\mathbf{R}$, correspondendo aos autovalores $\hat{\theta}_1^2, \hat{\theta}_2^2, \ldots, \hat{\theta}_p^2$;

A aproximação de **R** vista dessa maneira, é bastante simples e necessita apenas de um algoritmo com boa estabilidade numérica, como por exemplo o algoritmo de Francis (veja Lawson e Henson,1974).

Devemos agora, procurar algum meio de determinar o número p de componentes multiplicativos, que devem permanecer no modelo, de tal forma que representem a interação e que o modelo nos forneça um estimador para a variância σ^2.

3.4 UMA ANALOGIA COM ANÁLISE DE VARIÂNCIA

Seja $\mathbf{R}=(r_{ij})$ uma matriz mxn , onde

$$r_{ij} = y_{ij} - y_{i.} - y_{.j} - y_{..}$$

ou seja, r_{ij} é o resíduo do modelo (3.2.1) após o ajuste dos termos aditivos.

Seja

$$\mathbf{R} \simeq \sum_{i=1}^{p} \hat{\theta}_i \, \hat{\mathbf{u}}_i \, \hat{\mathbf{v}}_i^t \quad , \text{ para } \quad p \leq \text{posto}(\mathbf{R}) \qquad (3.4.1)$$

A soma de quadrados dos resíduos do modelo aditivo é dada por:

$$\| \mathbf{R} \|^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (r_{ij})^2$$

e nos interessa estudar a forma com que ela se decompõe quando os resíduos são aproximados por (3.4.1).

Denotaremos então, por $\mathbf{R}_i$ (i=1,...,p), os resíduos de (3.3.2), obtidos após o ajuste de i componentes $\theta \, \mathbf{u} \, \mathbf{v}^t$, ou seja:

$$\mathbf{R}_1 = \mathbf{R} - \hat{\theta}_1 \, \hat{\mathbf{u}}_1 \, \hat{\mathbf{v}}_1^t$$

$$\mathbf{R}_2 = \mathbf{R}_1 - \hat{\theta}_2 \, \hat{\mathbf{u}}_2 \, \hat{\mathbf{v}}_2^t = \mathbf{R} - (\hat{\theta}_1 \, \hat{\mathbf{u}}_1 \, \hat{\mathbf{v}}_1^t + \hat{\theta}_2 \, \hat{\mathbf{u}}_2 \, \hat{\mathbf{v}}_2^t)$$

$$\vdots$$

$$\mathbf{R}_k = \mathbf{R}_{k-1} - \hat{\theta}_k \, \hat{\mathbf{u}}_k \, \hat{\mathbf{v}}_k^t = \mathbf{R} - (\hat{\theta}_1 \, \hat{\mathbf{u}}_1 \, \hat{\mathbf{v}}_1^t + \ldots + \hat{\theta}_k \, \hat{\mathbf{u}}_k \, \hat{\mathbf{v}}_k^t)$$

onde k é um inteiro tal que $1 \leq k \leq p$.

Temos então,

$$\mathbf{R}_k = \mathbf{R} - \sum_{i=1}^{k} \hat{\theta}_i \, \hat{\mathbf{u}}_i \, \hat{\mathbf{v}}_i^t = \sum_{i=k+1}^{p} \hat{\theta}_i \, \hat{\mathbf{u}}_i \, \hat{\mathbf{v}}_i^t$$

e

$$\| \mathbf{R}_k \|^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (r_{ij}^{(k)})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\sum_{l=k+1}^{p} \hat{\theta}_l \hat{u}_i^{(l)} \hat{v}_j^{(l)} \right]^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=k+1}^{p} \sum_{l'=k+1}^{p} \hat{\theta}_l \hat{\theta}_{l'} \hat{u}_i^{(l)} \hat{u}_i^{(l')} \hat{v}_j^{(l)} \hat{v}_j^{(l')}$$

$$= \sum_{l=k+1}^{p} \sum_{l'=k+1}^{p} \hat{\theta}_l \hat{\theta}_{l'} \sum_{i=1}^{m} \hat{u}_i^{(l)} \hat{u}_i^{(l')} \sum_{j=1}^{n} \hat{v}_j^{(l)} \hat{v}_j^{(l')} \qquad (3.4.2)$$

Como os conjuntos $\{\hat{u}_i, i=1,\ldots,p\}$ e $\{\hat{v}_i, i=1,\ldots,p\}$ são dois conjuntos ortonormais de vetores,

$$\sum_{i=1}^{m} \hat{u}_i^{(l)} \hat{u}_i^{(l')} = \sum_{j=1}^{n} \hat{v}_j^{(l)} \hat{v}_j^{(l')} \quad \begin{cases} = 0 & \text{para } l \neq l' \\ = 1 & \text{para } l = l' \end{cases} \qquad (3.4.3)$$

logo de (3.4.2) e (3.4.3) resulta que

$$\| \mathbf{R}_k \|^2 = \sum_{l=k+1}^{p} \hat{\theta}_l^2 \quad , \; 1 \leq k \leq p \qquad (3.4.4)$$

De (3.4.4) podemos verificar que

$$\| \mathbf{R}_k \|^2 - \| \mathbf{R}_{k+1} \|^2 = \hat{\theta}_{k+1}^2$$

o que representa a redução na soma de quadrados obtida pela inclusão do componente $\theta_{k+1} \mathbf{u}_{k+1} \mathbf{v}_{k+1}^t$ no modelo.

Conseguimos, assim, uma fórmula geral para a soma de quadrados dos resíduos, dada por:

$$\| \mathbf{R} \|^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (r_{ij})^2 = \sum_{i=1}^{p} \hat{\theta}_i^2$$

e o quanto cada componente multiplicativo contribui para essa soma de quadrados.

Esse particionamento da soma de quadrados dos resíduos, em p componentes $\hat{\theta}_i^2$, nos induz a pensar numa análise do tipo análise de variância, para decidir se um componente é

ou não significante.

Surge, no entanto, a questão sobre a distribuição dos $\hat{\theta}_i^2$, que não são formas quadráticas das variáveis originais y_{ij}. Se supusermos que os erros e_{ij} são normalmente distribuídos, então os $\hat{\theta}_i^2$ são distribuídos como os autovalores de uma matriz com distribuição de Wishart. Isso pode ser verificado em Carvalho(1977).

Uma maneira de resolver o problema, é considerar algumas suposições feitas por Mandel(1969), que nos permitirão realizar uma análise de forma análoga à análise de variância usual.

Suponha que os r_{ij} são elementos de uma amostra aleatória, proveniente de uma população normalmente distribuida, com média zero e variância unitária.

Sejam as quantidades

$$M_1=E[\hat{\theta}_1^2] \quad , \quad M_2=E[\hat{\theta}_2^2] \quad , \quad M_3=E[\hat{\theta}_3^2] \quad , \ldots$$

os valores esperados dos resultados $\hat{\theta}_1^2$, $\hat{\theta}_2^2$, $\hat{\theta}_3^2$,... obtidos dos valores r_{ij} .

Se a variância da população for σ^2, temos

$$E[\hat{\theta}_1^2]= \sigma^2 * M_1, \quad E[\hat{\theta}_2^2]= \sigma^2 * M_2 \quad , \ldots$$

e as razões $\hat{\theta}_i^2/M_i$ serão, todas, estimadores não tendenciosos de σ^2.

Agora, se uma matriz $Y=(y_{ij})$ é dada, e se os elementos $r_{ij}= y_{ij}- y_{i.}- y_{.j}+ y_{..}$ são apenas erros aleatórios normais, então as razões $\hat{\theta}_i^2/M_i$ serão simplesmente estimativas de σ^2, onde os valores $\hat{\theta}_i^2$ são obtidos de $R=(r_{ij})$ e as quantidades M_i de uma matriz de elementos N(0,1) de mesmas

dimensões.

As quantidades M_i, obtidas dessa maneira, cumprem, em certo sentido, um papel análogo ao dos graus de liberdade na análise de variância usual e são chamados de "Pseudo graus de Liberdade".

Uma tabela do tipo da análise de variância pode então ser construida, para decidir se um componente deve ou não ser mantido no modelo (tabela 3.1).

Tabela 3.1 Análise de Variância

FONTE DE VARIAÇÃO	**GRAUS DE LIBERDADE**	**SOMA DE QUADRADOS**	**RAZÃO**
α	$m-1$	$n \sum_{i=1}^{m} (y_{i.} - y_{..})^2$	
β	$n-1$	$m \sum_{j=1}^{n} (y_{.j} - y_{..})^2$	
γ	$(m-1)(n-1)$	$\sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - y_{i.} - y_{.j} + y_{..})^2$	
$\theta_1 \mathbf{u}_1 \mathbf{v}_1^t$	M_1	$\hat{\theta}_1^2$	$\hat{\theta}_1^2/M_1$
$\theta_2 \mathbf{u}_2 \mathbf{v}_2^t$	M_2	$\hat{\theta}_2^2$	$\hat{\theta}_2^2/M_2$
$\theta_3 \mathbf{u}_3 \mathbf{v}_3^t$	M_3	$\hat{\theta}_3^2$	$\hat{\theta}_3^2/M_3$
.	.	.	.
.	.	.	.
.	.	.	.
TOTAL (corrigido)	$mn-1$	$\sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - y_{..})^2$	

Se as razões $\hat{\theta}_i^2/M_i$ forem todas aproximadamente iguais, não teremos argumentos para rejeitar a suposição de que os elementos da interação são erros aleatórios normais. Então, a análise poderá ser levada a cabo, utilizando-se a interação como resíduo.

Por outro lado, se $\hat{\theta}_1^2/M_1 >> \hat{\theta}_2^2/M_2 \simeq \hat{\theta}_3^2/M_3 \simeq \ldots$, podemos supor que existe um termo multiplicativo que representa a interação e que $\| R \|^2 - \hat{\theta}_1^2$ só estima a variância σ^2.

De maneira análoga, podemos supor que existem dois termos multiplicativos que representam a interação, se

$$\hat{\theta}_1^2/M_1 >> \hat{\theta}_2^2/M_2 >> \hat{\theta}_3^2/M_3 \simeq \hat{\theta}_4^2/M_4 \simeq \ldots,$$

e assim por diante.

Esse é o caso, em que, além de obtermos um estimador para a variância σ^2, podemos interpretar o comportamento da interação.

A variável resposta y_{ij} do experimento poderá, nesses casos, ser representada por

$$y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \sum_{l=1}^{k} \hat{\theta}_l \hat{u}_i^{(l)} \hat{v}_j^{(l)t}$$

onde k é o número de termos multiplicativos selecionados para representar a interação.

Pode ocorrer, também, que a interação esteja presente no experimento mas não apresente, nem ao menos aproximadamente, um comportamento multiplicativo. Nesse caso, as razões apresentarão um comportamento diferente dos mencionados acima, e a representação da interação através do modelo multiplicativo não será útil.

3.4.1 OS PSEUDO GRAUS DE LIBERDADE

As quantidades M_i podem ser encontradas tabeladas no apêndice do artigo de Mandel(1969), para i≤3 e alguns valores específicos de m e n ($4 \leq m \leq 100$ e $4 \leq n \leq 20$).

Além das quantidades M_i, o artigo nos fornece também seus desvios padrão e os valores M_i expressos como porcentagem do total de graus de liberdade da interação. Essas porcentagens prestam-se a uma interpolação mais precisa que os valores M_i.

Os valores M_1,M_2, e M_3, foram obtidos através de 625 experimentos de Monte Carlo, para cada combinação de m e n, e representam os valores esperados dos três primeiros autovalores de uma matriz de Wishart.

Essas tabelas estão apresentadas no apêndice B (Tabelas B1, B2 e B3).

3.5 O AJUSTE PELO MÉTODO DE MÁXIMA VEROSSIMILHANÇA

Johnson e Graybill(1972), deduziram os estimadores de máxima verossimilhança para os parâmetros no modelo:

$$y_{ij} = \mu + \alpha_i + \beta_j + \theta\, u_i v_j + e_{ij} \qquad \begin{matrix} i=1,\ldots,m \\ j=1,\ldots,n \end{matrix} \qquad (3.5.1)$$

onde os e_{ij} são independentes e identicamente distribuidos $N(0,\sigma^2)$. As seguintes restrições foram impostas:

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = \sum_{i=1}^{m} u_i = \sum_{j=1}^{n} v_j = 0$$

$$\sum_{i=1}^{m} u_i^2 = \sum_{j=1}^{n} v_j^2 = 1 \qquad \text{e} \qquad n \leq m$$

Note que essas restrições não implicam em perda de generalidade na parte aditiva. Na parte multiplicativa as restrições já são atendidas ($\sum_{i=1}^{m} u_i = \sum_{j=1}^{n} v_j = 0$). O parâmetro θ incorpora o fator escalar de **u** e **v**, de sorte que $\sum_{i=1}^{m} u_i^2 = \sum_{j=1}^{n} v_j^2 = 1$ tampouco é restritivo.

Novamente vamos considerar $\mathbf{R}=[r_{ij}]$, i=1,...,m e j=1,...,n, matriz mxn de resíduos do modelo aditivo, ou seja

$$r_{ij} = y_{ij} - y_{i.} - y_{.j} + y_{..}$$

Note que

$$\mathbf{R} = (I_m - \frac{1}{m} J_m)\, Y\, (I_n - \frac{1}{n} J_n) ,$$

onde **Y** é uma matriz mxn de observações; I_m e I_n são matrizes identidade mxm e nxn respectivamente; J_m e J_n são matrizes de 1's com dimensões mxm e nxn respectivamente.

Sejam $\lambda_1^2 > \lambda_2^2 > \ldots > \lambda_{n-1}^2$ os autovalores não nulos de $\mathbf{R}^t\mathbf{R}$. Note que $\mathbf{R}^t\mathbf{R}$ e $\mathbf{R}\mathbf{R}^t$ tem o mesmo conjunto de autovalores não nulos.

Apresentaremos a seguir 3 lemas que são utilizados para encontrar os estimadores de máxima verossimilhança dos parâmetros no modelo dado.

Lema 3.1: O valor máximo de $(\sum\sum_{ij} u_i v_j y_{ij})^2$ com respeito a u_i e v_j, sujeito às restrições

$$\sum_i u_i = \sum_j v_j = 0 \quad \text{e} \quad \sum_i u_i^2 = \sum_j v_j^2 = 1 \qquad (3.5.2)$$

é λ_1^2, sendo λ_1^2 o maior autovalor de $\mathbf{R}^t\mathbf{R}$. O valor máximo é atingido quando $\mathbf{u}=[u_i]$ é o autovetor normalizado de $\mathbf{R}\mathbf{R}^t$ e

$\mathbf{v}=[v_j]$ é o autovetor normalizado de $\mathbf{R}^t\mathbf{R}$, correspondentes ao autovalor λ_1^2.

Lema 3.2: Considere $\mathbf{Y}$, $\mathbf{R}$, $\mathbf{u}$, $\mathbf{v}$, λ_1 como definidos acima. O valor máximo de $(\sum_i \sum_j u_i v_j r_{ij})^2$ com respeito a u_i e v_j, sujeito às restrições em (3.5.2), é λ_1, e é atingido quando $\mathbf{u}=[u_i]$ é o autovetor normalizado de $\mathbf{RR}^t$ e $\mathbf{v}=[v_j]$ é o autovetor normalizado de $\mathbf{R}^t\mathbf{R}$, correspondentes ao autovalor λ_1.

Lema 3.3: A seguinte desigualdade é válida para todo real $x > -1$ e para todos os valores positivos y e n:

$$(y\, e^{-1})^n \geq [y\,(1+x)\, e^{-(x+1)}]^n$$

Teorema 3.1: No modelo expresso por (3.5.1), os estimadores de máxima verossimilhança dos parâmetros são:

$\hat{\mu} = y_{..}$

$\hat{\alpha}_i = y_{i.} - y_{..}$ $\quad i=1,\ldots,m$

$\hat{\beta}_j = y_{.j} - y_{..}$ $\quad j=1,\ldots,n$

$\hat{\theta}^2 = \lambda_1$ o maior valor singular de $\mathbf{R}^t\mathbf{R}$

$\hat{\mathbf{u}} = [u_i]$ o autovetor normalizado de $\mathbf{RR}^t$ correspondente ao autovalor λ_1^2

$\hat{\mathbf{v}} = [v_j]$ o autovetor normalizado de $\mathbf{R}^t\mathbf{R}$ correspondente ao autovalor λ_1^2

$\hat{\sigma}^2 = (\lambda_2^2+\ldots+\lambda_{n-1}^2)/mn$

PROVA:

Seja $\tau^t=(\mu,\alpha^t,\beta^t,\theta,\mathbf{u}^t,\mathbf{v}^t,\sigma^2)$. Após particionar a soma

de quadrados no expoente da maneira usual, a função de verossimilhança de τ é dada por:

$$L(\tau)=\left[\frac{1}{2\pi\sigma^2}\right]^{mn/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_i\sum_j r_{ij}^2+ mn\,(\mu - y_{..})^2 + n\sum_i(\alpha_i - y_{i.} + y_{..})^2 + m\sum_j(\beta_j - y_{.j} + y_{..})^2 + (\theta - \sum_i\sum_j u_i v_j r_{ij})^2 - (\sum_i\sum_j u_i v_j r_{ij})^2\right]\right\}$$

Pode-se mostrar que:

$$L(\tau)\leq\left[\frac{1}{2\pi\sigma^2}\right]^{mn/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_i\sum_j r_{ij}^2-(\sum_i\sum_j u_i v_j r_{ij})^2\right]\right\}$$

No lema 6.3 considere

$$x+1=\left[\sum_i\sum_j r_{ij}^2 - (\sum_i\sum_j u_i v_j r_{ij})^2\right]/\, mn\sigma^2$$

$$y = 1/[2\pi\ (x+1)\sigma^2] > 0$$

$$a = mn/2$$

Então, segue-se que $L(\tau)\leq(ye^{-1})^a$. Pelo lema 3.2 tem-se que

$$(\sum_i\sum_j u_i v_j r_{ij})^2\leq \lambda_1^2 \text{ , e consequentemente}$$

$$L(\tau) \leq \left[\frac{mn}{2\pi\ (\sum_{ij} r_{ij}^2-\lambda_1^2)}\right]^{mn/2} e^{-mn/2}$$

Para mostrar que este valor máximo é atingido para os estimadores de máxima verossimilhança dados neste teorema, basta substituir os parâmetros pelos estimadores na equação de verossimilhança. ■

3.7 TESTE DA RAZÃO DE VEROSSIMILHANÇA PARA $H_0:\theta=0$ vs $H_a:\theta\neq 0$

Apresentaremos a seguir o teste da razão de verossimilhança apresentado por Johnson e Graybill(1972), para a hipótese de aditividade ($H_0:\theta=0$), no modelo (3.5.1).

Sob o modelo aditivo, $y_{ij}= \mu + \alpha_i + \beta_j + e_{ij}$, o máximo da função de verossimilhança é dado por:

$$\sup L(\mu,\alpha,\beta) = \left[\frac{mn}{2\pi \sum_{i}^{m}\sum_{j}^{n} r_{ij}^2} \right]^{mn/2} e^{-mn/2}$$

A estatística do teste da razão de verossimilhança é dada por:

$$\Lambda^* = \left[\frac{\sum_{i}^{m}\sum_{j}^{n} r_{ij}^2 - \lambda_1^2}{\sum_{i}^{m}\sum_{j}^{n} r_{ij}^2} \right]^{mn/2} < K^*$$

onde $P_{H_0}[\Lambda^*< K^*]=\alpha$, α fixado. Para uma probabilidade de erro tipo I igual a α, a hipótese $H_0:\theta=0$ é rejeitada em favor de H_a se $\Lambda^*<K^*$.

Este teste é equivalente a rejeitar H_0 se:

$$\Lambda = \frac{\lambda_1^2}{\lambda_1^2+ \lambda_2^2+ \ldots + \lambda_{n-1}^2} > K \quad , \text{ onde}$$

$$P_{H_0}\left\{ \frac{\lambda_1^2}{\lambda_1^2+ \lambda_2^2+ \ldots + \lambda_{n-1}^2} > K \right\} = \alpha, \text{ para } \alpha \text{ fixado.}$$

Uma vantagem de se empregar o teste da razão de

verossimilhança é que, em muitos casos, pelo menos a distribuição assintótica de $-2\log\Lambda$ é conhecida. Neste caso, Johnson e Graybill conseguiram estudar a distribuição de Λ.

A distribuição conjunta de $\lambda_1^2, \lambda_2^2, \ldots, \lambda_{n-1}^2$ tem a distribuição das raízes não-nulas da matriz **W**, distribuída como uma Wishart $W_{n-1}(m-1, \sigma^2\mathbf{I}, \mathbf{K}_n^t\Gamma^t\Gamma\,\mathbf{K}_n)$, onde

$\mathbf{K}_n\mathbf{K}_n^t = \mathbf{I} - \frac{1}{n}\mathbf{J}$ (nota-se que $\mathbf{R}=\mathbf{K}_n\mathbf{K}_n^t\mathbf{Y}\mathbf{K}_m\mathbf{K}_m^t$) e Γ é uma matriz genérica representando a interação (isto é, com $\mathbf{1}^t\Gamma=\mathbf{0}^t$ e $\Gamma\mathbf{1}=\mathbf{0}$)

A distribuição de Λ, para $\theta=0$, pode ser estudada com o resultado acima, de maneira direta, através de transformações.

3.6.1 DISTRIBUIÇÃO DE Λ QUANDO $\theta=0$

Vamos encontrar a distribuição da estatística da razão de verossimilhança $\Lambda=\lambda_1^2/(\lambda_1^2 + \lambda_2^2 + \ldots + \lambda_{n-1}^2)$, sob a hipótese de inexistência de interação.

