APLICAÇÃO DA ÁLGEBRA LINEAR AOS MÉTODOS DE ANÁLISE DISCRIMINANTE

Turíbio José Gomes dos Santos

MONOGRAFIA SUBMETIDA À COORDENAÇÃO DO

CURSO DE PÓS-GRADUAÇÃO EM MATEMÁTICA, COMO REQUISITO PARACIAL

PARA OBTENÇÃO DO GRAU DE MESTRE

UNIVERSIDADE FEDERAL DO CEARÁ

FORTALEZA - 1985

"A MATEMÁTICA é a arte de dar o mesmo nome a coisas diferentes".

(H. Poincaré)

"Sim, era desse jeito outrora; mudamos, porém, tudo isso e agora fazemos Medicina com método inteiramente novo".

(Molière)

À minha esposa Ana e filhos

Cristhiano e Andréa

Aos meus pais Avelino e Lôide

Às minhas irmãs Lélia e Ana

AGRADECIMENTOS

À U.F.Pb, ao PICD e ao CNPq, pelo apoio financeiro.

Ao Professor Airton Fontenele Sampaio Xavier, meu orientador, pela incansável orientação e também pelo exemplo em Climatologia, juntamente com a Professorea Teresinha de M^{a} B. S. Xavier.

Ao Médico Waldir Pedrosa Dias de Amorim pelos dados reais para a realização do exemplo aplicado em Gastroenterologia.

Aos professores e colegas por sugestões que tornaram possível a realização deste trabalho.

Ao Sr. José Alves Ferreira, pelo trabalho datilografia.

página

APRESENTAÇÃO

Esta Monografia tem a finalidade de expor os fundamentos da Análise Discriminante com ênfase nas técnicas de Álgebra Linear.

De fato, aqui, nos restringimos ao caso dos métodos de discriminação com base em variáveis quantitativas. Assim, achamo-nos em presença de um conjunto de indivíduos ou elementos, repartidos num certo número k de classes definidas "a priori"; ademais, supõe-se que para cada indivíduo ou elemento dispõe-se dos resultados de p medidas ali efetuadas (ou equivalentemente, de observações ou valores ali assumidos por p variáveis)

Na discriminação com fins descritivos deseja-se evidenciar o possível poder discriminante das variáveis em causa; ou seja, verificar se as medidas ou observações realizadas justificam a separação segundo as distintas classes consideradas "a priori". Por exemplo, em uma aplicação típica em Medicina, interessa discernir se os resultados de certos exames clínicos ou laboratoriais (expressados quantitativamente) justificam, ou não, a separação de um grupo de pacientes em duas classes, de acordo com as medidas terapêuticas mais indicadas:

i) a classe dos pacientes para os quais está reservada uma conduta cirúrgica.

ii) a classe dos pacientes para os quais a melhor conduta envolve um tratamento medicamentoso.

A essa "etapa descritiva", por sua vez, pode seguir-se uma "etapa decisional", ou discriminação com fins decisionais ou de identificação, que se destina a se realizar a atribuição de cada novo indivíduo, a

uma das classes, sob o menor risco possível de atribuição incorreta.

Com relação ao exemplo precedente, que se relaciona com a indicação da melhor terapêutica, essa escolha se impõe desde que se apresenta um novo paciente.

Note-se que a Análise Discriminante contrapõe-se aos chamados "métodos de classificação", segundo os quais não existem classes determinadas "a priori".

Os métodos de Análise Discriminante têm sido utilizados em diversos domínios da pesquisa aplicada.

a) Antropologia

Discriminação entre as diversas castas da India com base em dados antropométricos, conforme referido por **Rao** (1952).

b) **Política**

Discriminação entre duas facções de parlamentares do Partido Liberal britânico no período 1874 - 1855 (facções radical e não-radical) com bases nos votos atribuídos pelos parlamentares a determinadas moções (+1 = voto a favor; 0 = voto branco; -1 = voto contra); este exemplo, devido a **Heyck & Kleck** (1973), encontra-se relatado por **Nil** et al. (1975). Um exemplo semelhante segundo dados do Laboratório do prof. Benzecri diz respeito a posições políticas de deputados da III^{a} República (França) sendo apresentado por **Romeder** (1972).

c) Psiquiatria

Discriminação entre três grupos: normais, nevrosados e esquizofrênicos, com bases numa escala de ansiedade (30 sintomas), conforme Nakache et al (1971).

d) Medicina Clínica

Um exemplo clássico diz respeito à discriminação de pacientes ictéricos, em dois grupos: ictéricos cirúrgicos (necessitando cirurgia, por exemplo, para extração de cálculos biliares) e icterícias médicas (pacientes que se beneficiam de terapêutica exclusivamente medicamentosa e/ou dietética), tendo como base os resultados de exames clínicos, laboratoriais, radiológicos, etc; encontra-se desenvolvido com detalhes em Romeder (1972).

Outro exemplo ainda no campo de medicina, refere-se à discriminação de doentes com infarto do miocárdio, em dois grupos: sobreviventes e mortos, em seguida ao tratamento, com base em exames realizados antes de ser instituído o tratamento específico, conforme Lorente & Nakache (1977), também referido por Nakache (1978).

e) Geografia Agrária

Discriminação entre áreas especializadas na produção de trigo e áreas de produção de grãos (milho, etc...) em consórcio com a criação de gado, no Estado de Dakota do Sul (EEUU), com base nos seguintes dados: X_1 = densidade da população rural; X_2 = precipitação média anual; X_3 = porcentual de terras aráveis; X_4 = tamanho médio das propriedades; vide King (1969).

Em nosso trabalho inicialmente (Capítulo 1), apresentamos critérios de discriminação com fins decisionais desenvolvidos por Sebestyen(1936). Conforme são apresentados por Romeder (1972), permitindo atribuir um elemento qualquer a cada uma dentre várias classes definidas "a priori", na dependência da consideração de funções que medem a semelhança de um novo indivíduo a cada classe. Outro critério (Capítulo 2) relaciona-se a uma abordagem com objetivos descritivos envolvendo a determinação dos chamados

eixos fatoriais discriminantes. Tais eixos fatoriais correspondem aos vetores próprios de certa matriz $T^{-1}B$, definida no texto.

Faz-se, outrossim, uma ligação com o método clássico das funções lineares discriminantes introduzidas por Fisher(1970). No apítulo 3, apresentamos a discriminação sob a hipótese de validade de leis normais, após considerações gerais sobre a Lei de Laplace-Gauss a duas dimensões (bem como sua generalização para o caso p-dimensional), sendo abordado o problema da classificação de novos indivíduos, mediante a definição de fronteiras separadoras (hiperplanos e hipersuperfícies), seja sem considerar "custos de erros de classificação", seja considerando tais custos (ou mais precisamente, via a minimização do custo de má classificação).

Por outro lado, no Capítulo 4, abordamos a técnica básica de "discriminação passo a passo". Para esse fim, consideramos a chamada "discriminação passo a passo ascendente"a qual consiste em, dado um certo conjunto de variáveis medidas sobre uma população, restringí-las à "melhor", em seguida às duas "melhores", etc..., no sentido de permitir de cada vez uma melhor discriminação entre elementos pertencentes a classes distintas; ademais, em cada passo, não se põe em causa as variáveis relacionadas nas etapas precedentes. Evidentemente, são considerados testes para julgar o grau de otimização alcançado e decidir em que momento se deve parar o processo (regra de parada).

Finalmente, no Capítulo 5, estudamos um teste multidimensional não-paramétrico para avaliação do poder discriminante de um hiperplano separador com base em técnicas de Análise Combinatória Linear, conforme descrito por Romeder (1972).

No Apêndice I apresentamos um estudo sucinto sobre os operadores E (valor esperado), Var (variância) e Cov (co-variância), para o caso particular de p-uplas e matrizes de dados (valores observados), cuja utilização é requerida no Capítulo 1.

No Apêndice II apresentamos um exemplo de aplicação de métodos discriminantes, conforme sugerido por Xavier & Xavier(1982), para a análise da validade de técnicas de Análise de Fourier no que concerne à identificação de "anos secos" e "anos chuvosos", a partir da série secular de precipitações pluviométricas de Fortaleza-Ceará. Apresentamos também um outro exemplo, aplicado em Gastroenterologia utilizando métodos de Análise Discriminante, com dados cedidos por **Amorim** (1984) oriundos de sua tese de Mestrado em Medicina.

Para concluir esta apresentação, advertimos que não esgotamos absolutamente o problema do estudo de métodos utilizáveis na discriminação com variáveis quantitativas, pois de fato esse é um domínio de estudos bastante rico e complexo.

CAPÍTULO 1

CRITÉRIOS DE DISCRIMINAÇÃO :

ABORDAGEM CONFORME SEBESTYEN

1.1 - INTRODUÇÃO

Neste capítulo, consideramos o problema da discriminação com fins decisionais, ou seja, procura-se estabelecer critérios discriminantes que permitam atribuir um elemento arbitrário a uma dentre várias classes definidas *a priori*.

Os critérios discriminantes a que nos referimos acima dependem de como se defina, convenientemente, funções permitindo medir a semelhança de um novo indivíduo ou elemento, relativamente aos diversos grupos ou classes. Assim, a regra de atribuição consiste em afirmar que o elemento pertence à classe com relação à qual sua semelhança é a maior possível.

O método aqui utilizado, essencialmente, é o descrito por SEBESTYEN (1962), vide ROMEDER (1972).

Uma vez definidas tais funções discriminantes, seria necessário dispor de meios para examinar a validade do método. Uma maneira de proceder é mediante a determinação de porcentagens de atribuições corretas, assunto que será objeto de capítulos subsequentes (4 e 5).

1.2 - NOTAÇÕES

Sempre que, para cada elemento ou indivíduo **x** são considerados os valores correspondentes a p variáveis quantitativas, então um tal elemento pode ser pensado como um vetor x no espaço $\mathbb{R}^p$, tal que $x' = (x_1, \ldots, x_p)$

Suponhamos que, a priori, se disponha de N indivíduos repartidos por k classes C_r, cada uma delas contendo N_r elementos $(r=1,2,\ldots,k)$. Evidentemente, tem-se $\sum N_r = N$, que é o total de elementos em causa.

Representaremos por x_i^r o i-ésimo elemento pertencente à classe C_r $(i=1,2,\ldots,N_r;\ r=1,2,\ldots,k)$. Assim, para cada um desses elementos

teremos:

$$(x_i^r)' = (x_{i1}^r, x_{i2}^r, \ldots, x_{ip}^r),$$

onde a componente x_{ij}^r é o valor assumido pela j-ésima variável (ou o resultado da j-ésima medida a ser efetuada no indivíduo x_i^r); $j=1,2,\ldots,p$.

Por outro lado, com relação a cada classe C_r, podemos considerar o vetor médio $\overline{x}^r$, tal que:

$$(1.2.1) \qquad (\overline{x}^r)' = (\overline{x}_1^r, \overline{x}_2^r, \ldots, \overline{x}_p^r) \quad ,$$

onde cada componente $\overline{x}_j^r$, $j = 1, 2, \ldots, p$, é o valor médio:

$$(1.2.1.1) \qquad \overline{x}_j^r = \left(\sum_{i=1}^{N} x_{ij}^r \right) / N_r \quad ,$$

este, calculado com relação a cada variável, a partir dos valores por ela assumidos nos N_r elementos ou indivíduos da classe C_r.

Ainda com relação a cada classe C_r, pode-se também considerar sua *matriz de variâncias-covariâncias*:

$$(1.2.2) \qquad \Sigma_r = \frac{1}{N_r} \sum_{i=1}^{N_r} \left[(x_i^r - \overline{x}^r)(x_i^r - \overline{x}^r)' \right],$$

$$(1.2.2.1) \qquad \sigma_{jk}^r = \frac{1}{N_r} \sum_{i=1}^{N_r} \left[(x_{ij}^r - \overline{x}_j^r)(x_{ik}^r - \overline{x}_k^r) \right] ;$$

$$j,k = 1,2, \ldots, p \quad ;$$

quando $j = k$, então σ_{kk}^r diz-se a *variância* da j-ésima variável na classe C_r; quando $j \neq k$, σ_{jk}^r diz-se a *covariância* entre as *j-ésima* e *k-ésima* variáveis.

Sendo o costume notar a variância de uma variável qualquer por σ^2, onde σ (raiz quadrada positiva da variância) é designado como o *desvio padrão*, então também podemos notar $\sigma_{jj}^r = (\sigma_j^r)^2$, sendo σ_j^r o desvio

padrão respectivo. Para maiores detalhes a respeito dos conceitos que conduzem a noção de matriz de variâncias-covariâncias remetemos ao Apêndice I.

1.3 - Q-DISTÂNCIA ENTRE OS INDIVÍDUOS NO ESPAÇO $\mathbb{R}^p$

Dados dois indivíduos a e b no espaço $\mathbb{R}^p$, tais que $a'=(a_1, a_2, \ldots, a_p)$ e $b'=(b_1, b_2, \ldots, b_p)$, estamos interessados em considerar sua distância mútua.

Para esse fim, em geral, consideramos uma matriz real $Q=(q_{jk})$; $j,k = 1,2,\ldots,p$ de sorte que a *Q-Distância* entre a e b seja expressada por $d_Q(a,b) = d(a,b)$, tal que:

$$(1.3.1) \qquad d^2(a,b) = \sum_{j,k=1}^{p} q_{jk}(a_j-b_j)(a_k-b_k) \; ;$$

ou em termos matriciais ,

$$(1.3.2) \qquad d^2(a,b) = (a-b)'Q(a-b)$$

Note-se que um vetor v (ou x) é sempre pensado como um vetor coluna, enquanto v' (ou x') representa o vetor linha correspondente, obtido por transposição; essa convenção permite melhor legibilidade das fórmulas matriciais.

Note-se que $d(a,b)$, de fato, deve gozar das propriedades de uma métrica, isto é:

(A_1) $d(a,b) \geq 0$

(A_2) $d(a,b) = 0 \iff a = b$

(A_3) $d(a,b) = d(b,a)$

(A_4) $d(a,c) \leq d(a,b) + d(b,c)$;

$a,b,c \in R^p$.

Portanto, $d^2(a,b)$ é definida através de uma forma quadrática positiva definida, onde Q é uma matriz real, simétrica e positiva definida; ver (1.3.2).

Quando $Q = I_p$ (matriz identidade $p \times p$), então a distância em consideração é a distância euclidiana usual, onde:

$$d^2(a,b) = \sum_{j=1}^{p} (a_j - b_j)^2 \qquad (1.3.3)$$

O seguinte lema é essencial para o que se segue e, em especial, para nos permitir uma interpretação transparente da Q-distância.

1.3.4 - LEMA

Se Q for uma matriz real, simétrica e positiva definida, então pode ser escrita como $Q = S'S$.

Demonstração

Sendo Q uma matriz real, simétrica e positiva definida, podemos encontrar uma matriz ortogonal P , tal que:

$$P^{-1}Q\,P = P'\,Q\,P = D = \begin{bmatrix} d_1 & & & 0 \\ & d_2 & & \\ & & \ddots & \\ 0 & & & d_p \end{bmatrix}$$

onde cada $d_i \geq 0$.

Consideremos a matriz diagonal D_1, cujos elementos sobre a diagonal principal são $\sqrt{d_1}$, $\sqrt{d_2}$, ..., $\sqrt{d_p}$, de modo que $D = D_1^2 = D_1 D_1'$. Portanto, $Q = PDP^{-1} = P\,D_1^2\,P' = P\,D_1 D_1' P' = (P\,D_1)(P\,D_1)'$,

isto é, pondo $S = (P\ D_1)'$; obtém-se

$$Q = S'S$$

Agora, substituindo $Q = S'S$ em (1.3.2) segue-se:

$$d^2(a,b) = (a-b)'\ S'S(a-b) = (S_a - S_b)'(S_a - S_b),$$

onde $S_a = Sa$ e $S_b = Sb$.

Assim, fica claramente evidenciado o fato contido na proposição abaixo.

1.3.5 PROPOSIÇÃO

A Q-*distância* entre dois vetores $a,b \in R^p$ pode ser considerada como a distância euclidiana usual entre os transformados S_a e S_b, sendo a matriz S obtida via a decomposição $Q = S'S$.

1.4 Q-SEMELHANÇA E Q-AGREGAÇÃO

No que se segue, introduzimos o conceito de Q-semelhança entre um indivíduo e uma classe (fazendo-se a distinção entre os casos em que o elemento pertença ou não a essa classe).

1.4.1 DEFINIÇÃO

a) Dada uma classe C_r e um indivíduo $a \notin C_r$, a Q-*semelhança* entre a e C_r é avaliada através da expressão:

(1.4.1.1) $$\pi(a, C_r) = \frac{1}{N_r} \sum_{i=1}^{N_r} d^2(a, x_i^r) ,$$

onde os x_i^r percorrem C_r.

b) Por outro lado, tomando em particular um indivíduo $x_h^r \in C_r$, a Q-*semelhança* entre x_h^r e C_r é avaliada mediante :

(1.4.1.2) $$\pi(x_h^r, C_r) = \frac{1}{N_r - 1} \sum_{i=1}^{N_r} d^2(x_h^r , x_i^r)$$

Note-se que, num caso e outro, a semelhança deve ser considerada tanto mais forte, quanto menor for $\pi(a, C_r)$ ou $\pi(x_h^r , C_r)$. Assim, a rigor, $\pi(a, C_r)$ ou $\pi(x_h^r , C_r)$ é uma medida de proximidade, variando inversamente com a semelhança.

Em seguida, estamos interessados em definir o que se chama a Q-*agregação* entre os indivíduos pertencentes a determinada classe; ou seja, de forma que se tenha uma avaliação numérica de quão se encontram agrupados (ou pelo contrário, dispersos) dentro da classe respectiva.

1.4.2 DEFINIÇÃO

A Q-agregação na classe C_r é avaliada através da expressão:

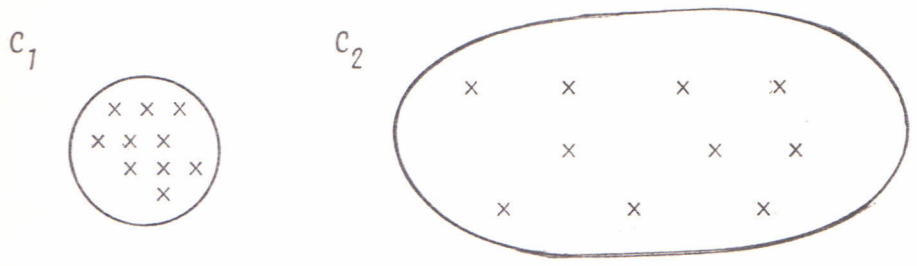
(1.4.2.1) $$D_r^2 = \frac{1}{N_r} \sum_{h=1}^{N_r} \pi(x_h^r , C_r)$$

Note-se que (1.4.2.1) também se escreve

(1.4.2.2) $$D_r^2 = \frac{1}{N_r(N_r - 1)} \sum_{h=1}^{N_r} \sum_{\substack{i=1 \\ i \neq h}}^{N_r} d^2(x_h^r , x_i^r) ,$$

bastando para isso, substituir em (1.4.2.1) a expressão para $\pi(x_h^r , C_r)$, dada em (1.4.1.2).

De forma análoga ao que ocorre com a Q-*semelhança*, a Q-agregação será tanto maior, quanto menor for a medida D_r^2 . Portanto, a rigor, D_r^2 é mais uma medida de dispersão, variando inversamente com a agregação. Vide esquema abaixo, para o caso de pontos no R^2 .



$$\begin{cases} C_1 : \text{grande agregação} \cong \text{pequena dispersão} \\ C_2 : \text{pequena agregação} \cong \text{grande dispersão} \end{cases}$$

1.5 ABORDAGEM DE SEBESTYEN

O fundamento da abordagem proposta por Sebestyen, para o problema de discriminação, é determinar a matriz Q que torne a agregação a maior possível para uma dada classe, sob certa restrição de normalidade, a saber: detQ = *1*. Note-se que a maior agregação possível corresponde ao menor valor possível para D_r^2 .

Com a decomposição Q = S'S, a condição detQ = *1* se escreve $(\text{detS})^2 = 1$, em geral escolhe-se S de sorte que detS = *1* .

Alguns resultados preliminares (lemas) serão estabelecidos, antes que se passe a resolver o problema de otimização acima proposto. No que se segue, supomos nossa atenção centrada em determinada classe, de sorte que nos permitimos omitir o sobre-índice r que a identifica. Com isso, a notação ficará mais aliviada.