Para simplificar a notação faremos $p=n-1$, $q=m-1$, com $p\leq q$, e $l_i=\lambda_i^2$, de forma que os autovalores de **W** são $l_1>l_2>\ldots>l_p$ e **W** é distribuido como $W_p(q,\sigma^2\mathbf{I},0)$.

A função densidade de probabilidade conjunta dos autovalores é dada por:

$$f(l_1,l_2,\ldots,l_p) = C\prod_{i=1}^{p} l_i^{(q-p-1)/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i}^{m} l_i\right] \prod_{i<j}(l_i - l_j)$$

$$\text{para } 0<l_p<l_{p-1}<\ldots<l_2<l_1<\infty$$

onde

$$C = \frac{\Pi^{p/2}}{(2\sigma^2)^{pq/2} \prod_{i=1}^{p} \left[\Gamma(\frac{q-i+1}{2}) \quad \Gamma(\frac{p-i+1}{2}) \right]}$$

Seja agora,

$$U_i = \frac{l_i}{l_1 + l_2 + \ldots + l_p} \qquad i=1,\ldots,p-1$$

$$U_p = l_1 + l_2 + \ldots + l_p$$

Portanto,

$$l_i = U_i U_p \quad \text{e} \quad l_p = \frac{1 - U_1 - \ldots - U_{p-1}}{U_p}$$

O Jacobiano da transformação é U_p^{p-1}, como é fácil de verificar.

Os u's variam em $a_i < u_i < b_i$ $\quad i=2,3,\ldots,p-1$ e

$$\frac{1}{p} < u_1 < 1, \quad 0 < u_p < \infty \qquad (3.6.1.1)$$

onde

$$a_i = \frac{1-u_1 - \ldots - u_{i-1}}{p + 1 - i} \qquad i=2,3,\ldots,p-1$$

$$b_i = \min (u_{i-1}, 1-u_1 - u_2 - \ldots - u_{i-1}) \qquad i=2,3,\ldots,p-1$$

Portanto, a densidade conjunta de $U_1, U_2, \ldots, U_p$ é dada por $g(u_1,u_2,\ldots,u_p) = \Pi^{p/2} \left[(1-u_1 - \ldots - u_{p-1}) \prod_{i=1}^{p-1} u_i \right]^{(q-p-1)/2} \cdot$

$$\cdot \frac{\prod_{i<j}^{p-1} (u_i - u_j) \prod_{i=1}^{p-1} (u_i - 1 + u_1 + \ldots + u_{p-1})}{\prod_{i=1}^{p} \left[\Gamma(\frac{q-i+1}{2}) \quad \Gamma(\frac{p-i+1}{2}) \right]} \cdot$$

$$\cdot \frac{1}{(2\sigma^2)^{pq/2}} \; u_p^{(pq/2)-1} \exp(-u_p / 2\sigma^2)$$

para os valores de u_i dados em (3.6.1.1).

Vemos que U_p é independente de U_i i=1,2,...,p-1 e que, portanto, a densidade conjunta de U_i i=1,2,...,p-1 é dada por:

$$h(u_1,\ldots,u_{p-1}) = C^* \left[u_1 u_2 \ldots u_{p-1}(1-u_1- \ldots -u_{p-1}) \right]^{(q-p-1)/2}.$$

$$. \prod_{i<j}^{p-1} (u_i - u_j) \prod_{i=1}^{p-1} (u_i - 1 + u_1 + \ldots + u_{p-1})$$

para os valores de u_i dados em (3.6.1.1) onde

$$C^* = \frac{\Pi^{p/2} \quad \Gamma(pq/2)}{\prod_{i=1}^{p} \left[\Gamma(\frac{q-i+1}{2}) \quad \Gamma(\frac{p-i+1}{2}) \right]}$$

A densidade de U_1 pode ser obtida, então, por integração:

$$g_1(u_1) = \int_{a_2}^{b_2} \int_{a_3}^{b_3} \ldots \int_{a_{p-1}}^{b_{p-1}} h(u_1,\ldots,u_{p-1}) \, du_{p-1} \ldots du_3 \, du_2$$

para $\frac{1}{p} < u_1 < 1$.

O cálculo desta integral é trabalhoso quando p não é pequeno. Uma aproximação da distribuição de U_1 pode ser dada pela distribuição beta como veremos na próxima seção.

3.6.2 APROXIMAÇÃO PARA A DISTRIBUIÇÃO DE U_1

Esta aproximação depende dos momentos de λ_1^2. Esses momentos não são conhecidos explicitamente, mas Mandel(1969) calculou a média e a variância de λ_1^2/σ^2 por Monte Carlo, e nesta seção denotaremos por ν_1 e ν_2, respectivamente.

A aproximação da distribuição de U_1 pode ser feita

utilizando-se a seguinte transformação:

$$V = \frac{p\,U_1 - 1}{p - 1} \quad \text{(para levar } U_1 \text{ de (1/p,1) a (0,1))}.$$

Aproximamos V por uma beta B(k,r).

Se μ_1=E(V) e μ_2=Var(V) são conhecidos, podemos aproximar os momentos da beta por:

$$\mu_1 = \frac{k}{k + r} \qquad \mu_2 = \frac{kr}{(k+r)^2\,(k+r+1)^2}$$

Resolvendo para k e r resulta

$$k = \frac{\mu_1\,(\mu_1-\mu_2-\mu_1^2)}{\mu_2} \qquad \text{e} \qquad r = \frac{(1-\mu_1)\,(\mu_1-\mu_2-\mu_1^2)}{\mu_2}$$

Temos ainda:

$$E(\lambda_1^2/\sigma^2) = E[(U_p/\sigma^2)\,U_1] = E(U_p/\sigma^2)E(U_1) = pq\,E(U_1)$$

donde segue que:

$$E(U_1) = \frac{\nu_1}{pq}\ , \text{ onde } \nu_1 = E(\lambda_1^2/\sigma^2)$$

Mais ainda:

$$E(\lambda_1^2/\sigma^4) = E\,[(U_p^2/\sigma^4)\,U_1^2] = E(U_p^2/\sigma^4)\,E(U_1^2) =$$

$$= (p^2q^2 + 2pq)\,E(U_1^2)$$

então

$$Var(U_1) = \frac{\nu_2 + \nu_1^2}{p^2q^2 + 2pq} - \frac{\nu_1^2}{p^2q^2}$$

donde

$$\mu_2 = Var(V) = Var[(pU_1-1)/(p-1)] = [p^2/(p-1)^2]/Var(U_1)$$

$$= \frac{pq\nu_2 - 2\nu_1^2}{q^2(pq+2)\,(p-1)^2}\ , \text{ onde } \nu_2 = Var(\lambda_1^2/\sigma^2).$$

Podemos usar os resultados obtidos por Mandel(1969)

para ν_1 e ν_2 e calcular valores críticos para Λ.

Johnson e Graybill(1972) fizeram este trabalho, que resultou na tabela que está no apêndice B (Tabela B4).

Vemos, assim, que a estatística recomendada para testar a hipótese de aditividade no caso particular do modelo de Mandel, com só um termo multiplicativo como alternativa, é

$$\frac{\lambda_1^2}{\lambda_1^2 + \lambda_2^2 + \ldots + \lambda_{n-1}^2}$$

É sugestivo o fato de que este resultado, como outros pertinentes à estimação, sejam iguais quando se emprega máxima-verossimilhança sob erros normais ou mínimos quadrados com as condições de Gauss-Markov.

3.7 UM EXEMPLO DO AJUSTE DE MODELOS MULTIPLICATIVOS

Com a finalidade de tornar mais clara a utilização da técnica apresentada neste capítulo, faremos uma aplicação da mesma a partir de dados obtidos em John e Quenouille(1977). Trata-se de um experimento onde avaliou-se o peso total, em gramas, de pés de alface, que foram cultivados para a combinação de 12 tempos diferentes de semeadura e 3 tratamentos de água. Os dados[1] estão listados na tabela 3.2.

Vamos supor que o objetivo do experimento seja avaliar os efeitos dos fatores **tempo de semeadura** e **tratamento**

---

[1] Os dados originais estão apresentados em Woodman,R.M. e Johnson,D.A.(1974). The effect of time of sowing and water supply on the bolting and growth of lettuce. J.Agric.Sci., 37, 95-112.

de água, bem como possível interação entre eles.

Como mencionado anteriormente, os métodos usuais não se prestam à determinação da presença de interação e não obtêm uma estimativa para a variância do erro aleatório.

Tabela 3.2 Peso Total de Pés de Alface (gramas)

TEMPO DE SEMEADURA	TRATAMENTO DE ÁGUA 1	2	3
1	183	269	207
2	166	364	214
3	115	418	270
4	116	278	213
5	202	224	168
6	1037	1331	1158
7	953	1145	932
8	912	1375	735
9	1079	1706	909
10	735	1405	616
11	847	1775	774
12	642	1182	467

Vamos fazer uso da suposição proposta por Mandel, a qual nos permite estudar a interação, ou seja, vamos analisar os resíduos do modelo após o ajuste dos termos aditivos, e verificar se estes contêm algum componente sistemático. Para tanto, consideremos inicialmente o ajuste do modelo aditivo, ou seja

$$PESO_{ij} = TEMP\_SEM_i + TRAT\_AG_j + e_{ij} \qquad (3.7.1)$$

$$i=1,\ldots,12$$

$$j=1,\ldots,3$$

onde cada termo do modelo representa:

$PESO_{ij}$ peso dos pés de alface associados ao i-ésimo tempo de semeadura e j-ésimo tratamento de água

$TEMP\_SEM_i$ efeito do i-ésimo tempo de semeadura

$TRAT\_AG_j$ efeito do j-ésimo tratamento de água

e_{ij} erro aleatório não mensurável, com média zero e variância σ^2

A análise de variância referente a este ajuste está apresentada na tabela 3.3. A soma de quadrados devida aos resíduos equivale a interação TEMP_SEM*TRAT_AG. Os resultados obtidos para os componentes desse modelo só são válidos se a suposição de inexistência de interação for verdadeira.

Tabela 3.3 Análise de Variância do Modelo Aditivo

FONTE DE VARIAÇÃO	GRAUS DE LIBERDADE	SOMA DE QUADRADOS	QUADRADO MÉDIO
TEMP_SEM	11	6176577,67	561507,06
TRAT_AG	2	1204074,50	602037,25
RESíDUO	22	786584,83	35753,86
TOTAL (corrigido)	35	8167237,00	

Vamos representar por r_{ij} os resíduos do modelo (3.7.1), e prosseguir a análise considerando agora o ajuste dos componentes multiplicativos, ou seja:

$$r_{ij} = \theta\, u_i v_j + \theta' u'_i v'_j + \ldots + e^*_{ij} \qquad (3.7.2)$$

A análise de variância decorrente do modelo (3.7.2) está apresentada na tabela 3.4. A quantidade M_1, que assume o

papel dos graus de liberdade, para m=12 e n=3, não consta da tabela B1 (apêndice B), a qual foi estabelecida por Mandel. Jonhson e Graybill(1972) apresentam os valores exatos de M_1, para n=3 e vários valores de m, os quais estão apresentados na tabela B5. Os graus de liberdade do resíduo foram obtidos por diferença.

A análise das razões na tabela 3.4, indica que a interação TEMP_SEM*TRAT_AG não é desprezível, e que esta pode ser explicada por um termo multiplicativo. Desta forma, os outros dois componentes representam apenas erro aleaório, e podem ser utilizados para estimar a variância σ^2. Da tabela 3.4 tem-se que a estimativa de σ^2 é $\hat{\sigma}^2$=5441,72.

Tabela 3.4 Análise de Variância do Ajuste do Componente Multiplicativo

FONTE DE VARIAÇÃO	GRAUS DE LIBERDADE	SOMA DE QUADRADOS	RAZÃO
θ	15,06	748819,28	49722,40
RESÍDUO	6,94	37765,55	5441,72

A hipótese de aditividade pode também ser avaliada pelo teste da razão de máxima verossimilhança. A estatística do teste é dada por: $U_1 = \lambda_1^2/(\lambda_1^2+\lambda_2^2)$ = 748819,28 / (748819,28+37765,55) = 0,9520. Da tabela B4, o ponto crítico ao nível de significância de 1% é 0,8037. Como U_1 é maior que o ponto crítico, a conclusão é a mesma obtida com a análise feita pelo método de Mandel, ou seja, a hipótese de aditividade não deve ser aceita.

O estimador de máxima verossimilhança de σ^2 é $\hat{\sigma}^2=\lambda_2^2/mn$. Neste exemplo, $\hat{\sigma}^2=37765,55/36 = 1049,04$. Observa-se que esta estimativa é menor que a obtida pelo ajuste de Mandel.

A partir das tabelas 3.3 e 3.4 podemos construir a tabela 3.5. Não podemos comparar as razões de variância obtidas, com valores tabelados, pois não se tem uma distribuição definida para essas razões. Mas, intuitivamente, essas indicam que ambos os fatores contribuem para explicar o peso total dos pés de alface.

Tabela 3.5 Análise de Variância Resultante do Ajuste dos Componentes Aditivos e Multiplicativo

FONTE DE VARIAÇÃO	GRAUS DE LIBERDADE	SOMA DE QUADRADOS	QUADRADO MÉDIO	RAZÃO DAS VARIÂNCIAS
TEMP_SEM	11	6176577,67	561507,06	103,19
TRAT_AG	2	1204074,50	602037,25	110,63
θ	15,06	748819,28	49722,40	
RESÍDUO	6,94	37765,55	5441,72	

Para facilidade de interpretação do modelo, em geral, calculam-se as estimativas padronizadas dos parâmetros. Cada estimativa é dividida pela raiz quadrada da correspondente soma de quadrados. Considerando:

$$S = \sqrt{\sum_{i=1}^{12} \widehat{TEMP\_SEM}_i} \qquad A = \sqrt{\sum_{j=1}^{3} \widehat{TRAT\_AG}_j}$$

$$s_i = \frac{1}{S}\,\widehat{TEMP\_SEM}_i \qquad a_j = \frac{1}{A}\,\widehat{TRAT\_AG}_j$$

o modelo que emerge dessa padronização é dado pela seguinte equação:

$$PESO_{ij} = \mu + Ss_i + Aa_j + \theta\, u_i v_j^t + e_{ij}$$

substituindo os valores de S,A e θ temos:

$$PESO_{ij} = 697,83 + 1435,61s_i + 316,76a_j + 865,34\, u_i v_j^t + e_{ij}$$

A partir dos parâmetros padronizados, apresentados na tabela 3.6, pode-se notar que, existem subgrupos de tempo de semeadura, o que significa que este fator pode ser representado por um número menor de níveis. Observa-se que os tempos de 1 a 5, agem de forma semelhante no peso total dos pés de alface, o mesmo ocorrendo com os tempos 6,9 e 11, e também 7 e 8. Com relação ao fator tratamento de água, observa-se que os tratamentos 1 e 3 são semelhantes.

Tabela 3.6 Parâmetros Estruturais e Parâmetros Padronizados

TEMPO DE SEMEADURA			**TRATAMENTO DE ÁGUA**		
i	$\hat{u}_i$	s_i	j	$\hat{v}_j$	a_j
1	0,2971	-0,3331	1	0,2305	-0,3649
2	0,2100	-0,3133	2	-0,7936	0,8150
3	0,1833	-0,2996	3	0,5631	-0,4501
4	0,2749	-0,3451			
5	0,3180	-0,3482			
6	0,1694	0,3326			
7	0,1706	0,2174			
8	-0,1795	0,2156			
9	-0,3253	0,3716			
10	-0,3316	0,1538			
11	-0,5382	0,3024			
12	-0,2488	0,0459			

4. O Biplot

4. O BIPLOT

Uma avaliação visual de dados multivariados seria muito simples se envolvesse apenas duas variáveis; cada observação bivariada seria representada como um ponto num espaço bidimensional definido pelas duas variáveis. Na prática são frequentes situações em que o número de variáveis investigadas é superior a dois. Em geral, os dados são representados por uma matriz cujas linhas referem-se as unidades amostrais e as colunas representam as variáveis. Várias técnicas gráficas foram desenvolvidas visando obter uma análise exploratória de dados multivariados, possibilitando melhor compreensão da estrutura dos dados. Gabriel(1971,1980,1981), descreve um procedimento, denominado de Biplot, que permite uma avaliação visual da estrutura de grandes matrizes de dados. Graficamente pode-se exibir a relação entre unidades amostrais, quando indicadas por certas medidas de distâncias, e entre variáveis, quando indicadas pelas suas covariâncias e correlações. O biplot também mostra as variâncias das variáveis e permite ao investigador visualizar as observações individuais e suas diferenças. Muitas variações do Biplot estão presentes na literatura, mas não iremos abordá-las neste trabalho.

4.1 O BIPLOT PARA MATRIZES DE POSTO 2

Esta técnica é baseada no resultado que qualquer matriz **Y** mxn, de posto r, pode ser fatorada como:

$$\mathbf{Y} = \mathbf{G}\mathbf{H}^t \qquad (4.1.1)$$

onde **G** é uma matriz mxr e **H** é uma matriz nxr, ambas necessariamente de posto r. Essa fatoração não é única. A transformação pode ser reescrita como:

$$y_{ij} = \mathbf{g}_i^t \mathbf{h}_j \qquad (4.1.2)$$

para cada i e j, onde y_{ij} é o elemento da i-ésima linha e j-ésima coluna da matriz **Y**, $\mathbf{g}_i^t$ é a i-ésima linha de **G** e $\mathbf{h}_j$ é a j-ésima linha de **H**. Desta forma, a fatoração acima designa um dos vetores $\mathbf{g}_1,\ldots,\mathbf{g}_m$ para cada linha de **Y**, e um dos vetores $\mathbf{h}_1,\ldots,\mathbf{h}_n$, para cada coluna de **Y**. A dimensão de cada um desses vetores é r. A equação (4.1.2) dá uma representação da matriz baseada em m+n vetores, num espaço r-dimensional.

Para uma matriz de posto 1, os vetores se reduzem a escalares $g_1,\ldots,g_m$, associados às linhas, e $h_1,\ldots,h_n$, associados às colunas, e y_{ij} é simplesmente o produto escalar $g_i h_j$. Neste caso, é dito que **Y** tem uma estrutura multiplicativa.

No caso de matrizes com posto 2, os vetores $\mathbf{g}_1,\ldots,\mathbf{g}_m$ e $\mathbf{h}_1,\ldots,\mathbf{h}_n$ são vetores de ordem 2. Consequentemente, esses vetores podem ser desenhados num plano, dando a representação de mn elementos de **Y**, por meio do produto escalar dos correspondentes vetores linhas e colunas. Tal gráfico foi referido por Gabriel como biplot, uma vez que permite representar graficamente linhas e colunas de uma matriz,

conjuntamente.

O biplot representa uma matriz de posto 2 exatamente. O produto escalar de dois vetores pode ser avaliado visualmente por considerá-lo como o produto do comprimento de um dos vetores pelo comprimento da projeção do outro vetor sobre ele. Isto possibilita a visualização de quais linhas ou colunas são proporcionais (mesma direção), quais elementos são nulos (ortogonalidade de linhas e colunas).

Como mencionado anteriormente, a fatoração (4.1.1) não é única. A disparidade entre fatorações diferentes de $\mathbf{Y}$, e consequentemente entre biplots resultantes, depende quase que inteiramente da particular fatoração escolhida. Entretanto, para que o biplot seja uma ferramenta útil para a inspeção de relações entre linhas e/ou colunas da matriz $\mathbf{Y}$, é necessário impor uma métrica para que a fatoração resultante e o biplot sejam únicos.

Vamos supor que se deseje representar relações entre linhas de $\mathbf{Y}$ pelas correspondentes relações dos vetores $\mathbf{g}_1,\ldots,\mathbf{g}_m$. Então, a seguinte condição é exigida:

$$\mathbf{H}^t\mathbf{H} = \mathbf{I}_2 \qquad (4.1.3)$$

e em consequência tem-se:

$$\mathbf{Y}\mathbf{Y}^t = \mathbf{G}\mathbf{G}^t$$

Da condição (4.1.3) decorre

$$\mathbf{Y}^t(\mathbf{Y}\mathbf{Y}^t)^-\mathbf{Y} = \mathbf{H}\mathbf{H}^t$$

para qualquer inversa generalizada $(\mathbf{Y}\mathbf{Y}^t)^-$ de $\mathbf{Y}\mathbf{Y}^t$, e esta é a matriz de projeção no espaço linha de $\mathbf{Y}$. Os produtos escalares dos vetores $\mathbf{h}_1,\ldots,\mathbf{h}_n$ são portanto aqueles das colunas de $\mathbf{Y}$,

tomados através de qualquer métrica $(YY^t)^-$, isto é:

$$n_j^t(YY^t)^-n_g = h_j^t h_g$$

onde n_j e n_g indicam a j-ésima e g-ésima colunas de **Y**.

Por outro lado, para que os produtos internos das colunas de **Y** sejam representados pelos vetores $h_1,\ldots,h_n$, sem que o mesmo ocorra com as linhas de **Y** e vetores $g_1,\ldots,g_m$, a condição a ser imposta é:

$$G^tG = I_2 \qquad (4.1.4)$$

e obtém-se

$$Y(Y^tY)^-Y^t = GG^t$$

onde $(Y^tY)^-$ é qualquer inversa generalizada de Y^tY, e

$$Y^tY = HH^t$$

De um modo geral, se uma métrica **M** é usada para as linhas, isto é:

$$YMY^t = GG^t$$

H pode ser escolhida de forma a satisfazer a condição:

$$H^tMH = I_2 \qquad (4.1.5)$$

e qualquer inversa generalizada $(YMY^t)^-$ pode servir como métrica para as colunas, resultando em:

$$Y^t(YMY^t)^-Y = HH^t \qquad (4.1.6)$$

A prova da expressão (4.1.6) se dá a partir das equações (4.1.1) e (4.1.5) e do uso do resultado $G^t(GG^t)^-G=I_2$, pois $G^t(GG^t)^-G$ é a matriz projeção no espaço coluna de G^t, o qual é o espaço Euclideano $\mathbb{R}^2$, cuja matriz de projeção é I_2.

Analogamente, para qualquer métrica **N**, usada para as

colunas, G deve ser escolhida de forma que:

$$G^tNG = I_2$$

então

$$Y^tNY = HH^t$$

e

$$Y(Y^tNY)^-Y^t = GG^t$$

para qualquer inversa generalizada $(Y^tNY)^-$.

A partir de rotações e reflexões, operações estas que não mudam as relações entre os vetores, o biplot torna-se único pela introdução de uma particular métrica para comparação das linhas e colunas.

4.2 O BIPLOT PARA MATRIZES DE POSTO MAIOR QUE 2

Matrizes de posto maior que 2 não podem ser representadas exatamente por um biplot. Entretanto, se a matriz Y pode ser satisfatoriamente aproximada por uma matriz de posto 2, $Y_{(2)}$, o biplot de $Y_{(2)}$ pode resultar numa inspeção visual aproximada de Y. Nesta situação, os produtos internos das linhas e colunas de $Y_{(2)}$ serão aproximações para os elementos de Y.

A aproximação de uma matriz por matriz de posto menor, é uma consequência da decomposição em valores singulares, resultado já abordado no capítulo 2. Essa aproximação é no sentido de mínimos quadrados, e produz

$$Y_{(k)} = \sum_{i=1}^{k} \lambda_i \, u_i \, v_i^t$$

como sendo a aproximação de posto k da matriz Y, onde λ_i, u_i e

$\mathbf{v}_i$, i=1,...,r são os valores singulares e vetores singulares esquerdos e direitos, respectivamente, satisfazendo

$$\mathbf{u}_i^t\mathbf{u}_j = \mathbf{v}_i^t\mathbf{v}_j = \begin{cases} 0 \ , \ i\neq j \\ 1 \ , \ i=j \end{cases} \qquad (4.2.1)$$

A adequação da aproximação pode ser avaliada por:

$$\rho_k^{(2)} = 1 - \|\mathbf{Y}-\mathbf{Y}_{(k)}\|^2 \ / \ \|\mathbf{Y}\|^2 = \sum_{i=1}^{k} \lambda_i^2 \ / \ \sum_{i=1}^{r} \lambda_i^2 \qquad (4.2.2)$$

De particular importância na interpretação do critério de mínimos quadrados é o fato de que este é equivalente ao critério de mínimos quadrados sobre as diferenças entre todas as linhas, bem como entre todas as colunas. Portanto, $\mathbf{Y}_{(k)}$ é a matriz de posto k, cujas diferenças das linhas melhor aproximam as diferenças das linhas da matriz **Y**, e esta aproximação se dá com adequação $\rho_k^{(2)}$. O mesmo ocorre com relação as colunas.

O biplot aproximado de **Y** é então dado pelo biplot exato de $\mathbf{Y}_{(2)}$,

$$\mathbf{Y}_{(2)} = [\mathbf{u}_1,\mathbf{u}_2]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\begin{bmatrix} \mathbf{v}_1^t \\ \mathbf{v}_2^t \end{bmatrix} = \mathbf{U}_{(2)}\mathbf{D}_{\lambda(2)}\mathbf{V}_{(2)}^t \qquad (4.2.3)$$

A adequação da aproximação é medida por:

$$\rho_2^{(2)} = (\lambda_1^2+\lambda_2^2) \ / \ \sum_{i=1}^{r} \lambda_i^2$$

Se $\rho_2^{(2)}$ é próximo de 1, o biplot de $\mathbf{Y}_{(2)}$ dará uma boa aproximação para o biplot de **Y**.

Na escolha dos fatores **G** e **H** de $\mathbf{Y}_{(2)}$, para a construção do biplot, uma das possibilidades é utilizar a própria decomposição em valores singulares de $\mathbf{Y}_{(2)}$, conforme (4.2.3). Podemos considerar:

$$\mathbf{u}_i^t = (u_{i1},\ldots,u_{im}) \quad \text{e} \quad \mathbf{v}_i^t = (v_{i1},\ldots,v_{in})$$

e obtemos:

$$\mathbf{Y}_{(2)} = \begin{bmatrix} u_{11} & u_{21} \\ \vdots & \vdots \\ u_{1m} & u_{2m} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1n} \\ v_{21} & \cdots & v_{2n} \end{bmatrix}$$

Algumas possibilidades de escolha para **G** e **H** são:

$$\begin{aligned} \textbf{(a)}\ \mathbf{g}_i^t &= (\sqrt{\lambda_1}u_{1i}, \sqrt{\lambda_2}u_{2i}) && i=1,\ldots,m \\ \mathbf{h}_j^t &= (\sqrt{\lambda_1}v_{1j}, \sqrt{\lambda_2}v_{2j}) && j=1,\ldots,n \end{aligned} \qquad (4.2.4)$$

$$\begin{aligned} \textbf{(b)}\ \mathbf{g}_i^t &= (u_{1i}, u_{2i}) && i=1,\ldots,m \\ \mathbf{h}_j^t &= (\lambda_1 v_{1j}, \lambda_2 v_{2j}) && j=1,\ldots,n \end{aligned} \qquad (4.2.5)$$

a qual satisfaz (4.1.4).

$$\begin{aligned} \textbf{(c)}\ \mathbf{g}_i^t &= (\lambda_1 u_{1i}, \lambda_2 u_{2i}) && i=1,\ldots,m \\ \mathbf{h}_j^t &= (v_{1j}, v_{2j}) && j=1,\ldots,n \end{aligned} \qquad (4.2.6)$$

a qual satisfaz (4.1.3).

4.3 O BIPLOT PARA A MATRIZ DE VARIÂNCIA E COVARIÂNCIA E DISTÂNCIAS PADRONIZADAS ENTRE UNIDADES AMOSTRAIS

Vamos considerar Y mxn, matriz das observações, cuja média das variáveis tenha sido subtraida de cada observação. Então a matriz de variância e covariância estimada, correspondente às n variáveis é dada por:

$$\mathbf{S} = (1/m)\mathbf{Y}^t\mathbf{Y} \qquad (4.3.1)$$

Uma medida padronizada de distância, denominada de distância de Mahalanobis, entre a i-ésima e e-ésima unidade amostral é dada por:

$$d_{i,e} = (\mathbf{y}_i - \mathbf{y}_e)^t \mathbf{S}^{-1} (\mathbf{y}_i - \mathbf{y}_e) \qquad (4.3.2)$$

Como visto anteriormente, decorrente da decomposição em valores singulares, a matriz Y pode ser fatorada como:

$$\mathbf{Y} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)(\lambda_1 \mathbf{v}_1, \ldots, \lambda_r \mathbf{v}_r)^t \qquad (4.3.3)$$

Esta fatoração, em vista de (4.2.1), tem as seguintes propriedades:

$$(\mathbf{u}_1, \ldots, \mathbf{u}_r)^t (\mathbf{u}_1, \ldots, \mathbf{u}_r) = \mathbf{I}_r \qquad (4.3.4)$$

$$(\mathbf{u}_1, \ldots, \mathbf{u}_r)(\mathbf{u}_1, \ldots, \mathbf{u}_r)^t = (1/m)\mathbf{Y}\mathbf{S}^{-1}\mathbf{Y}^t \qquad (4.3.5)$$

$$(\lambda_1 \mathbf{v}_1, \ldots, \lambda_r \mathbf{v}_r)^t (\lambda_1 \mathbf{v}_1, \ldots, \lambda_r \mathbf{v}_r) = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_r^2) \qquad (4.3.6)$$

$$(\lambda_1 \mathbf{v}_1, \ldots, \lambda_r \mathbf{v}_r)(\lambda_1 \mathbf{v}_1, \ldots, \lambda_r \mathbf{v}_r)^t = m\mathbf{S} \qquad (4.3.7)$$

Vamos considerar a aproximação de posto 2 de Y, $\mathbf{Y}_{(2)}$, conforme (4.2.3), e para propósito do biplot, escolher a seguinte fatoração:

$$\begin{aligned} \mathbf{G} &= (\mathbf{u}_1, \mathbf{u}_2)\sqrt{m} \\ \mathbf{H} &= (1/\sqrt{m})\,(\lambda_1 \mathbf{v}_1, \lambda_2 \mathbf{v}_2) \end{aligned} \qquad (4.3.8)$$

a qual, a menos de constante, consiste em considerar a condição (4.1.4). Das propriedades (4.3.3) até (4.3.7) decorrem as seguintes aproximações, com respeito à aproximação de posto 2:

$$\mathbf{Y} \sim \mathbf{G}\mathbf{H}^t \qquad (4.3.9)$$

$$\mathbf{Y}\mathbf{S}^{-1}\mathbf{Y}^t \sim \mathbf{G}\mathbf{G}^t \qquad (4.3.10)$$

$$\mathbf{S} \sim \mathbf{H}\mathbf{H}^t \qquad (4.3.11)$$

Qualquer biplot de Y, ou biplot exato de $Y_{(2)}$, permite aproximações, todas decorrentes de (4.3.9), para:

(a) as observações individuais

$$y_{ij} \sim \mathbf{g}_i^t \mathbf{h}_j$$

(b) a diferença entre a i-ésima e e-ésima unidades amostrais, com respeito a variável j

$$y_{ij} - y_{ej} \sim (\mathbf{g}_i - \mathbf{g}_e)^t \mathbf{h}_j$$

(c) a diferença entre as variáveis j e g, com relação a i-ésima unidade

$$y_{ij} - y_{ig} \sim \mathbf{g}_i^t (\mathbf{h}_j - \mathbf{h}_g)$$

(d) a interação das unidades i e e com as variáveis j e g

$$y_{ij} - y_{ej} - y_{ig} + y_{eg} \sim (\mathbf{g}_i - \mathbf{g}_e)^t (\mathbf{h}_j - \mathbf{h}_g)$$

O biplot de Y com a particular escolha da fatoração (4.3.8), permite aproximações adicionais. De (4.3.10) a distância padronizada apresentada em (4.3.2) pode ser aproximada por:

$$d_{i,e} \sim \|\mathbf{g}_i - \mathbf{g}_e\| \qquad (4.3.12)$$

De (4.3.11), aproximações das covariâncias, variâncias e correlações das n variáveis são dadas por:

$$s_{j,g} \sim \mathbf{h}_j^t \mathbf{h}_g \qquad (4.3.13)$$

$$s_j \sim \|\mathbf{h}_j\|^2 \qquad (4.3.14)$$

$$r_{j,g} \sim \cos(\mathbf{h}_j, \mathbf{h}_g) \qquad (4.3.15)$$

onde $s_{j,g}$ é o elemento da j-ésima linha e g-ésima coluna da

matriz S, e $r_{j,g} = s_{j,g}/\sqrt{s_{j,j}s_{g,g}}$. A expressão:

$$\frac{1}{m}\sum_{i=1}^{m}(y_{ij}-y_{ig})^2 \sim \|h_j-h_g\|^2$$

dá uma aproximação para a diferença quadrática média entre variáveis.

Em resumo, o biplot permite aproximação para os elementos da matriz Y, para diferenças padronizadas entre unidades amostrais e para variâncias, covariâncias e correlações das variáveis. O biplot é portanto uma ferramenta gráfica muito útil para interpretar matrizes de observações multivariadas, desde que essas possam ser adequadamente aproximadas por matrizes de posto 2.

Como visto anteriormente, a adequação da aproximação para os elementos de Y é medida por:

$$\rho_2^{(2)} = (\lambda_1^2+\lambda_2^2) \;/\; \sum_{i=1}^{r}\lambda_i^2$$

Para os elementos de S, as aproximações conforme (4.3.13) e (4.3.14), tem adequação ainda melhor:

$$\rho_2^{(4)} = (\lambda_1^4+\lambda_2^4) \;/\; \sum_{i=1}^{r}\lambda_i^4$$

Por outro lado, para as distâncias padronizadas $d_{i,e}$, tem-se:

$$\rho_2^{(0)} = (\lambda_1^0+\lambda_2^0) \;/\; \sum_{i=1}^{r}\lambda_i^0 = 2/r$$

Pode-se verificar que, enquanto os elementos das matrizes Y e S podem ser excelentemente representados num biplot, as distâncias padronizadas não são adequadamente aproximadas.

Um biplot alternativo pode ser obtido pela escolha:

$$\mathbf{G} = (\lambda_1 \mathbf{u}_1, \lambda_2 \mathbf{u}_2)$$

$$\mathbf{H} = (\mathbf{v}_1, \mathbf{v}_2)$$

Esta fatoração seria apropriada se considerássemos a distância entre indivíduos como sendo a Euclideana ao invés da distância de Mahalanobis. A desvantagem dessa escolha é que as aproximações para as variâncias e covariâncias como as apresentadas em (4.3.13) e (4.3.15) não são válidas.

4.4 UM EXEMPLO DO BIPLOT

Com o intuito de ilustrar a técnica Biplot, vamos utilizar um conjunto de dados obtido junto ao Hospital Universitário Regional do Norte do Paraná - Centro de Ciências da Saúde, Universidade Estadual de Londrina. Trata-se de um estudo realizado em 6 cidades do estado do Paraná , a saber: Cascavel, Curitiba, Londrina, Maringá, Paranavaí e Ponta Grossa, envolvendo 427 pacientes com insuficiência renal crônica, submetidos à diálise peritonial e/ou hemodiálise. Uma amostra de sangue de cada paciente foi colhida para a realização de marcadores do vírus das hepatites A,B e C. São estes:

Anti-VHA anticorpo contra o antígeno da hepatite A

AgHBs antígeno de superfície da hepatite B

Anti-HBc anticorpo contra o antígeno central da hepatite B

Anti-HBs anticorpo contra o antígeno de superfície da hepatite B

Anti-VHC anticorpo contra o antígeno da hepatite C

Na tabela 4.1, apresentamos a porcentagem de positividade de cada marcador, em relação ao total de pacientes pesquisados em cada cidade.

Vamos, através do Biplot, tentar entender a estrutura da matriz de dados apresentada na tabela 4.1, e consequentemente,tentar caracterizar as cidades.

Tabela 4.1 Porcentagem de Positividade dos Marcadores em Cada Cidade

	MARCADOR				
CIDADE	Anti-VHA	AgHBs	Anti-HBc	Anti-HBs	Anti-VHC
Cascavel	98,00	4,00	56,00	52,00	4,00
Curitiba	100,00	18,75	75,00	46,87	43,75
Londrina	98,66	7,58	44,19	35,71	17,85
Maringá	100,00	6,66	51,11	40,00	13,33
Paranavaí	100,00	7,69	46,15	30,76	7,69
P. Grossa	100,00	22,22	80,95	50,79	17,46

A porcentagem média em cada cidade indica o padrão de infecção da referida cidade, enquanto que a porcentagem média para cada marcador indica a prevalência global do marcador. No modelo multiplicativo tais médias são ajustadas, por mínimos quadrados, pela primeira componente singular. Para estudar as prevalências diferenciais dos marcadores nas diferentes

cidades, esta primeira componente singular foi subtraida, e buscamos a aproximação de posto 2 para a matriz $\mathbf{Y}-\mathbf{Y}_{(1)}$, ou seja:

$$\mathbf{Y}-\mathbf{Y}_{(1)}=\sum_{i=2}^{4}\lambda_i^2\mathbf{u}_i\mathbf{v}_i^t$$

que representa a decomposição em valores singulares da matriz residual, cujos resíduos correspondem a interação depois de ajustar um modelo aditivo na análise de variância com dois fatores. Isto se deve ao fato de que o primeiro componente singular no modelo multiplicativo é equivalente ao ajuste do modelo aditivo.

Os valores de $\mathbf{Y}-\mathbf{Y}_{(1)}$ estão apresentados na tabela 4.2. Assim, por exemplo, os valores positivos para Anti-VHA associados às cidades de Londrina, Maringá e Paranavaí não apontam para uma maior prevalência de Anti-VHA nestas cidades. Estes valores são indicativos de que frente a baixa prevalência dos outros marcadores nestas cidades, o anticorpo contra o antígeno da hepatite A não é tão raro como os outros.

Tabela 4.2 Resíduos após Subtração da Primeira Componente Singular

	MARCADOR				
CIDADE	Anti-VHA	AgHBs	Anti-HBc	Anti-HBs	Anti-VHC
Cascavel	0,76	-7,40	-2,75	9,77	-13,64
Curitiba	-9,86	5,87	8,63	-0,84	23,82
Londrina	8,06	-3,04	-10,55	-3,64	1,41
Maringá	5,36	-4,44	-6,07	-1,10	-3,84
Paranavaí	10,31	-2,83	-8,04	-8,20	-8,58
P. Grossa	-10,44	9,27	14,22	2,82	-2,58

Os resultados da decomposição em valores singulares de $\mathbf{Y}-\mathbf{Y}_{(1)}$ estão expressos na tabela 4.3, onde se tem os dois primeiros valores singulares da matriz residual, denotados por λ_2 e λ_3, e os correspondentes autovetores $\mathbf{u}_2$,$\mathbf{u}_3$,$\mathbf{v}_2$ e $\mathbf{v}_3$. A adequação da aproximação de $\mathbf{Y}-\mathbf{Y}_{(1)}$ pelas segunda e terceira componentes singulares pode ser avaliada por:

$$(\lambda_2^2+\lambda_3^2) \;/\; \sum_{i=2}^{4} \lambda_i^2 = 0{,}9312$$

o que revela que a matriz de posto 2 está muito próxima da matriz em questão.

Tabela 4.3 Valores Singulares e Autovetores Decorrentes da Decomposição em Valores Singulares da Matriz Residual

CIDADE	$\mathbf{u}_2$	$\mathbf{u}_3$	**MARCADOR**	$\mathbf{v}_2$	$\mathbf{v}_3$
Cascavel	1,99	-2,45	Anti-VHA	2.98	1.65
Curitiba	-4,21	1,95	AgHBs	-2,08	-0,35
Londrina	1,55	2,11	Anti-HBc	-3,15	-2,14
Maringá	1,56	0,54	Anti-HBs	-0,31	-2,14
Paranavaí	2,56	1,01	Anti-VHC	-3,86	3,38
Ponta Grossa	-2,26	-2,80			
		λ_2=38,15		λ_3=23,41	

O biplot da figura 4.1 foi construido a partir dos valores singulares e autovetores obtidos, considerando-se a fatoração (4.2.4).

A inspeção da figura 4.1 sugere que as cidades de Londrina, Paranavaí e Maringá, constituem um agrupamento, e estão em posição oposta à cidade de Ponta Grossa. Esta por sua vez está ligeiramente ortogonal às cidades de Curitiba e Cascavel, que também estão em posição oposta entre si. Observa-se que existe uma alta prevalência de hepatite por vírus A associada às cidades de Londrina, Paranavaí e Maringá. O gráfico evidencia alta prevalência de hepatite por vírus C associado à cidade de Curitiba. Os marcadores AgHBs, Anti-HBc e Anti-HBs são marcadores da hepatite B, e são analisados conjuntamente. O perfil viral AgHBs positivo, Anti-HBc positivo e Anti-HBs negativo, apontam para uma infecção pelo vírus B que está presente, enquanto que o perfil viral AgHBs negativo, Anti-HBc positivo e Anti-HBs positivo apontam para uma infecção pelo vírus B que ocorreu no passado. O resultado positivo para o marcador Anti-HBc isoladamente não tem significado clínico. O resultado negativo para os 3 marcadores indica ausência de infecção. Observa-se que os vetores relativos a esses 3 marcadores estão na mesma direção do vetor que representa a cidade de Ponta Grossa, apontando para o fato de que a infecção pelo vírus B, presente ou passada, está associada a esta cidade. Observa-se ainda que, infecção presente pelo vírus B também está associada à cidade de Curitiba, enquanto que a infecção passada está associada à cidade de Cascavel.