1.5.1 LEMA

Seja C uma classe e $\sum$ sua matriz de variâncias-covariâncias. Então:

(1.5.1.1) $$2N^2 \sum = \sum_{h=1}^{N} \sum_{i=1}^{N} (x_h - x_i)(x_h - x_i)'$$

Demonstração

Conforme (1.2.2.1), o termo de ordem (j,k) de $\sum$ escreve-se :

$$\sigma_{jk} = \frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Desenvolvendo esta última expressão, teremos:

$$\sigma_{jk} = \frac{1}{N} \sum_{i=1}^{N} (x_{ij}\, x_{ik} - x_{ij}\, \bar{x}_k - \bar{x}_j\, x_{ik} + \bar{x}_j\, \bar{x}_k) =$$

$$= \frac{1}{N} \sum_{i=1}^{N} (x_{ij}\, x_{ik} - \bar{x}_j\, \bar{x}_k) \qquad [\,A\,]$$

Por outro lado, quanto ao termo de ordem (j,k) da matriz que comparece no lado direito de (1.5.1.1), podemos escrever:

$$\alpha_{jk} = \sum_{h=1}^{N} \sum_{i=1}^{N} (x_{hj} - x_{ij})(x_{hk} - x_{ik})$$

Desenvolvendo esta expressão, vem :

$$\alpha_{jk} = \sum_{h=1}^{N} \Big(\sum_{i=1}^{N} x_{hj}\, x_{hk} - \sum_{i=1}^{N} x_{hj}\, x_{ik} - \sum_{i=1}^{N} x_{ij}\, x_{hk} + \sum_{i=1}^{N} x_{ij}\, x_{ik} \Big)$$

$$= \sum_{h=1}^{N} (N\, x_{hj}\, x_{hk} - N\, x_{hj}\, \bar{x}_k - N\, x_{hk}\, \bar{x}_j + \sum_{i=1}^{N} x_{ij}\, x_{ik})$$

$$= N \sum_{h=1}^{N} x_{hj}\, x_{hk} - N^2\, \bar{x}_j\, \bar{x}_k - N^2\, \bar{x}_k\, \bar{x}_j + N \sum_{i=1}^{N} x_{ij}\, x_{ik}$$

$$= 2N^2 \left[\frac{1}{N} \sum_{i=1}^{N} (x_{ij}\, x_{ik} - \bar{x}_j\, \bar{x}_k \right] \qquad [\,B\,]$$

Comparando-se $[^{A}]$ e $[^{B}]$, conclui-se que

$$2N^2 \sigma_{jk} = \alpha_{jk} \; ; \qquad \forall \; j,k = 1, 2, \ldots, p \; .$$

5.2 LEMA

Seja $\sum$ uma matriz real, simétrica e positiva definida (supomos, no que se segue, a matriz de variâncias-covariâncias de uma classe C satisfaz a esta condição). Então:

$$\sum = C \Lambda C'$$

onde Λ designa a matriz diagonal dos seus valores próprios e C a matriz dos vetores próprios normalizados e escritos em colunas.

Demonstração

Seja c_j o vetor próprio normalizado, (isto é, $||c_j|| = 1$), correspondente a cada valor próprio λ_j de $\sum$.

Sabe-se que:

$$\sum c_j = \lambda_j c_j$$

Por outro lado, como $\sum$ é simétrica e positiva definida, ela admite p valores próprios ($j = 1,2,\ldots,p$).

Seja C a matriz dos vetores próprios normalizados correspondentes, escritos em colunas. Então, a equação $\sum c_j = \lambda_j c_j$ escreve-se, matricialmente:

$$\sum C = C \Lambda$$

Por outro lado, sabe-se que a matriz C é ortogonal, isto é, $C' = C^{-1}$.

Portanto

$$\sum = C \Lambda C^{-1} = C \Lambda C' \qquad \blacksquare$$

Note-se que, em consequência, $||\sum|| = \prod_{j=1}^{p} \lambda_j$ e $\sum^{-1} = C \Lambda^{-1} C'$

No que se segue, notaremos Tr(A) o traço de uma matriz quadrada **A**, isto é, a soma dos elementos sobre sua diagonal principal.

1.5.3 LEMA

Seja C uma classe e $\sum$ sua matriz de variâncias-covariâncias. Então, a expressão (1.4.2.1) utilizada para avaliar a Q-*agregação*, passa a escrever-se

(1.5.3.1) $$D^2 = \frac{2N}{N-1} \text{ Tr } (Q \sum)$$

Demonstração

Tem-se que $D^2 = \frac{1}{N(N-1)} \sum_{h=1}^{N_r} \sum_{\substack{i=1 \\ i \neq h}}^{N_r} d^2(x_h ; x_i)$, conforme a expressão (1.4.2.2). Conforme (1.3.2), vem

$$D^2 = \frac{1}{N(N-1)} \sum_{h=1}^{N_r} \sum_{\substack{i=1 \\ i \neq h}}^{N_r} (x_h - x_i)' \, Q(x_h - x_i)$$

Por outro lado, $(x_h - x_i)' \, Q(x_h - x_i)$, representa o produto de duas matrizes, a saber: $A = (x_h - x_i)'$ e $B = Q(x_h - x_i)$. Donde, sendo A.B uma matriz de ordem 1×1, tem-se :

$$(x_h - x_i)' \, Q(x_h - x_i) = \text{Tr}\left[(x_h - x_i)' Q(x_h - x_i)\right]$$

$$= \text{Tr}\left[Q(x_h - x_i).(x_h - x_i)'\right]$$

Assim sendo, utilizando (1.5.1.1) (vide Lema 1.5.1), vem:

$$D^2 = \frac{1}{N(N-1)} \sum_{h=1}^{N_r} \sum_{\substack{i=1 \\ i \neq h}}^{N_r} \mathrm{Tr}[Q(x_h - x_i)(x_h - x_i)'] =$$

$$= \frac{1}{N(N-1)} \mathrm{Tr}(Q\ 2N^2 \Sigma)$$

$$= \frac{2N}{N-1} \mathrm{Tr}(\ Q\ \Sigma\ )$$

■

De posse dos resultados precedentes, temos condições de determinar a matriz Q, para a qual a agregação é a maior possível, relativamente a uma dada classe, conforme o Teorema a seguir.

1.5.4 TEOREMA

Seja uma Q-*distância* (ou métrica) no R^p, definida por $Q = S'S$, que se supõe representar uma transformação a volume constante (isto é, $\det Q = 1$) e que minimiza a média dos quadrados das distâncias entre indivíduos de uma dada classe. Então,

$$Q = (\det \Sigma)^{1/p}\ \Sigma^{-1}$$

onde Σ é a matriz de variâncias-covariâncias das p variáveis, com relação à classe considerada.

Demonstração

Note-se que Σ e Q são matrizes quadradas reais, de ordem p, supostas simétricas e positivas definidas; sejam σ_{jk} e q_{jk} $(j,k=1,2,\ldots,p)$ os termos gerais dessas matrizes, respectivamente.

Ora, para minimizar D^2 sob a restrição de normalidade $\det Q = 1$,

corresponde a minimizar $Tr(Q\sum)$, sob a mesma condição, conforme a expressão (1.5.3.1). Para tal fim, utilizaremos o método de *multiplicadores de Lagrange*; a respeito desse método, vide por exemplo, Kaplan (1969). Dessa maneira, obtem-se a seguinte equação

$$\frac{\partial}{\partial q_{jk}} \left[\sum q_{jk}\, \sigma_{jk} - \lambda(\det Q) \right] = 0 \quad ,$$

$$j,k = 1,2,\ldots,p \quad ,$$

donde se segue:

$$\sigma_{jk} = \lambda Q_{jk} \quad ; \qquad j,k = 1,2,\ldots,p \qquad [\,A\,]$$

expressão na qual Q_{jk} designa o co-fator do termo q_{jk}, isto é, o determinante menor correspondente a q_{jk}, afetado de seu sinal $(-1)^{j+k}$.

Lembremos que a matriz Q^* dos cofatores Q_{jk} dividida por $\det Q$ é, exatamente, a matriz inversa Q^{-1}. Logo sob forma matricial, a expressão $[\,A\,]$ passa a se escrever :

$$\sum = \lambda\, Q^* = \lambda \frac{Q^*}{\det Q} = \lambda\, Q^{-1} \qquad [\,B\,]$$

Segue-se que $\det\sum = \det(\lambda Q^{-1}) = \lambda^p (\det Q)^{-1} = \lambda^p$

donde: $$\lambda = (\det\textstyle\sum)^{1/p} \qquad [\,C\,]$$

Substituindo $[\,C\,]$ em $[\,B\,]$, obtém-se :

(1.5.4.1) $$Q = (\det\textstyle\sum)^{1/p} \sum^{-1}$$

Assim, esta matriz, constitue um "ponto crítico" (no $R^{p\times p}$) para $Tr(Q\sum)$. Resta mostrar que, de fato é um "ponto de mínimo".

Pelo Lema 1.5.2, a matriz Q se escreve $Q = H\,D\,H'$, onde D é uma matriz diagonal, com elementos todos positivos na diagonal principal;

e H é uma matriz ortogonal, isto é, $H' = H^{-1}$. Evidentemente ,

$$\det D = (\det H)(\det H')(\det D) = \det(HDH') = \det Q = 1$$

Por outro lado, podemos definir uma matriz M , positiva definida, tal que $M = H' \sum H$ onde $\sum = H M H'$. Dessa maneira, minimizar $Tr(Q, \sum)$ sob a condição $\det Q = 1$, corresponde a minimizar $Tr(DM)$, com $\det D = 1$.

Deve-se observar que, embora sendo M positiva definida, não necessariamente é diagonal; ao passo que D é diagonal, com elementos $d_1 > 0$, $d_2 > 0$, ... , $d_p > 0$ dispostos na diagonal principal. Então :

$$Tr(DM) = \sum_{j=1}^{p} d_j m_{jj} \qquad \text{e}$$

$$\det D = \prod_{j=1}^{p} d_j$$

sendo m_{jk} o termo geral da matriz M .

Desde que M é positiva definida, tem-se $m_{11} > 0$; em seguida, observa-se que se poderia escolher d_1 arbitrariamente grande e os demais $d_j (j > 1)$ arbitrariamente pequeno; porém, de sorte que $\prod_{j=1}^{p} d_j = 1$. Assim, $Tr(Q \sum) = Tr(D M)$ pode se tornar arbitrariamente grande . Portanto, a quantidade encontrada como o único extremo de $Tr(Q \sum)$ não podendo ser um máximo, é um mínimo.

Uma vez determinada a forma da matriz Q , apresentaremos as expressões gerais das medidas de proximidades utilizadas para avaliar a Q-*agregação* e a Q-*semelhança*.

Para o caso da Q-*agregação*, basta substituir em (1.5.3.1) o valor (1.5.4.1); donde ;

(1.5.4.2) $$D^2 = \frac{2\,N\,p}{N-1}\,(\det\Sigma)^{1/p}$$

Por outro lado, no caso de Q-*semelhança* de um indivíduo a $\notin \varphi$ com esta classe C, teremos:

(1.5.4.3) $$\pi(a\ ,\ C) = (\det\Sigma)^{1/p}\,[p + (a-\bar{x})'\,\Sigma^{-1}(a-\bar{x})]$$

Já no caso em que o indivíduo $x_h \in C$, então $d(x_h,x_h) = 0$; sendo assim, são apenas N-1 indivíduos na classe para serem relacionados com x_h. Portanto, basta tomar a média sobre os $N-1$ indivíduos, ou seja, multiplicar o segundo membro da expressão (1.5.4.3) por $N/N-1$; donde:

(1.5.4.4) $$\pi(x_h\ ,\ C) = \frac{N}{N-1}\,(\det\Sigma)^{1/p}\,[p + (x_h-\bar{x})'\,\Sigma^{-1}(x_h-\bar{x})]$$

Consideremos, em seguida, o caso particular em que a matriz Q é diagonal, constituindo-se num corolário do Teorema 1.5.4.

1.5.5 COROLÁRIO

Se Q é diagonal, a quantidade D^2 é minimizada na classe C, desde que:

$$w_j = \frac{1}{\sigma_j}\left(\prod_{k=1}^{p}\sigma_k\right)^{1/p}\ ,$$

onde os w_j^2 ; $j = 1,2,\ldots,p$ são os elementos da diagonal principal de Q e os σ_j são os desvios-padrão das j-ésimas variáveis; $j = 1,2,\ldots,p$.

Demonstração

Observe-se que, neste caso, $Q = S'S$; denotando-se por $w_j > 0$ os elementos situados na diagonal principal da matriz S .

Por outro lado, é óbvio que $\sum^{-1}$ é a matriz diagonal cujos elementos na diagonal principal têm a forma $1/\sigma_j^2$, onde $\sigma_j^2 = \sigma_{jj}$; bem como:

$$(\det\sum)^{1/p} = \left(\prod_{j=1}^{p} \sigma_j^2 \right)^{1/p} = \left[\left(\prod_{j=1}^{p} \sigma_j \right)^{1/p} \right]^2 .$$

Portanto, a expressão $Q = ||\sum||^{1/p} \sum^{-1}$ é equivalente a :

(1.5.5.1) $$w_j = \frac{1}{\sigma_j} \left(\prod_{k=1}^{p} \sigma_k \right)^{1/p} ;$$

$$j = 1, 2, \ldots, p$$

■

Neste caso particular, as expressões definidas em (1.5.4.2), (1.5.4.3) e (1.5.4.4), tomam as seguintes formas:

(1.5.5.2) $$D^2 = \frac{2N}{N-1} \left(\prod_{j=1}^{p} \sigma_j \right)^{1/p}$$

(1.5.5.3) $$\pi(a, C) = \left(\prod_{j=1}^{p} \sigma_j^2 \right)^{1/p} \left[\sum_{j=1}^{p} \left(\frac{a_j - \bar{x}_j}{\sigma_j} \right)^2 + p \right]$$

(1.5.5.4) $$\pi(x_h, C) = \frac{N}{N-1} \left(\prod_{j=1}^{p} \sigma_j^2 \right)^{1/p} \left[\sum_{j=1}^{p} \left(\frac{x_{hj} - x_j}{\sigma_j} \right)^2 + p \right]$$

1.5.6 LEMA

Os valores próprios λ_j de $\sum$ são iguais às variâncias $(\sigma_j)'^2$ das variáveis transformadas $y^{(j)} = C'x^{(j)}$, $j = 1, \ldots, p$,

onde cada $x^{(j)} = (x_{ij})$, sendo C uma matriz ortogonal e $\sum = C \Lambda C'$.

Demonstração

Como $Q = (\det\sum)^{1/p} \sum^{-1}$ é simétrica e positiva definida, segue-se que $\sum^{-1}$ também o é, bem como $\sum$. Por outro lado, conforme lema(1.5.2) tem-se $\sum = C \Lambda C'$, onde C é ortogonal.

Sejam $X = (x^{(j)})$ e $Y = (y^{(j)}) = C'X$; tem-se (vide proposição (I.6) do Apêndice I):

$$\sum = E\left[(X-\overline{X})(X-\overline{X})'\right] ,$$

donde, em virtude de (I.7.3):

$$E\left[(Y-\overline{Y})(Y-\overline{Y})'\right] =$$

$$= E\left[(C'X - \overline{C'X})(C'X - \overline{C'X})'\right] =$$

$$= E\left[C'(X - \overline{X})(X - \overline{X})'C \right] =$$

$$= C' E\left[(X - \overline{X})(X - \overline{X})'\right] C =$$

$$= C' \sum C = \Lambda .$$

Portanto, vê-se que os elementos na diagonal principal da matriz diagonal Λ , que são os valores próprios λ_j de $\sum$, coincidem com as variâncias das variáveis associadas aos vetores $y^{(j)} = C'x^{(j)}$.

■

1.5.7 TEOREMA

A métrica Q definida por $Q = (\det\sum)^{1/p} \sum^{-1}$ equivale a uma transformação linear S sobre as variáveis iniciais, e que se

escreve como um produto $S = WC'$, onde C' é uma rotação (definida pela matriz C' dos vetores próprios ortonormais de $\sum$, escritos em linhas) e W é uma transformação diagonal (definida da mediante a expressão (1.5.5.1)).

Demonstração

Sabemos que; $Q = S'S$, $\sum \; C \Lambda C'$, bem como, $Q = (\det\sum)^{1/p} \sum^{-1}$ segue-se :

$$(\det\sum)^{1/p} \sum^{-1} = ||C \Lambda C'||^{1/p} \cdot (C \Lambda C')^{-1}$$

$$= ||\Lambda||^{1/p} C \Lambda^{-1} C'$$

$$= \left(\prod_{j=1}^{p} \lambda_j\right)^{1/p} C \Lambda^{-1} C'$$

$$= C\left[\left(\prod_{j=1}^{p} \lambda_j\right)^{1/p} \Lambda^{-1}\right]C'$$

$$= C\left[\left(\prod_{j=1}^{p} \lambda_j\right)^{1/p} \begin{pmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & 0 \\ & & \ddots & \\ 0 & & & 1/\lambda_p \end{pmatrix}\right]$$

$$= C\left[\left(\prod_{j=1}^{p} \sigma_j'^{2}\right)^{1/p} \begin{pmatrix} 1/\sigma_1'^{2} & & & \\ & 1/\sigma_2'^{2} & & 0 \\ & & \ddots & \\ 0 & & & 1/\sigma_p'^{2} \end{pmatrix}\right]$$

$$= C\, W^2\, C' \quad ,$$

onde $$W = \left(\prod_{j=1}^{p} \sigma_1'\right) \begin{pmatrix} 1/\sigma_1' & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_p' \end{pmatrix} .$$

Sendo $(\det\Sigma)^{1/p}\,\Sigma^{-1} = C\,W^2\,C'$, tem-se $S = W\,C'$,

pois $(\det\Sigma)^{1/p}\,\Sigma^{-1} = Q = S'S = (WC')'\,(WC') = CW'\,WC'$

Capítulo 2

DISCRIMINAÇÃO LINEAR :

PESQUISA DE EIXOS FATORIAIS DISCRIMINANTES

2.1 INTRODUÇÃO

Neste capítulo são indicados os aspectos básicos relacionados a uma abordagem do problema de discriminação com finalidade descritiva. Assim, devem ser determinados os chamados eixos fatoriais discriminantes , permitindo uma melhor separação entre duas classes. Esses eixos correspondem a certos vetores unitários os quais, mais exatamente, são vetores próprios de uma matriz $T^{-1}B$, onde T é uma "*matriz de variâncias-covariâncias* total" e B é uma "*matriz de variâncias-covariâncias* inter-classes", as quais serão definidas no texto subsequente.

Faz-se, por outro lado, uma ligação com o método clássico das funções lineares discriminantes, introduzido por *R.A. Fisher* em 1936. Tal método consiste em comparar as distâncias de um indivíduo arbitrário aos centros das classes, distâncias essas que se medem através de certa métrica, com o objetivo de serem evidenciados determinados fatores, definidos como combinações lineares das variáveis originais, de sorte os valores respectivos sejam tanto quanto possível mais diferentes, para indivíduos pertencentes a classes distintas.

2.2 NUVENS DE PONTOS NO R^p

Procuramos manter, quanto possível, as notações já introduzidas no capítulo anterior.

Em cada classe C_r , os indivíduos x_i^r poderão encontrar-se afetados de pesos ou massas $p_r(x_i^r)$, $i = 1,2,\ldots, N_r$. Isso corresponde a se considerar uma aplicação

$$(2.2.1) \qquad p_r : C_r \longrightarrow [0,1] \ ,$$

$$x \rightsquigarrow p_r(x)$$

sujeita à "condição de normalização":

(2.2.2) $$\sum_{x \in C_r} p_r(x) = 1 \quad .$$

(2.2.3) DEFINIÇÃO

A nuvem C_r é o conjunto dos pontos $x \in C_r$, afetados de seus pesos $p_r(x)$.

A todo rigor, observa-se que a nuvem C_r é de fato o par (C_r , p_r). Diz-se que os pesos ou massas são equidistribuídas em C_r, quando $p_r(x) = 1/N_r$, para todo x pertencente à referida classe.

Por sua vez, consideraremos as classes C e $\tilde{C}$, tais que :

$$C = \bigcup_{r=1}^{k} C_r \quad ,$$

$$\tilde{C} = \{C_r ; r = 1, 2, \ldots, k\} ;$$

ou seja, a classe C é formada por todos os elementos das classes C_r, enquanto $\tilde{C}$ possui como elementos aquelas classes.

Se, para cada $y = C_r \in \tilde{C}$, atribuirmos um peso ou massa $q(y)$, de sorte que:

$$\sum_{y \in \tilde{C}} q(y) = 1 \quad ,$$

então o par $(\tilde{C} , q)$ será ainda uma nuvem. Em consequência, pode-se igualmente considerar a nuvem (C , π); onde:

(2.2.4) $$\pi(x_i^r) = q(C_r) \quad p_r(x_i^r) \quad ;$$

note-se que, de fato:

$$\sum_{x \in C} \pi(x) = 1 \quad .$$

Na hipótese dos pesos serem equidistribuídos em cada classe C_r

e, se além disso, $q(C_r) = N_r/N$, então os pesos em C também serão equi_ distribuídos, com $\pi(x) = 1/N$, para todo x pertencente a C .

2.3 BARICENTROS DE NUVENS DE PONTOS

Na presente secção, considera-se o conceito de *baricentro* (*ou centro de gravidade*) de uma nuvem de pontos.

(2.3.1) DEFINIÇÃO

Dada a nuvem (C_r , p_r) no R^p, seu *baricentro* (*ou centro de gravidade*) $g^r \in R^p$ é definido por :

$$(2.3.1.1) \qquad g^r = \sum_{x \in C_r} p_r(x)\ x$$

No caso de equidistribuição de pesos, observe-se que g^r é, nada mais, que o vetor médio introduzido no capítulo anterior.

Analogamente, no caso da nuvem (C , π), tem-se o baricentro $g \in R^p$, dado por :

$$(2.3.1.2) \qquad g = \sum_{x \in C} \pi(x)\ x \quad .$$

Por outro lado, ao se considerar a nuvem $(\tilde{C} , q)$ cada classe $y = C_r$, poderá ser identificada ao seu baricentro $\dot{y} = g^r$; em tal circunstância, o baricentro de $\tilde{C}$ será :

$$(2.3.1.3) \qquad \tilde{g} = \sum_{y \in \tilde{C}} q(y)\ \dot{y} \quad .$$

(2.3.2) PROPOSIÇÃO

Sejam g e $\tilde{g}$ os baricentros das nuvens C e $\tilde{C}$, respectivamente. Então: $g = \tilde{g}$.

Demonstração

$$\tilde{g} = \sum_{y \in \tilde{C}} q(y)\, g^r = \sum_{y \in \tilde{C}} q(y) \sum_{x \in C_r} p_r(x)\, x =$$

$$= \sum_{y \in \tilde{C}} \sum_{x \in C_r} q(y)\, p_r(x)\, x = \sum_{x \in C} \pi(x)\, x = g \quad .$$

2.4 INÉRCIA DE NUVENS DE PONTOS

Na secção precedente considerou-se o conceito de baricentro de uma nuvem de pontos; que é o equivalente, em têrmos "mecânicos", ao conceito estatístico de média (ou vetor médio).