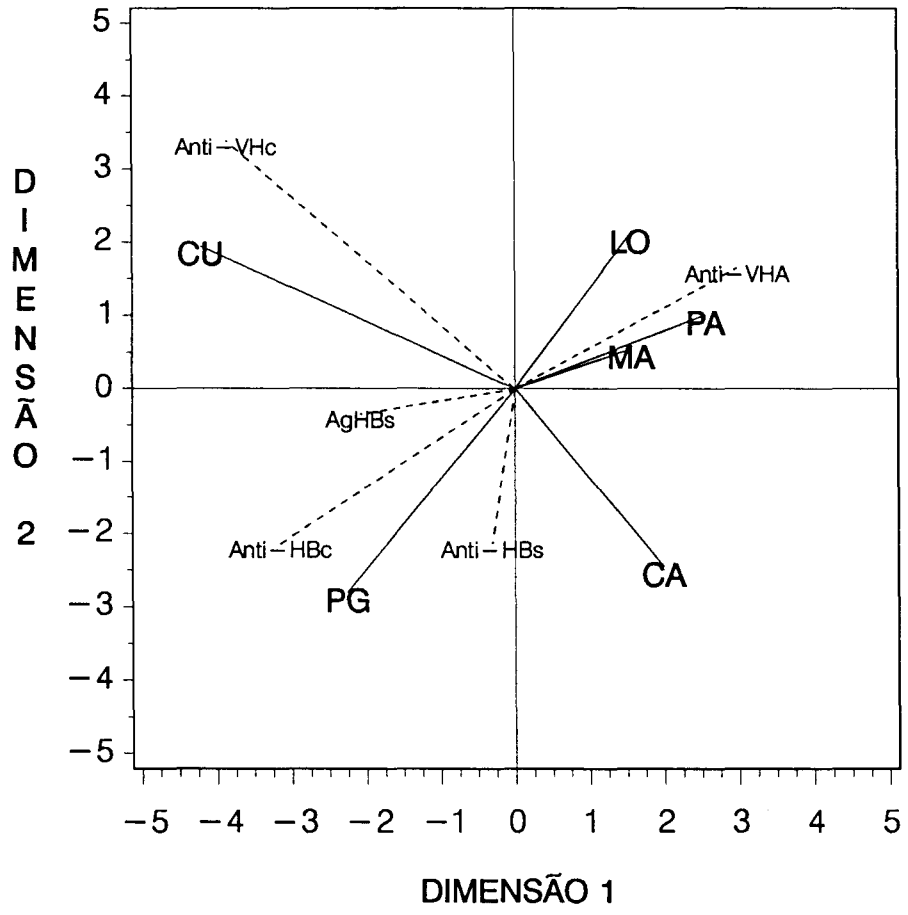

Figura 4.1 *Biplot para a matriz residual das porcentagens de positividade dos marcadores em relação ao total de pacientes de cada cidade. Os vetores* **u** *foram identificados pelas cidades e os vetores* **v** *pelos marcadores.*

LEGENDA

CA	Cascavel
CU	Curitiba
LO	LOndrina
MA	Maringá
PA	Paranavaí
PG	Ponta Grossa

5. Análise de Correspondência

5. ANÁLISE DE CORRESPONDÊNCIA

Uma descrição gráfica de dados multivariados é mais facilmente assimilada e interpretada do que um valor numérico e pode sumarizar uma massa de dados, simplificar o aspecto dos dados por apelar para nossa habilidade de absorver imagens visuais, e dar uma visão global da informação, e, em consequência, estimular possíveis explicações. Apesar dessas vantagens, apenas nos últimos anos o valor e o potencial dos gráficos estatísticos tem sido explorados.

A análise de correspondência, método que iremos abordar neste capítulo, é uma técnica exploratória multivariada, que converte uma matriz de dados não negativos em um particular tipo de gráfico no qual linhas e colunas são apresentadas por pontos. Essa técnica aplica-se às tabelas de contingência, ou seja, aplica-se basicamente para variáveis discretas. Entretanto, pode também ser aplicada para variáveis contínuas desde que estas estejam codificadas de maneira apropriada.

A forma mais básica de análise de correspondência é sua aplicação a tabelas de contingência de dupla entrada, e é denominada de análise de correspondência simples. O caso de tabelas de contingência de múltiplas entradas é denominado de análise de correspondência múltipla, e não iremos tratar nesse trabalho.

O objetivo fundamental desse método é obter a melhor representação simultânea de dois ou mais conjuntos de variáveis através de gráficos que representem cada variável nos planos de projeção formados pelos primeiros eixos fatoriais cruzados dois a dois.

O substrato de uma matriz de dados adequada para análise de correspondência é uma tabela de dupla entrada, a qual expressa a associação observada entre duas variáveis qualitativas. Em 1935, H.O.Hartley publicou um artigo , em seu nome original alemão(Hirschfeld,1935), o qual dá uma formulação algébrica da correlação entre linhas e colunas de uma tabela de contingência. Pode-se atribuir a origem matemática da análise de correspondência a esse artigo, embora Richardson e Kuder (1933) e Horst (1935), independentemente, tenham sugerido idéias similares. Posteriormente, Fisher deduziu a mesma teoria em forma de análise discriminante sobre uma tabela de contingência (Fisher,1940). Paralela e independentemente, Louis Guttman, publicou um método para construção de escalas para dados categóricos (Guttman,1941), novamente a mesma teoria em diferente contexto. Guttman tratou o caso geral de mais de duas variáveis qualitativas. Como esses dois famosos autores, Fisher e Guttman, apresentaram essencialmente a mesma teoria dentro dos contextos da biometria e da psicometria, respectivamente, são frequentes as citações de Fisher pelos biometristas e as de Guttman pelos psicometristas.

Nos anos 40 e 50, novas abordagens matemáticas foram desenvolvidas. Isto deu-se principalmente por Guttman e seus

colaboradores.

Análise de correspondência, a forma geométrica do método, originou-se na França com Jean-Paul Benzécri. No início dos anos 60, um grupo de analistas da França estudou grandes tabelas de dados obtidas de várias fontes bibliográficas. O termo francês "correspondence" foi empregado para expressar "sistema de associações entre elementos de dois conjuntos", no caso linhas e colunas. O grupo liderado por Benzécri adquiriu extensiva experiência prática em análise de correspondência e também em outras técnicas descritivas multivariadas. Vários trabalhos estão presentes na literatura, a maioria deles publicados no jornal "Les Cahiers de l'Analyse des Données". Entretanto, o estilo matemático dotado de uma notação algébrica extremamente rigorosa, impossibilitou a comunicação com as escolas Americanas, que por sua vez, utilizavam um estilo muito mais pragmático. Mallows e Tukey(1982) e Gifi(1981) abordam o assunto fazendo menção às idéias de Benzécri. Da literatura americana, os livros textos sobre o assunto são devidos a: Nishisato(1980), Greenacre(1984), Lebart, Morineau e Warwick(1984). Em nosso meio veja Souza(1982).

Apresentaremos nesse capítulo um tratamento formal da álgebra da análise de correspondência. Os conceitos geométricos mais importantes para a compreensão dessa técnica estão apresentados no apêndice A.

5.1 CONSTRUÇÃO DAS NUVENS DE PONTOS E ESCOLHA DAS DISTÂNCIAS

Vamos considerar **Y** uma matriz mxn de números não-negativos, tais que a soma das linhas e das colunas sejam não-nulas, e que seus elementos y_{ij} representem o número de observações pertencentes à linha i e coluna j, onde i=1,...,m e j=1,...,n.

A matriz de correspondência é definida por:

$$\mathbf{P} = (1/y_{..})\mathbf{Y}$$

onde

$$\mathbf{Y} = [y_{ij}] \quad , \quad y_{ij} \geq 0$$

$$y_{..} = \mathbf{1}^t \mathbf{Y}\, \mathbf{1}$$

$$\mathbf{1} = [1...1]^t$$

Os vetores das somas das linhas e das colunas de **P** serão denotados por **r** e **c**, respectivamente, e as correspondentes matrizes diagonais dessas somas por $\mathbf{D}_r$ e $\mathbf{D}_c$. Em notação matricial,

$$\mathbf{r} = \mathbf{P1} \quad \text{e} \quad \mathbf{c} = \mathbf{P}^t\mathbf{1}$$

com $r_i > 0$ i=1,...,m e $c_j > 0$ j=1,...,n

$$\mathbf{D}_r = \text{diag}(\mathbf{r}) \quad \text{e} \quad \mathbf{D}_c = \text{diag}(\mathbf{c})$$

Cabe observar que **P** nada mais é que uma matriz de frequências relativas, **r** contém as proporções marginais das linhas e **c** as proporções marginais das colunas.

No espaço $\mathbb{R}^m$ temos n vetores, cada um com m coordenadas (cada coluna constitui um vetor de $\mathbb{R}^m$). Se as

componentes dos vetores em $\mathbb{R}^m$ forem os próprios valores y_{ij} , as proximidades entre os elementos podem ficar deturpadas pela falta de padronização dos dados. Não são os valores brutos que interessam na análise de correpondência e sim os perfis das linhas, que serão dados pelas probabilidades condicionais da observação aparecer na coluna j, dado que pertence a linha i, ou seja, y_{ij} deve ser dividido pelo total da linha i.

Consideremos agora o espaço $\mathbb{R}^n$, temos m vetores, cada um com n coordenadas(cada linha constitui um vetor de $\mathbb{R}^n$). Usando o mesmo raciocínio anterior, devemos usar na análise os perfis das colunas, que serão dados pelas probabilidades condicionais da observação aparecer na linha i, dado que pertence a coluna j, ou seja y_{ij} deve ser dividido pelo total da coluna j.

Os **perfis das linhas (R)** e **das colunas (C)** de **P** são portanto definidos como os vetores das linhas e colunas de **P** divididos por suas respectivas somas, isto é:

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P} = \begin{bmatrix} \tilde{\mathbf{r}}_1^t \\ \vdots \\ \tilde{\mathbf{r}}_m^t \end{bmatrix} \quad \text{e} \quad \mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}^t = \begin{bmatrix} \tilde{\mathbf{c}}_1^t \\ \vdots \\ \tilde{\mathbf{c}}_n^t \end{bmatrix}$$

Os perfis das linhas $\tilde{r}_i (i=1,\ldots,m)$ e das colunas $\tilde{c}_j (j=1,\ldots,n)$ são escritos nas linhas de **R** e **C** respectivamente. Esses perfis são iguais às linhas e colunas de **Y** divididas pelas respectivas somas, bem como **r** e **c** são iguais às somas das linhas e colunas de **Y**, divididas pelo total geral $y_{..}$. É notacionalmente mais fácil trabalhar com **P** ao invés de **Y**, uma vez que a análise de correspondência lida com as frequências relativas dos dados, e assim se torna

invariante com respeito ao total de observações.

Os perfis das linhas e das colunas definem duas nuvens de pontos nos espaços Euclideanos ponderados, n e m-dimensional. Os perfis das linhas, $\tilde{r}_1 \ldots \tilde{r}_m$, definem os m pontos no espaço n-dimensional. Os n perfis das colunas, $\tilde{c}_1 \ldots \tilde{c}_n$, definem os n pontos no espaço m-dimensional.

A escolha dos perfis das linhas e das colunas como coordenadas dos pontos no espaço $\mathbb{R}^m$ e $\mathbb{R}^n$ dão a todos os pontos-linhas e pontos-colunas a mesma importância. No estudo da geometria de um conjunto de pontos, a atribuição de diferentes massas aos vetores determina diferentes graus de importância às posições dos pontos no espaço. É natural que cada ponto tenha um peso proporcional à sua frequência, para que não haja uma falsa idéia da repartição real da população. Assim, aos m pontos-linhas atribuiremos como **massas** os m elementos de **r**. Aos n pontos-colunas atribuiremos os n elementos de **c**. O objetivo é determinar um subespaço de menor dimensão que melhor se aproxime de todos os pontos. Quando os pontos têm diferentes massas, o subespaço deveria estar mais próximo dos pontos de maiores massas. No ajustamento das nuvens de pontos num subespaço vetorial, a quantidade a ser minimizada será uma soma de quadrados ponderada por essas massas.

Uma vez definida as nuvens de pontos, o interesse é definir uma **medida de distância ou métrica** entre eles. Conforme abordado no apêndice A, é desejável que a distância não dependa da escala de medida utilizada. Consequentemente, a distância escolhida é a **distância Euclideana ponderada**, ao

invés da Euclideana usual. Os **pesos das dimensões** serão definidos pelas inversas dos elementos de c (métrica de quiquadrado), para a nuvem linha, isto é $\mathbf{D}_c^{-1}$, e as inversas dos elementos de r para a nuvem coluna, isto é $\mathbf{D}_r^{-1}$. Portanto definimos:

em $\mathbb{R}^n$: $d^2(i,i')=(\tilde{r}_i-\tilde{r}_{i'})^t\mathbf{D}_c^{-1}(\tilde{r}_i-\tilde{r}_{i'})$

distância Euclideana ponderada ao quadrado entre os pontos-linhas i e i'.

em $\mathbb{R}^m$: $d^2(j,j')=(\tilde{c}_j-\tilde{c}_{j'})^t\mathbf{D}_r^{-1}(\tilde{c}_j-\tilde{c}_{j'})$

distância Euclideana ponderada ao quadrado entre os pontos-colunas j e j'.

A distância escolhida verifica uma propriedade importante, denominada de "Princípio da Equivalência Distributiva"(Benzecri e cols., 1973), a qual expressa que, se dois pontos-linhas ocupam posições idênticas no espaço multidimensional, então eles podem ser juntados num único ponto, que terá massa igual à soma das duas massas. Essa junção não afeta as massas e as distâncias entre pontos, dos pontos-colunas. Similarmente, a linha de dados pode ser subdividida em duas ou mais linhas cada uma proporcional a linha original, mantendo a geometria dos pontos-colunas invariante.

Essa propriedade é importante, pois garante uma invariância nos resultados em relação à codificação escolhida para construir as classes das variáveis. Do ponto de vista

técnico, é lógico que dois pontos confundidos no espaço, por estarem muito próximos, possam ser considerados como um só correspondendo a um valor total igual à soma dos valores de dois pontos. Assim, juntar pontos ou subdividir em mais pontos não provoca perda de informação.

Esse princípio tem o seguinte enunciado:

PRINCÍPIO DA EQUIVALÊNCIA DISTRIBUTIVA:

Se dois perfis de linhas (perfis de colunas) são idênticos, então as duas linhas correspondentes na matriz de dados original podem ser substituidas pela soma das linhas, sem afetar a geometria dos perfis das colunas (perfis das linhas).

PROVA:

Vamos supor, sem perda de generalidade, que a primeira e a segunda linha de **Y** tem o mesmo perfil, isto é, $y_{1j}/y_{1.} = y_{2j}/y_{2.}$, j=1,...,n. Vamos remover as linhas 1 e 2 de **Y** e criar uma nova linha cujos elementos são: $y_{1j} + y_{2j}$, j=1,...,n. A nova matriz $\tilde{\mathbf{Y}}$ terá uma linha a menos do que a matriz original **Y**, e o perfil da primeira linha de $\tilde{\mathbf{Y}}$ tem massa igual à soma das massas da primeira e segunda linha de **Y**. As massas do perfil coluna não são afetadas por esta substituição. A distância ao quadrado entre dois perfis coluna j e l é dada por :

$$\sum_{i=2}^{m} \{(\tilde{y}_{ij}/y_{.j}) - (\tilde{y}_{il}/y_{.l})\}^2 \; / \; (\tilde{y}_{i.}/y_{..})$$

Previamente esta distância era dada por:

$$\sum_{i=1}^{m} \{(y_{ij}/y_{.j}) - (y_{i1}/y_{.1})\}^2 / (y_{i.}/y_{..})$$

Os termos a partir de i=3 são idênticos nestas duas expressões. Assim, necessitamos mostrar que o primeiro termo da primeira expressão é igual aos dois primeiros termos da segunda expressão. Isto é facilmente provado por substituir $\tilde{y}_{2j} = y_{1j} + y_{2j}$ e usando a igualdade $y_{1j}/y_{1.} = y_{2j}/y_{2.}$, tem-se:

$$(y_{1j} + y_{2j}) / (y_{1.} + y_{2.}) = \tilde{y}_{2j}/\tilde{y}_{2.} \quad \blacksquare$$

5.2 AJUSTAMENTO DAS NUVENS DE PONTOS

Na seção anterior definiram-se duas nuvens de pontos constituídas pelos perfis das linhas e das colunas nos espaços Euclideanos ponderados m e n-dimensionais. Pretende-se encontrar o subespaço de dimensão k, sendo k de dimensão inferior à dimensão do subespaço original, que melhor se aproxime de todos os pontos, considerando-se que esses tem diferentes massas, e usando como distância entre eles a distância Euclideana ponderada, com a métrica de quiquadrado (vide apêndice A).

Para reforçar a distinção entre ponderação de pontos e de dimensões, iremos utilizar o termo massa quando nos referirmos a quantidade que ponderam pontos, e o termo peso a quantidades que ponderam dimensões de um espaço.

Os **centróides** (pontos médios) dos pontos-linhas e colunas em seus respectivos espaços são definidos por:

$$\mathbf{c} = \mathbf{R}^t\mathbf{r} \quad \text{e} \quad \mathbf{r} = \mathbf{C}^t\mathbf{c}$$

Esses resultados são facilmente verificados: o j-ésimo elemento de cada perfil linha é p_{ij}/r_i, onde r_i é o i-ésimo elemento de **r**. Assim, o j-ésimo elemento do centróide é $\sum_i r_i(p_{ij}/r_i)/\sum_i r_i = \sum_i p_{ij}$ (pois $\sum_i r_i=1$), o qual é c_j, o j-ésimo elemento de **c**. Em notação matricial, o centróide linha (como um vetor linha) é dado por:

$$\mathbf{r}^t\mathbf{R} \,/\, \mathbf{r}^t\mathbf{1} = \mathbf{r}^t\mathbf{R} = \mathbf{r}^t\mathbf{D}_r^{-1}\mathbf{P} = \mathbf{1}^t\mathbf{P} = \mathbf{c}^t$$

(pois $\mathbf{r}^t\mathbf{D}_r^{-1}=\mathbf{1}^t$ e $\mathbf{r}^t\mathbf{1}=1$).

Similarmente pode-se verificar que o centróide dos perfis das colunas é **r**. ■

É importante observar que o centróide dos perfis das linhas, com massas definidas pelos elementos de **r**, é o perfil dos totais das colunas, ou seja, o vetor das massas dos pontos-colunas.

Os respectivos subespaços k-dimensionais relativos às nuvens linhas e colunas, os quais estão mais próximos dos pontos em termos de soma ponderada de distâncias ao quadrado, são definidos pelos k vetores singulares generalizados (direitos e esquerdos, respectivamente), de $\mathbf{P}-\mathbf{rc}^t$, nas métricas $\mathbf{D}_c^{-1}$ e $\mathbf{D}_r^{-1}$, correspondendo aos k maiores valores singulares. Em outras palavras, os autovetores (direitos e esquerdos) definem os **eixos principais das nuvens linhas e colunas**, respectivamente. Portanto, em símbolos:

$$\mathbf{P} - \mathbf{rc}^t = \mathbf{A}\,\mathbf{D}_\lambda\mathbf{B}^t \quad \text{onde} \quad \mathbf{A}^t\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{I}_m \quad \text{e} \quad \mathbf{B}^t\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}_n \qquad (5.2.1)$$

sendo $\lambda_1 \geq \ldots \geq \lambda_k > 0$ os valores singulares de $\mathbf{P}-\mathbf{rc}^t$. As

colunas de **A** e **B** definem os eixos principais das nuvens colunas e linhas, respectivamente. Uma abordagem suscinta da decomposição em valores singulares generalizada é dada no capítulo 2.

A verificação em relação aos eixos principais é da seguinte forma:

Vamos considerar primeiramente a nuvem de pontos-linhas definida pelos perfis das linhas em $\mathbf{R}=\mathbf{D}_r^{-1}\mathbf{P}$, com massas iguais à diagonal de $\mathbf{D}_r$, e num espaço Euclideano ponderado definido pela métrica diagonal $\mathbf{D}_c^{-1}$. Os eixos principais, bem como as coordenadas dos perfis das linhas com respeito a esses eixos, são obtidos da decomposição em valores singulares generalizada de $\mathbf{R}-\mathbf{1c}^t$ (perfis das linhas centrados), sendo que os vetores singulares direitos e esquerdos são ortonormalizados em relação a $\mathbf{D}_r$ e $\mathbf{D}_c^{-1}$, respectivamente, isto é, se

$$\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c}^t = \mathbf{L}\,\mathbf{D}_\phi\mathbf{M}^t \quad \text{onde} \quad \mathbf{L}^t\mathbf{D}_r\,\mathbf{L} = \mathbf{I}_m \quad \text{e} \quad \mathbf{M}^t\mathbf{D}_c^{-1}\mathbf{M} = \mathbf{I}_n \tag{5.2.2}$$

então as colunas de M definem os eixos principais e as linhas de $\mathbf{LD}_\phi$ definem as coordenadas (relativas aos eixos principais), dos pontos-linhas (vide seção 2.4, capítulo 2).

Se multiplicarmos à esquerda da equação (5.2.2) por $\mathbf{D}_r$, obtemos:

$$\mathbf{P} - \mathbf{rc}^t = (\mathbf{D}_r\mathbf{L})\,\mathbf{D}_\phi\,\mathbf{M}^t$$

Note que

$$(\mathbf{D}_r\mathbf{L})^t\,\mathbf{D}_r^{-1}\,(\mathbf{D}_r\mathbf{L}) = \mathbf{I}_m \qquad \text{e} \qquad \mathbf{M}^t\mathbf{D}_c^{-1}\mathbf{M} = \mathbf{I}_n$$

o qual está na forma da expressão (5.2.1) e mostra que as

colunas de **M** (os eixos principais) são idênticas às de **B**.

De maneira análoga e simétrica, tem-se que os eixos principais da nuvem de pontos-colunas, os quais são definidos no espaço m-dimensional pelos vetores singulares direitos de $\mathbf{C-1r}^t$, considerando a seguinte decomposição:

$$\mathbf{D}_c^{-1}\mathbf{P}^t - \mathbf{1r}^t = \mathbf{W}\,\mathbf{D}_\psi\mathbf{Z}^t \quad \text{onde} \quad \mathbf{W}^t\mathbf{D}_c\mathbf{W} = \mathbf{I}_n \quad \text{e} \quad \mathbf{Z}^t\mathbf{D}_r^{-1}\mathbf{Z} = \mathbf{I}_m \tag{5.2.3}$$

são idênticos às colunas de **A**.

Cabe observar que os valores singulares $\lambda_1,\ldots,\lambda_r$ em (5.2.1), $\phi_1,\ldots,\phi_r$ em (5.2.2) e $\psi_1,\ldots,\psi_r$ em (5.2.3) são idênticos. Se forem diferentes, diferirão apenas pela ordem. Neste caso, deveríamos dizer que os eixos principais em **M** da nuvem linha são idênticos às colunas de **B** a menos de transformações ortogonais. Entretanto, o subespaço definido por **M** é o mesmo que o definido por **B**. ■

A decomposição em valores singulares de $\mathbf{P-rc}^t$, desprezando o último valor singular (o qual é zero) e o último vetor singular direito e esquerdo, é exatamente a decomposição em valores singulares de **P**, desprezando o primeiro valor singular (o qual é igual a um), o primeiro vetor singular esquerdo, **r**, e o primeiro vetor singular direito **c**.