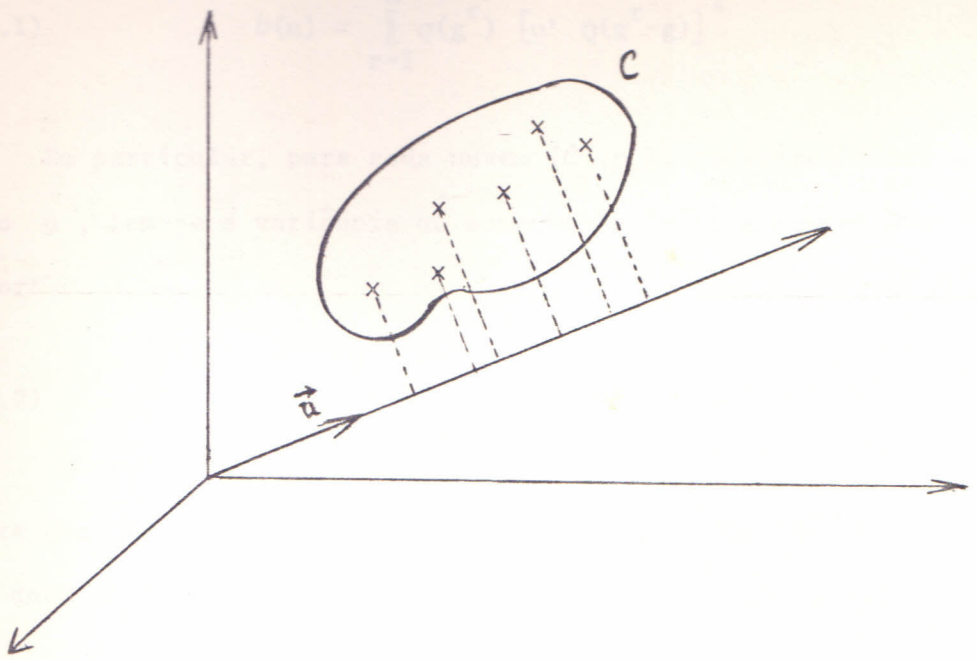
De maneira análoga introduz-se, aqui, o conceito de *inércia*, equivalente ao de variância e de covariância do Capítulo *1*.

O espaço onde trabalhamos é, ainda, o R^p. No caso unidimensional ($p = 1$), a nuvem de pontos $(C\ ,\ \pi)$ distribui-se sobre uma linha reta e sua variância ou momento de inércia, calculado com relação ao seu centro de gravidade g, é dado por:

$$(2.4.1) \qquad v = \sum_{x \in C} \pi(x)\ (x-g)^2$$

Já no caso multidimensional ($p > 1$), teremos uma variância ou momento de inércia para cada uma das p variáveis; ou seja, trata-se da *variância* ou *momento de inércia* da nuvem, quando projetada ortogonalmente sobre o eixo de coordenadas correspondente.

De maneira mais geral, desde que seja fixado um vetor $\mu \in R^p$, pode-se considerar a variância ou momento de inércia da nuvem, projetada ortogonalmente sobre a direção definida pelo mesmo vetor unitário μ (vide gráfico e definição a seguir).



(2.4.2) DEFINIÇÃO

A *variância total* ou *momento de inércia total* $t(u)$ da nuvem (C, π), projetada ortogonalmente sobre a direção u (onde $u \in R^p$ é um vetor unitário), define-se como:

$$(2.4.2.1) \qquad t(u) = \sum_{x \in C} \pi(x) \left[u' Q (x-g) \right]^2$$

Note-se que $u' Q (x-g)$ é o produto interno (ou produto escalar) do vetor unitário u, pelo vetor $(x-g)$; por outro lado, $(x-g)$ denota o afastamento ou desvio do vetor x com relação ao vetor médio ou baricentro g da nuvem C.

Analogamente, para a nuvem $(\tilde{C}, q)$ tem-se:

(2.4.3) DEFINIÇÃO

A *variância inter-classes* ou *momento de inércia inter-classes* $b(u)$ da nuvem $(\tilde{C}, q)$, projetada sobre a direção u, define-se como:

(2.4.3.1) $$b(u) = \sum_{r=1}^{k} q(g^r) \left[u' \, Q(g^r - g)\right]^2$$

Em particular, para cada nuvem (C_r, p_r), projetada sobre a mesma direção u, tem-se a variância ou momento de inércia dentro da classe C_r, dada por:

(2.4.3.2) $$v^r(u) = \sum_{x \in C_r} p_r(x) \left[u' \, Q(x - g^r)\right]^2$$

de sorte que, a variância intra-classes ou momento de inércia intra-classes, é dada conforme a definição seguinte.

(2.4.4) DEFINIÇÃO

A *variância intra-classes* ou *momento de inércia intra-classes* $w(u)$, segundo a direção u, define-se como:

(2.4.4.1) $$w(u) = \sum_{r=1}^{k} q(g^r) \; v^r(u) \; .$$

Finalmente, apresenta-se o conceito de covariância ou produto de inércia, segundo duas direções u_1 e u_2 (onde $u_1, u_2 \in R^p$ são vetores unitários).

No caso da nuvem (C, π), tem-se a covariância total, ou produto de inércia total, segundo as referidas direções, através da definição abaixo.

(2.4.5) DEFINIÇÃO

A *covariância total* ou *produto de inércia total* $t(u_1, u_2)$ da nuvem (C, π) é dada por:

(2.4.5.1) $$t(u_1, u_2) = \sum_{x \in C} \pi(x) \left[u_1' \, Q(x - g)\right] \left[u_2' Q(x - g)\right]$$

Por outro lado, considerando-se a nuvem $(\tilde{C}, q)$, tem-se:

(2.4.6) DEFINIÇÃO

A *covariância inter-classes* ou *produto de inércia inter-classes* $b(u_1,u_2)$, da nuvem $(\tilde{C}, q)$, segundo as direções u_1 e u_2, define-se como:

$$(2.4.6.1) \qquad b(u_1,u_2) = \sum_{r=1}^{k} q(g^r) \left[u_1' Q(g^r-g)\right] \left[u_2' Q(g^r-g)\right]$$

Finalmente, tem-se:

(2.4.7) DEFINIÇÃO

A *covariância intra-classes* ou *produto de inércia intra-classes* $w(u_1,u_2)$, segundo as direções u_1 e u_2, é dada por:

$$(2.4.7.1) \qquad w(u_1,u_2) = \sum_{r=1}^{k} q(g^r) \sum_{x \in C_r} p_r(x) \left[u_1' Q(x-g^r)\right] \left[u_2' Q(x-g^r)\right]$$

Se u_1 e u_2 forem os vetores canônicos e_i e e_j (vetores unitários sobre os eixos coordenados), então $t(u_1)$, $t(u_1,u_2)$, $b(u_1)$, $b(u_1,u_2)$, $w(u_1)$ e $w(u_1,u_2)$ notar-se-ão t_{ii}, t_{ij}, b_{ii}, b_{ij}, w_{ii} e w_{ij}; observe-se que as variâncias ou momentos de inércia t_{ii}, b_{ii} e w_{ii} são covariâncias ou produtos de inércia, onde $i = j$ (isto é, são variâncias ou momentos de inércia).

Os t_{ij}, b_{ij} e w_{ij} constituem os termos gerais de matrizes de inércia T, B e W, conforme a definição seguinte:

(2.4.8) DEFINIÇÃO

As *matrizes de inércia total* (T), *inter-classes* (B) e *intra-classes* (W), são as matrizes de termos gerais t_{ij} , b_{ij} , w_{ij}.

Note-se que o produto interno $e_i' Q(x-g)$ nos dá a i-ésima coordenada do vetor (x-g), isto é, $x_i - g_i$; analogamente, $e_i' Q(g^r-g) = (g_i^r - g_i)$.

Como:

$$t_{ij} = \sum_{x \in C} \pi(x)(x_i-g_i)(x_j-g_j), \qquad (2.4.8.1)$$

então T também se escreve:

$$T = \sum_{x \in C} \pi(x)(x-g)(x-g)' \quad . \qquad (2.4.8.1a)$$

Por outro lado, como:

$$b_{ij} = \sum_{r=1}^{k} q(g^r)(g_i^r-g_i)(g_j^r-g_j), \qquad (2.4.8.2)$$

segue-se que B se escreve:

$$B = \sum_{r=1}^{k} q(g^r)(g^r-g)(g^r-g)' \quad . \qquad (2.4.8.2a)$$

Analogamente:

$$w_{ij} = \sum_{r=1}^{k} \sum_{r \in C_r} q(g^r)p_r(x)(x_i-g_i^r)(x_j-g^r{}_j) \qquad (2.4.8.3)$$

e

$$W = \sum_{r=1}^{k} \sum_{r \in C_r} q(g^r)p_r(x)(x-g^r)(x-g^r)' \quad . \qquad (2.4.8.3a)$$

Em consequências das definiçõoes apresentadas no texto, podemos ter em função de T, B e W , os valores das variâncias total, inter-classes e intra-classes com relação a qualquer direção u, respectivamente. Para esse fim, consideremos:

2.5 LEMA

(2.5.1) $\quad t(u) = v' \, T \, v$

(2.5.2) $\quad b(u) = v' \, B \, v$

(2.5.3) $\quad w(u) = v' \, W \, v \quad ,$

onde v é a imagem do vetor unitário u pelo isomorfismo do $\mathbb{R}^p$ no $(\mathbb{R}^p)^*$ cuja matriz em relação à base (e_i) do $\mathbb{R}^p$ e à base dual (e^i) de $(\mathbb{R}^p)^*$, é precisamente a matriz Q .

Demonstração da expressão (2.5.1)

$$t(u) = \sum_{x \in C} \pi(x) \left[u'Q(x-g)\right]^2 =$$

$$= \sum_{x \in C} \pi(x) \left[u'Q(x-g)\right] \left[(x-g)'Q\, u\right] =$$

$$= (Qu)' \left[\sum_{x \in C} \pi(x-g)(x-g)' \right] Qu \; ;$$

como $v = Qu$ e utilizando (2.4.8.1a), segue-se o resultado desejado, i.e,

$$t(u) = v'T \, v$$

Demonstração da expressão (2.5.2)

$$b(u) = \sum_{r=1}^{k} q(g^r) \left[u'Q(g^r-g)\right]^2 =$$

$$= \sum_{r=1}^{k} q(g^r) \left[u'Q(g^r-g)\right] \left[(g^r-g)'Qu\right] =$$

$$= (Qu)' \left[\sum_{r=1}^{k} q(g^r)(g^r-g)(g^r-g)' \right] (Qu) \; ;$$

desde que $v = Qu$ e utilizando (2.4.8.2a), segue-se: $b(u) = v'B \, v$.

Demonstração da expressão (2.5.3)

De forma análoga, usando (2.4.8.3a), tem-se:

$$b(u) = \sum_{r=1}^{k} q(g^r)\, v^r(u) =$$

$$= \sum_{r=1}^{k} q(g^r) \sum_{x \in C_r} p_r(x) \left[u'Q(x-g^r)\right]^2 =$$

$$= \sum_{r=1}^{k} q(g^r) \sum_{x \in C_r} p_r(x) \left[u'Q(x-g^r)\right]\left[(x-g^r)'Qu\right] =$$

$$= (Qu)'\left[\sum_{r=1}^{k} \sum_{x \in C_r} q(g^r) p_r(x)(x-g^r)(x-g^r)'\right](Qu) =$$

$$= v'\, W\, v\ .$$

Com relação às matrizes de inércia, vale a seguinte importante relação:

2.6 TEOREMA (HUYGHENS)

Sejam T , W e B as matrizes de inércia ou de covariâncias total, intra-classes e inter-classes, respectivamente. Então:

$$T = W + B$$

Demonstração

Tem-se:

$$W = \sum_{r=1}^{k} \sum_{r \in C_r} q(g^r) p_r(x)(x-g^r)(x-g^r)' =$$

$$= \sum_{r=1}^{k} \sum_{x \in C_r} q(g^r) p_r(x)(x\,x' - g^r x' - x\,g^{r'} + g^r\, g^{r'}) =$$

$$= \sum_{r=1}^{k} \sum_{r \in C_r} q(g^r) p_r(x) x\, x' - \sum_{r=1}^{k} q(g^r) g^r \Big(\sum_{x \in C_r} p_r(x) x \Big)' -$$

$$- \sum_{r=1}^{k} q(g^r) \Big(\sum_{x \in C_r} p_r(x)\, x \Big) g^{r'} + \sum_{r=1}^{k} (g^r) \Big(\sum_{x \in C_r} p_r(x) \Big) g^r\, g^{r'} =$$

$$= \sum_{r=1} \sum_{x \in C_r} q(g^r) p_r(x) x\, x' - \sum_{r=1}^{k} q(g^r) g^r\, g^{r'} \; ; \qquad [\,A\,]$$

por outro lado,

$$B = \sum_{r=1}^{k} q(g^r)(g^r - g)(g^r - g)' =$$

$$= \sum_{r=1}^{k} q(g^r)(g^r g^{r'} - g^r g' - g g^{r'} + g g') =$$

$$= \sum_{r=1}^{k} f(g^r) g^r g^{r'} - \Big(\sum_{r=1}^{k} q(g^r) g^r \Big) g' - g \Big(\sum_{r=1}^{k} q(g^r) g^r \Big)' +$$

$$+ \Big(\sum_{r=1}^{k} q(g^r) \Big) g\, g' =$$

$$= \sum_{r=1}^{k} q(g^r)\, g^r\, g^{r'} + g\, g' \, . \qquad [\,B\,]$$

Ora, de $[\,A\,]$ e $[\,B\,]$, segue-se:

$$B + W = \sum_{r=1}^{k} \sum_{x \in C_r} q(g^r) p_r(x) x x' + g g' \, . \qquad [\,C\,]$$

Ademais,

$$T = \sum_{x \in C} \pi(x)(x - g)(x - g)' =$$

$$= \sum_{x \in C} \pi(x)(x x' - g x' - x g' + g g') =$$

$$= \sum_{r=1}^{k} \sum_{x \in C_r} q(g^r) p_r(x) x\, x' -$$

$$- g \Big(\sum_{x \in C} \pi(x) x \Big)' - \Big(\sum_{x \in C} \pi(x) x \Big) g' + g\, g' =$$

$$= \sum_{r=1}^{k} \sum_{x \in C_r} q(g^r)\, p_r(x) x\, x' + g\, g' \, . \qquad [\,D\,]$$

Comparando $[\,C\,]$ e $[\,D\,]$ chega-se ao resultado desejado, i.e,

$T = B + W$.

A partir do resultado desse teorema, obtem-se:

(2.6.1) $\mathbf{v}' \mathbf{T} \mathbf{v} = \mathbf{v}' \mathbf{W} \mathbf{v} + \mathbf{v}' \mathbf{B} \mathbf{v}$; onde $\mathbf{v} = Qu, \forall\, \mathbf{u} \in R^p$

e conforme o lema (2.5), segue-se:

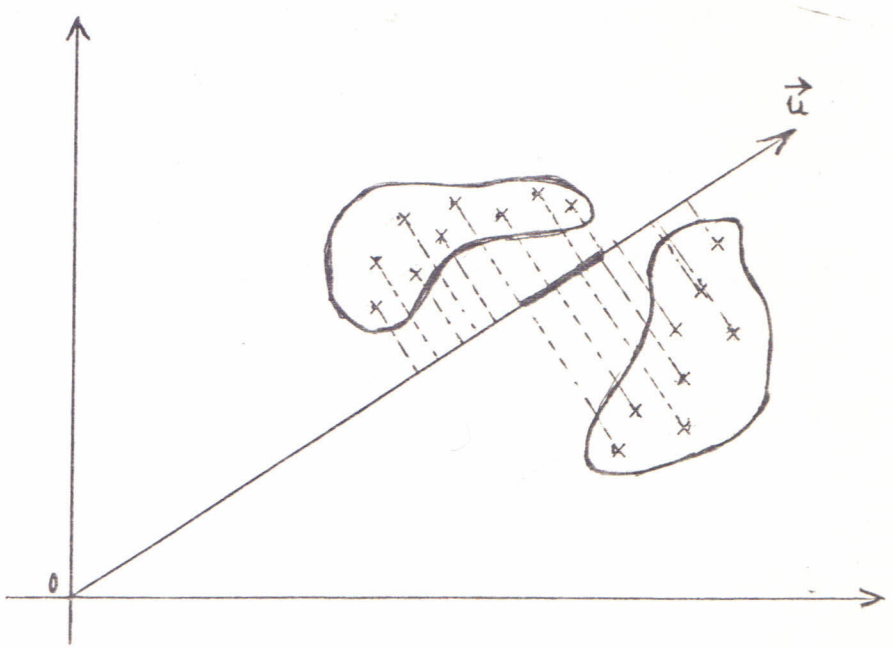
(2.6.2) $t(\mathbf{u}) = w(\mathbf{u}) + b(\mathbf{u})$,

que é o teorema de Huyghens (clássico em mecânica), aplicado à nuvem C .

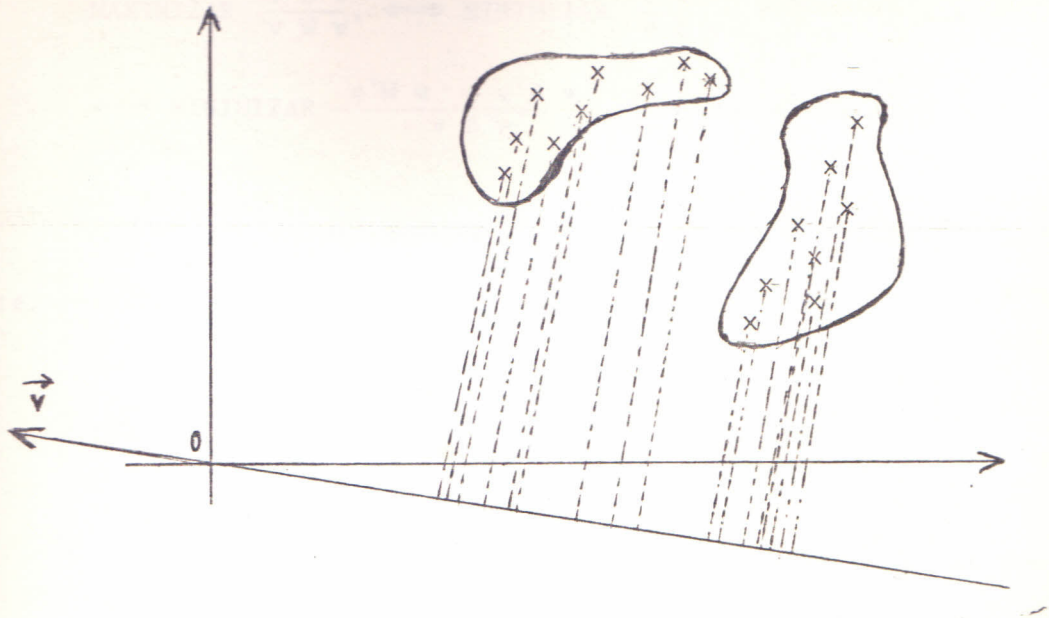
2.7 EIXOS FATORIAIS DISCRIMINANTES

O objetivo deste parágrafo é escolher um eixo, isto é, um vetor unitário $\mathbf{u}$ que melhor permite discriminar as classes. Iniciamos por dois exemplos no R^2 que nos encaminham de maneira mais objetiva para a compreensão das idéias aí subjacentes.

(2,7.1) EXEMPLO (a): (má discriminação)



(2.7.2) EXEMPLO (b): (boa discriminação)



Nota-se no exemplo (a), de fato, o vetor unitário **u** não permite discriminar as classes C_1 e C_2 tão bem quanto **v** o faz no exemplo (b).

Obviamente, a discriminação será tanto mais facilmente alcançada quanto as classes se encontrem mais distanciadas uma das outras (variância inter-classes grande) e, concomitantemente, quanto os indivíduos (ou elementos) de uma mesma classe se encontram o mais próximo possível entre si, (variância intra-classes pequena).

Dessa forma, encontraremos o eixo fatorial, ou seja, o vetor unitário **u**, que melhor discrimina não somente o conjunto dos indivíduos de C, mas o conjunto $\tilde{C}$ das classes de indivíduos de C.

Para esse fim, podemos dizer que o primeiro eixo fatorial discriminante $\mathbf{u}^1$ será o elemento **u** que maximiza o quociente da variância inter-classes de **u** pela variância intra-classes de **u**, isto é, trata-se de :

(2.7.3) MAXIMIZAR $\dfrac{\mathbf{v}'B\,\mathbf{v}}{\mathbf{v}'W\,\mathbf{v}}$, ou equivalentemente

(2.7.4) MAXIMIZAR $\dfrac{\mathbf{v}'B\,\mathbf{v}}{\mathbf{v}\,T\,\mathbf{v}'}$; onde $v = Qu$

De fato, tem-se a seguinte cadeia de equivalências.

$$\text{MAXIMIZAR } \frac{v'B\,v}{v\,W\,v'} \iff \text{MINIMIZAR } \frac{v'W\,v}{v\,B\,v'} + 1 \iff$$

$$\iff \text{MINIMIZAR } \frac{v'W\,v + v'B\,v}{v\,B\,v'} \iff \text{MAXIMIZAR } \frac{v'B\,v}{v\,T\,v'} \ .$$

Com finalidade de determinar o primeiro eixo fatorial discriminante, considera-se o seguinte teorema:

2.8 TEOREMA

O primeiro eixo fatorial discriminante u^1 é tal que $v^1 = Qu_1$ é o vetor próprio de $T^{-1}B$, correspondente ao maior valor próprio λ_1.

Demonstração

Observa-se que o problema formulado através da expressão(2.7.4) corresponde ao problema clássico de se obter o máximo do quociente de duas formas quadráticas. E para esse fim, utilizaremos o método dos multiplicadores de Lagrange.