As respectivas coordenadas dos perfis das linhas com relação a seus próprios eixos principais (isto é, coordenadas principais), estão relacionadas aos eixos principais dos perfis colunas através de um simples reescalonamento. O reverso também é verdadeiro. Vamos considerar as coordenadas

dos perfis das linhas com respeito aos eixos principais **B** (na métrica quiquadrado $\mathbf{D}_c^{-1}$). Note que, como os eixos principais são ortonormais ($\mathbf{B}^t\mathbf{D}_c^{-1}\mathbf{B}=\mathbf{I}_n$), essas coordenadas são dadas pelo produto escalar dos perfis centrados, $\mathbf{R}-\mathbf{1c}^t$, com **B**. Consequentemente, pode-se escrever:

$$_m\mathbf{F}_K = (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c}^t)_{mn}\mathbf{D}_c^{-1}\mathbf{B}_{nK}$$

Então:

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{A}\,\mathbf{D}_\lambda \qquad (5.2.4)$$

Nós podemos mostrar (5.2.4), reescrevendo a equação acima da seguinte forma:

$$\mathbf{F} = \mathbf{D}_r^{-1}(\mathbf{P}-\mathbf{rc}^t)\mathbf{D}_c^{-1}\mathbf{B} \qquad (5.2.5)$$

(usando $\mathbf{1}=\mathbf{D}_r^{-1}\mathbf{r}$). Multiplicando à direita da SVD generalizada de $\mathbf{P}-\mathbf{rc}^t$ por $\mathbf{D}_c^{-1}\mathbf{B}$, obtemos:

$$(\mathbf{P}-\mathbf{rc}^t)\,\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{A}\,\mathbf{D}_\lambda\,\mathbf{B}^t\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{A}\,\mathbf{D}_\lambda$$

daí a expressão (5.2.5) torna-se $\mathbf{F}=\mathbf{D}_r^{-1}\mathbf{A}\mathbf{D}_\lambda$, que é o resultado desejado. ■

O resultado simétrico, ou seja, se considerarmos as coordenadas dos perfis das colunas com respeito aos eixos principais em **A** (na métrica quiquadrado $\mathbf{D}_r^{-1}$), é:

$$_n\mathbf{G}_K = (\mathbf{D}_c^{-1}\mathbf{P}^t - \mathbf{1r}^t)_{nm}\mathbf{D}_{rm}^{-1}\mathbf{A}_{mK}$$

$$\mathbf{G} = \mathbf{D}_c^{-1}\mathbf{B}\,\mathbf{D}_\lambda \qquad (5.2.6)$$

As expressões (5.2.4) e (5.2.6) definem as coordenadas dos perfis das linhas e das colunas, com respeito a todos os eixos principais (as coordenadas dos pontos

individuais estão contidas nas linhas de **F** e **G**). As coordenadas dos pontos com respeito a um subespaço ótimo K-dimensional estão contidas nas linhas das primeiras K colunas de **F** e **G**. Por exemplo, se representarmos por $\mathbf{F}_{(2)}$ e $\mathbf{G}_{(2)}$ as duas primeiras colunas de **F** e **G** respectivamente, então as linhas de $\mathbf{F}_{(2)}$ e $\mathbf{G}_{(2)}$ definem as projeções dos perfis das linhas e das colunas, respectivamente, sobre os planos ótimos.

Como consequência imediata de (5.2.1), (5.2.4) e (5.2.6), temos que os dois conjuntos de coordenadas **F** e **G** estão relacionados pela seguinte fórmula:

Fórmula de transição das linhas (F) para as colunas (G):

$$\mathbf{G} = \mathbf{D}_c^{-1}\,\mathbf{P}^t\,\mathbf{F}\,\mathbf{D}_\lambda^{-1} = \mathbf{C}\,\mathbf{F}\,\mathbf{D}_\lambda^{-1} \qquad \text{isto é,}$$

$$\mathbf{G}\,\mathbf{D}_\lambda = \mathbf{D}_c^{-1}\,\mathbf{P}^t\,\mathbf{F} \qquad (5.2.7)$$

Fórmula de transição das colunas (G) para as linhas (F):

$$\mathbf{F} = \mathbf{D}_r^{-1}\,\mathbf{P}\,\mathbf{G}\,\mathbf{D}_\lambda^{-1} = \mathbf{R}\,\mathbf{G}\,\mathbf{D}_\lambda^{-1} \qquad \text{isto é,}$$

$$\mathbf{F}\,\mathbf{D}_\lambda = \mathbf{D}_r^{-1}\,\mathbf{P}\,\mathbf{G} \qquad (5.2.8)$$

5.3 INTERPRETAÇÃO DOS RESULTADOS

No uso da análise de correspondência é usual a representação gráfica planar dos pontos-linhas e colunas obtidos através dos primeiros eixos principais tomados dois a dois. Teremos m pontos para as linhas e n pontos para as colunas. A interpretação dos gráficos representando as nuvens de pontos nos planos de projeção é o mais difícil objeto da

análise. Greenacre e Hastie(1987) apresentam a interpretação geométrica da Análise de Correspondência. Temos que verificar quais são os eixos da análise mais representativos e interpretar as proximidades entre elementos de uma mesma nuvem. Não é possível interpretar a proximidade entre pontos de nuvens diferentes (um de $\mathbb{R}^m$ outro de $\mathbb{R}^n$), uma vez que nenhuma medida de distância entre nuvens diferentes foi estabelecida. Lembre-se que apenas a distância entre pontos de uma mesma nuvem foi explicitamente definida. A análise de pontos pertencentes à nuvens diferentes se dá através de ângulos entre os vetores que vão da origem do gráfico (ponto (0,0)), até cada um dos pontos.

Definimos a seguir três coeficientes que auxiliam na interpretação dos resultados: taxa de inércia, contribuição absoluta e contribuição relativa.

(i) Inércia

A variação espacial total de cada nuvem de pontos é quantificada pela sua **inércia total**, ou seja, pela soma ponderada das distâncias ao quadrado dos pontos aos seus centróides. A **inércia total dos pontos-linhas** é dada por:

$$in(m) = \sum_i r_i(\tilde{r}_i - c)^t \, D_c^{-1} \, (\tilde{r}_i - c) \qquad (5.3.1)$$

Uma outra forma de expressar a inércia acima é:

$$in(m) = \text{traço}\,[D_r(R-1c^t)\, D_c^{-1}\, (R-1c^t)^t]$$

Analogamente, a **inércia total dos pontos-colunas** é:

$$in(n) = \sum_j c_j(\tilde{c}_j - r)^t \, D_r^{-1} \, (\tilde{c}_j - r) \qquad (5.3.2)$$

A inércia total é a mesma em ambas as nuvens e

pode-se verificar que:

$$in(m) = in(n) = \sum_i \sum_j (p_{ij} - r_i c_j)^2 / r_i c_j = \chi^2 / y_{..}$$

$$= \text{traço } [\mathbf{D}_r^{-1}(\mathbf{P}-\mathbf{r}\mathbf{c}^t)\ \mathbf{D}_c^{-1}\ (\mathbf{P}-\mathbf{r}\mathbf{c}^t)^t]$$

onde $\chi^2 = \sum_i \sum_j (y_{ij}-e_{ij})^2/e_{ij}$, e_{ij} é o valor esperado da casela (i,j), ($e_{ij} = y_{i.}y_{j.}/y_{..}$).

Vamos demonstrar a igualdade in(m)=in(n):

Das equações (5.3.1) e (5.3.2) tem-se:

$$in(m) = \sum_i r_i \sum_j (p_{ij}/r_i - c_j)^2 / c_j$$

$$= \sum_i \sum_j (p_{ij} - r_i c_j)^2 / r_i c_j$$

e

$$in(n) = \sum_j c_j \sum_i (p_{ij}/c_j - r_i)^2 / r_i$$

$$= \sum_i \sum_j (p_{ij} - r_i c_j)^2 / r_i c_j$$

Portanto in(m)=in(n).

Na fórmula de χ^2 apresentada anteriormente, $y_{ij}=y_{..}p_{ij}$, e consequentemente o valor esperado pode ser escrito como:

$$e_{ij} = (\sum_j y_{ij})\ (\sum_i y_{ij})\ /\ y_{..}$$

$$= (y_{..}r_i)\ (y_{..}c_j)\ /\ y_{..}$$

$$= y_{..}r_i c_j$$

então $\chi^2 = \sum_i \sum_j (y_{ij}-y_{..}r_i c_j)^2/y_{..}r_i c_j$

$$= y_{..} \sum_i \sum_j (p_{ij}-r_i c_j)^2/r_i c_j$$

o que implica que $\chi^2 = y_{..}\, in(m) = y_{..}\, in(n)$. ■

(ii) Contribuição Absoluta

Como já vimos, a inércia ao longo de um eixo principal é dada pela soma ponderada das distâncias ao quadrado dos pontos aos seus centróides. Cada termo nessa soma (correspondente à distância de um ponto i), pode ser expresso como uma porcentagem da inércia principal (eixo principal), a qual denominamos de contribuição do ponto para o eixo principal ou contribuição absoluta. Se imaginarmos os pontos-linhas e os pontos-colunas nas suas posições de origem, nos dois espaços correspondentes, exercendo forças de atração para os eixos principais em virtude de suas posições e de suas massas, então são os pontos com alta contribuição que estabelecem a orientação final dos eixos. Em geral observa-se que são os pontos de maiores massas que tem maiores contribuições.

(iii) Contribuição Relativa

A avaliação de quão próximo de cada ponto está o subespaço ótimo também é de grande interesse. Para tanto pode-se averiguar o ângulo θ entre o ponto observado e o eixo principal.

É conveniente examinar o co-seno ao quadrado do ângulo, pois, para cada ponto, os co-senos ao quadrado do ângulo com o conjunto completo de eixos principais ortogonais somam 1. Uma abordagem equivalente vem do fato de que a inércia de um ponto $r_i d_i^2$ (isto é, o i-ésimo perfil linha com massa r_i e distância d_i do centróide) é decomposta ao longo dos eixos principais. A parte dessa inércia ao longo do

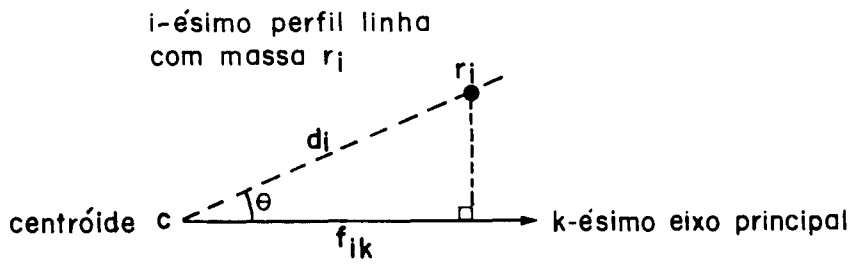

Figura 5.1 *Coordenada do i-ésimo perfil-linha relativo ao k-ésimo eixo principal, o qual está a uma distância* d_i *do centróide* c. $Cos^2\theta=(f_{ik}/d_i)^2$ *é denominado de contribuição (relativa) do eixo para o k-ésimo ponto.*

primeiro eixo é $r_i f_{i1}^2$, onde f_{i1}^2 é a coordenada do ponto sobre este eixo (figura 5.1). Expressando como uma proporção da inércia total dos pontos tem-se: $r_i f_{i1}^2 / r_i d_i^2 = (f_{i1}/d_i)^2 = \cos^2\theta$. O $\cos^2\theta$ é designado de contribuição do eixo para a inércia do ponto. Se $\cos^2\theta$ é um valor alto então o eixo explica a inércia do ponto muito bem; equivalentemente θ é um ângulo pequeno e o vetor perfil é dito estar na direção do eixo, ou estar relacionado com o eixo. Os valores de $\cos^2\theta$ são também chamados de contribuições relativas, pois são independentes da massa dos pontos.

Geralmente uma alta contribuição do ponto para a inércia do eixo implica em alta contribuição relativa do eixo para a inércia do ponto, mas o reverso não é necessariamente verdadeiro.

Vamos abordar esses 3 coeficientes com relação aos pontos expressos pelas coordenadas dos eixos principais.

Com respeito aos eixos principais, as respectivas

nuvens dos perfis das linhas e das colunas tem centróides na origem. A soma ponderada dos quadrados das coordenadas dos pontos, ao longo do k-ésimo eixo principal em cada nuvem é igual a λ_k^2 , a qual é denominada de **k-ésima inércia principal.** A soma ponderada dos produtos cruzados das coordenadas é zero. Portanto, o centróide das linhas de **F** e de **G** são, respectivamente:

$$\mathbf{r}^t\mathbf{F} = \mathbf{0}^t \quad \text{e} \quad \mathbf{c}^t\mathbf{G} = \mathbf{0}^t \tag{5.3.3}$$

As linhas de **F** e **G** são simplesmente os respectivos conjuntos de perfis centrados com respeito a um novo sistema de eixos. Por exemplo:

$$\mathbf{r}^t\mathbf{F} = \mathbf{r}^t(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^t)\mathbf{D}_c^{-1}\mathbf{B} = (\mathbf{1}^t\mathbf{P} - \mathbf{c}^t)\mathbf{D}_c^{-1}\mathbf{B}$$

$$= (\mathbf{c}^t - \mathbf{c}^t)\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{0}^t$$

Em notação matricial, as inércias principais das nuvens linha e coluna são:

$$\begin{aligned} \mathbf{F}^t\mathbf{D}_r\mathbf{F} &= \mathbf{D}_\lambda^2 \\ \mathbf{G}^t\mathbf{D}_c\mathbf{G} &= \mathbf{D}_\lambda^2 \end{aligned} \tag{5.3.4}$$

Os resultados em (5.3.4) são pertinentes à soma de quadrados e produtos cruzados das coordenadas principais, e seguem diretamente da padronização dos eixos em (5.2.1), (5.2.4) e (5.2.6). Como uma consequência de (5.3.1), (5.3.2) e (5.3.4), a inércia total de cada nuvem de pontos é decomposta ao longo dos eixos principais e entre os pontos de uma forma similar e simétrica. Temos uma decomposição da inércia para cada nuvem de pontos, a qual é análoga à decomposição da variância.

A tabela (5.1) dá um suporte numérico para a representação gráfica. Nas colunas dessa tabela tem-se a contribuição das nuvens linhas e colunas respectivamente à inércia de um eixo. Como vimos anteriormente, é possível expressar cada uma dessas contribuições como proporções da respectiva inércia λ_k^2 para interpretar o eixo (contribuições absolutas). Cada linha dessa tabela contém as contribuições dos eixos para a inércia dos respectivos perfis. Novamente, nós podemos expressar cada uma dessas como proporções das inércias dos pontos para interpretar quão bem o ponto está representado nos eixos (contribuições relativas).

Tabela 5.1 Decomposição da Inércia

		E I X O S				
		1	2	...	K	Total
LINHAS	1	$r_1 f_{11}^2$	$r_1 f_{12}^2$	...	$r_1 f_{1k}^2$	$r_1 \sum_k f_{1k}^2$
	2	$r_2 f_{21}^2$	$r_2 f_{22}^2$	...	$r_2 f_{2k}^2$	$r_2 \sum_k f_{2k}^2$
	⋮	⋮	⋮	...	⋮	⋮
	m	$r_m f_{m1}^2$	$r_m f_{m2}^2$	...	$r_m f_{mk}^2$	$r_m \sum_k f_{mk}^2$
	total	λ_1^2	λ_2^2	...	λ_k^2	in(m)=in(n)
COLUNAS	1	$c_1 g_{11}^2$	$c_1 g_{12}^2$	...	$c_1 g_{1k}^2$	$c_1 \sum_k g_{1k}^2$
	2	$c_2 g_{21}^2$	$c_2 g_{22}^2$	...	$c_2 g_{2k}^2$	$c_2 \sum_k g_{2k}^2$
	⋮	⋮	⋮	...	⋮	⋮
	n	$c_n g_{n1}^2$	$c_n g_{n2}^2$	...	$c_n g_{nk}^2$	$c_n \sum g_{nk}^2$

5.4 UM EXEMPLO DE ANÁLISE DE CORRESPONDÊNCIA

Nesta seção apresentaremos uma aplicação da técnica de análise de correspondência. Não temos a intenção de fazer uma análise exaustiva dos dados, mas apenas ilustrar a utilização da técnica.

Consideremos os dados da Tabela 5.2 os quais foram extraídos de Everitt(1977,pag 95). Um total de 1237 indivíduos sem doença arterial coronária foram classificados com respeito a pressão arterial(mmHg) e nível sérico de colesterol total(mg/100cc).

Tabela 5.2 Frequência de Indivíduos classificados por Nível de Colesterol Total e Pressão Arterial

COLESTEROL TOTAL	PRESSÃO ARTERIAL <127	127-146	147-166	≥167	TOTAL
<200	117	121	47	22	307
200-219	85	98	43	20	246
220-259	119	209	68	43	439
≥260	67	99	46	33	245
TOTAL	388	527	204	118	1237

Vamos, através da análise de correspondência, tentar avaliar eventuais associações entre níveis de colesterol e pressão arterial.

Os resultados que serão apresentados foram obtidos por meio do procedimento CORRESP do "software" SAS. O programa utilizado está no apêndice C.

As frequências relativas dos níveis pressóricos em

cada intervalo de colesterol definem os perfis das linhas, e as frequências relativas dos níveis de colesterol em cada intervalo de pressão definem os perfis das colunas. Esses perfis estão apresentados na tabela 5.3.

Tabela 5.3 Perfis das Linhas e das Colunas

COLESTEROL TOTAL	**PRESSÃO ARTERIAL** <127	127-146	147-166	≥167
		PERFIS DAS LINHAS		
<200	0,38	0,39	0,15	0,07
200-219	0,35	0,40	0,40	0,08
220-259	0,27	0,48	0,48	0,10
≥260	0,27	0,40	0,40	0,13
		PERFIS DAS COLUNAS		
<200	0,30	0,23	0,23	0,17
200-219	0,22	0,19	0,21	0,17
220-259	0,31	0,40	0,33	0,36
≥260	0,17	0,19	0,23	0,28

Considerando os perfis como pontos num espaço 4-dimensional, a ponderação de cada perfil é feita pelo número de indivíduos que constitui o perfil, dividido pelo total de indivíduos pesquisados. Assim, as massas alocadas aos 4 pontos-linhas foram: 0,248, 0,199, 0,355 e 0,198, e aos 4 pontos colunas foram: 0,314, 0,426, 0,165 e 0,095. Os pesos das dimensões foram considerados iguais aos inversos dos valores esperados, estabelecendo-se consequentemente a distância de qui-quadrado entre os perfis.

A tabela 5.4 apresenta as coordenadas dos

pontos-linhas e dos pontos-colunas no subespaço bidimensional ótimo, dadas pelas linhas de $\mathbf{F}_{(2)}$ e de $\mathbf{G}_{(2)}$, respectivamente.

Tabela 5.4 Coordenadas dos Pontos-Linhas e dos Pontos-Colunas

	DIMENSÃO 1	DIMENSÃO 2
PONTOS-LINHAS		
<200	-0,153	-0,001
200 - 219	-0,080	0,025
219 - 259	0,084	-0,072
≥260	0,120	0,106
PONTOS-COLUNAS		
>127	-0,151	0,019
127 - 146	0,055	-0,068
147 - 166	0,029	0,074
≥167	0,201	0,112

A tabela 5.5 mostra que a inércia total é 0,01686 (soma dos valores singulares) e as porcentagens das inércias associadas a cada um dos eixos são: 73,92%, 24,79%, e 1,28%. Assim os dois primeiros eixos explicam 98,71% da variância total. Em particular neste exemplo, obtivemos uma representação quase exata dos pontos, uma vez que apenas 1,28% da inércia total dos pontos não está representada no sub-espaço bidimensional.

Tabela 5.5 Valores Singulares e Inércias

VALORES SINGULARES	INÉRCIAS PRINCIPAIS	PORCENTAGEM	15 30 45 60 75
			----+----+----+----+----
0.11163	0.01246	73.92%	**********
0.06465	0.00418	24.79%	********
0.01470	0.00022	1.28%	
	0.01686		

A tabela 5.6 traz a contribuição dos pontos-linhas e dos pontos-colunas na construção dos eixos (contribuições absolutas), bem como a contribuição dos eixos para explicação da dispersão dos pontos (contribuições relativas). Observa-se, a partir das contribuições absolutas, que o primeiro eixo é fortemente influenciado pelo ponto-linha <200 e pelo ponto-coluna <127, enquanto que para a formação do segundo eixo destacam-se os pontos-linhas 220-259 e ≥260, e o ponto-coluna 127-146. Quanto as contribuições relativas, observa-se que o primeiro eixo é muito mais importante na explicação da dispersão dos pontos-linhas e colunas, do que o segundo eixo, a exceção dos pontos-colunas 127-146 e 147-166.

Tabela 5.6 Contribuições Relativas e Absolutas dos Pontos-Linhas e dos Pontos-Colunas em Relação aos Dois Primeiros Eixos Principais

	CONTRIBUIÇÃO ABSOLUTA (%)		**CONTRIBUIÇÃO RELATIVA (%)**	
	dimensão 1	dimensão 2	dimensão 1	dimensão 2
PONTOS-LINHAS				
<200	46,58	0,01	98,94	0,01
200-219	10,12	3,03	82,28	8,27
220-259	20,33	44,18	57,84	42,16
≥260	22,97	52,78	56,37	43,44
PONTOS-COLUNAS				
<127	57,64	2,75	98,19	1,57
127-146	10,43	46,91	39,86	60,14
147-166	1,12	21,52	11,93	76,83
≥167	30,81	28,82	75,13	23,57

A representação gráfica dos pontos-linhas e dos

pontos-colunas, com relação aos dois primeiros eixos principais estão na figura 5.2, a qual na verdade corresponde à figura resultante da superposição de dois gráficos, um de cada variável categórica. É importante ressaltar que distâncias entre pontos de uma mesma variável tem significado, enquanto que distâncias entre pontos de diferentes variáveis não podem ser interpretadas. Um modo seguro de interpretar o gráfico é através dos ângulos entre os vetores que vão da origem do gráfico (ponto(0,0)), até cada um dos pontos. Nessa figura pode-se observar que os ângulos entre os vetores 1,2 e A são pequenos, sugerindo que indivíduos com colesterol total até 200mg/100cc tendem a ter valores de pressão menores que 127mmHg. O fato que o ponto 1 e o ponto A estão próximos deve ser ignorado, uma vez que pertencem a variáveis diferentes. A análise é feita pela proximidade de vetores e não de pontos. Da análise do gráfico pode-se ainda observar que indivíduos com níveis de colesterol entre 220 e 259mg/100cc tendem a apresentar pressão entre 127 e 146mmHg, e para valores de colesterol igual ou acima de 260mg/100cc tem-se valores de pressão acima de 147mmHg.

Embora esse exemplo tenha sido construido para propósitos ilustrativos e não como uma aplicação séria de análise de correspondência, em resumo, pode-se comentar que no grupo de indivíduos pesquisados, a análise sugere que aos maiores valores de colesterol total tem-se associados os maiores valores de pressão arterial.

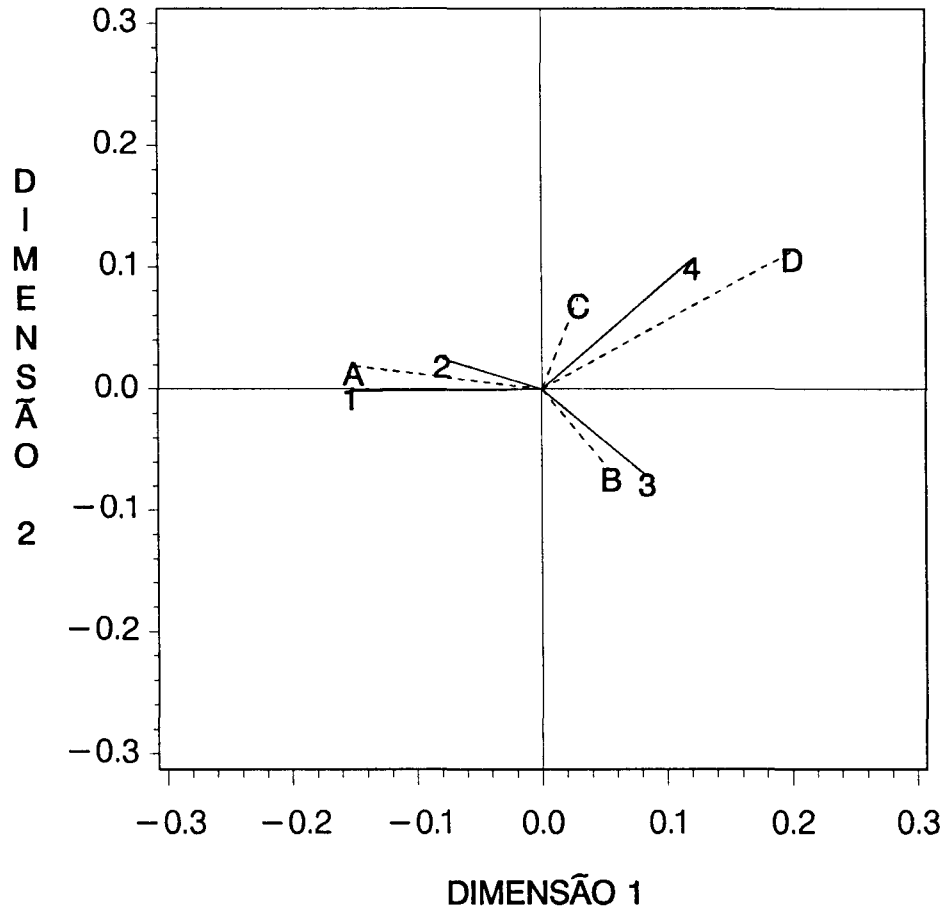

Figura 5.2 *Representação dos pontos-linhas e dos pontos-colunas no plano principal.*

LEGENDA

1	< 200	**A**	< 127
2	200-219	**B**	127-146
3	220-259	**C**	147-166
4	≥ 260	**D**	≥ 167

6. Referências Bibliográficas

REFERÊNCIAS BIBLIOGRÁFICAS

BEN-ISRAEL,A. & GREVILLE,T.N.E. **Generalized Inverses: Theory and Applications.** New York, John Wiley,1974.

BENZÉCRI,J.P. et al. **"L'Analyse des Données".** 1: La Taxinomie. Tome 2: L'Analyse des Correspondances. Paris,Dunod,1973.

CARVALHO,J.F. **Data Analysis of Multiway Tables.** Tese de Doutoramento. Ames, Iowa,1977.

CHAMBERS,J.M. **Computacional Methods for Data Analysis.** New York, John Wiley,1977.

ECKART,C. & YOUNG,G. The Aproximation of One Matrix by Another of Lower Rank. **Psychometrika.** 1:211-218,1936.

EVERITT,B.S. **The Analysis of Contingency Tables.** New York. John Wiley ,1977,p.95

FISHER,R.A. The precision of discriminant functions. **Ann. Eugen.** 10,422-429,1940.

FREITAS,M.A. **Teste de Não-Aditividade para Experimentos a Dois Fatores Não Replicados: Aplicação de Um Modelo Multiplicativo Geral.** Tese de Mestrado, Campinas,1886.

GABRIEL,K.R. Least Squares Aproximation of Matrices by Additive and Multiplicative Models. **J. Am. Stat. Soc.**, 40, 2:186-196,1978.

GABRIEL,K.R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis, **Biometrika**, 58, 3:453-467,1971.

GABRIEL,K.R. Biplot. In: **Encyclopedia of Statistical Science.** (Johnson,N.L and Kotz,S,eds), New York, John Wiley,1980. p.263-271.

GABRIEL,K.R. Biplot Display of Multivariate Matrices for Inspection of Data and diagnosis. In: **Interpeting Multivariate Data.** New York, John Wiley,1981.p.147-174.

GIANOTTI,I. **Análise de Interações em Experimentos Fatoriais Não Replicados.** Tese de Mestrado, São Carlos,1982.

GIFI,A. **Nonlinear Multivariate Analysis.** Department of Data Theory, University of Leiden, The Netherlands,1981.

GOOD,I.J. Some Applications of the Singular Decomposition of a Matrix, **Technometrics** 11:823-831,1969.

GREENACRE,M.J. **Theory and Applications of Correspondence Analysis**. London, Academic Press,1984.

GREENACRE,M.J. & HASTIE,T. **The Geometric Interpretation of Correspondence Analysis**, JASA, 82:437-447,1987.

GREENACRE,M.J. & UNDERHILL,L.G. Scaling a Data Matrix in Low-Dimensional Euclidean Space. In: **Topics in Applied Multivariate Analysis**(Hawkins,D.M.,ed). Cambridge, Cambrige University Press,1982.p.183-268.

GUTTMAN,L. The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In: **The Prediction of Personal Adjustment** (Horst,P.,ed).New York, Social Science Research Council,1941.p.319-348.

HIRSCHFELD,H.O. A Connection Between Correlation and Contingency. **Cambridge Philosophical Soc. Proc.** (Math. Proc.). 31:520-524,1935.

HORST,P. Measuring Complex Attitudes. **J. Social Psychol.** 6:369-374,1935.

JOHN,J.A., QUENOVILLE,M.H. **Experiments: Design and Analysis.** New York, Macmilan Publishing Co.,1977,p.79.

JOHNSON,D. & GRAYBILL,F.A. An Analysis of a Two-Way Model with Interaction and No Replication. **JASA.** 67:862-868,1972.

LAWSON,C.L. & HANSON,R.J. **Solving Least Squares Problems.** Englewoods Cliffs, Prentice-Hall,1974.

LEBART,L., MORINEAU,A. & WARWICK,K.M. **Multivariate Discriptive Statistical Analysis.** New York, Jonh Wiley,1984.

MACDUFFEE,C.C. **The Theory of Matrices.** New York, Chelsea,1956.

MALLOWS,C.L. & TUKEY,J.W. An Overview of Techniques of Data Analysis, Emphasizing its Exploratory Aspects. In: **Some Recent Advances in Statistcs** (Olivina,J.T. & Epstein,B., eds). London, Academic Press,1982,p.111-172.

MANDEL,J. Use of the Singular Value Decomposition in Regression Analysis, **Am. Statitician**, 36:15-24,1982.

MANDEL,J. The Partitioning of Interaction in Analysis of Variance. **Journal of the Research of the National Bureau of Standards**, 73:309-328,1969.

MARSHALL,A.W. & Olkin,I. **Inequalities: Theory of Majorization and its Applications.** New York, Academic Press,1979.

MILLIKEN,G.A. & JOHNSON,D.E. **The Analysis of Mess Data.** New York, Van Nostrand Reinhold,1989.

NISHISATO,S. **Analysis of Categorical Data : Dual Scaling and Its Applications.** New York, University of Toronto Press, 1980.

RAO,C.R. Matrix Approximations and Redution of Dimensionality in Multivariate Statistical Analysis. In: **Multivariate Analysis** (Krishnaiah,P.R.,ed) Vol. 5, North Holland, Amsterdam,1980.

RICHARDSON,M. & KUDER,G.F. Making a rating scale that measures. **Personnel J.** 12:36-40,1933.

SAS Institute Inc. SAS Language Guide, Release 6.03 Edition. Cary, NC: SAS Institute Inc.,1988.

SAS Institute Inc. SAS/GRAPH Guide for Personal Computers, Version 6 Edition. Cary, NC: SAS Institute Inc.,1978.

SAS Institute Inc. SAS Procedure Guide, Release 6.03 Edition. Cary, NC: SAS Institute Inc.,1988.

SAS Institute Inc. SAS/STAT Guide for Personal Computers, Version 6 Edition. Cary, NC: SAS Institute Inc.,1987.

SAS Institute Inc. SAS Technical Report P-179, Additional SAS/STAT Procedures, Release 6.03, Cary, NC: SAS Institute Inc.,1988.

SEARLE,S.R. **Matrix Algebra Useful for Statistics.** New York, John Wiley,1982.

SOUZA,A.M.R. **Análise de Correspondência.** Tese de Mestrado, São Paulo,1982.

TUKEY,J.W. One Degree of Freedom for Non-Additivity. **Biometrics** 5:232-242,1949.

APÊNDICE A . Conceitos Geométricos no Espaço Multidimensional

APÊNDICE A. CONCEITOS GEOMÉTRICOS NO ESPAÇO MULTIDIMENSIONAL

Para facilitar o entendimento das técnicas que abordamos neste trabalho, apresentamos um resumo dos principais conceitos geométricos, comuns aos métodos de análises multidimensionais. Definimos distância, ângulo e produto escalar entre pontos num espaço multidimensional; atribuimos pesos às dimensões do espaço e aos pontos individuais, e identificamos sub-espaço de menor dimensionalidade, que melhor se aproxime de um conjunto de pontos dados.

A.1 DISTÂNCIA, ÂNGULO E PRODUTO ESCALAR

A definição de uma medida de distância (ou métrica) entre pontos em um espaço multidimensional de dados é crucial. A distância e o ângulo são ambos quantidades escalares (números reais), definidos em termos de 2 pontos: **distância** é um valor que quantifica a proximidade de um ponto **a** a um ponto **b**, e **ângulo** é um valor que quantifica quão rapidamente 2 vetores estão divergindo de uma origem comum. A representação gráfica desses conceitos, para quaisquer dois pontos **a** e **b** é dada na figura A.1.

Se conhecemos as distâncias de **a** e de **b** até a origem (comprimento do vetor **a** e **b** respectivamente), e o ângulo entre **a** e **b**, então podemos extrair a informação da distância de **a** até **b**.

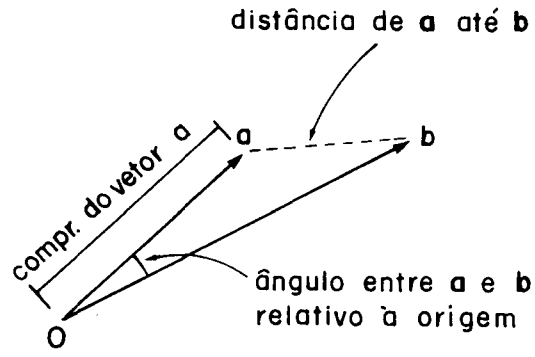

Figura A.1 - *Distância e ângulo entre os vetores* **a** *e* **b**.

Ambos os conceitos, de distância e ângulo, estão englobados num único conceito que é fundamental quando se trabalha com espaço multidimensional, denominado de produto escalar (ou produto interno).

Primeiramente, vamos abordar essas definições no espaço Euclideano bidimensional. Consideremos dois pontos $\mathbf{a}=[a_1\ a_2]^t$ e $\mathbf{b}=[b_1\ b_2]^t$ como mostrados na figura A.2.

Os comprimentos (ou normas) dos vetores **a** e **b** são:

$$\|\mathbf{a}\| = (a_1^2 + a_2^2)^{1/2} \qquad \|\mathbf{b}\| = (b_1^2 + b_2^2)^{1/2}$$

A distância entre os pontos **a** e **b**, denotada por $d(\mathbf{a},\mathbf{b})$ é:

$$d(\mathbf{a},\mathbf{b}) = ((a_1 - b_1)^2 + (a_2 - b_2)^2)^{1/2}$$

O ângulo θ entre **a** e **b** tem co-seno:

$$\cos\theta = (a_1 b_1 + a_2 b_2) / ((a_1^2 + a_2^2)(b_1^2 + b_2^2))^{1/2}$$

Todas as fórmulas acima podem ser expressas em termos de produto escalar de **a** e **b**, denotado por $(\mathbf{a},\mathbf{b})=a_1 b_1 + a_2 b_2$. Em notação vetorial $(\mathbf{a},\mathbf{b})= \mathbf{a}^t\mathbf{b}$. As fórmulas acima levam a:

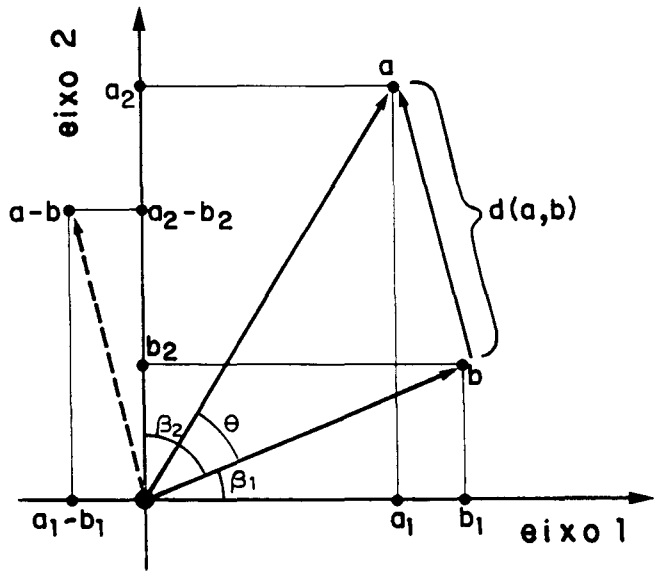

Figura A.2 - *Pontos* **a** *e* **b** *num espaço bidimensional, vetor resultante da diferença, e distância entre os pontos.*

$$\|\mathbf{a}\| = (\mathbf{a},\mathbf{a})^{1/2} = (\mathbf{a}^t\mathbf{a})^{1/2}$$
$$\|\mathbf{b}\| = (\mathbf{b},\mathbf{b})^{1/2} = (\mathbf{b}^t\mathbf{b})^{1/2} \qquad \text{(A.1.1)}$$

ou seja, o comprimento ao quadrado de um vetor é o produto escalar do vetor por ele mesmo.

$$d(\mathbf{a},\mathbf{b}) = (\mathbf{a}-\mathbf{b},\mathbf{a}-\mathbf{b})^{1/2} = ((\mathbf{a}-\mathbf{b})^t(\mathbf{a}-\mathbf{b}))^{1/2} \qquad \text{(A.1.2)}$$

$$\cos\theta = (\mathbf{a},\mathbf{b})/((\mathbf{a},\mathbf{a})(\mathbf{b},\mathbf{b}))^{1/2} = \mathbf{a}^t\mathbf{b}/(\mathbf{a}^t\mathbf{a}\mathbf{b}^t\mathbf{b})^{1/2} \qquad \text{(A.1.3)}$$

Como mencionado, $d(\mathbf{a},\mathbf{b})$ pode ser avaliado em termos de $\|\mathbf{a}\|, \|\mathbf{b}\|$ e $\cos\theta$:

$$\begin{aligned} d^2(\mathbf{a},\mathbf{b}) &= (\mathbf{a}-\mathbf{b})^t(\mathbf{a}-\mathbf{b}) \\ &= \mathbf{a}^t\mathbf{a} + \mathbf{b}^t\mathbf{b} - 2\mathbf{a}^t\mathbf{b} \\ &= \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|.\|\mathbf{b}\|.\cos\theta \end{aligned}$$

Tal expressão é conhecida como regra do co-seno.

É importante ressaltar que esses resultados dependem da perpendicularidade do sistema de coordenadas, ou seja, os vetores bases devem ser ortogonais. Dois vetores são **ortogonais** se o produto escalar entre eles é zero, ou seja, eles não têm um componente na direção do outro. Se adicionalmente os vetores têm comprimentos unitários, dizemos que eles são **ortonormais**. Então eles constituem uma base ortonormal para o espaço Euclideano. Todas as definições acima são válidas se os vetores são expressos num sistema de coordenadas ortonormais.

É imediata a extensão das definições mencionadas acima para o espaço Euclideano J-dimensional. O produto escalar de quaisquer dois vetores $\mathbf{a}=[a_1 \ldots a_j]^t$ e $\mathbf{b}=[b_1 \ldots b_j]^t$ é definido como:

$$(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{J} a_i b_i = \mathbf{a}^t\mathbf{b}$$

As definições de comprimento, distância e ângulo seguem exatamente as equações (A.1.1 - A.1.3). Novamente as definições acima requerem que **a** e **b** sejam expressos relativos a uma base ortonormal.

A.2 ESPAÇO EUCLIDEANO PONDERADO

Vamos ilustrar o conceito de ponderação de eixos num espaço bidimensional. Considere os vetores **x** e **y** contendo as informações de peso e altura de dois indivíduos respectivamente, e as unidades de medidas utilizadas sendo centímetro para altura e kilograma para peso. A medida de altura tem um valor numérico maior que a medida de peso.

Portanto, a diferença entre as alturas de dois indivíduos é, em geral, maior que a diferença entre os pesos. Assim, a medida de altura contribuirá relativamente mais para a distância Euclideana, a qual depende da soma de quadrados dessa distância. Por outro lado, se a altura fosse expressa em metros, a medida peso teria maior contribuição para a distância Euclideana. Certamente, não é desejável que a distância dependa diretamente da escala de medida escolhida. Uma forma comum de se contornar esse problema é efetuar a divisão da medida pelo seu respectivo desvio-padrão, antes de calcular a distância Euclideana. Esta forma padronizada de medida permanece a mesma para qualquer unidade escolhida originalmente. O produto escalar entre **x** e **y** conterá um fator de ponderação em cada termo, isto é:

$$(\mathbf{x},\mathbf{y}) = x_1y_1/\ s_1^2 + x_2y_2/s_2^2 = \mathbf{x}^t\ \mathbf{D}_{s^2}^{-1}\ \mathbf{y} \qquad (A.2.1)$$

onde

$$\mathbf{D}_{s^2}^{-1} = \begin{bmatrix} 1/s_1^2 & 0 \\ 0 & 1/s_2^2 \end{bmatrix}$$

é a matriz diagonal das inversas das variâncias.

Geometricamente os vetores **x** e **y** são desenhados nas suas unidades originais, mas o produto escalar, e consequentemente, distâncias e comprimentos, neste espaço são calculados usando (A.2.1). Este espaço é chamado de Espaço Euclideano Ponderado, cujos pesos, neste exemplo, foram considerados como sendo as inversas das variâncias.

Vamos abordar a ponderação num espaço Euclideano multidimensional. Em geral, o espaço Euclideano ponderado é

definido pelo produto escalar:

$$\mathbf{x}^t \mathbf{D}_q \mathbf{y} = \sum_{i=1}^{J} q_i x_i y_i$$

onde $q_1 \ldots q_j$ são números reais positivos definindo os pesos relativos atribuidos às J dimensões, respectivamente. A distância ao quadrado entre dois pontos **x** e **y** neste espaço é a soma **ponderada** das diferenças das coordenadas ao quadrado:

$$d^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^t \, \mathbf{D}_q (\mathbf{x}-\mathbf{y}) = \sum_{i=1}^{J} q_i (x_i - y_i)^2$$

Este tipo de função distância é frequentemente referida como métrica diagonal.