Verifica-se primeiramente que é possível escolher $v'T\,v$ igual a uma constante k prefixada, pois os vetores próprios são dados a menos de uma constante. Com efeito, suponhamos que se obtivesse um vetor próprio w, tal que $w'T\,w = k_1 \neq k$; bastaria, nesse caso, em lugar de w, escolher $v = \sqrt{\frac{k}{k_1}}\ w$, donde $v'T\,v = k$.

Assim, tem-se que maximizar $v'B\,v$ sob a condição $v'T\,v = k$ (constante). Derivando em relação a v e igualando a zero, segue-se:

$$\frac{\partial}{\partial v} - [v'B\,v - \lambda\, v'T\,v] = 0 \ ;$$

isto é, $$2Bv - \lambda\, 2T\,v = 0 \ ,$$

donde $\mathbf{B}\,\mathbf{v} = \lambda\,\mathbf{T}\,\mathbf{v}$.

Como $\mathbf{T}$ é suposta inversível, temos:

$$\mathbf{T}^{-1}\mathbf{B}\,\mathbf{v} = \lambda\,\mathbf{v} .$$

Dessa forma, $\mathbf{v}$ é o vetor próprio de $\mathbf{T}^{-1}\mathbf{B}$, corresponde ao valor próprio λ .

Por outro lado, tomando a equação $\mathbf{Bv} = \lambda\mathbf{Tv}$ e multiplicando por $\mathbf{v}'$ à esquerda dos dois membros, obtem-se

$$\mathbf{v}'\,\mathbf{B}\,\mathbf{v} = \lambda\,\mathbf{v}'\mathbf{T}\,\mathbf{v}$$

donde $$\lambda = \frac{\mathbf{v}'\mathbf{B}\,\mathbf{v}}{\mathbf{v}'\mathbf{T}\,\mathbf{v}}$$

que é exatamente a quantidade que desejávamos maximizar

■

(2.8.1) COROLÁRIO

Os valores próprios λ de $\mathbf{T}^{-1}\mathbf{B}$, são todos positivos e e inferiores a 1 (isto é, $0 \leq \lambda \leq 1$).

Demonstração

Como $\mathbf{v}'\mathbf{B}\,\mathbf{v}$ e $\mathbf{v}'\mathbf{T}\,\mathbf{v}$ são variâncias, então

$\lambda = \frac{\mathbf{v}'\mathbf{B}\,\mathbf{v}}{\mathbf{v}'\mathbf{T}\,\mathbf{v}} > 0$.

Por outro lado

$$\lambda = \frac{\mathbf{v}'\mathbf{B}\,\mathbf{v}}{\mathbf{v}'\mathbf{T}\,\mathbf{v}} = \frac{\mathbf{v}'\mathbf{T}\,\mathbf{v} - \mathbf{v}'\mathbf{W}\,\mathbf{v}}{\mathbf{v}'\mathbf{T}\,\mathbf{v}} = 1 - \frac{\mathbf{v}'\mathbf{W}\,\mathbf{v}}{\mathbf{v}'\mathbf{T}\,\mathbf{v}} < 1$$

■

De maneira análoga, definiremos o segundo eixo fatorial discriminante $\mathbf{u}^2$, como sendo o segundo vetor próprio de $\mathbf{T}^{-1}\mathbf{B}$ e Q-ortogonal a $\mathbf{u}^1$; o qual, constitue o melhor eixo fatorial discriminante independente do primeiro.

De agora em diante chamaremos de PODER DISCRIMINANTE DO VETOR $\mathbf{u}^1$ (ou eixo fatorial $\mathbf{u}^1$) à quantidade λ_1 .

Observe-se que considerando a variância total ao longo da direção definida por $\mathbf{u}^1$, o poder discriminante λ_1 varia entre zero e hum, pois, se for igual a hum (*1*) a variância intra-classes será nula, isto é, os pontos da mesma classe têm as mesmas abscissas sobre o eixo fatorial discriminante correspondente; sendo essas abscissas diferentes para classes distintas. Por outro lado, se for igual a *0* (zero), os pontos médios de cada classe têm as mesmas abscissas sobre o eixo fatorial.

2.9 EIXOS FATORIAIS DISCRIMINANTES SUCESSIVOS

Conforme foi visto anterioremente, o vetor próprio de $\mathbf{T}^{-1}\mathbf{B}$ relativo ao segundo valor próprio $\lambda_2 (\lambda_2 \leq \lambda_1)$, define o segundo eixo fatorial discriminante e, a seguir, para cada um dos vetores próprios sucessivos, duas questões se apresentam:

A_1) Quantos vetores próprios independentes existem?

A_2) Quantos eixos fatoriais podem se extrair de tal forma que sejam significativamente discriminantes?

Primeiramente, observe-se que dentro da maior parte dos problemas de discriminação, o número total de indivíduos N é superior ao número de variáveis p consideradas, que por sua vez é superior ao número de classes k.

De um modo geral, os indivíduos $\mathbf{x}_i^r$ geram o espaço $\mathbb{R}^p$, ao passo que os elementos $\mathbf{g}^r$ (pontos médios de cada uma das k classes)

geram uma variedade linear afim V de dimensão :

$$dim(\text{variedade linear gerada pelos } C_r) \leq min\ \{p, k-1\}.$$

Observe-se que V é gerado pelos k pontos correspondentes aos centros das classes e por isso não deverá ser confundido com o subespaço de dimensão k gerado pelos k vetores correspondentes.

Por outro lado, temos que a forma quadrática $\mathbf{u}\,\mathbf{T}\,\mathbf{u}'$ é definida positiva e, portanto define sobre o R^p uma estrutura euclidiana o produto escalar associado sendo definido mediante a matriz inversa $\mathbf{T}^{-1}$. Assim sendo, a distância entre dois pontos $\mathbf{x}_1$, $\mathbf{x}_2 \in R^p$, é dado por:

$$(2.9.1) \qquad d(x_1,x_2) = (x_1-x_2)'\ T^{-1}(x_1-x_2)$$

Dessa maneira as formas lineares discriminantes podem ser definidas geometricamente da seguinte maneira:

Em V , munido da métrica induzida pela matriz $\mathbf{T}^{-1}$ ficam determinados os vetores diretores $\mathbf{f}_1,\mathbf{f}_2,\ldots,\mathbf{f}_{k-1}$ dos eixos principais de inércia do sistema de pontos $\mathbf{x}_i^r$ afetado das massas $p_r(\mathbf{x}_i^r)$. As formas lineares discriminante são as formas,

$$(2.9.2) \qquad \mathbf{v}_i = \mathbf{T}^{-1}\,\mathbf{f}_i$$

Já no caso de duas classes de centros $\mathbf{g}^1$ e $\mathbf{g}^2$, tem-se $\mathbf{f}_1 = \mathbf{g}^1 - \mathbf{g}^2$ e a forma linear discriminante será:

$$(2.9.3) \qquad \mathbf{v}_1 = \mathbf{T}^{-1}(\mathbf{g}^1 - \mathbf{g}^2) \quad .$$

Portanto, observa-se que sob as hipóteses $N > p > k$, existem exatamente $(k-1)$ vetores próprios de $\mathbf{T}^{-1}\mathbf{B}$, isto é, $(k-1)$ eixos fatoriais discriminantes.

2.10 FUNÇÃO LINEAR DISCRIMINANTE DE FISHER E D^2 DE MAHALANOBIS

Em 1936 Fisher introduziu a função linnear discriminante, para duas classes, como sendo: "A função linear das variáveis iniciais, tais que: a razão do quadrado da diferença das médias (para cada uma das duas classes) desta função à variância desta função (variância calculada a partir da matriz de covariância intra-classe), seja máxima. É bem assim, que formulamos em (2.7.3), o problema da procura do primeiro eixo fatorial discriminante, e não será surpresa que recaimos sobre a função de Fisher, no caso de duas classes.

Os coeficientes desta função linear, podemos denotar como sendo um vetor linha

$$(g^1 - g^2)' \; T^{-1}$$

Se denotarmos por $\lambda_1, \lambda_2, \ldots, \lambda_p$ as componentes deste vetor, o valor desta função 1 discriminante para um indivíduo **x** escreve-se como:

(2.10.1) $$f(x) = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_p x_p$$

De posse dos fatos apresentados, podemos mostrar que o vetor v_1 é o vetor próprio de $T^{-1}B$. Para tal fim escreve-se conforme a definição *(2.4.8.2), o termo geral da matriz de covariância inter-classes*, para o caso de duas classes, com centros g^1 e g^2. *Isto é;*

$$b_{ij} = \sum_{r=1}^{k} \frac{N_r}{N} (g_i^r - g_i)(g_j^r - g_j)$$

donde:

$$B = \frac{N_1}{N} (g^1 - g)(g^1 - g)' + \frac{N_2}{N} (g^2 - g)(g^2 - g)',$$

com $$g = \frac{N_1 g^1 + N_2 g^2}{N} \quad \text{e} \quad N = N_1 + N_2$$

a qual se transforma em:

$$(2.10.2) \qquad B = \frac{N_1 N_2}{N^2}(g^1 - g^2)(g^1 - g^2)'$$

Por outro lado, se λ e v designam o úncio valor próprio e o único vetor próprio de $T^{-1}B$, então:

$$(2.10.3) \qquad T^{-1}B\, v = \lambda\, v \quad ,$$

a qual se transforma em:

$$(2.10.4) \qquad T^{-1}\, \frac{N_1 N_2}{N^2}\, (g^1 - g^2)(g^1 - g^2)'\, v = \lambda\, v$$

quando substituimos o valor da expressão (2.10.2).

Afirmamos que $T^{-1}(g^1 - g^2) = v$ é o valor próprio

De fato:

$$T^{-1}\, \frac{N_1 N_2}{N^2}\, (g^1 - g^2)(g^1 - g^2)'\, v =$$

$$T^{-1}\, \frac{N_1 N_2}{N^2}\, (g^1 - g^2)(g^1 - g^2)'\, T^{-1}(g^1 - g^2) =$$

$$\frac{N_1 N_2}{N^2}\, T^{-1}(g^1 - g^2)(g^1 - g^2)'\, T^{-1}(g^1 - g^2) =$$

$$T^{-1}(g^1 - g^2)\, \frac{N_1 N_2}{N^2}\, (g^1 - g^2)'\, T^{-1}(g^1 - g^2)$$

Como $(g^1 - g^2)'\, T^{-1}(g^1 - g^2)$ é um escalar, então $\frac{N_1 N_2}{N^2}\, (g^1 - g^2)'\, T^{-1}(g^1 - g^2)$ também será um escalar, portanto

$$T^{-1}\, \frac{N_1 N_2}{N^2}(g^1 - g^2)(g^1 - g^2)'\, T^{-1}(g^1 - g^2) = \theta\, T^{-1}(g^1 - g^2) \quad ,$$

onde

$$\theta = \frac{N_1 N_2}{N^2}\, (g^1 - g^2)'\, T^{-1}(g^1 - g^2)$$

Como λ é o único valor próprio e $\mathbf{T}^{-1}(g^1 - g^2) = v$ é o único vetor próprio, então $\theta = \lambda$.

Dessa forma, $\frac{N_1 N_2}{N^2}(g^1 - g^2)' \mathbf{T}^{-1}(g^1 - g^2)$ é um escalar e exatamente o valor próprio λ.

Sendo assim:

$$\mathbf{T}^{-1}\frac{N_1 N_2}{N^2}(g^1 - g^2)(g^1 - g^2)' \mathbf{T}^{-1}(g^1 - g^2) = \lambda \mathbf{T}^{-1}(g^1 - g^2)$$

Nota-se que o valor próprio encontrado $\lambda = \frac{N_1 N_2}{N^2}(g^1 - g^2)' \mathbf{T}^{-1}(g^1 - g^2)$, o qual indica o poder discriminante da função discriminante encontrada , não é senão o D^2 de Mahalanobis a menos do coeficiente $\frac{N_1 N_2}{N^2}$. Com efeito; o D^2 de Mahalanobis se escreve como:

(2.10.4) $$D^2 = (g^1 - g^2)' \mathbf{T}^{-1}(g^1 - g^2) \quad ,$$

a qual não é uma maneira de medir a distância entre duas classes, mas é precisamente a distância entre os centros das classes, pela métrica definida por $\mathbf{T}^{-1}$.

2.11 MÉTODO DE CLASSIFICAÇÃO

Os tópicos que apresentamos nas secções anteriores, concernentes à discriminação, tinham objetivo descritivo. A partir de agora passaremos a nos interessar pelo objetivo de natureza decisional, isto é, pelo problema de classificação, propriamente dito.

Sendo assim, para um problema de identificação ou classificação, dispomos de um novo indivíduo (ou elemento) "anônimo" que designaremos pela letra **a** , que também corresponde a um vetor do R^p , de coordenadas $a_1, a_2, \ldots, a_p$. Para tal finalidade, teremos:

(2.11.1) DEFINIÇÃO DO PROCESSO

Sendo dado um novo indivíduo $\underline{a}$, desejamos saber a que classe ele pertence. De fato, supomos que o indivíduo $\underline{a}$ considerado pertence a uma das classes definidas inicialmente e queremos decidir de qual delas se trata. Para esse fim, procuraremos uma partição do R^p em k regiões, correspondentes às k classes.

Primeiramente, calculamos a distância pela métrica T^{-1}, de $\underline{a}$ ao centro g^r da classe C_r, conforme expressão (2.9.1). Logo:

$$(2.11.1.1) \qquad d(a,g^r) = (a - g^r)' T^{-1}(a - g^r) .$$

Logo após, decide-se afetar $\underline{a}$ à classe C_o, tal que:

$$(2.11.1.2) \qquad d(a,g_o) = \min \{d(a,g^r);\ g^r \in C_r\} ,$$

onde g_o é o centro da classe C_o.

De fato, é nula a probabilidade de serem iguais as distâncias de um ponto a duas classes distintas, quando as variáveis observadas assumem um conjunto não discreto de valores, (isto é, no caso de variáveis contínuas).

Observa-se que as quantidades $d(a,g^r)$ definida em (2.11.1.1), são funções quadráticas de $\underline{a}$, possuindo em comum o termo quadrado $a' T^{-1} a$; portanto, poderemos comparar as funções lineares de $\underline{a}$, relativa a cada classe C_r.

Para tal fim, considere-se

$$(2.11.1.3) \qquad v_{C_r}(a) = d(a,g^r) - a' T^{-1} a ,$$

donde:

$$(2.11.1.4) \qquad v_{C_r}(a) = g^{r'} T^{-1}(g^r - 2a) .$$

Nestas condições, a regra de decisão definida em (2.11.1.2) tor-

na-se:

"Decide-se afetar $\underline{a}$ à classe C_o ; tal que :

(2.11.1.5) $$v_o(a) = \min \{v_{C_r}(a) \; ; \; C_r \subset C\} \quad ,$$

onde $v_{C_r}(a)$ é definida em (2.11.1.4).

Para melhor compreensão, vejamos uma aplicação, para o caso de duas classes C_1 e C_2 .

Neste caso, existem somente duas funções $v_{C_1}(a)$ e $v_{C_2}(a)$ a comparar.

A regra de decisão é a seguinte: "Tomamos à classe C_1 o indivíduo $\underline{a}$ se :

$$v_{C_1}(a) < v_{C_2}(a) \text{ "}$$

Utilizando os valores de $v_{C_r}(a)$ indicada em (2.11.1.4), tem-se:

$$v_{C_2}(a) = g^{2'} T^{-1}(g^2 - 2a)$$

$$v_{C_1}(a) = g^{1'} T^{-1}(g^1 - 2a)$$

donde: $$v_{C_2}(a) - v_{C_1}(a) = g^{2'} T^{-1}(g^2 - 2a) - g^{1'} T^{-1}(g^1 - 2a)$$

$$= g^{2'} T^{-1} g^2 - 2g^{2'} T^{-1} a - g^{1'} T^{-1} g^2 + 2g^1 T^{-1} a$$

$$= g^{2'} T^{-1} g^2 - g^{1'} T^{-1} g^1 + 2(g^1 - g^2)' T^{-1} a > 0$$

Por outro lado, observa-se que:

$$g^{1'} T^{-1} g^1 - g^{2'} T^{-1} g^2 = (g^1 - g^2)' T^{-1}(g^1 + g^2) \quad ,$$

pois:

$$g^{1'} T^{-1} g^2 = g^{2'} T^{-1} g^1 \quad ,$$

Portanto, a desigualdade anterior, escreve-se:

$$= (g^2 - g^1)' T^{-1}(g^2 + g^1) + 2(g^1 - g^2)' T^{-1} a > 0$$

$$= \frac{1}{2}(g^2 - g^1)' T^{-1}(g^2 + g^1) + (g^1 - g^2)' T^{-1} a > 0$$

$$= (g^1 - g^2)' T^{-1}a - \frac{1}{2}(g^1 - g^2)' T^{-1}(g^1 + g^2) > 0 ,$$

e a regra de decisão, torna-se: "AFETA-SE a à classe C_1 se :

(2.11.1.6) $$(g^1 - g^2)' T^{-1}a > \frac{1}{2}(g^1 - g^2)' T^{-1}(g^1 + g^2)$$ "

e; "AFETA SE a à classe C_2 se :

(2.11.1.7) $$(g^1 - g^2)' T^{-1}a < \frac{1}{2}(g^1 - g^2)' T^{-1}(g^1 + g^2)$$ "

Nota-se que, o termo da esquerda das expressões (2.11.1.6) e (2.11.1.7) é exatamente a função linear discriminante introduzida por R. A. Fisher (que é a função linear maximizando a razão do quadrado da diferença das médias de duas classes, pela variância global).

CAPÍTULO 3

DISCRIMINAÇÃO SOB A HIPÓTESE DE LEIS NORMAIS

3.1 INTRODUÇÃO

Neste capítulo consideraremos a distribuição dos valores assumidos por cada variável, relativamente aos indivíduos de uma classe, como sendo uma distribuição aleatória. Assim, para cada classe e cada variável supomos existir uma lei probabilística, o que de certa forma nos traz dificuldades para conhecê-la com exatidão.

Além disso, as variáveis tratadas no presente contexto são de caracter quantitativo, podendo em geral assumir um conjunto contínuo de valores reais. Lembramos que uma variável podendo assumir distintos valores, sujeitos a certas probabilidades, chama-se uma variável aleatória.

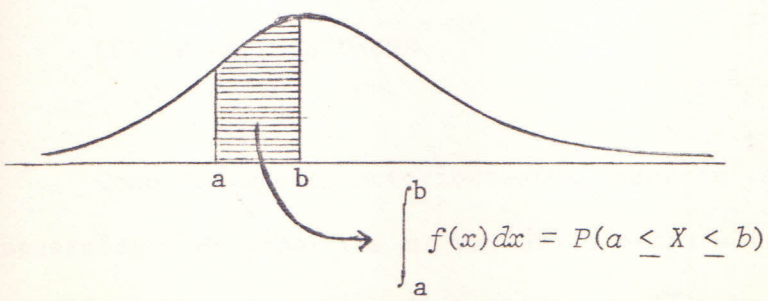
No caso em que a variável X é contínua e unidimensional, supomos existir certa função $f(x)$, ou densidade probabilística, de sorte que a probabilidade de X estar entre dois valores quaisquer a e b é dada por:

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad ;$$

Note-se que uma função densidade $f(x)$ está sempre sujeita às seguintes condições:

(i) $f(x) \geq 0 \qquad \forall \; x \in \mathbb{R}$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1 \quad .$



No caso bidimensional, dispomos de um vetor aleatório $\xi=(X_1,X_2)$, sendo $f(x_1,x_2)$ sua densidade, tal que:

(i) $\quad f(x_1,x_2) \geq 0 \qquad \forall \ (x_1,x_2) \in R^2$

(ii) $\quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1,x_2)\,dx_1 dx_2 = 1 \ .$

Então, a probabilidade de $\xi = (X_1,X_2)$ se encontra numa dada região N do plano, será:

$$P(\xi \in N) = \int\int_N f(x_1,x_2)\,dx_1 dx_2 \ ;$$

se

$N = [a, b] \times [c, d]$, então :

$$P(a \leq X_1 \leq b,\ c \leq X_2 \leq d) = \int_a^b \int_c^d f(x_1,x_2)\ dx_1 dx_2 \ .$$

O conceito de função densidade pode ser estendida evidentemente, para o caso de um vetor aleatório $\xi = (X_1,\ldots,X_p)$ de qualquer dimensão $p > 1$. Assim, determinar a lei de ξ corresponde a determinar sua função densidade $f(\xi) = f(x_1,x_2,\ldots,x_p)$.

Vamos nos restringir a situações em que a lei associada ao vetor ξ pertence à chamada família de leis multinormais (ou leis de Laplace-Gauss), à cujo respeito nos deteremos no parágrafo seguinte.

3.2 LEI DE LAPLACE-GAUSS

Como foi visto anteriormente, ocorrem situações em que há necessidade de supor que as variáveis estão sujeitas a certas leis probabilísticas. As leis multinormais, também chamadas de leis normais

multivariadas (ou leis de Laplace-Gauss), são aquelas supostas mais comumente. Isso é decorrência, em parte, de sua universalidade; ou seja, são aquelas que mais frequentemente ocorrem na prática. Além disso, os testes usuais de significância estatística exigem de regra a normalidade como pré-requisito para sua aplicação.