Se o pré-requisito de bases ortonormais é mantido, não há problema em como definir o produto escalar. A definição de produto escalar Euclideano pode ser aplicada às coordenadas dos vetores expressos em termos da base ortonormal no espaço Euclideano ponderado. Se considerarmos Q qualquer matriz de pesos, positiva definida, o produto escalar entre **x** e **y** no espaço J-dimensional é definido por:

$$\mathbf{x}^t \mathbf{Q}\ \mathbf{y} = \sum_i \sum_{i'} q_{i\,i'}\ x_i\ y_{i'}$$

Vamos expressar **x** e **y** em relação a qualquer base $\mathbf{b}_1 \ldots \mathbf{b}_j$:

$$\mathbf{x} = \sum_i u_i \mathbf{b}_i \qquad\qquad \mathbf{y} = \sum_{i'} v_{i'} \mathbf{b}_{i'}$$

então o produto escalar é dado por:

$$\mathbf{x}^t \mathbf{Q}\ \mathbf{y} = \left(\sum_i u_i \mathbf{b}_i \right)^t \mathbf{Q} \left(\sum_{i'} v_{i'} \mathbf{b}_{i'} \right)$$

$$= \sum_i \sum_{i'} u_i\ v_{i'} \mathbf{b}_i^t\ \mathbf{Q}\ \mathbf{b}_{i'}$$

Se a base $\mathbf{b}_1 \ldots \mathbf{b}_j$ é ortonormal então por definição

$$\mathbf{b}_j^t \mathbf{Q}\, \mathbf{b}_{j'} = 0 \qquad \text{se } j \neq j'$$

$$\mathbf{b}_j^t \mathbf{Q}\, \mathbf{b}_{j'} = 1 \qquad \text{se } j = j'$$

Em geral, diz-se que a base $b_1 \ldots b_j$ é ortonormal na métrica **Q**, e, consequentemente, tem-se:

$$\mathbf{x}^t \mathbf{Q}\, \mathbf{y} = \sum_i u_i v_i$$

ou seja, o produto escalar Euclideano ponderado é simplesmente o produto escalar Euclideano das coordenadas relativas a qualquer base ortonormal (ortonormal no mesmo espaço ponderado).

Algumas técnicas de análise de dados multidimensionais, como a análise de correspondência, lidam com os vetores de frequências relativas (**p**), como pontos num espaço multidimensional. Tais vetores são conhecidos como perfis. Um dos exemplos mais comuns de distâncias Euclideanas ponderadas entre vetores de frequências é a estatística de quiquadrado (χ^2):

$$\chi^2 = (\mathbf{o}-\mathbf{e})^t\, \mathbf{D}_e^{-1}\, (\mathbf{o}-\mathbf{e})$$

o e **e** são frequências observadas e esperadas respectivamente, e $\mathbf{D}_e^{-1}$ é a matriz diagonal das inversas das frequências esperadas.

Se definirmos $\mathbf{p}=(1/n)\mathbf{o}$ e $\bar{\mathbf{p}}=(1/n)\mathbf{e}$ como sendo as frequências relativas observadas e esperadas, respectivamente, onde n é a frequência observada total, então a estatística χ^2 acima é dada por:

$$\chi^2 = n(\mathbf{p}-\bar{\mathbf{p}})^t\, \mathbf{D}_{\bar{p}}^{-1} (\mathbf{p}-\bar{\mathbf{p}}) = n\sum_i (p_i - \bar{p}_i)^2 / \bar{p}_i \qquad \text{(A.2.2)}$$

A distância ao quadrado entre **p** e $\bar{\mathbf{p}}$ é dada por

$(\mathbf{p}-\bar{\mathbf{p}})^t \mathbf{D}_{\bar{p}}^{-1} (\mathbf{p}-\bar{\mathbf{p}})$, uma distância Euclideana ponderada, onde os pesos são as inversas das frequências relativas esperadas. Por causa da proporcionalidade desta distância à estatística de χ^2, esta função distância é denominada de distância de quiquadrado. O fator de proporcionalidade é o tamanho da amostra.

A.3 ATRIBUINDO MASSAS AOS VETORES

Existem muitos métodos de análise estatística que permitem a ponderação de determinadas observações por razões justificáveis. Por exemplo, vamos supor que em uma pesquisa de opinião pública teve-se dificuldades em obter respostas de pessoas do sexo feminino, por razões inerentes à pesquisa. Consequentemente, na análise dos dados a opinião das mulheres estará sub-representada e em qualquer sumário geral dos dados predominará a opinião do sexo masculino. Neste caso, poder-se-ia atribuir pesos maiores às respostas provenientes do sexo feminino para equalizar a contribuição de ambos os sexos.

É necessário termos bem claro quando estamos nos referindo à ponderação de pontos e quando estamos nos referindo à ponderação de dimensões. Para tanto, iremos utilizar o termo massa quando estivermos nos referindo a quantidades que ponderam pontos, e o termo peso a quantidades que ponderam dimensões de um espaço.

No estudo da geometria de um conjunto de vetores, a atribuição de diferentes massas aos vetores é equivalente à

atribuição de diferentes massas aos vetores é equivalente à designação de diferentes graus de importância às posições dos pontos no espaço. Na grande maioria das vezes o objetivo é identificar um subespaço de menor dimensão que fique o mais próximo possível de todos os pontos. Quando estes têm diferentes massas então o subespaço deveria estar mais próximo dos pontos de maiores massas, enquanto que um desvio dos pontos de menores massas seria tolerável.

O centróide de um conjunto de pontos $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_I$ com diferentes massas $w_1, w_2, \ldots, w_I$ é o ponto médio ponderado, isto é:

$$\bar{\mathbf{x}} = \sum_i w_i \mathbf{x}_i / \sum_i w_i \qquad \text{(A.3.1)}$$

Portanto $\bar{\mathbf{x}}$ tende em direção aos pontos de maiores massas.

A estatística χ^2 apresentada em (A.2.2) descreve a distância ao quadrado entre o vetor $\mathbf{p}$ de frequências relativas observadas e o vetor $\bar{\mathbf{p}}$ de frequências relativas esperadas, multiplicada pela frequência observada total n.

Vamos supor que existam s subpopulações de tamanho n_i i=1,...,s com perfil $\mathbf{p}_i = [p_{i1}\ p_{i2} \ldots p_{ik}]^t$. Então para cada subpopulação podemos calcular a estatística χ_i^2:

$$\chi_i^2 = n_i\ (\mathbf{p}_i - \bar{\mathbf{p}})^t\ \mathbf{D}_{\bar{\mathbf{p}}}^{-1}\ (\mathbf{p}_i - \bar{\mathbf{p}}) \qquad \text{(A.3.2)}$$

e para a soma das subpopulações tem-se:

$$\chi^2 = \sum_{i=1}^{s} \chi_i^2 \qquad \text{(A.3.3)}$$

A estatística χ^2 em (A.3.3) pode ser descrita como a soma ponderada de distâncias ao quadrado entre $\mathbf{p}_i$ e $\bar{\mathbf{p}}$, com

Comparando (A.3.3) com (A.3.1), verifica-se que o vetor $\bar{p}$ é o centróide dos vetores de frequências relativas das subpopulações, sendo que cada um é ponderado pelo tamanho da subpopulação. O termo perfil médio é empregado para denominar $\bar{p}$.

Vamos introduzir as seguintes definições:

$n = \sum_i n_i$

$w_i = n_i/n$

$d_i^2 = (p_i - \bar{p})^t \, D_p^{-1} (p_i - \bar{p})$ (distância ao quadrado entre $\bar{p}_i$ e $\bar{p}$, na métrica definida por D_p^{-1})

$in(I) = \chi^2/n$ (inércia total do conjunto de I vetores perfis)

O centróide $\bar{p}$ e a inércia in(I) podem ser expressos como médias ponderadas:

$\bar{p} = \sum_i w_i p_i$

$in(I) = \sum_i w_i d_i^2$

Assim o perfil médio $\bar{p}$ é um vetor que representa o centróide dos perfis individuais, enquanto que a inércia é uma medida de quanto os perfis individuais estão espalhados ao redor do centróide.

A.4 IDENTIFICANDO SUBESPAÇOS ÓTIMOS

O objetivo é identificar o subespaço de menor dimensionalidade que melhor contenha o conjunto de pontos,

dimensionalidade que melhor contenha o conjunto de pontos, isto é, o subespaço que mais se aproxime do conjunto de pontos.

A primeira questão a ser abordada é como definir a proximidade de um conjunto de pontos a um subespaço dado. A definição de distância entre dois pontos quaisquer já foi apresentada. Intuitivamente, a distância entre um ponto e um subespaço dado é a menor distância entre o ponto e todos os pontos contidos no subespaço. Assim, a proximidade de um conjunto de pontos a um subespaço pode ser definida como a média, ou a média ponderada, do correspondente conjunto das menores distâncias. Por razões de simplicidade algébrica, bem como por conveniências geométricas, a medida de proximidade é baseada nas distâncias ao quadrado.

A figura (A.3) mostra uma nuvem de pontos num espaço Euclideano ponderado J-dimensional, com um subespaço de menor dimensionalidade k,desenhado esquematicamente como um plano cortando o espaço. Para um ponto típico $\mathbf{y}_i$, $\hat{\mathbf{y}}_i$ representa o ponto no subespaço o qual é o mais próximo de $\mathbf{y}_i$; a distância mínima entre eles sendo igual a d_i. Se $\mathbf{y}_i$ é ponderado pela massa w_i (i=1,...,I), então a definição de proximidade do conjunto inteiro de pontos ao subespaço S é:

$$\psi\ (S;\mathbf{y}_1 \ldots \mathbf{y}_I) = \sum_i w_i d_i^2 \qquad \text{(A.4.1)}$$

onde

$$d_i^2 = \| \mathbf{y}_i - \hat{\mathbf{y}}_i \|^2 \mathbf{D}_q = (\mathbf{y}_i - \hat{\mathbf{y}}_i)^t \mathbf{D}_q (\mathbf{y}_i - \hat{\mathbf{y}}_i)$$

e $\mathbf{D}_q$ é uma matriz diagonal de pesos (da dimensão) positivos. A distância ao quadrado, d_i^2, depende do subespaço S, e o

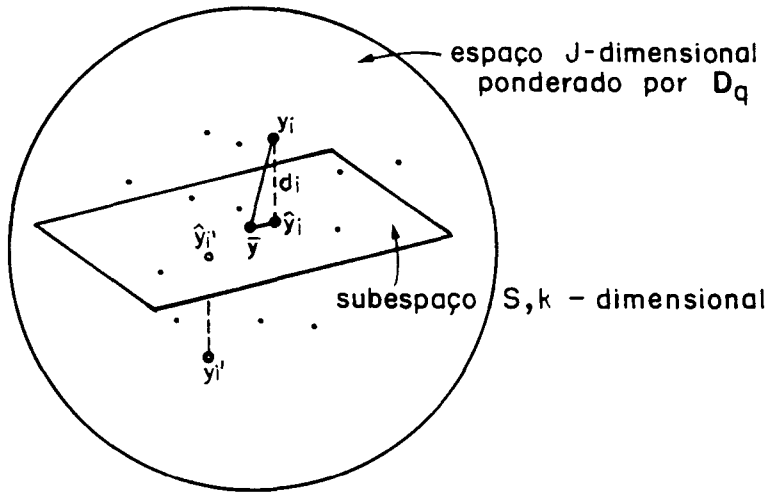

Figura A.3 - *Pontos num espaço multidimensional e suas projeções num subespaço, representado por um plano.*

objetivo é, assim, encontrar o subespaço S^* o qual minimiza a função ψ em (A.4.1).

Vamos imaginar um ponto s como um subespaço zero dimensional. A função (A.4.1) torna-se:

$$\psi\ (\mathbf{s};\mathbf{y}_1 \ldots \mathbf{y}_I) = \sum_i w_i (\mathbf{y}_i - \mathbf{s})^t\ \mathbf{D}_q\ (\mathbf{y}_i - \mathbf{s})$$

já que $\hat{\mathbf{y}}_i$ é igual a s para todo i.

O centróide $\bar{\mathbf{y}}$ é o ponto que minimiza essa função (resultado facilmente verificado por colocar cada derivada da função em relação aos elementos de s igual a zero). Assim, o centróide é, neste sentido, o ponto mais próximo de todos os pontos $\mathbf{y}_1 \ldots \mathbf{y}_I$ dados.

Pode-se mostrar que, na busca por um subespaço k -dimensional "ótimo", necessitamos apenas considerar os subespaços S que contenham $\bar{\mathbf{y}}$, e consequentemente podemos desenhar $\bar{\mathbf{y}}$ no subespaço candidato da figura (A.4.1). Portanto,

qualquer subespaço S que é ótimo no sentido de minimizar (A.4.1) deve incluir o centróide, com o resultado que nós podemos restringir as aproximações $\hat{\mathbf{y}}_i$ dos pontos $\mathbf{y}_i$ como sendo da seguinte forma:

$$\hat{\mathbf{y}}_i = \bar{\mathbf{y}} + \sum_{s=1}^{k} f_{is}\mathbf{v}_s$$

onde $\mathbf{v}_1 \ldots \mathbf{v}_k$ são vetores bases do subespaço. A função (A.4.1) para ser minimizada pode ser escrita como:

$$\psi\ (S;\mathbf{y}_1 \ldots \mathbf{y}_I) = \sum_i w_i(\mathbf{y}_i - \bar{\mathbf{y}} - \sum_{s=1}^{k} f_{is}\mathbf{v}_s)^t\ \mathbf{D}_q\ (\mathbf{y}_i - \bar{\mathbf{y}} - \sum_{s=1}^{k} f_{is}\mathbf{v}_s) \qquad (A.4.2)$$

As variáveis desta função objetivo são os k eixos $\mathbf{v}_1 \ldots \mathbf{v}_k$, implicando num total de Jk variáveis escalares.

A solução teórica completa para o problema de minimizar (A.4.2) para uma dimensionalidade k especificada está embutida nos conceitos de decomposição em valores singulares e aproximação de uma matriz por matriz de posto menor, apresentados no capítulo 2.

APÊNDICE B . Tabelas

Tabela B1 Valores Esperados de Autovalores, para Desvios Aleatórios Normais (Valores M_i)

i-m	n								
	4	5	6	7	8	10	12	16	20
1-4	6,45	8,47	9,86	11,61	12,88	15,08	17,96	23,33	28,20
1-5	8,47	10,37	11,82	13,59	14,75	18,11	21,01	26,29	31,87
1-6	9,86	11,82	13,35	15,44	16,92	20,42	22,85	28,87	36,78
1-7	11,61	13,59	15,44	17,18	18,91	22,89	25,59	31,83	37,46
1-8	12,88	14,75	16,92	18,91	20,72	24,15	27,67	34,52	40,70
1-10	15,08	18,11	20,42	22,89	24,15	27,81	31,88	38,81	45,11
1-12	17,96	21,01	22,85	25,59	27,67	31,88	35,39	42,76	50,00
1-16	23,33	26,29	28,87	31,83	34,52	38,81	42,76	51,08	58,11
1-20	28,20	31,87	36,78	37,46	40,70	45,11	50,00	58,11	66,34
1-32	42,86	46,77	50,83	53,77	57,50	63,21	68,60	78,62	88,71
1-50	63,85	68,93	73,74	77,62	81,68	88,66	95,11	107,41	117,36
1-100	120,90	126,58	132,94	138,45	143,64	153,00	161,58	176,32	192,08
2-4	1,97	3,04	4,01	5,06	6,06	7,81	9,98	14,08	17,90
2-5	3,04	4,25	5,45	6,72	7,78	10,10	12,50	16,80	21,27
2-6	4,01	5,45	6,96	8,21	9,68	12,15	14,60	19,51	25,54
2-7	5,06	6,72	8,21	9,77	11,21	14,26	16,77	22,18	27,02
2-8	6,06	7,78	9,68	11,21	12,73	15,70	18,72	24,26	29,77
2-10	7,81	10,10	12,15	14,26	15,70	19,21	22,12	28,60	33,86
2-12	9,98	12,50	14,60	16,77	18,72	22,12	25,28	32,43	38,32
2-16	14,08	16,80	19,51	22,18	24,26	28,60	32,43	39,68	46,05
2-20	17,90	21,27	25,54	27,02	29,77	33,86	38,32	46,05	53,76
2-32	29,99	34,01	38,38	41,15	44,63	50,22	55,54	65,32	74,28
2-50	48,10	53,54	58,35	62,61	66,57	73,16	80,12	91,31	101,51
2-100	98,31	105,69	112,21	118,49	123,48	133,58	142,62	156,95	171,26
3-4	0,27	0,67	1,15	1,72	2,26	3,39	4,87	7,74	10,66
3-5	0,67	1,38	2,22	3,08	3,72	5,42	7,11	10,65	14,34
3-6	1,15	2,22	3,17	4,15	5,23	7,28	9,15	13,18	18,06
3-7	1,72	3,08	4,15	5,44	6,48	8,95	11,05	15,55	19,69
3-8	2,26	3,72	5,23	6,48	7,76	10,39	12,80	17,61	22,35
3-10	3,39	5,42	7,28	8,95	10,39	13,38	16,03	21,33	26,30
3-12	4,87	7,11	9,15	11,05	12,80	16,03	18,85	25,03	30,50
3-16	7,74	10,65	13,18	15,55	17,61	21,33	25,03	31,79	37,73
3-20	10,66	14,34	18,06	19,69	22,35	26,30	30,50	37,73	45,04
3-32	20,14	25,11	29,34	32,34	35,65	41,24	46,31	55,84	64,41
3-50	35,12	41,62	47,57	51,70	55,72	62,49	68,76	80,16	90,24
3-100	79,50	89,40	97,12	103,02	109,26	119,49	127,74	143,20	156,38

Fonte: Mandel,1969.

Tabela B2 Desvios Padrão de Autovalores, para Desvios Aleatórios Normais

i-m	n								
	4	5	6	7	8	10	12	16	20
1-4	3,24	3,83	4,05	4,36	4,52	4,72	5,26	5,71	6,25
1-5	3,83	4,18	4,12	4,28	4,71	5,12	5,45	5,84	6,40
1-6	4,05	4,12	4,30	4,88	4,71	5,64	5,51	6,09	6,92
1-7	4,36	4,28	4,88	4,80	5,16	5,28	5,64	6,03	6,51
1-8	4,52	4,71	4,71	5,16	5,26	5,30	6,02	6,49	6,88
1-10	4,72	5,12	5,64	5,28	5,30	5,50	6,41	6,25	6,89
1-12	5,26	5,45	5,51	5,64	6,02	6,41	6,65	6,57	7,53
1-16	5,71	5,84	6,09	6,03	6,49	6,25	6,57	7,10	7,60
1-20	6,25	6,40	6,92	6,51	6,88	6,89	7,53	7,60	7,87
1-32	7,38	7,59	7,96	7,89	7,98	8,02	8,26	8,37	8,77
1-50	9,55	9,55	9,44	8,88	9,04	9,36	9,77	9,86	10,16
1-100	12,96	12,22	12,50	11,84	11,96	11,67	11,59	11,95	12,19
2-4	1,39	1,77	2,08	2,25	2,62	2,77	3,17	3,56	4,11
2-5	1,77	2,04	2,29	2,48	2,70	3,11	3,43	3,95	4,32
2-6	2,08	2,29	2,61	2,83	2,94	3,36	3,68	4,00	4,46
2-7	2,25	2,48	2,83	3,04	3,04	3,36	3,77	4,20	4,72
2-8	2,62	2,70	2,94	3,04	3,17	3,43	3,82	4,26	4,75
2-10	2,77	3,11	3,36	3,36	3,43	3,99	4,09	4,66	5,00
2-12	3,17	3,43	3,68	3,77	3,82	4,09	4,10	4,85	4,93
2-16	3,56	3,95	4,00	4,20	4,26	4,66	4,85	5,30	5,55
2-20	4,11	4,32	4,46	4,72	4,75	5,00	4,93	5,55	5,83
2-32	5,46	5,24	5,60	5,44	5,52	5,82	5,88	6,27	6,20
2-50	6,93	7,09	6,79	6,96	6,61	6,92	7,08	7,36	7,20
2-100	10,15	9,59	9,08	9,56	9,04	9,23	8,90	9,27	9,41
3-4	0,37	0,69	0,92	1,20	1,35	1,71	2,13	2,80	3,14
3-5	0,69	0,97	1,23	1,56	1,66	1,86	2,30	2,90	3,36
3-6	0,92	1,23	1,53	1,75	2,00	2,21	2,46	3,08	3,44
3-7	1,20	1,56	1,75	2,00	2,07	2,43	2,66	3,32	3,62
3-8	1,35	1,66	2,00	2,07	2,31	2,62	2,80	3,19	3,56
3-10	1,72	1,86	2,21	2,43	2,62	2,95	3,11	3,53	3,68
3-12	2,13	2,30	2,46	2,66	2,80	3,11	3,34	3,60	3,87
3-16	2,80	2,90	3,08	3,32	3,19	3,53	3,60	4,17	4,45
3-20	3,14	3,36	3,44	3,62	3,56	3,68	3,87	4,45	4,76
3-32	4,59	4,29	4,53	4,55	4,60	4,52	4,98	5,09	5,38
3-50	6,54	5,64	5,68	5,67	5,84	5,78	5,77	5,87	6,01
3-100	9,21	8,27	8,12	7,77	7,98	7,83	7,36	8,11	8,05

Fonte: Mandel,1969.

Tabela B3 Valores de M_i Expressos como Porcentagem do Total de Graus de Liberdade da Interação

i-m	n								
	4	5	6	7	8	10	12	16	20
1-4	74,54	69,42	65,66	62,91	60,68	57,64	54,80	51,63	49,72
1-5	69,42	63,83	59,06	55,83	53,67	50,46	47,74	44,18	41,89
1-6	65,66	59,06	54,05	51,57	48,48	44,96	42,26	38,85	36,74
1-7	62,91	55,83	51,57	47,70	45,50	41,47	38,84	35,07	32,96
1-8	60,68	53,67	48,48	45,50	42,79	38,66	36,07	32,61	30,30
1-10	57,64	50,46	44,96	41,47	38,66	34,55	32,36	28,73	26,54
1-12	54,80	47,74	42,26	38,84	36,07	32,36	29,96	25,90	24,08
1-16	51,63	44,18	38,85	35,07	32,61	28,73	25,90	22,59	20,62
1-20	49,72	41,89	36,74	32,96	30,30	26,54	24,08	20,62	18,27
1-32	46,09	38,12	32,71	29,41	26,42	22,77	20,13	16,96	14,97
1-50	43,39	35,36	29,98	26,36	23,73	20,14	17,69	14,53	12,66
1-100	40,44	32,14	26,96	23,40	20,82	17,15	14,83	11,90	10,21
2-4	22,39	25,06	26,69	27,67	28,61	29,56	30,38	31,21	31,51
2-5	25,06	26,34	27,25	27,75	28,13	28,13	28,40	28,17	27,96
2-6	26,69	27,25	28,09	27,41	27,72	26,85	26,86	26,17	25,55
2-7	27,67	27,75	27,41	27,13	26,86	26,04	25,44	24,50	23,77
2-8	28,61	28,13	27,72	26,86	26,17	25,05	24,36	22,99	22,24
2-10	29,56	28,13	26,85	26,04	25,05	23,83	22,41	21,18	19,88
2-12	30,38	28,41	26,86	25,44	24,36	22,41	21,16	19,65	18,40
2-16	31,21	28,17	26,17	24,50	22,99	21,18	19,65	17,58	16,28
2-20	31,51	27,96	25,55	23,77	22,24	19,88	18,40	16,28	14,83
2-32	32,25	27,62	24,73	22,37	20,53	18,06	16,30	14,08	12,56
2-50	32,72	27,44	23,76	21,28	19,36	16,61	14,89	12,38	10,93
2-100	32,93	26,80	22,72	20,01	17,87	14,98	13,09	10,59	9,10
3-4	3,07	5,52	7,65	9,42	10,71	12,80	14,82	17,16	18,77
3-5	5,52	8,56	11,08	12,75	13,41	15,09	16,15	17,83	18,86
3-6	7,65	11,08	12,76	13,86	14,96	16,10	16,77	17,66	18,08
3-7	9,42	12,75	13,86	15,12	15,50	16,38	16,76	17,19	17,30
3-8	10,71	13,41	14,96	15,50	15,94	16,57	16,65	16,71	16,72
3-10	12,80	15,09	16,10	16,38	16,57	16,58	16,23	15,79	15,42
3-12	14,82	16,15	16,77	16,76	16,65	16,23	15,76	15,16	14,63
3-16	17,16	17,83	17,66	17,19	16,71	15,79	15,16	14,09	13,32
3-20	18,77	18,86	18,08	17,30	16,72	15,42	14,63	13,32	12,44
3-32	21,66	20,38	18,90	17,56	16,41	14,82	13,59	12,03	10,90
3-50	23,89	21,31	19,37	17,57	16,21	14,18	12,77	10,88	9,71
3-100	26,63	22,65	19,66	17,38	15,81	13,40	11,73	9,66	8,31

Fonte: Mandel,1969.

Tabela B4 Valores Críticos para $\lambda_1^2 / \sum_{i=1}^{n-1} \lambda_i^2$

m	n									
	3	4	5	6	7	8	10	12	16	20
					10%					
3	,9975*									
4	,9743*	,8349*								
5	,9429*	,8458*	,8021							
6	,9135*	,8130*	,6975	,6398						
7	,8879*	,7631*	,6548	,6358	,5687					
8	,8660*	,7435*	,6487	,5725	,5462	,5098				
10	,8308*	,6749*	,6057	,5570	,4972	,4489	,3982			
12	,8037*	,6594*	,5695	,5001	,4563	,4289	,3830	,3506		
16	,7647*	,6022*	,5131	,4550	,4137	,3815	,3268	,2952	,2562	
20	,7376*	,5737*	,4843	,4329	,3778	,3502	,3023	,2765	,2326	,2049
32	,6886*	,5161*	,4306	,3740	,3329	,2992	,2556	,2257	,1881	,1654
50	,6512*	,4887*	,3978	,3362	,2913	,2624	,2226	,1960	,1598	,1386
100	,6071*	,4421*	,3490	,2942	,2533	,2255	,1815	,1595	,1279	,1001
					5%					
3	,9994*									
4	,9873*	,8567*								
5	,9648*	,8811*	,8407							
6	,9406*	,8505*	,7294	,6681						
7	,9168*	,8003*	,6823	,6703	,5957					
8	,8974*	,7811*	,6815	,5985	,5733	,5345				
10	,8630*	,7043*	,6361	,5901	,5096	,4680	,4143			
12	,8357*	,6936*	,5979	,5242	,4774	,4501	,4016	,3665		
16	,7950*	,6295*	,5356	,4760	,4227	,3991	,3390	,3064	,2656	
20	,7661*	,6290*	,5054	,4542	,3932	,3052	,3139	,2876	,2408	,2117
32	,7127*	,5349*	,4469	,3894	,3454	,3105	,2644	,2335	,1938	,1702
50	,6713*	,5078*	,4127	,3482	,3002	,2706	,2296	,2020	,1643	,1423
100	,6218*	,4610*	,3583	,3024	,2595	,2311	,1892	,1629	,1306	,1116
					1%					
3	,99997*									
4	,9975*	,8930*								
5	,9883*	,9303*	,9004							
6	,9743*	,9082*	,7825	,7194						
7	,9587*	,8619*	,7325	,7325	,6457					
8	,9429*	,8446*	,7407	,6470	,6243	,5809				
10	,9135*	,7575*	,6924	,6516	,5523	,5044	,4452			
12	,8879*	,7411*	,6514	,5702	,5170	,4911	,4372	,3969		
16	,8472*	,6256*	,5788	,5167	,4560	,4331	,3628	,3372	,2837	
20	,8164*	,5966*	,5462	,4955	,4229	,3945	,3364	,3095	,2567	,2249
32	,7571*	,5367*	,4788	,4198	,3700	,3326	,2818	,2485	,2049	,1796
50	,7089*	,5043*	,4423	,3722	,3178	,2864	,2430	,2139	,1731	,1495
100	,6498*	,4463*	,3771	,3189	,2720	,2421	,1977	,1698	,1359	,1159

*:indica que o ponto crítico é exato.

Fonte: Johnson e Graybill,1972.

Tabela B5 Valores Esperados do Primeiro Autovalor - Exatos e Aproximados por Monte Carlo

n	m	Média[a]		Variância[b]	
		Exata	Estimada	Exata	Estimada
3	3	3,57	3,55	6,68	7,18
3	4	5,00	5,03	9,68	9,49
3	5	6,36	6,34	11,68	9,99
3	6	7,33	7,53	13,68	12,67
3	7	8,94	8,96	15,68	15,37
3	8	10,20	10,15	17,68	16,08
3	10	12,66	12,77	20,68	21,53
3	12	15,06	15,09	24,68	25,00
3	16	19,77	19,58	31,68	31,92
3	20	24,39	24,10	38,68	36,60
3	32	37,92	37,88	58,68	52,13
3	50	57,73	57,85	88,68	82,26
3	100	111,44	111,22	167,68	168,48
4	5	8,33	8,33	13,68	14,67
4	7	11,32	11,32	17,68	19,01

[a] cada valor deve ser multiplicado por σ^2.

[b] cada calor deve ser multiplicado por σ^4.

Fonte: Johnson e Graybill, 1972.

APÊNDICE C . Programas

C1. PROGRAMA UTILIZADO PARA O AJUSTE DO MODELO DE MANDEL

```
OPTIONS LS=80 PS=60 ;
************************************************************
*                                                          *
* ESTE PROGRAMA FAZ USO INICIALMENTE DO PROC GLM PARA AJUSTAR*
* UM MODELO ADITIVO E POSTERIORMENTE DO PROC IML PARA AJUSTAR*
* AOS RESIDUO DO MODELO ADITIVO UM MODELO MULTIPLICATIVO    *
*                                                          *
*                                                          *
* INFORMACOES QUE DEVEM SER FORNECIDAS:                    *
*           M = NUMERO DE NIVEIS DO PRIMEIRO FATOR         *
*           N = NUMERO DE NIVEIS DO SEGUNDO FATOR          *
*           DSN = SAS DATA SET                             *
*           FATOR1 = VARIAVEL QUE REPRESENTA O PRIMEIRO FATOR*
*           FATOR2 = VARIAVEL QUE REPRESENTA O SEGUNDO FATOR *
*           Y = VARIAVEL RESPOSTA                          *
*                                                          *
*__________________________________________________________*
;
%MACRO MANDEL(M,N,DSN,FATOR1,FATOR2,Y);
LIBNAME DISCO 'A:\';
*__________________________________________________________*
   A J U S T E   D O   M O D E L O   A D I T I V O
*__________________________________________________________*
;
DATA &DSN ; SET DISCO.&DSN ;
PROC GLM DATA = &DSN ORDER=DATA ;
     CLASSES  &FATOR1 &FATOR2 ;
     MODEL    &Y = &FATOR1 &FATOR2 ;
     OUTPUT   OUT = RES  R = R1 ;
TITLE1 " AJUSTE DO MODELO ADITIVO " ;
DATA RESIDUO ; SET RES   /* RESIDUOS DO MODELO ADITIVO */ ;
     KEEP R1 ;
*__________________________________________________________*
 A J U S T E   D O   M O D E L O   M U L T I P L I C A T I V O
*__________________________________________________________*
;
PROC IML ;
```

```
USE RESIDUO ;
READ ALL INTO R1 ;
A = R1[1:&N,1] ;      /* MATRIZ R1(MN,1) TRANSFORMADA EM */
R = A`;               /* R (M,N) */
DO I=2 TO &M ;
   C =R1[(I-1)*&N+1:I*&N,1]`;
   R =R // C ;
END ;
PRINT "MATRIZ DOS RESIDUOS" , R ;
CALL SVD(U,Q,V,R) ;   /* DECOMPOSICAO EM VALORES SINGULARES */
;                     /* DA MATRIZ R */
D = Q # Q ;
PRINT "AUTO VETORES DE RR`" , U ;
PRINT "VALORES SINGULARES DE R " , Q ;
PRINT "AUTO VALORES DE R  " , D ;
PRINT "AUTO VETORES DE R`R" , V ;
*_____________________________________________________________*
  TESTE DA RAZAO DE MAXIMA VEROSSIMILHANCA PARA A HIPOTESE DE
  ADITIVIDADE DO MODELO
*_____________________________________________________________*
;
D1= SUM(D) ;
D2= SUM(D[2:3,1]);
U1= D[1,1]/D1 ;    /* ESTATISTICA PARA HIPOTESE H0:TETA=0 */
PRINT "ESTATISTICA MV PARA H0:TETA=0" , U1 ;
QUIT ;
RUN ;
%MEND MANDEL ;
%MANDEL(12,3,ALFACE,TEMP_SEM,TRAT_AG,PESO)
RUN ;
```

C2. PROGRAMA UTILIZADO PARA O BIPLOT

```
OPTIONS LS=80 PS=60 ;
FILENAME GRAFICO 'C:\BIPLOT';
GOPTIONS DEVICE=HPLJS2  NODISPLAY GSFNAME=GRAFICO NOCHARACTERS
GSFMODE=REPLACE ;
*************************************************************
*                                                           *
* ESTE PROGRAMA FAZ A DECOMPOSICAO EM VALORES SINGULARES DE  *
* UMA MATRIZ (Y=UQV) E POSTERIORMENTE REPRESENTA GRAFICAMENTE*
* AS LINHAS E COLUNAS DA MATRIZ, UTILIZANDO A FATORACAO      *
* (4.2.4)                                                   *
*                                                           *
* INFORMACOES A SEREM FORNECIDAS :                          *
*        M = NUMERO DE LINHAS DA MATRIZ Y                   *
*        N = NUMERO DE COLUNAS DA MATRIZ Y                  *
*        DSN = SAS DATA SET                                 *
*        LINHA = VARIAVEL QUE CONTEM IDENTIFICACAO DA LINHA  *
*        COLUNA = VARIAVEL QUE CONTEM IDENTIFICACAO DA COLUNA*
*        Y = VARIAVEL QUE CONTEM OS ELEMENTOS DE Y          *
*                                                           *
* OBSERVACOES:                                              *
*   1.AS VARIAVEIS LINHA E COLUNA DEVEM SER DEFINIDAS COMO   *
*     CARACTER                                              *
*   2.O SAS DATA SET DEVE CONTER APENAS AS VARIAVEIS: LINHA, *
*     COLUNA E Y, E ESTAR ORDENADO PELAS VARIAVEIS LINHA  E  *
*     COLUNA                                                *
*                                                           *
*************************************************************
;
%MACRO BIPLOT(M,N,DSN,LINHA,COLUNA,Y);
LIBNAME DISCO 'C:\';
*___________________________________________________________*
 DECOMPOSICAO EM VALORES SINGULARES DE Y
*___________________________________________________________*
;
PROC IML ;
```

```
USE DISCO.&DSN ;
   READ ALL INTO A ;
B = A[1:&N,1] ;      /* MATRIZ A(MN,1) TRANSFORMADA EM */
Y = B`;              /* Y (M,N) */
DO I=2 TO &M ;
   C =A[(I-1)*&N+1:I*&N,1]`;
   Y =Y // C ;
END ;
CALL SVD(U,Q,V,Y) ;   /* DECOMPOSICAO EM VALORES SINGULARES */
;                     /* DA MATRIZ Y */
D = Q # Q ;
Q1 = Q[1,1] ;
U1 = U[1:&M,1] ;
V1 = V[1:&N,1] ;
RE = Y-Q1#U1#(V1)`; /* MATRIZ DOS RESIDUOS APOS SUBTRACAO  */
                    /* DA PRIMEIRA COMPONENTE SINGULAR    */ ;
PRINT "AUTO VETORES DE YY`" , U ;
PRINT "VALORES SINGULARES DE Y " , Q ;
PRINT "AUTO VALORES DE Y  " , D ;
PRINT "AUTO VETORES DE Y`Y" , V ;
PRINT "MATRIZ DOS RESIDUOS" , RE ;
PRINT "MATRIZ Y" , Y ;
*_____________________________________________________________*
 FORMACAO DO SAS DATA SET QUE CONTEM AS COORDENADAS
*_____________________________________________________________*
;
E = SQRT(Q) ;            /* FATORACAO (4.2.4)     */;
F1 = E[1,1]#U[1:&M,2] ;
F2 = E[2,1]#U[1:&M,3] ;
G1 = E[1,1]#V[1:&N,2] ;
G2 = E[2,1]#V[1:&N,3] ;
W = F1||F2 ;
Z = G1||G2 ;
CREATE BIPLOT1 FROM W ;
USE BIPLOT1 ;
APPEND FROM W ;
CREATE BIPLOT2 FROM Z ;
USE BIPLOT2 ;
```

```
APPEND FROM Z ;
QUIT ;
DATA LABELL ; SET DISCO.&DSN ;
   IDENTIF=&LINHA;
PROC SORT   ; BY IDENTIF;
DATA LABELL ; SET LABELL ;
     BY IDENTIF;
     IF FIRST.IDENTIF THEN OUTPUT ;
     KEEP IDENTIF ;
DATA LABELC ; SET DISCO.&DSN ;
   IDENTIF=&COLUNA;
PROC SORT   ; BY IDENTIF ;
DATA LABELC ; SET LABELC ;
     BY IDENTIF;
     IF FIRST.IDENTIF THEN OUTPUT ;
     KEEP IDENTIF ;
DATA COOR1 ; MERGE BIPLOT1 LABELL ;
     FUNCTION='DRAW' ;
     COL1=-COL1 ;
     LINE=1 ; SIZE=2 ;
DATA COOR2 ; MERGE BIPLOT2 LABELC ;
     FUNCTION='DRAW' ;
     COL1=-COL1 ;
     LINE=2 ; SIZE=2 ;
DATA LABEL1 ; MERGE BIPLOT1 LABELL ;
     FUNCTION='LABEL' ;
     SIZE=1 ; STYLE="SWISS" ;
DATA LABEL2 ; MERGE BIPLOT2 LABELC ;
     FUNCTION='LABEL' ;
     SIZE=0.7 ; STYLE="SWISSL";
DATA LABEL; SET LABEL1 LABEL2;
     COL1=-COL1;
     SIZE=1 ; STYLE="SWISS";
     X=COL1 ; Y=COL2;
     TEXT=IDENTIF;
     DROP COL1 COL2 IDENTIF;
DATA COOR3 ; SET LABELL LABELC;
     FUNCTION='MOVE' ;
```

```
     COL1=0 ; COL2=0 ;
RUN ;
*_____________________________________________________________*
 CONSTRUCAO DO GRAFICO
*_____________________________________________________________*
;
DATA COOR ; SET COOR3 COOR1 COOR2 ;
     Y = COL2 ;  /* VARIAVEL DO EIXO Y */
     X = COL1 ;  /* VARIAVEL DO EIXO X */
     LABEL Y = 'DIMENSAO 2'
           X = 'DIMENSAO 1';
     DROP COL1 COL2 ;
PROC SORT; BY IDENTIF;
DATA COOR;
     LENGTH TEXT $ 8 FUNCTION $ 5;
     SET COOR LABEL;
     XSYS = '2' ;
     YSYS = '2' ;
     DROP IDENTIF;
PROC PRINT DATA=COOR ;
RUN ;
PROC GPLOT DATA=COOR ;
     SYMBOL1 V=NONE ;
     AXIS1  LENGTH=10 CM  WIDTH=2 ORDER=-5 TO 5 BY 1
            VALUE=(F=SWISSL) LABEL=(F=SWISS);
     AXIS2  LENGTH=10 CM  WIDTH=2 ORDER=-5 TO 5 BY 1
            VALUE=(F=SWISSL) LABEL=(F=SWISS);
     PLOT Y*X=1 / ANNOTATE=COOR FRAME HAXIS=AXIS1 VAXIS=AXIS2
     VREF=0  HREF=0 ;
%MEND BIPLOT ;
%BIPLOT(6,5,DIALISE,CIDADE,MARCADOR,PORC)
RUN ;
```

C3. PROGRAMA UTILIZADO PARA A ANÁLISE DE CORRESPONDÊNCIA

```
OPTIONS LS=80 PS=60 ;
GOPTIONS DEVICE=FX85;
*****************************************************************
*                                                               *
*  ESTE PROGRAMA FAZ USO  DO  PROC  CORRESP  PARA  REALIZAR     *
*  ANALISE DE CORRESPONDENCIA SIMPLES DE DADOS EXTRAIDOS DE     *
*  EVERITT(1977, PAG 95)                                        *
*                                                               *
*****************************************************************
;
DATA DADOS ;
     INPUT COL$  PRES$  FREQ @@ ;
     CARDS ;
     1  A 117   1  B 121   1  C   47    1  D   22
     2  A  85   2  B  98   2  C   43    2  D   20
     3  A 119   3  B 209   3  C   68    3  D   43
     4  A  67   4  B  99   4  C   46    4  D   33
     ;
PROC CORRESP ALL DATA=DADOS OUTF=FREQ OUTC=COOR ;
*
* OUTC=COOR CRIA UM SAS DATA SET CONTENDO  AS COORDENADAS
* DOS PONTOS-LINHAS E DOS PONTOS-COLUNAS
* OUTF=FREQ  CRIA  UM  SAS  DATA  SET  CONTENDO TABELA DE
* CONTINGENCIA, PERFIL DAS LINHAS E DAS COLUNAS E VALORES
* ESPERADOS
;
     TABLES COL,PRES ;
     WEIGHT FREQ ;
     RUN ;
*---------------------------------------------------------------*
* PREPARACAO DOS RESULTADOS DA ANALISE DE CORRESPONDENCIA       *
* PARA A CONSTRUCAO DO GRAFICO - USO  DO  PROC  GPLOT  E        *
* ANNOTATE FACILITY                                             *
*---------------------------------------------------------------*
;
```

```
DATA COOR1 ; SET COOR ;
     IF _TYPE_='OBS';
     FUNCTION='DRAW';
     LINE=1 ; SIZE=2 ;
DATA COOR2 ; SET COOR ;
     IF _TYPE_='VAR';
     FUNCTION='DRAW';
     LINE=2 ; SIZE=2 ;
DATA COOR3 ; SET COOR ;
     IF _TYPE_='OBS' OR _TYPE_='VAR' ;
     FUNCTION='MOVE' ;
     DIM1=0 ; DIM2=0 ;
DATA COOR4 ; SET COOR3 COOR1 COOR2;
PROC SORT DATA=COOR4 ; BY _NAME_ ;

DATA LABEL ; SET COOR ;
     IF _TYPE_='OBS' OR _TYPE_='VAR' ;
     FUNCTION='LABEL';
     SIZE=1 ; STYLE='SWISS' ;

DATA COOR4 ;
     LENGTH FUNCTION $ 5 ;
     SET   COOR4 LABEL ;
     LABEL X='DIMENSAO 1'
           Y='DIMENSAO 2' ;
     X=DIM1    ;  Y=DIM2 ;
     XSYS='2'  ;  YSYS='2' ;
     TEXT=_NAME_;
     KEEP FUNCTION TEXT X Y SIZE STYLE XSYS YSYS LINE ;
RUN ;
PROC GPLOT DATA=COOR4 ;
     SYMBOL1 V=NONE;
     AXIS1  LENGTH=10 CM ORDER=-.30 TO .30  BY .10 WIDTH=2
            VALUE=(F=SWISSL) LABEL=(F=SWISS);
     PLOT Y*X=1 / ANNOTATE=COOR4 FRAME HAXIS=AXIS1 VAXIS=AXIS1
                  HREF=0 VREF=0 ;
RUN ;
```