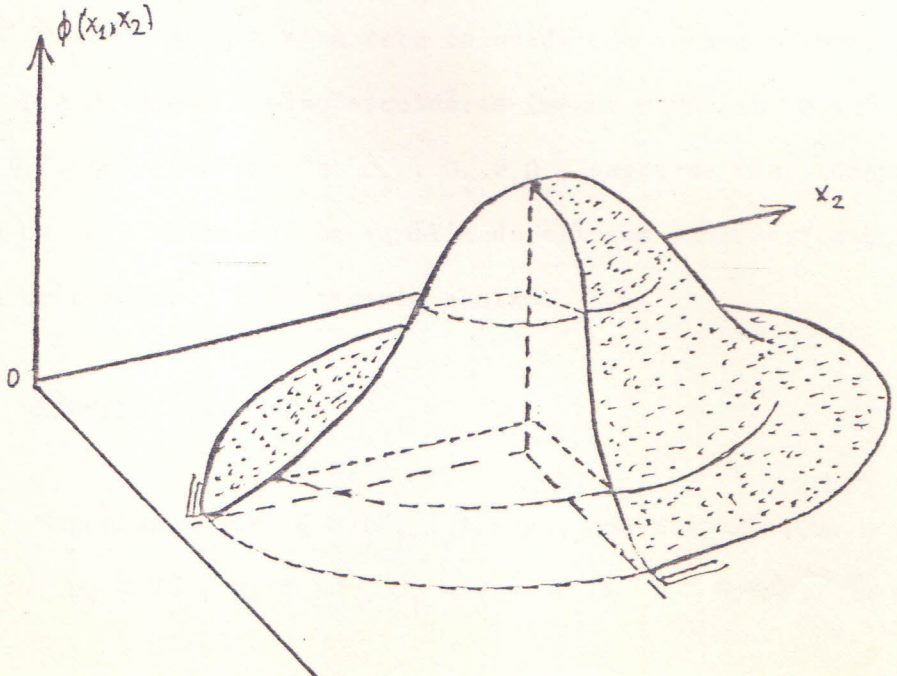
Assim sendo, para o caso de um vetor aleatório $\xi = (X_1, X_2)$ bidimensional a função densidade normal bivariada correspondente é dada por:

$$(3.2.1) \qquad f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \right.\right.$$

$$\left.\left. - 2\rho\,\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right], \qquad \forall\ (x_1,x_2) \in R^2\ ;$$

onde μ_i e σ_i^2 são as médias e variâncias de X_i $(i = 1,2)$, respectivamente, e ρ é o coeficiente de correlação entre X_1 e X_2.

Sabemos que uma função de duas variáveis pode ser representada graficamente por uma superfície no espaço a três dimensões. Assim, a superfície representada por (3.2.1) assemelha-se a um sino, como revela a figura seguinte (para $\rho = 0,6$ e $\sigma_1/\sigma_2 = 1$) :



Por outro lado podemos também descrever funções de duas variáveis por meio de curvas ou linhas de nível, obtidas seccionando a superfície $z = f(x_1,x_2)$ com os planos $z = c$ (constante) e, em seguida projetando as curvas resultantes de tais intersecções no plano $x_1\ 0\ x_2$.

No nosso caso, isso corresponde a considerar curvas no plano $x_1\ 0\ x_2$, de equação:

$$(3.2.1.1) \qquad \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\,\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} = c \quad .$$

Cada uma dessas equações representa uma elípse com centro no ponto (μ_1,μ_2), o qual é chamado de centroide da população bivariada. Além disso, cada elípse tem um dos eixos (o principal ou o secundário) coincidindo com uma reta passando pelo ponto (μ_1,μ_2) e fazendo ângulo θ com o eixo positivo $x_1\ 0\ x_2$, tal que:

$$\theta = \begin{cases} \frac{1}{2}\ arctg\ \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2-\sigma_2^2} , & \text{se} \quad \sigma_1 \neq \sigma_2 \\ 45^o , & \text{se} \quad \sigma_1 = \sigma_2 \end{cases}$$

Observa-se que essa reta coincide com o eixo principal (maior eixo) se $\rho > 0$ e com o eixo secundário (menor eixo) se $\rho < 0$. Como o ângulo θ depende somente de σ_1 , σ_2 e ρ , segue-se que, tomando vários valores de c , obtem-se uma família de elípses concêntricas, todas com a mesma orientação. Vide exemplo abaixo.

EXEMPLO (a)

Suponhamos que $\xi = (X_1,X_2)$ segue uma distribuição normal bivariada com $\mu_1 = 15$, $\mu_2 = 20$, $\sigma_1 = \sigma_2 = 5$ e $\rho = 0,60$. Então a expressão

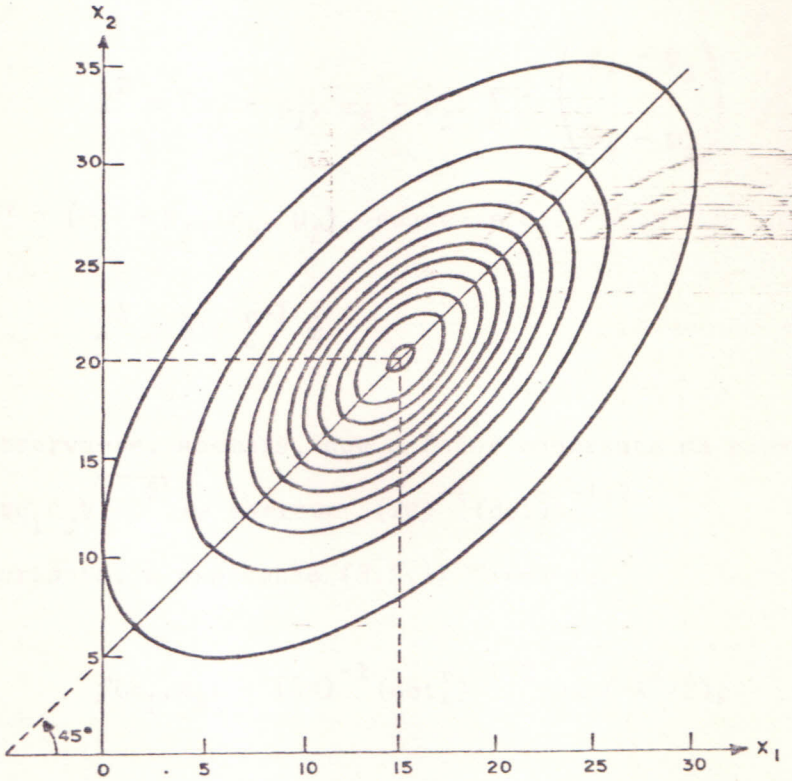
(3.2.1.1) torna-se:

$$\frac{(x_1-15)^2}{5^2} + \frac{(x_2-20)^2}{5^2} - 2.0,60\,\frac{(x_1-15)(x_2-20)}{5.5} = C\ ,$$

ou ainda,

$$(x_1-15)^2 + (x_2-20)^2 - 1,2\ (x_1-15)(x_2-20) = 25C\ ;$$

ou seja, é uma equação definindo contornos elípticos de equidensidade, ou uma família de elípses concêntricas, cujo centro é o ponto $(15,20)$ e tendo o eixo maior fazendo um ângulo de 45^o com o eixo $x_1 0 x_2$, uma vez que $\sigma_1 = \sigma_2$ e $\rho > 0$. A figura abaixo mostra as elípses de equidensidades correspondentes a determinados valores da constante C, a saber: $C = 5,89$; $C = 2,95$; $C = 2,06$ (para as três elípses mais internas) e $C = 0,29$; $C = 0,13$; $C = 0,01$ (para as três mais externas).



Veremos, agora, ser conveniente que se escreva a equação(3.2.1) na forma matricial.

Primeiramente definimos a matriz de variâncias-covariâncias ou matriz de dispersão para uma população bivariada, como sendo:

$$(3.2.2) \qquad \sum = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix} ,$$

de sorte que: $\det \sum = \sigma_1^2\sigma_2^2(1 - \rho^2)$;

consequentemente, supondo a inversibilidade de $\sum$ temos:

$$(3.2.2.1) \qquad \sum^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1/\sigma_1^2 & -\rho/\sigma_1\sigma_2 \\ -\rho/\sigma_2\sigma_1 & 1/\sigma_2^2 \end{pmatrix}$$

Note-se que a expressão no expoente da equação (3.2.1), a menos do fator $(-1/2)$, é equivalente à forma quadrática.

$$(3.2.3) \qquad X^2 = (x_1 - \mu_1, x_2 - \mu_2) \sum^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} ;$$

pondo $\tilde{x}' = (x_1 - \mu_1, x_2, \mu_2)$, segue-se :

$$(3.2.3.1) \qquad X^2 = \tilde{x}' \sum^{-1} \tilde{x}$$

Observa-se, ademais, que o fator constante da expressão (3.2.1), que é $1/2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}$ se escreve $(2\pi)^{-1}(\det\sum)^{-1/2}$

Portanto, a expressão (3.2.1) torna-se:

$$(3.2.4) \qquad f(x_1,x_2) = (2\pi)^{-1}(\det\sum)^{-1/2} \exp(-X^2/2),$$

que é a forma compacta para a função densidade normal bivariada.

No caso p-dimensional, definimos a matriz de variâncias-covariâncias $\sum$, como :

(3.2.5) $$\sum = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots\cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots\cdots & \rho_{2p}\sigma_2\sigma_p \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1}\sigma_p\sigma_1 & \rho_{p2}\sigma_p\sigma_2 & \cdots\cdots & \sigma_p^2 \end{pmatrix}$$

onde σ_i^2 é a variância de X_i e ρ_{ij} $(i \neq j)$ é o coeficiente de correlação entre X_i e X_j. Seja:

(3.2.5.1) $$\mathbf{x}^2 = \tilde{x}' \sum^{-1} \tilde{x} \ ,$$

com:

$$\tilde{x}' = (x_1-\mu_1,\ x_2-\mu_2,\ \ldots\ldots,x_p-\mu_p) \ ;$$

então, a função densidade normal p-variada é dada por:

(3.2.6) $$f(x_1,x_2,\ldots,x_p) = k\ exp(-X^2/2) \ ,$$

onde:

$$k = (2\pi)^{-p/2}(\det\sum)^{-1/2} \ ,$$

ademais, μ é o vetor médio de dimensão p correspondente às coordenadas do centro da distribuição, enquanto $\sum$ é uma matriz quadrada, simétrica e positiva definida de ordem p (matriz de variâncias-covariâncias).

3.3 MÉTODOS DE CLASSIFICAÇÃO DE NOVOS INDIVÍDUOS E FUNÇÕES DISCRIMINANTES

No presente parágrafo abordamos o problema de atribuição que aparece, naturalmente, quando se dispõe de certas medidas (ou observações) sobre cada indivíduo e, em decorrência, desejamos classificá-lo em uma dentre várias classes, com base nessas medidas. Como foi estabelecido

anteriormente, neste capítulo, supomos que todas as variáveis envolvidas seguem leis normais.

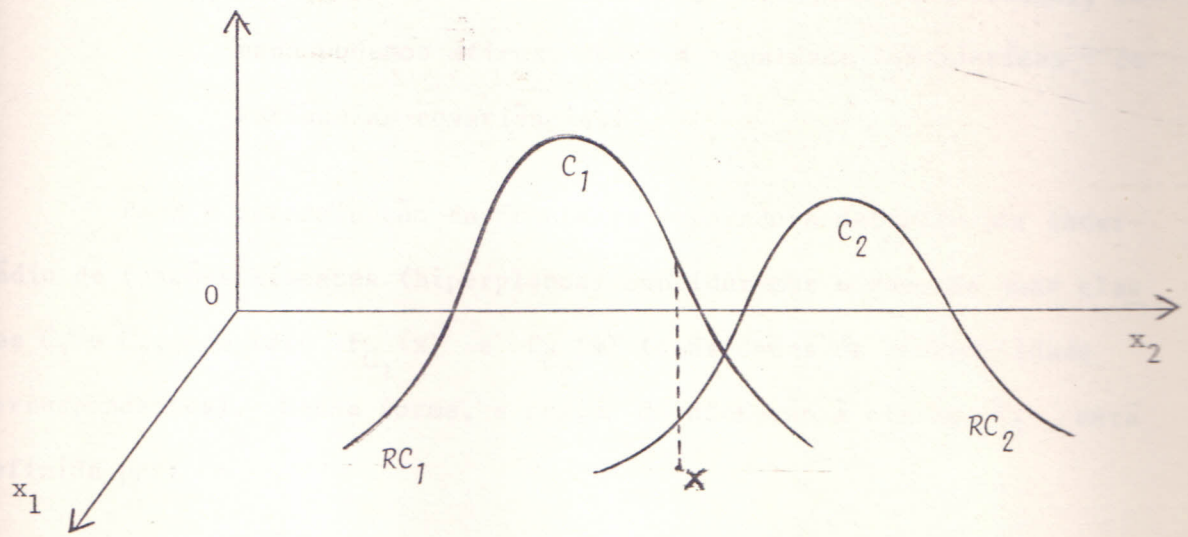
De um modo geral, na classificação de novos indivíduos, põe-se a questão de se considerar custos de erros, interpretados como perdas ou como multas a pagar pela classificação incorreta dos indivíduos.

Dessa forma, estudaremos os métodos de classificação ligados à Análise Discriminante, seja sem considerar custos de erros, que é o caso mais simples; seja numa etapa seguinte, ao se considerar esses custos.

3.3.1 CLASSIFICAÇÃO SEM "CUSTOS DE ERROS"

Consideraremos duas classes C_1 e C_2 e variáveis X_1 e X_2. Não levaremos em conta os custos da ocorrência de erros.

Seja a figura abaixo, onde as densidades de probabilidade relativas às classes C_1 e C_2, respectivamente, são representadas por duas superfícies em forma de sino (distribuições normais bivariadas).



Assim sendo, parece natural adotarmos o seguinte método ou critério de classificação:

"Afetar um novo indivíduo $x = (x_1, x_2)$ à classe para a qual a

densidade de probabilidade respectiva é mais forte; ou em outras palavras, à classe para a qual se tenha maior probabilidade a posteriori de se obter os valores x_1 e x_2."

Em consequência, a fronteira de separação das classes C_1 e C_2 é definida pela projeção, sobre o plano $x_1\ 0\ x_2$, da intersecção das duas superfícies.

É claro que o método assim descrito se generaliza, de imediato, a várias classes; observando-se que tal procedimento não concede privilégio a nenhuma classe e tampouco considera os custos de possíveis erros.

O problema consiste em definir de maneira precisa as fronteiras separadoras, as quais são de dois tipos, a saber:

(i) **Hiperplanos** (definidos por meio de funções lineares), se supomos que as leis multinormais correspondentes às várias classes possuem a mesma matriz de variâncias-covariâncias.

(ii) **Hipersuperfícies** (definidos por funções quadráticas), se nada podemos afirmar sobre a igualdade das matrizes de variâncias-covariâncias.

Para a determinação da fronteira separadora definida por intermédio de funções lineares (hiperplanos) consideremos o caso de duas classes C_1 e C_2, com leis $f_{C_1}(\mathbf{x})$ e $f_{C_2}(\mathbf{x})$ (densidades de probabilidade correspondentes). Dessa forma, a região de afetação à classe C_1 será definida por:

(3.3.1.1) $$R_{C_1} = \{x\ ;\ f_{C_1}(x) > f_{C_2}(x)\}\ ,$$

supondo $$\Sigma_{C_1} = \Sigma_{C_2} = \Sigma\ .$$

Portanto, a condição $f_{C_1}(x) > f_{C_2}(x)$ traduz-se por:

$$(2\pi)^{-p/2}(\det\Sigma_{C_1})^{-1/2}\, exp[-1/2(x-\mu_{C_1})'\Sigma_{C_1}^{-1}(x-\mu_{C_1})] >$$

$$> (2\pi)^{-p/2}(\det\Sigma_{C_2})^{-1/2}\, exp[-1/2(x-\mu_{C_2})'\Sigma_{C_2}^{-1}(x-\mu_{C_2})]$$

donde:

(3.3.1.2) $$(x-\mu_{C_1})'\Sigma^{-1}(x-\mu_{C_1}) < (x-\mu_{C_2})'\Sigma^{-1}(x-\mu_{C_2}) .$$

Observa-se que o método de classificação apresentado no capítulo 2 é análogo a este, bastanto substituir $\underline{x}$ por $\underline{a}$ e μ_{C_1}, μ_{C_2} por y; sendo T, ali, a matriz de covariância total.

Por outro lado, para a determinação da fronteira separadora definida através de funções quadráticas, consideremos novamente o caso de duas classes C_1 e C_2, de leis $f_{C_1}(x)$ e $f_{C_2}(x)$, respectivas. Nesse caso, a região de afetação à classe C_1, será definida por

(3.3.1.3) $$R_{C_1} = \{ x ; f_{C_1}(x) > f_{C_2}(x) \} ;$$

supondo $$\Sigma_{C_1} \neq \Sigma_{C_2} .$$

Portanto, a condição $f_{C_1}(x) > f_{C_2}$ traduz-se agora por:

$$(2\pi)^{-p/2}(\det\Sigma_{C_1})^{-1/2} exp[-1/2(x-\mu_{C_1})'\Sigma_{C_1}^{-1}(x-\mu_{C_1})] >$$

$$> (2\pi)^{-p/2}(\det\Sigma_{C_2})^{-1/2} exp[-1/2(x-\mu_{C_2})'\Sigma_{C_2}^{-1}(x-\mu_{C_2})] ;$$

donde:

(3.3.1.4) $$\frac{1}{2}(x-\mu_{C_1})'\Sigma_{C_1}^{-1}(x-\mu_{C_1}) + \frac{1}{2}\ell n \det\Sigma_{C_1} <$$

$$< \frac{1}{2}(x-\mu_{C_2})'\Sigma_{C_2}^{-1}(x-\mu_{C_2}) + \frac{1}{2}\ell n \det\Sigma_{C_2} .$$

Observamos aqui, a analogia desta expressão com o método de classificação de Sebestyen dada pela expressão (1.5.4.3), Capítulo 1.

3.3.2 CLASSIFICAÇÃO COM "CUSTOS DE ERROS"

Vejamos, agora, o problema da classificação de indivíduos, quando intervêm "custos de erros".

Para tal fim, consideremos o caso de duas classes C_1 e C_2, estando o espaço R^p particionado em regiões R_1 e R_2, designadas como as "regiões críticas" para a afetação de um indivíduo arbitrário na classe C_1 ou C_2, respectivamente. Assim, se a p-upla (vetor no R^p) que representa o indivíduo cai na região R_1, ele estará classificado na classe C_1; caso contrário, isto é, se ele cai em R_2, estará classificado em C_2.

Evidentemente, existem possibilidades de má classificação (ou erros de classificação), isto é, de um indivíduo pertencer a uma classe e ser classificado na outra. A partir desse erro de classificação é que surge a questão do "custo de erro"; o qual será notado por C(i/j), significando dizer que o indivíduo é classificado em C_i quando, na verdade, se encontra em C_j.

Pode-se apreciar na tabela abaixo a indicação dos custos de correta e incorreta classificação, para o caso de duas classes C_1 e C_2.

CLASSES	CLASSIFICAÇÃO C_1	CLASSIFICAÇÃO C_2
C_1	C(1/1) = 0	C(2/1) > 0
C_2	C(1/2) > 0	C(2/2) = 0

modo tem-se de minimizar o custo da má classificação; para isso considera-se distintas maneiras de definir o "custo mínimo" conforme dois casos a saber:

(i) probabilidades a priori conhecidas;

(ii) probabilidades a priori não conhecidas.

– PROCESSO DE CLASSIFICAÇÃO COM PROBABILIDADES "A PRIORI" CONHECIDAS (caso de duas classes)

Aqui, consideraremos q_i como sendo a probabilidade de um indivíduo arbitrário provir da classe C_i $(i = 1,2)$. Além disso, suponhamos que as distribuições envolvidas, relativamente a cada classe, tenham densidades; ou seja, $f_{C_i}(x)$ é a densidade associada a C_i; $i = 1,2$.

Se R_i é a região crítica (ou região de decisão) associada a C_i, então:

(3.3.2.1) $$P(1/1,R) = \int_{R_1} f_{C_1}(x)\,dx$$

representa a probabilidade de um indivíduo pertencente a classe C_1 ser bem classificado; enquanto

(3.3.2.2) $$P(2/1,R) = \int_{R_2} f_{C_1}(x)\,dx$$

é a probabilidade da má classificação. Na notação $P(j/i,R)$, tem-se $R = \{R_1,R_2\}$, que se refere a uma dada partição do espaço em duas regiões de decisão.

De maneira análoga, consideram-se as probabilidades de correta classificação e de má classificação, respectivamente, de um indivíduo proveniente de C_2.

Visto que a probabilidade de se extrair um indivíduo de C_1 é q_1, então a probabilidade de um indivíduo ser extraído de C_1 e, em seguida ser bem classificado, torna-se $q_1\ P(1/1,R)$; no caso de má classificação tem-se $q_1\ P(2/1,R)$. Toma-se o produto de probabilidades, tendo em vista a hipótese de independência entre a escolha do indivíduo e sua classificação.

Analogamente, para a classe C_2, nos casos de boa e de má classificação tem-se as probabilidades $q_2\ P(2/2,R)$ e $q_2\ P(1/2,R)$.

Sendo assim, pode-se definir o custo esperado, como a soma dos produtos de cada custo da má classificação, pelas probabilidade de sua respectiva ocorrência, isto é,

$$\overline{C}(R) = C(2/1)\ q_1\ P(2/1,R) + C(1/2)\ q_2\ P(1/2,R)\ , \tag{3.3.2.5}$$

onde q_1 e q_2 são conhecidas.

Este custo esperado (ou custo médio de má classicação) é aquele que se deseja minimizar. Os procedimentos utilizados para tal, dizem-se procedimentos de Bayes, os quais dependem da escolha de uma partição $R^* = \{R_1\ ,\ R_2\}$, adequada (conforme a definição seguinte).

(3.3.2.6) DEFINIÇÃO

R^* determina um procedimento de Bayes (portanto um processo ótimo), se $\overline{C}(R^*) \leq \overline{C}(R)$; $\forall\ R$.

Em outras palavras, para que se possa minimizar o custo esperado $\overline{C}(R)$, devem ser escolhidas regiões apropriadas $R_1 = R_1^*$ e $R_2 = R_2^*$.

Em seguida, podemos definir probabilidades associadas às classes e ao conjunto de variáveis observadas. Seja $y = (y_1, y_2, \ldots, y_p) \in \mathbb{R}^p$; por outro lado, seja $x = (x_1, \ldots, x_p)$ um indivíduo arbitrário. Então, a

probabilidade desse indivíduo provir da classe C_1 e ser tal que $x_i \leq y_i$, $i = 1,2,\ldots,p$ é dada por:

$$(3.2.2.7) \qquad P_{C_1}(y_1,\ldots,y_p) = \int_{-\infty}^{y_1}\int_{-\infty}^{y_2}\cdots\int_{-\infty}^{y_p} q_1 \, f_{C_1}(x)\, dx_1 dx_2,\ldots,dx_p \ .$$

Por outro lado, podemos definir a probabilidade condicional de um indivíduo $x = (x_1,x_2,\ldots,x_p)$ provir de uma dada classe, sendo conhecidos os valores $x_1,x_2,\ldots,x_p$; isto é,

$$(3.2.2.8) \qquad \frac{q_1 f_{C_1}(x)}{q_1 f_{C_1}(x) + q_2 f_{C_2}(x)} \ .$$

Ora, relativamente à expressão (3.3.2.5) e supondo $C(1/2) = C(2/1) = 1$, vem :

$$(3.2.2.9) \qquad \overline{C}(R) = q_1 \int_{R_2} f_{C_1}(x)dx + q_2 \int_{R_1} f_{C_2}(x)dx \ ,$$

para o custo da má-classificação; o que não deixa de ser uma probabilidade de má-classificação (tendo em vista a hipótese $C(1/2) = C(2/1) = 1$).

Para um dado indivíduo x observado, minimizamos sua probabilidade de má-classificação, atribuindo-o à classe à qual corresponde a maior probabilidade condicional. Isto é, se

$$(3.2.2.10) \qquad \frac{q_1 f_{C_1}(x)}{q_1 f_{C_1}(x) + q_2 f_{C_2}(x)} \geq \frac{q_2 f_{C_2}(x)}{q_1 f_{C_1}(x) + q_2 f_{C_2}(x)} \ ,$$

classificamos o indivíduo x na classe C_1 . Caso contrário, será classificado na classe C_2 .

Uma vez que se minimizou a probabilidade de máclassificação em cada ponto, então o mesmo é feito para o espaço todo, donde a regra de decisão pode ser escrita:

$$(3.3.2.11)\qquad \begin{cases} R_1 : q_1 f_{C_1}(x) \geq q_2 f_{C_2}(x) \\ R_2 : q_2 f_{C_2}(x) > q_1 f_{C_1}(x) . \end{cases}$$

Observa-se que, se $q_1 f_{C_1}(x) = q_2 f_{C_2}(x)$, o indivíduo x poderia ser classificado indistintamente, em C_1 ou C_2; mas preferimos colocá-lo em C_1 . Contudo, se $q_1 f_{C_1}(x) + q_2 f_{C_2}(x) = 0$, o indivíduo x não poderia ser classificado en nenhuma das classes (em vista de (3.3.2.10)).

Mostraremos, em seguida, que (3.3.2.11) de fato nos dá o melhor procedimento de classificação.

Ora, para qualquer partição $R^* = \{R_1^* , R_2^*\}$, a probabilidade de má-classificação é dada por:

$$(3.3.2.12)\qquad \overline{C}(R^*) = \int_{R_2^*} q_1 f_{C_1}(x)\,dx + \int_{R_1^*} q_2 f_{C_2}(x)\,dx =$$

$$= q_1 \int_{R_2^*} f_{C_1}(x)\,dx - q_2 \int_{R_2^*} f_{C_2}(x)\,dx +$$

$$+ q_2 \int_{R_2^*} f_{C_2}(x)\,dx + q_2 \int_{R_1^*} f_{C_2}(x)\,dx =$$

$$= \int_{R_2^*} [q_1 f_{C_1}(x) - q_2 f_{C_2}(x)]\,dx + \int q_2 f_{C_2}(x)\,dx$$

Note-se que no último membro desta cadeia de igualdades o termo $q_2 \int f_{C_2}(x)\,dx$ é um valor constante (se $R_1 \cup R_2 = \mathbb{R}^p$, então fica apenas q_2); assim, $\overline{C}(R^*)$ será minimizado se R_2^* incluir pontos tais que $q_1 f_{C_1}(x) - q_2 f_{C_2}(x) < 0$ e excluir aqueles para os quais $q_1 f_{C_1}(x) - q_2 f_{C_2}(x) > 0$.

Com isso, mostramos que de fato a expressão (3.3.2.11) é o melhor procedimento possível.

Ademais, se tivermos garantia de que:

$$(3.3.2.13) \qquad Prob\,\{q_1 f_{C_1}(\mathbf{x}) - q_2 f_{C_2}(\mathbf{x}) = 0 \mid C_i\} = 0 \; ; \qquad i = 1,2 \; ,$$

então o procedimento de Bayes é único, a menos de conjuntos de probabilidade nula.

Se $C(1/2)$ e $C(2/1)$ são quaisquer (isto é, levantamos a restrição $C(1/2) = C(2/1) = 1$), então o custo se escreve:

$$(3.3.2.14) \qquad \overline{C}(R) = C(2/1)\; q_1 \int_{R_2} f_{C_1}(x)dx + C(1/2)\; q_2 \int_{R_1} f_{C_2}(x)dx \; ,$$

e escolheremos R_1 e R_2, conforme :

$$(3.3.2.15) \qquad \begin{cases} R_1 : C(2/1)\; q_1\; f_{C_1}(\mathbf{x}) \geq C(1/2)\; q_2\; f_{C_2}(\mathbf{x}) \quad ; \\ \\ R_2 : C(1/2)\; q_2\; f_{C_2}(\mathbf{x}) > C(2/1)\; q_1\; f_{C_1}(\mathbf{x}) \quad , \end{cases}$$

desde que $C(2/1)\; q_1\; C(1/2)\; q_2$ sejam constantes não-negativas.

— PROCESSO DE CLASSIFICAÇÃO COM PROBABILIDADES "A PRIORI" CONHECIDAS (caso de $p > 2$ classes)

No caso de haver p classes C_i com densidades $f_{C_i}(\mathbf{x})$ respectivas $(i=1,2,\ldots,p)$, sendo o espaço R^p particionado em regiões de decisão $R_1, R_2, \ldots, R_p$, as probabilidades de má-classificação serão:

$$(3.3.2.16) \qquad P(j/i,R) = \int_{R_j} f_{C_i}(x)dx \quad , \qquad i \neq j$$

Por outro lado, supondo as probabilidades a priori $q_1, q_2, \ldots, q_p$ conhecidas, o custo esperado define-se como:

(3.3.2.17) $$\overline{C}(R) = \sum_{i=1}^{p} q_i \left\{ \sum_{\substack{j=1 \\ j=i}}^{p} C(j/i)\ P(j/i,R) \right\} \ ;$$

as probabilidades condicionais análogas às definidas em (3.3.2.8), serão dadas por

(3.3.2.18) $$\frac{q_i\ f_{C_i}(x)}{\sum_{j=1}^{p} q_j\ f_{C_j}(x)} \ .$$

Obtem-se, portanto, um resultado análogo àquele já encontrado no caso $p = 2$.

(3.3.2.19) TEOREMA

Se q_i é a probabilidade a priori de se extrair um indivíduo de classe C_i, cuja densidade é $f_{C_i}(x)$; $(i=1,2,\ldots,p)$ e se o custo da má-classificação de um indivíduo de C_i como sendo de C_j é $C(j/i)$ $(i \neq j)$, então as regiões de decisão $R_1, R_2, \ldots, R_p$ que permitem minimizar o custo esperado de má-classificação são dadas pela condição $x \in R_k$, quando:

$$\sum_{\substack{i=1 \\ i\neq k}}^{p} q_1 f_{C_i}(x) C(k/i) < \sum_{\substack{i=1 \\ i\neq j}}^{q} q_i f_{C_i}(x) C(j/i) \ ,$$

para $j = 1,2,\ldots,p$ e $j \neq k$.

Demonstração

Consideremos $h_j(x) = \sum_{\substack{i=1 \\ i\neq j}}^{p} q_i\ f_{C_i}(x)\ C(j/i)$, então o custo esperado de um procedimento baseado em R é dado por:

$$\sum_{j=1}^{p} \int_{R_j} h_j(x)dx = \int h(x)dx$$

onde $h(x) = h_j(x)$ para $x \in R_j$.

Para o procedimento descrito no teorema, tem-se que $h(x)$ é $h^*(x) = \min_i h_i(x)$. Portanto:

$$\int [h(x) - h^*(x)]dx = \sum_j \int_{R_j} [h_j(x) - \min_i h_i(x)]dx \geq 0$$

A igualdade pode considerar-se somente quando $h_j(x) = \min_i h_i(x)$ para $x \in R_j$, exceto para conjuntos de probabilidade nula.

Vejamos como este método se aplica quando $C(j/i) = 1$ para todo i e j $(i \neq j)$. Neste caso, em R_k, tem-se :

(3.3.2.20) $$\sum_{\substack{i=1 \\ i\neq k}}^{p} q_i f_{C_i}(x) < \sum_{\substack{i=1 \\ i\neq j}}^{p} q_i f_{C_i}(x) \quad ;$$

subtraindo $\sum_{\substack{i=1 \\ i\neq k,j}}^{p} q_i f_{C_i}(x)$ de ambos os lados, obtem-se

(3.3.2.21) $$q_j f_{C_j}(x) < q_k f_{C_k}(x) \quad .$$

Neste caso, o indivíduo x está em R_k , se k é o índice para o qual $q_i f_{C_i}(x)$ é um máximo; ou seja, C_k é a classe mais provável.

3.3.3 PROCESSO DE CLASSIFICAÇÃO COM PROBABILIDADES A PRIORI DESCONHECIDAS

Suponhamos duas classes C_1 e C_2. As probabilidades a priori q_1

e q_2 são supostas não conhecidas. Então, o custo esperado de má-classificação, sob a hipótese de que o indivíduo $x \in C_1$, é dado por:

(3.3.3.1) $$r(1,R) = C(2/1) \quad P(2/1,R) \quad ;$$

enquanto, se $x \in C_2$, é dada por uma expressão análoga.

Consideremos dois preocedimentos R e R^*, este último suposto um procedimento de Bayes. Neste caso, diremos que "R é pelo menos tão bom quanto R^*" se:

(3.3.3.2) $$\begin{cases} r(1,R) \leq r(1,R^*) \\ \\ r(2,R) \leq r(2,R^*) \end{cases} ;$$

ao passo que "R é melhor que R^*, se ao menos uma das desigualdades precedentes vale estritamente.

Para o caso de várias classes e supondo ainda desconhecidas as probabilidades a priori, não é possível definir um custo esperado incondicional para um processo de classificação. No entanto, podemos definir um custo esperado sob a condição de que o elemento provém de uma dada classe, como foi feito linhas atrás. Assim, *o custo esperado condicional* de má-classificação se o elemento provém de C_i, define-se por :

(3.3.3.3) $$r(i,R) = \sum_{\substack{j=1 \\ j \neq i}}^{p} C(j/i) \quad P(j/i,R) \quad .$$

Podemos dizer que "R é pelo menos tão bom quanto R^*" se :

(3.3.3.4) $$r(i,R) \leq r(i,R^*) \; ; \qquad i = 1,2,\ldots,p \quad ,$$

ao passo que "R é melhor que R^*", se pelo menos uma das desigualdades vale estritamente.

A obtenção de um processo de Bayes, no caso das probabilidades a priori serem desconhecidas é, contudo, um problema um pouco mais complexo (exigindo a introdução do conceito de "procedimentos admissíveis de Bayes"), o que nos obrigaria a divergir bastante dos objetivos do nosso trabalho, caso tivéssemos que aí nos deter. Os detalhes, nesse caso, poderão ser encontrados em ANDERSON (1958).

Finalmente, note-se que para a discriminação sob a hipótese de leis normais, todas as densidades $f_{C_i}(x)$ supõe-se ser densidades de Laplace-Gauss.

CAPÍTULO 4

DISCRIMINAÇÃO PASSO A PASSO

4.1 INTRODUÇÃO

Neste capítulo apresentaremos os processos de discriminação passo a passo e analisaremos as vantagens dos possíveis critérios utilizáveis.

4.2 DEFINIÇÃO

A técnica básica envolvida num processo de discriminação passo a passo consiste em, dado certo conjunto de variáveis medidas sobre uma população, sucessivamente restringí-las à melhor, em seguida às duas melhores, às três melhores, etc., no sentido de assim permitir, de cada vez, uma melhor discriminação entre elementos pertencentes a classes distintas.

No caso da chamada regressão linear múltipla tem-se uma variável y a prever, com a ajuda de certo número de outras variáveis $x_1, x_2, \ldots, x_p$. A variável que se deseja prever é frequentemente chamada endógena, dependente ou "a explicar", enquanto que as demais são ditas exógenas, independentes ou "explicativas". Ora, no que concerne à regressão linear múltipla, podemos estar interessados em selecionar, dentre as variáveis exógenas $x_1, x_2, \ldots, x_p$, aquelas que mais contribuem para a previsão ou explicação da variável endógena y. Para tal fim, podem ser utilizados processos passo a passo (stepwise), sabendo distinguir entre processos "stepwise" ascendentes e descendentes.

Nos processos ascendentes, as variáveis são introduzidas uma a uma, de sorte a serem construídos subconjuntos de variáveis, de porte crescente; evidentemente, a variável a ser introduzida em cada etapa é aquela que melhor contribue para o aumento do "índice ou percentual de explicação" (onde R^2 é o "coeficiente de correlação múltipla"). Nos

processos descendentes, pelo contrário, partimos do conjunto de todas as variáveis exógenas $x_1, x_2, \ldots, x_p$, sendo sucessivamente eliminadas aquelas de pequeno poder explicativo com relação à variável endógena y .

Procede-se de maneira análoga com relação aos problemas de discriminação. Aqui, vamos nos restringir a processos passo a passo do tipo ascendente. Dessa forma, procura-se construir subconjuntos de variáveis garantindo a melhor discriminação possível, onde em cada etapa se acrescenta uma variável suplementar ao subconjunto retido no passo anterior. Assim, o objeto deste capítulo será estudar os diferentes critérios para a escolha da nova variável. Note-se que em cada passo não se colocará em causa o subconjunto considerado no passo anterior.

A vantagem do método passo a passo ascendente descrito acima é duplo, pois permite:

i) diminuição do "custo operacional", que se liga ao volume de cálculos a serem realizados.

ii) melhoria da confiabilidade do método.

No que concerne ao "custo operacional", podemos nos referir, tanto ao tempo necessário para efetuar esses cálculos como, de maneira equivalente, ao custo financeiro correspondente. Com efeito, em Análise Discriminante, os cálculos envolvidos são em geral impraticáveis sem a ajuda do computador; ora, quanto maior o volume de cálculos, maior será o tempo de processamento e o preço a ser pago pelo usuário.

Suponhamos que se escolha, ao azar, q variáveis; vamos chamar de "unidade de operação" ao conjunto de cálculos a realizar quando se deseja testar a discriminação proporcionada pela escolha de tais variáveis. No processo passo a passo em pauta, onde não se coloca em causa o subconjunto de variáveis previamente escolhidas, teremos a seguinte quantidade de "unidades de operação" em cada passo, a saber:

1º passo ⟶ p unidades de operação 1

2º passo ⟶ $p-1$ unidades de operação 2

⋮

q-ésimo passo ⟶ $p-q+1$ unidades de operação q

⋮

p-ésimo passo ⟶ 1 unidade de operação

donde o total de unidades de operação será:

(4.2.1) $$N_1 = p(p+1)/2$$

Pelo contrário, se em cada passo tivéssemos de por em causa a escolha anterior, o número total de unidades de operação seria drasticamente aumentado, com:

(4.2.2) $$N_2 = \sum_{q=1}^{p} \binom{p}{q} = 2^p - 1 \quad .$$

A seguinte tabela mostra a diferença entre os valores de N_1 e N_2 nas duas alternativas, conforme o valor de P (número de variáveis consideradas).

P	N_1	N_2
5	15	31
10	55	1.023
20	210	1048.575

Evidentemente, podemos atingir uma etapa a partir da qual as variáveis a serem sucessivamente acrescentadas pouco contribuem para a melhoria da discriminação entre as classes. Daí, a necessidade de se dispor

de critérios para dizer em que momento se deve parar. Note-se que o fato de nos determos num dado ponto em que apenas q variáveis são retidas ($q < p$), também contribue para a diminuição do "custo operacional".

Por fim, quanto à confiabilidade do método, reside exatamente no fato de se poder escolher um conjunto de variáveis que nos proporcione uma discriminação aceitável.

Com relação aos possíveis critérios de discriminação, consideram-se os seguintes, dentre outros:

i) critério da "porcentagem de bem classificados".

ii) critério do "traço da matriz $\mathbf{T}^{-1}\mathbf{B}$".

iii) critério do "Λ de Wilks"

iv) critério da "maximização das diferenças entre as médias condicionais para as diferentes classes".

No presente capítulo, estudaremos com detalhes os critérios (i) e (ii); quanto aos demais, não serão considerados em profundidade. De uma parte, eles constituem testes clássicos baseados na hipótese de multinormalidade, donde sua aplicabilidade se torna mais restrita dentro do ponto de vista da moderna Análise de Dados Multidimensionais. Por outro lado, seu estudo exige forte embasamento de Estatística Matemática, o que foge aos propósitos do nosso trabalho, mais dirigido para aspectos de Álgebra Linear e de Topologia Métrica.

4.3 PORCENTAGEM DE BEM CLASSIFICADOS (PRIMEIRO CRITÉRIO)

(4.3.1) INTRODUÇÃO E DEFINIÇÃO

O critério baseado em porcentagens de ítens bem classificados é intuitivo, apresentando-se nitidamente ao espírito, quando se queira avaliar a validade de um método de discriminação. Tal critério é de caráter

bem geral, aplicando-se indistintamente às várias técnicas de discriminação estudados nos capítulos precedentes.

Como foi visto, essas diversas técnicas, em última instância, a se obter uma partição do $\mathbb{R}^p$ em k regiões $R_1, R_2, \ldots, R_k$, as quais são supostas corresponder aproximadamente às k classes $C_1, C_2, \ldots, C_k$ já definidas "a priori".

No caso do Capítulo *1*, onde se descreve a técnica introduzida por Sebestyen, decide-se afetar cada indivíduo à classe com relação à qual o mesmo for mais próximo (sem olhar para o fato de pertencer ou não, "a priori", à referida classe). Obtem-se dessa maneira, uma partição do $\mathbb{R}^p$, nas k regiões $R_1, R_2, \ldots, R_k$; donde:

(4.3.1.1) $$R_r = \{\mathbf{x} \; ; \; \pi(x \; , \; C_r) \leq \pi(x \; , \; C_j)\} \; ;$$

$$r, \; j = 1, 2, \ldots, k \; .$$

Observe-se, sem dúvida que existe ambiguidade quanto à classificação dos pontos fronteiras, ou seja, os pontos tais que $\pi(\mathbf{x}, C_r) = \pi(\mathbf{x}, C_j)$; porém essa eventualidade não possue nenhuma chance de ocorrer, na prática.

Sendo esse particionamento efetuado a partir das amostras disponíveis, para as diversas classes, é natural que se pergunte sobre a possibilidade de alguns elementos ou ítens serem afetados a classes às quais de fato não pertençam.

A seguir, apresentamos um exemplo bastante simples, a título de esclarecimento.

(4.3.1.2) EXEMPLO

Por hipótese, dispomos de um total de *56* (cinqüenta e seis) indivíduos, os quais são, "a priori", supostos pertencer a classes C_I ,

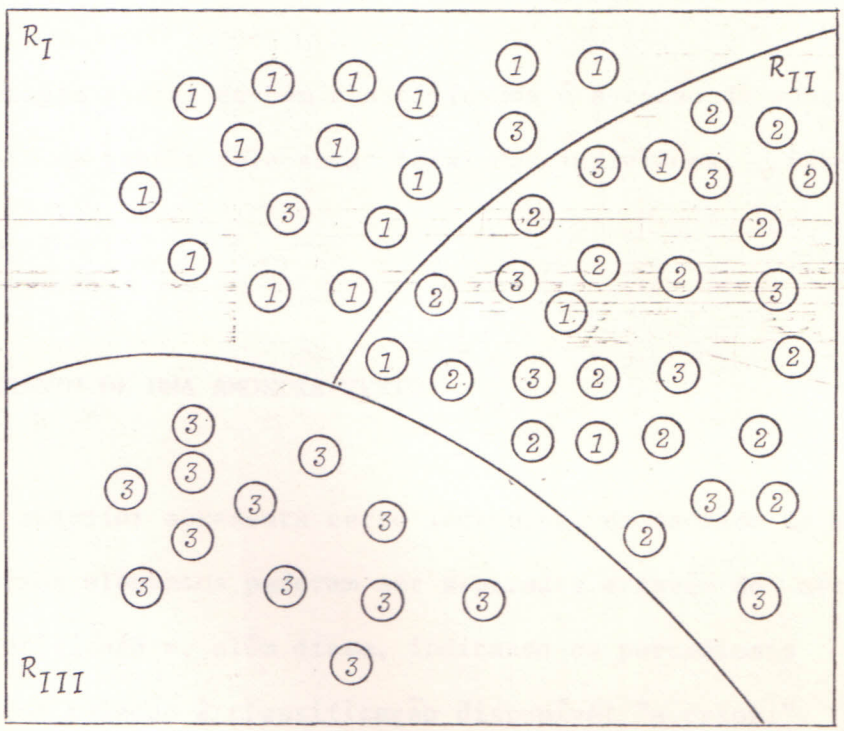
C_{II} e C_{III}, sendo que *18* (dezoito) indivíduos encontram-se na classe C_I, *16* (dezesseis) na classe C_{II} e *22* (vinte e dois) na classe C_{III}. (Na figura que acompanha este exemplo, esses indivíduos são representados pelos algarismos *1*, *2* e *3*, respectivamente).

Suponhamos, além disso, que o método de discriminação permitiu particionar o espaço em três regiões (ou subconjuntos) R_I, R_{II} e R_{III}. (também representados na figura). Ora, se um dado indivíduo é classificado na região R_α de mesmo índice da classe C_β à qual pertence "a priori" (neste caso $\alpha = \beta$), então esse elemento está bem classificado; caso contrário (isto é, $\alpha \neq \beta$) então o indivíduo terá sido, "a posteriori", classificado erroneamente.

CLASSE I : *1*

CLASSE II : *2*

CLASSE III: *3*



A partir do exame dessa figura, não é difícil construir a seguinte tabela:

(4.3.1.3) TABELA DE CLASSIFICAÇÃO DE AMOSTRA INICIAL

GRUPO DE PERTINÊNCIA

GRUPO DE CLASSIFICAÇÃO	I	II	III
I	14	0	2
II	4	16	8
III	0	0	12
	77,8	100,0	54,5

% de bem classificados em cada classe

Na tabela (4.3.1.3) observamos que a classe II é a mais homogênea, porque é encontrada uma porcentagem de bem classificados de $100,0\%$, ao passo que a classe III é a mais heterogênea, com uma porcentagem de apenas $54,5\%$.

A porcentagem global de bem classificados é a razão da soma dos elementos diagonais da tabela pela soma total dos indivíduos, isto é,

$\frac{42}{56} = 75,0\%$.

(4.3.2) COMPORTAMENTO DE UMA AMOSTRA-TESTE

A tabela anterior apresenta certo interesse, no sentido de mostrar como os diversos elementos puderam ser separados através do método de discriminação utilizado e, além disso, indicando os porcentuais dos erros cometidos, com relação à classificação disponível "a priori". Contudo, ela nada nos diz sobre a forma como virão a se comportar novos indivíduos frente ao nosso método de discriminação, ou seja, se os porcentuais de erro se manterão estáveis.

Assim, suponhamos que o número de indivíduos em cada classe C_h seja suficientemente grande, de sorte que possamos separar ao acaso a

classe C_r em duas sub-classes C'_r e C''_r; evidentemente, C'_r e C''_r constituem uma partição dicotômica de C_r $(r = 1, 2, \ldots, k)$.

Nessas circunstâncias, podemos utilizar a amostra $\{C'_1, C'_2, \ldots, C'_k\}$ como aquela de que nos servimos para a determinação das regiões de discriminação $R_1, R_2, \ldots, R_k$; esta, é chamada de Amostra-de-Base (ou inicial). Enquanto a segunda amostra, $\{C''_1, C''_2, \ldots, C''_k\}$ servirá de Amostra-Teste, a partir da qual se constroi nova tabela de classificação.

A tabela de classificação da amostra-teste é aquela de que nos serviremos para a estimação das "probabilidades a posteriori" de pertinência a cada uma das classes C_r $(r = 1, 2, \ldots, k)$. A vantagem é que os indivíduos da "amostra-teste" não estão comprometidos na determinação das regiões de discriminação R_r $(r = 1, 2, \ldots, k)$.

Assim, suponhamos que além da amostra-de-base (ou amostra inicial) constituída de *56* indivíduos, dispomos ainda de uma amostra-teste com *28* indivíduos. Além disso, supomos que os efetivos das classes nesta amostra são proporcionais aos efetivos das classes respectivas na amostra inicial, ou seja, às probabilidades "a priori" de pertinência às classes distintas.

(4.3.2.1) TABELA DE CLASSIFICAÇÃO DA AMOSTRA-TESTE

CLASSE DE PERTINÊNCIA

CLASSE DE CLASSIFICAÇÃO	I	II	III
I	6	1	2
II	2	7	4
III	1	0	5

(a)

PROBABILIDADES A POSTERIORI

	I	II	III
I	6/9 0,666	1/9 0,111...	2/9 0,222...
II	2/13 0,154	7/13 0,539	4/13 0,308
III	1/6 0,166...	0/6 0,000	5/6 0,833...

(b)

Note-se que a tabela (4.3.2.1)(a) incorpora as frequências absolutas enquanto na tabela (4.3.2.1)(b) comparecem as frequências relativas, as quais funcionam como estimativas das probabilidades de erro e acerto, "a posteriori".

Para melhor compreensão sobre a forma de construir a tabela do lado direito a partir da tabela do lado esquerdo, consideremos a primeira linha de ambas. Do total de *9* indivíduos que foram classificados na classe C_I (por terem caídos na região R_I de classificação), apenas *6* foram classificados corretamente, isto é, de fato pertencem à classe C_I, donde a probabilidade "a posteriori" de classificação correta é $2/3 \cong 0{,}666...$. Por outro lado *1* (hum) indivíduo foi classificado erroneamente na classe C_I, quando de fato ele pertence à classe C_{II}; nesse caso, a probabilidade "a posteriori" é estimada como sendo $1/9 \cong 0{,}111...$. Finalmente *2* indivíduos foram classificados erroneamente na classe C_I, quando de fato está na classe C_{III}; donde a probabilidade "a posteriori" estimada é $2/9 \cong 0{,}222...$. Procede-se analogamente para obter os valores nas demais linhas da tabela (b) do lado direito.

Resta nos convencermos de que as frequências relativas que comparecem na tabela são, de fato, estimativas adequadas das "probabilidades a posteriori". Para isso consideremos o teorema seguinte:

4.3.3 TEOREMA

Se n_{ij} são os elementos da tabela (4.3.1.3)(a) (n_{ij} = número de indivíduos da classe j classificados em i), e se os $n_{.j}$ (totais das colunas respectivas) são proporcionais às probabilidades a priori de pertinências às classes, então $n_{ij}/n_{i.}$ designa a probabilidade "a posteriori" de um dado indivíduo de fato pertencer à classe j, sabendo que foi classificado na classe i.

Demonstração

É preciso indicar com clareza as notações de sorte a evidenciar que o problema constitui um caso particular do Teorema de Bayes.

consideremos:

$n_{.j} = \sum_i n_{ij}$ = somatório dos termos da coluna j .

$n_{i.} = \sum_j n_{ij}$ = somatório dos termos da linha i .

Evidentemente:

$$n = \sum_{i,j} n_{ij} = \sum_i n_{i.} = \sum_j n_{.j} .$$

Supomos existirem k classes, donde $i,j = 1,2,\ldots,k$; sendo assim, distinguiremos dois tipos de acontecimentos:

A_j = {pertencer à classe j} ; $j = 1,2,\ldots,k$

B_i = {ser classificado na classe i}; $i = 1,2,\ldots,k$.

Nestas condições, a probabilidade a posteriori de um elemento pertencer à classe j , sabendo que i é a classe onde foi classificado, é dado por:

$$p_{ij} = Prob(A_j/B_i) .$$

Mas o *Teorema de Bayes* nos garante que:

$$Prob(A_j/B_i) = \frac{Prob(B_j/A_i) \cdot Prob\ (A_j)}{\sum_j Prob(B_i/A_j) \cdot Prob(A_j)}$$

ademais, sabemos que:

$$Prob\left(B_i/A_j\right) = n_{ij} / n_{.j}$$

Por outro lado, a hipótese de que os efetivos das classes da amostra-teste são proporcionais as probabilidades a priori de pertinência a estas classes, se formula como:

$$Prob\ (A_j) = n_{.j} / n$$

Donde se conclui que:

$$Prob\ (A_j\ /\ B_i) = n_{ij} / n_{i.}$$

■

Este resultado nos permite afirmar, por exemplo; que um indivíduo classificado na classe C_I tem *66,6* chances sobre 100 de pertencer a esta classe; que um indivíduo classificado na classe C_{II} não tem mais que *53,9* sobre 100 de pertencer a ela, ao passo que um indivíduo classificado na classe C_{III} tem uma possibilidade ainda maior de pertencer de fato a esta classe.

No que se segue, indicam-se os problemas que aparecem no que concerne à utilização do critério da "porcentagem de bem classificados", para realizar a discriminação passo a passo, no caso dos métodos estudados no Capítulo *1* (abordagem de Sebestyen, caso particular e caso geral).

4.3.4 PROCEDIMENTOS PASSO A PASSO PARA OS MÉTODOS DE SEBESTYEN

(a) Caso Particular (Matriz Diagonal)

Para aplicar o método de Sebestyen, no caso particular em que a matriz é diagonal (conforme Capítulo *1*; fórmulas (1.5.5.3) e (1.5.5.4)), tem-se expressões semelhantes (substituindo p por q, onde q é a ordem do passo que está sendo considerado). Assim a expressão (1.5.5.3) se escreve:

$$(4.3.4.1) \qquad \pi(a, \mathbb{C}) = \left(\prod_{j=1}^{q} \sigma_j^2\right)^{1/q} \left[\sum_{j=1}^{q} \left(\frac{a_j - x_j}{\sigma_j}\right)^2 + q\right]$$

Para o cálculo das porcentagens de bem classificados no passo $(q+1)$, pode-se proceder de duas maneiras:

i) calcular diretamente as novas distâncias no $\mathbb{R}^{q+1}$;

ii) conservar na memória do computador todas as informações utilizadas no cálculo das distâncias obtidas no passo anterior para auxiliar no cálculo das novas distâncias.

Os programas de computador desenvolvidos por ROMEDER utilizam a segunda forma em 53 passos para um dado problema num teste utilizado no Centro de Cálculo e de Estatística das Faculdades de Medicina de Paris , aquele pesquizador necessitou de *3* minutos e *13* segundos de tempo de CPU, contra *30* minutos utilizando a primeira forma de proceder.

(b) CASO GERAL

Consideremos, em seguida, os problemas que surgem ligados aos cálculos através das seguintes expressões (1.5.4.3) e (1.5.4.4) do capítulo 1 quando se aplica um procedimento passo a passo. No passo de ordem q , a expressão (1.5.4.3) torna-se:

$$(4.3.4.2) \qquad \pi(\mathbf{a}, \mathbb{C}) = (\det \Sigma_q)^{1/q} \left[q + (\mathbf{a}-\bar{\mathbf{x}})' \, \Sigma_q^{-1} (\mathbf{a}-\bar{\mathbf{x}})\right] ,$$

onde $(\mathbf{a}-\bar{\mathbf{x}})$ designa o vetor projeção no $\mathbb{R}^q$ do vetor de mesmo nome no $\mathbb{R}^p$.

No passo $(q+1)$, e necessário calcular a matriz Σ_{q+1}^{-1} , contudo, ela pode ser calculada a partir da matriz inversa obtida no passo anterior. Da mesma maneira, o determinante $\det \Sigma_{q+1}$ se calcula a partir do determinante $\det \Sigma_q$ no passo anterior. Para isso, utiliza-se o teorema seguinte:

4.3.5 TEOREMA

Seja A uma matriz quadrada, inversível de ordem p, que se completa por um vetor coluna $\mathbf{u}$, um vetor linha $\mathbf{v}$ e um escalar α de tal sorte que tenhamos uma matriz de ordem $p + 1$. Consideremos $\mathbf{Z} = \begin{pmatrix} A & u \\ v & \alpha \end{pmatrix}$. Se $\begin{pmatrix} C & x \\ y & a \end{pmatrix}$ designa a inversa da matriz $\mathbf{Z}$; então:

i) $a = \dfrac{1}{\alpha - vA^{-1}u}$

ii) $x = \dfrac{-A^{-1}u}{\alpha - vA^{-1}u}$

iii) $y = \dfrac{-vA^{-1}}{\alpha - vA^{-1}u}$

iv) $C = A^{-1}\left(I_p + \dfrac{uvA^{-1}}{\alpha - vA^{-1}u}\right)$

v) $\det \begin{pmatrix} A & u \\ v & \alpha \end{pmatrix} = \det A\,(\alpha - vA^{-1}u)$

Demonstração

Como $\begin{pmatrix} C & x \\ y & a \end{pmatrix}$ é inversa de $\begin{pmatrix} A & u \\ v & \alpha \end{pmatrix}$, então:

$$\begin{pmatrix} C & x \\ y & a \end{pmatrix}\begin{pmatrix} A & u \\ v & \alpha \end{pmatrix} = \begin{pmatrix} I_p & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{donde:}$$

$$\begin{cases} CA + xv = I \\ Cu + x\alpha = 0 \\ yA + av = 0 \\ yu + a\alpha = 1 \end{cases}, \quad \text{logo :}$$

$$\begin{cases} yA + av = 0 \\ yu + a\alpha = 1 \end{cases} \Longrightarrow \begin{cases} yAA^{-1} + avA^{-1} = 0 \\ yu + a\alpha = 1 \end{cases} \Longrightarrow$$

$$\begin{cases} y + a\, vA^{-1} = 0 \\ yu + a\,\alpha = 1 \end{cases} \Longrightarrow \begin{cases} -yu - a\, vA^{-1}u = 0 \\ yu + a\,\alpha = 1 \end{cases} \Longrightarrow$$

$$a = \frac{1}{\alpha - a\, vA^{-1}u}\ , \text{ donde } \quad y = \frac{-\,vA^{-1}}{\alpha - vA^{-1}u}\ .$$

Por outro lado ;

$$\begin{cases} CA + xv = I \\ Cu + x\alpha = 0 \end{cases} \Longrightarrow \begin{cases} CAA^{-1} + xvA^{-1} = A^{-1} \\ Cu + x\,\alpha = 0 \end{cases} \Longrightarrow$$

$$\begin{cases} -Cu - xvA^{-1}u = A^{-1}u \\ Cu + x\,\alpha = 0 \end{cases} \Longrightarrow x = \frac{-\,A^{-1}u}{\alpha - vA^{-1}u}\ ,$$

$$\text{donde} \qquad C = A^{-1}\left(I_p + \frac{u\, v\, A^{-1}}{\alpha - vA^{-1}u}\right)\ .$$

Para mostrar a última igualdade, basta considerar:

$$\begin{pmatrix} I_p & 0 \\ -vA^{-1} & 1 \end{pmatrix} \begin{pmatrix} A & u \\ v & \alpha \end{pmatrix} = \begin{pmatrix} A & u \\ 0 & -vA^{-1}u + \alpha \end{pmatrix}\ ;$$

portanto,

$$\det \left[\begin{pmatrix} I_p & 0 \\ -vA^{-1} & 1 \end{pmatrix} \begin{pmatrix} A & u \\ v & \alpha \end{pmatrix} \right] = \det A\ (-vA^{-1}u + \alpha)$$

i.e, $$\det \begin{pmatrix} A & u \\ v & \alpha \end{pmatrix} = \det A\ (-vA^{-1}u + \alpha)\ .$$

De fato, além da inversa da matriz Σ_q obtida no passo anterior, certos cálculos anteriores também devem ser conservados, de forma a permitir que não se seja obrigado a recalcular totalmente as formas quadráticas definidas pela expressão (4.3.4.2), as quais são necessárias com relação a cada indivíduo, relativamente a cada classe para fim da obtenção das porcentagens de bem classificados.

4.4 TRAÇO DA MATRIZ $T^{-1}B$ (Segundo Critério)

Este critério, contrário ao anterior, não necessita da definição de um processo de classificação. Iremos utilizar para justificar os resultados do Capítulo 2 , e em particular a interpretação geométrica do conteúdo do parágrafo (2.9).

Vimos que a métrica definida por T^{-1}, onde T é a matriz de covariância total se introduziria naturalmente. Com isso, procuraremos verificar, em cada passo, qual é o conjunto de variáveis que maximiza a a inércia da nuvem C_r, calculada com a métrica T^{-1}, relativamente a seu centro de gravidade.

Precisamente, no passo q , procura-se qual o melhor subconjunto de q variáveis que maximiza:

$$(4.4.1) \qquad \sum_{r=1}^{k} \left\{ \frac{N_r}{N} (g^r - g)' T_q^{-1} (g^r - g) \right\} \quad ;$$

onde T_q designa a matriz de ordem q , deduzida de T colocando zeros nas colunas e linhas correspondente às variáveis diferentes das q variáveis consideradas. Da mesma maneira, a matriz B_q é deduzida da matriz de covariância inter-classes; então (4.4.1) torna-se:

$$(4.4.1.2)\qquad \mathrm{Tr}\left[T_q^{-1} \sum_{r=1}^{k} \sum_{g_i^r \in C_r} \frac{N_r}{N} (g^r - g)'(g^r - g)\right] = \mathrm{Tr}(T_q^{-1} \cdot B_q)$$

Sendo assim, o critério passa a ser:

$$(4.4.2)\qquad \text{" MAXIMIZAR } \mathrm{Tr}(T_q^{-1} B_q)\text{"}$$

Como exemplo, podemos ter o caso de duas classes C_1 e C_2, e veremos que, o critério proposto é igual ao D^2 de MAHALANOBIS, definido no Capítulo 2 , a menos de um fator.

Com efeito; na expressão (4.4.2) a matriz B se escreve como:

$$(4.4.3)\qquad \frac{N_1}{N}(g^1 - g)(g^1 - g)' + \frac{N_2}{N}(g^2 - g)(g^2 - g)',$$

$$\text{onde}\quad g = \frac{N_1 g^1 + N_2 g^2}{N}\ ; \qquad N = N_1 + N_2\ ,$$

dessa forma, a expressão (4.4.3) transforma-se em:

$$(4.4.3.1)\qquad \frac{N_1 N_2}{N^2}(g^1 - g^2)(g^1 - g^2)'$$

e, se levarmos o valor obtido em (4.4.3.1) para (4.4.2) o critério torna-se

$$(4.4.4)\quad \text{MAXIMIZAR}\quad \mathrm{Tr}(T_q^{-1} B_q) = \frac{N_1 N_2}{N^2}(g^1 - g^2)'\, T^{-1}(g^1 - g^2)$$

De fato:

$$\text{MAX } \mathrm{Tr}(T_q^{-1} B_q) = \text{MAX } \mathrm{Tr}\left[T_q^{-1} \frac{N_1 N_2}{N^2}(g^1 - g^2)(g^1 - g^2)'\right]$$

$$= \text{MAX}\left\{\frac{N_1 N_2}{N^2}\,\mathrm{Tr}\left[\underbrace{T_q^{-1}(g^1 - g^2)}_{q\times 1}\ \underbrace{(g^1 - g^2)'}_{1\times q}\right]\right\}$$

$$= \text{MAX}\left\{\frac{N_1 N_2}{N^2}\, \text{Tr}\underbrace{[(g^1 - g^2)'}_{1\times q}\, T^{-1}\underbrace{(g^1 - g^2)}_{q\times 1}]\right\}$$

$$= \text{MAX}\left\{\frac{N_1 N_2}{N^2}\, (g^1 - g^2)'\, T^{-1}(g^1 - g^2)\right\}$$

$$= \frac{N_1 N_2}{N^2}\, (g^1 - g^2)'\, T^{-1}(g^1 - g^2) \quad ,$$

que é o D^2 de Mahalanobis menos do fator $N_1 N_2 / N$.

A quantidade determinada em (4.4.2) constitue uma generalização do D^2 de Mahalanobis e pode ser considerado como índice de separação entre várias classes no espaço $\mathbb{R}^p$.

Para o método passo a passo, os cálculos serão simples: dispõe-se inicialmente das matrizes **T** e **B** ambas de ordem p e no primeiro passo, calcula-se para cada variável, a quantidade $\text{Tr}(T_1^{-1}B_1)$, a qual se reduz ao quociente dos termos diagonais de **B** e **T** correspondendo à variável considerada; no segundo passo, utiliza-se T_2 e B_2 relativas à variável anterior e a uma nova variável acrescentada, e calcula-se então o $\text{Tr}(T_2^{-1}\, B_2)$ e assim sucessivamente.

Observa-se que não dispomos de teste de parada natural, como no caso do critério de porcentagem de bens classificados. Com efeito; a quantidade $\text{Tr}(T_q^{-1}B_q)$ poderá crescer na passagem do passo q ao passo $q + 1$ sem que a discriminação seja melhorada.

Nota-se que, para o cálculo de T_{q+1}^{-1} e B_{q+1} em função de T_q^{-1} e B_q, aplica-se o teorema (4.3.5).

4.5 CRITÉRIO DO Λ DE WILKS

Este critério é baseado no valor da expressão $\Lambda = \det W / \det T$ (que é o chamado Λ de Wilks). No nosso caso, lembremos que **W** e **T** são

as matrizes de covariância intra-classes e total, respectivamente. Trata-se de um teste com base estatística para detectar a existência de uma possível diferença significativa entre os vetores médios das diversas classes. Se, de fato essa diferença for significativa, é uma indicação da boa separação (ou discriminação) entre as classes; nesse caso, Λ é pequeno. Por outro lado, se a diferença revelar-se não significativa, é uma indicação de má discriminação, quando Λ é grande.

Tratando-se do procedimento passo a passo, considera-se em cada etapa $(q+1)$, o conjunto de $(q+1)$ variáveis que minimizam $\Lambda_{q+1} = \det W_{q+1} / \det T_{q+1}$, onde W_{q+1} e T_{q+1} são as matrizes correspondentes às $q+1$ variáveis consideradas. Observe-se que nesse passo de ordem $(q+1)$ são retidas as q variáveis relacionadas no passo anterior, de ordem q, de sorte que na verdade se trata de selecionar uma nova variável.

Não entraremos em detalhes nas bases estatísticas do teste em causa, conforme foi mencionado anteriormente. Contudo é fácil estabelecer uma relação entre este novo critério e o anterior (do traço da matriz $T^{-1}B$).

Para tal fim, sabemos que, o critério anterior é maximizar $\mathrm{Tr}(T_q^{-1} B_q)$, enquanto o novo critério baseia-se em $\Lambda = \det W / \det T$. Para que se possa estabelecer uma relação entre eles, mostraremos primeiramente que:

$$\Lambda_q = \prod_{i=1}^{q} \beta_i \qquad (4.5.1)$$

onde β_i é valor próprio de $T_q^{-1} W_q$. Com efeito,

$$\Lambda_q = \frac{\det W_q}{\det T_q} \det(T_q^{-1} W_q) = \prod_{i=1}^{q} \beta i$$

Portanto, é natural reter no passo q o conjunto de variáveis

que minimiza $\Lambda_q = \dfrac{\det W_q}{\det T_q}$. Assim, o critério torna-se:

$$\text{MINIMIZAR} \quad \prod_{i=1}^{q} \beta_i \ .$$

Por outro lado, o critério anterior é

$$\text{MAXIMIZAR} \quad Tr(T_q^{-1} B_q) \ ,$$

onde:

$$Tr(T_q^{-1} B_q) = \sum_{i=1}^{q} \lambda_i \ ,$$

sendo os λ_i os valores próprios de $T_q^{-1} B_q^{-1}$.

De fato, tem-se a relação:

$$\lambda_i = 1 - \beta_i \ .$$

Com efeito, se u_i designa o vetor de $T_q^{-1} W_q$ relativamente a β_i, então

$$T_q^{-1} W_q u_i = \beta_i u_i \ ;$$

porém (Cap. 2), $B_q = T_q - W_q$, donde:

$$T_q^{-1} B_q u_i = T_q^{-1}(T_q - W_q)u_i$$

$$= (T_q^{-1} T_q - T_q^{-1} W_q)u_i$$

$$= (1 - \beta_i)u_i = \lambda_i u_i$$

Dessa forma:

$$Tr(T_q^{-1} B_q) = \sum_{i=1}^{q} \lambda_i \ ,$$

torna-se:

$$Tr(T_q^{-1} B_q) = q - \sum_{i=1}^{q} \beta_i$$

e o critério que consiste em maximizar $Tr(T_q^{-1} B_q)$, corresponde a:

$$\text{MINIMIZAR} \quad \sum_{i=1}^{q} \beta_i \quad .$$

Portanto, observa-se que os dois critérios são semelhantes; um minimiza a soma e o outro minimiza o produto dos λ_i .

4.6 CRITÉRIO DA MAXIMIZAÇÃO DAS DIFERENÇAS ENTRE AS MÉDIAS CONDICIONAIS PARA AS DIFERENTES CLASSES

Este critério também possui uma forte fundamentação em termos estatísticos, donde não nos deteremos sobre a maneira de proceder, que foge a nossos objetivos. Na verdade, ele consiste num "teste F" exato, clássico em Estatística Matemática.

Apenas, adiantamos que oferece uma vantagem, que é o de detectar quanto à introdução uma de $(q+1)$-ésima variável no passo $(q+1)$, com relação as q variáveis já selecionadas até o passo anterior de ordem q, se de fato melhora significativamente a discriminação.

TESTE MULTIDIMENSIONAL NÃO-PARAMÉTRICO PARA O PODER DISCRIMINANTE DE UM HIPERPLANO

5.1 INTRODUÇÃO

Neste capítulo estudaremos um teste multidimensional não-paramétrico, utilizável para avaliar o poder discriminante de um hiperplano separando duas classes. Como se trata de um teste não-paramétrico, ele não deve depender de qualquer hipótese a respeito das distribuições de probabilidades envolvidas, ou seja, aquelas que eventualmente descrevam o comportamento das variáveis relacionadas às duas classes. Assim, tem-se a vantagem de nos colocarmos num plano de hipóteses menos rígidas que a dos testes baseados sobre a hipótese de normalidade (que são os testes paramétricos).

O teste a que nos referimos foi introduzido por ROMEDER, com base em trabalhos de COVER (1950) e BENZECRI (1969). Recalculamos as tabelas originariamente apresentadas por ROMEDER, cuja finalidade é a seguinte: determinar em função do número p de variáveis, o tamanho mínimo **N** do número total de indivíduos, para se ter garantia de alcançar uma discriminação significativa, e vice-versa, dado **N**, determinar o p máximo.

Note-se que a questão básica aí envolvida, sendo fornecidos N pontos repartidos em duas classes de efetivos N_1 e N_2 (tais que $N = N_1 + N_2$), no espaço $\mathbb{R}^p$, consiste em encontrar um hiperplano separando totalmente as duas classes e estimar a probabilidade da existência de um tal ente geométrico.

Para esse fim, são necessários alguns resultados prévios, os quais são indicados sem exaustivos detalhes, porém precedido das definições indispensáveis e de alguns exemplos esclarecedores.

5.2 UM TEOREMA DE ANÁLISE COMBINATÓRIA LINEAR

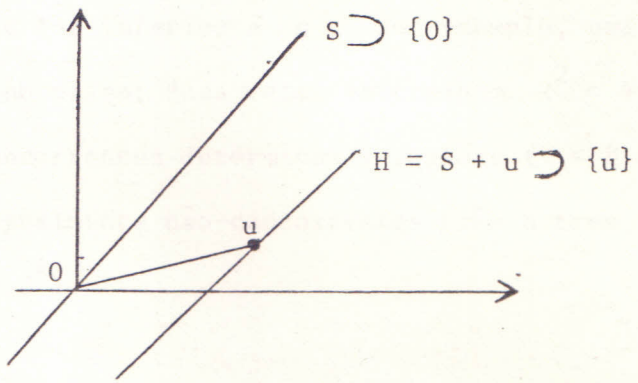
Começamos revendo o conceito básico de hiperplano (ou variedade afim de dimensão $p-1$).

(5.2.1) **DEFINIÇÃO**

Num espaço linear $E = \mathbb{R}^p$, um hiperplano é uma variedade afim H de dimensão p-1 (portanto de dimensão imediatamente inferior à do espaço). Ou seja, existe um subespaço S de dimensão p-1, tal que $H = S + u$, onde **u** é um vetor constante.

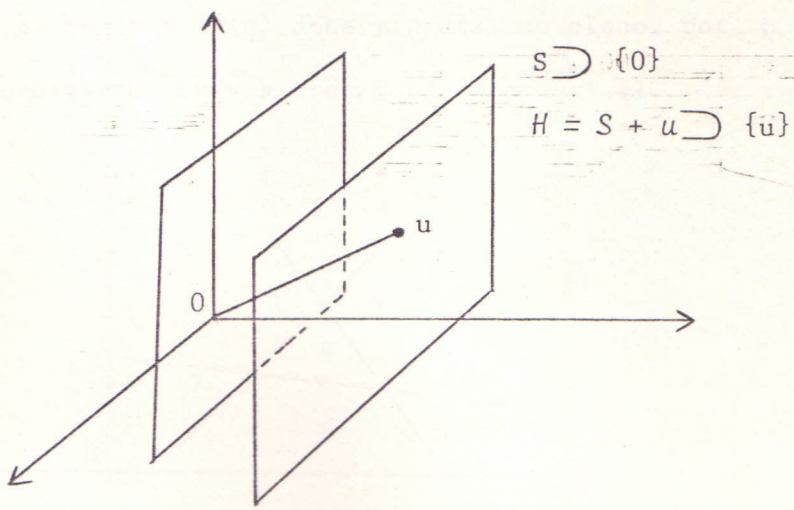
(5.2.1.1) **EXEMPLO:**

$E = \mathbb{R}^2$



(5.2.1.2) **EXEMPLO**

$E = \mathbb{R}^3$



Lembremos que $H = S + u$ interpreta-se como:

$$H = S + u = \{x + u \; ; \; x \in S\} =$$

$$= \{y \; ; \; y - u \in S\} .$$

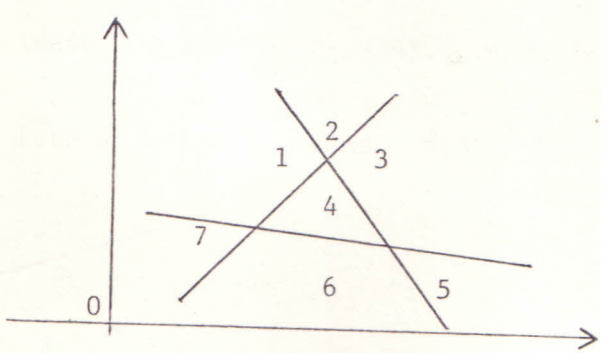
Ademais, é fácil ver que $H = S + u = S + v$, se e só se $v - u$ está em H. Dado um conjunto A não-vazio no espaço, da mesma maneira como faz sentido referir ao subespaço $< A >$ gerado por A, também se pode considerar a variedade afim $<<A>>$ gerada por A, como sendo a "menor variedade afim" contendo tal conjunto. (Vide Dieudomé).

Uma questão relevante diz respeito ao número de regiões delimitadas por q hiperplanos. De fato, esse número é sempre inferior ou igual a 2^q; podendo ser estritamente inferior à última quantidade se a dimensão p do espaço for inferior a q. Por exemplo, uma reta determina $2^1 = 2$ regiões no plano; duas retas determinam $2^2 = 4$ regiões; porém três retas não-concorrentes determinam 7 regiões ($7 < 8 = 2^3$), enquanto quatro retas igualmente não-concorrentes três a três, determinam 11 regiões ($11 < 16 = 2^4$).

(5.2.2) EXEMPLO

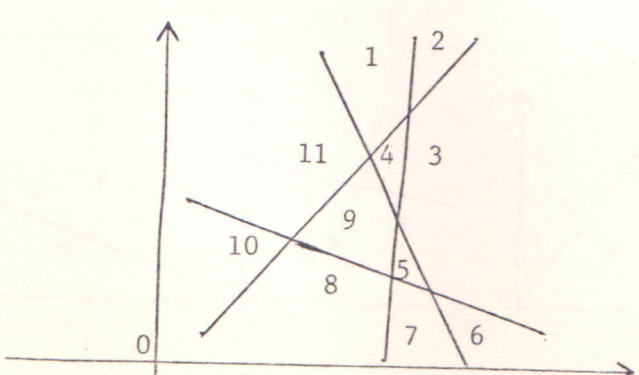
Número de regiões $\eta(q)$ determinadas, no plano, por q retas não-concorrentes três a três ($q = 3$ e $q = 4$).

a)



$\eta(3) = 7$

b)



$\eta(4) = 11$

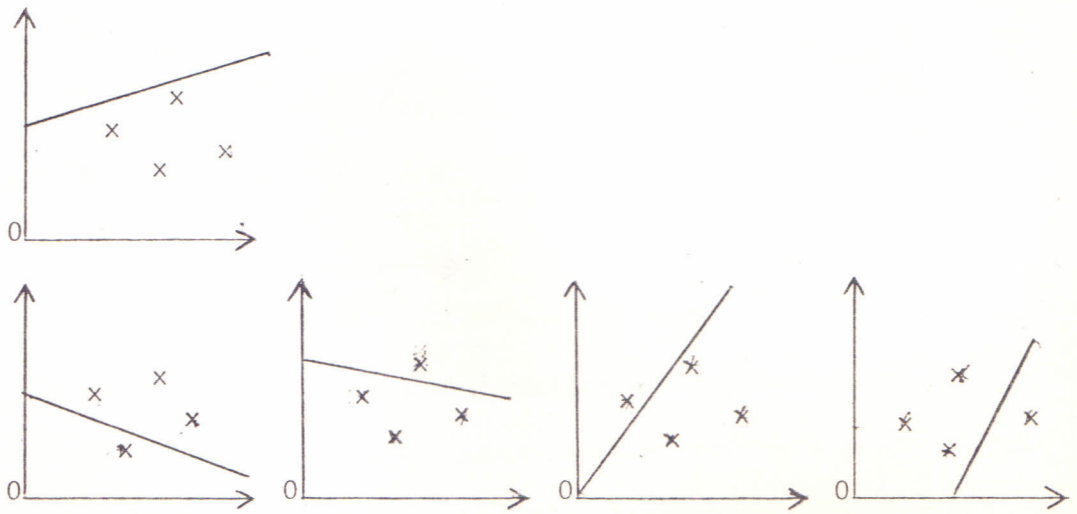
Em questão, é equivalente a uma outra, que se refere às diferentes maneiras de se conseguir separar $q+1$ pontos no espaço em duas classes (uma delas eventualmente vazia), por intermédio de um hiperplano. Assim, dados $4 = 3 + 1$ pontos situados no plano, não alinhado três a três, há 7 maneiras distintas de separá-los em duas classes, por intermédio de uma reta (sendo irrelevante qual a reta utilizada em cada possível corte); analogamente, para $5 = 4 + 1$ pontos em idênticas condições, o número de distintas maneiras de fazê-lo é igual a 11. Aqui, são reproduzidos os resultados (7 e 11) já encontrados anteriormente que se relacionam ao número de regiões determinados no plano por três ou quatro retas, respectivamente não concorrentes três a três.

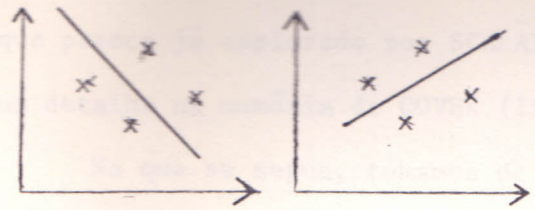
Assim, no exemplo (5.2.2) e no exemplo (5.2.3) a seguir, tem-se situações de todo equivalentes, em termos combinatórios.

(5.2.3) **EXEMPLO**

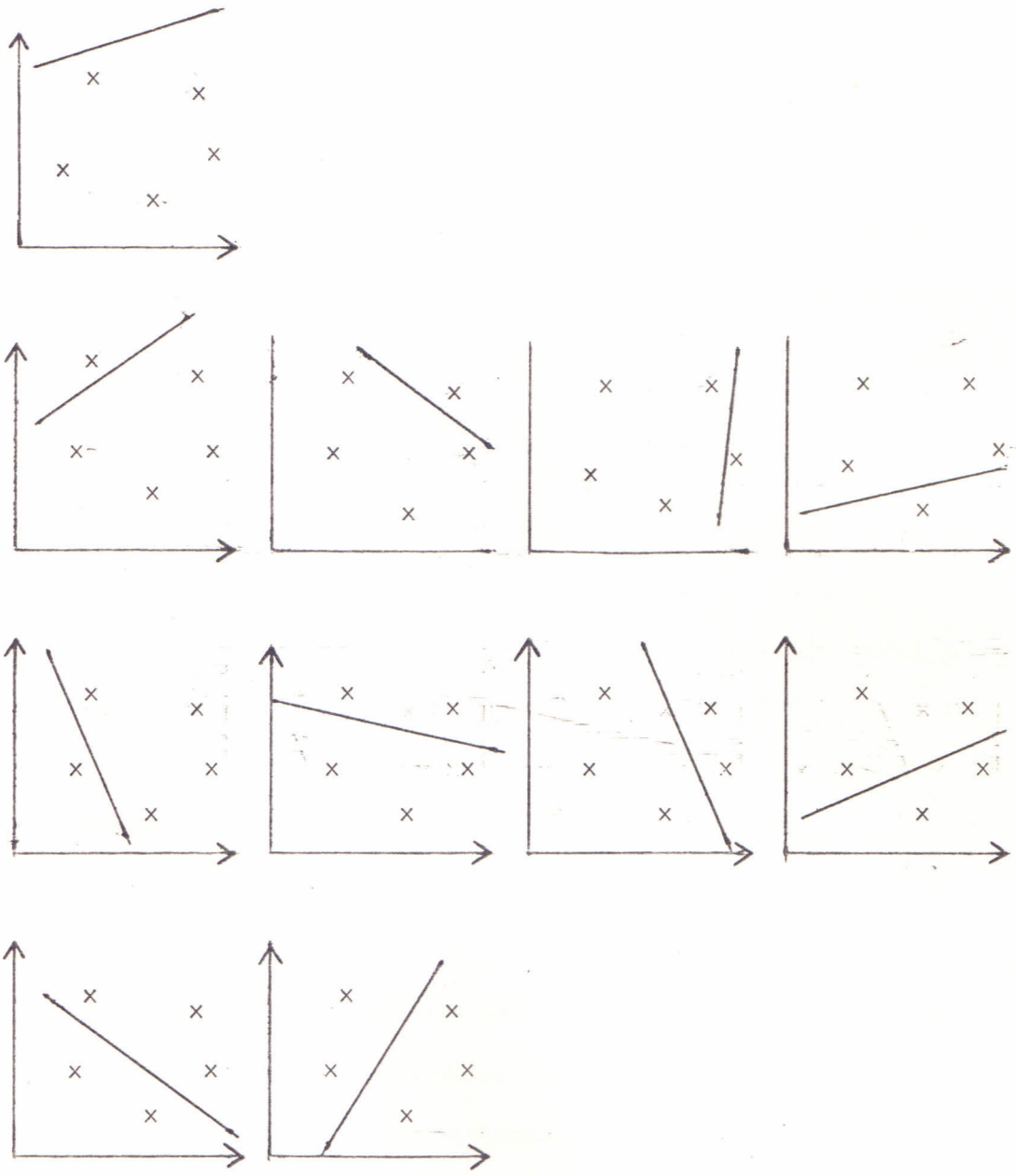
Distintas maneiras de separar em duas classes, por meio de uma reta, um conjunto de $q+1$ pontos no plano, não colineares três a três (casos $q = 3$ e $q = 4$).

a) $q = 3$, isto é $q+1 = 4$, donde $\eta(3) = 7$





b) $q = 4$, isto é $q+1 = 5$ donde $\eta(4) = 11$



De fato, as coincidências observadas nos exemplos(5.2.2) e (5.2.3) não são meramente casuais, pois se enquadram num resultado geral de Análise Combinatória Linear (válido para qualquer dimensão finita $p \geq 1$) ;

ao que parece já explorado por SCHLÄFFLI (1950), porém desenvolvido com algum detalhe na memória de COVER (1965) e no trabalho de BENZECRI(1969).

No que se segue, tomamos de empréstimo a BENZECRI (op. cit.), os aspectos que a esse respeito nos interessam de forma direta. Começamos pela introdução do conceito de "situação geral afim" para o caso de um conjunto finito de pontos no $\mathbb{R}^p$.

(5.2.4) DEFINIÇÃO

Seja A um conjunto de N pontos, no $\mathbb{R}^p$. Diz-se que A está em situação geral afim, nesse espaço, se todo subconjunto B de A, com M pontos ($M \leq N$), gerar uma variedade afim de dimensão igual a $inf(p\ ,\ M-1)$.

(5.2.4.1) EXEMPLO:

No $\mathbb{R}^3$, todo subconjunto B, com três pontos, de um conjunto A em situação geral afim, gera uma variedade afim de dimensão $inf(3,3-1)=2$; isto é, um plano, jamais uma reta.

(5.2.5) DEFINIÇÃO

O subconjunto $B \subset A$ é uma parte afinamente separável de A no $\mathbb{R}^p$, se existir no $\mathbb{R}^p$ um hiperplano separando totalmente B de $A - B$.

No seguinte resultado (apresentado sem demonstrar) fica determinado o número de partes afinamente separáveis de um conjunto em situação geral afim no $\mathbb{R}^p$.

(5.2.6) TEOREMA

Seja A um conjunto de N pontos em situação geral afim no

$\mathbb{R}^p$. O número das partes B de A, afinamente separáveis é dado por:

$$\phi(N , p + 1) ,$$

onde:

$$\phi(N , p) = 2 \sum_{0 \leq k < p} \binom{N-1}{k} \tag{5.2.6.1}$$

Note-se que, sendo $p = 2$ (pontos situados no plano) e $q = 3$ (donde $N = q + 1 = 4$), tem-se:

$$2\eta(3) = \phi(4,3) = 2 \sum_{x=o}^{2} \binom{3}{x} = 14$$

e, portanto $\eta(3) = 7$; analogamente se $q = 4$ (donde $N = 5$), segue-se:

$$2\eta(4) = \phi(5,3) = 2 \sum_{x=o}^{2} \binom{4}{x} = 22 \quad ,$$

isto é, $\eta(4) = 11$. Assim, são reproduzidos os resultados referentes ao exemplo (5.2.3).

A expressão (5.2.6.1) oferece certas dificuldades do ponto de vista de sua utilização para fins computacionais, uma vez que o cálculo dos números combinatórios $\binom{y}{x} = y! / [x!(y-x)!]$ pode rapidamente envolver valores (fatoriais) que exaurem de muito a capacidade de armazenamento de inteiros, mesmo num computador de grande porte. Assim, o lema seguinte nos propicia uma fórmula recorrente de cálculo, permitindo obter os $\phi(N + 1,p)$ em função dos $\phi(N,p)$ e $\phi(N,p-1)$.

(5.2.7) LEMA:

$$\phi(N+1,p) = \phi(N,p) + \phi(N,p-1)$$

Não demonstraremos o lema acima, pois tal nos desviaria consideravelmente dos nossos objetivos; deve-se mencionar, não obstante, que as linhas gerais para sua prova encontram-se bem delineadas em BENZECRI, op. cit. Contudo, mostraremos a equivalência entre as expressões (5.2.6) e (5.2.7).

Que (5.2.6) implica (5.2.7), não oferece dificuldades, pois se trata de aplicação corriqueira de conhecida identidade.

Quanto à implicação de (5.2.7) em (5.2.6), oferece mais resistência para sua prova, o que exige a formulação de dois lemas preliminares (para os quais, apresentam-se esboços de suas demonstrações).

(5.2.8) LEMA:

Seja φ uma função de duas variáveis inteiras, N e p, definida para $N \geq N_o > 0$ e p inteiro qualquer. Além disso, suponhamos que φ satisfaz à equação do lema (5.2.7), isto é

$$\varphi(N+1,p) = \varphi(N,p) + \varphi(N,p-1) .$$

Então, quaisquer que sejam $N \geq N_o$ e $p \in \mathbb{Z}$, φ coincide com a função ψ, dada por :

$$\psi(N,p) = \sum_{p_o \in \mathbb{Z}} \varphi(N_o,p_o) \binom{N-N_o}{p-p_o} \qquad (5.2.8.1)$$

Demonstração

Inicialmente, não será difícil concluir que ψ satisfaz à equação do lema (5.2.7), isto é,

$$\psi(N+1,p) = \psi(N,p) + \psi(N,p-1) .$$

Para esse fim, deve-se observar que (5.2.8.1) é somável pois cada número

combinatório $\binom{N-N_o}{p-p_o}$ é suposto não-nulo somente para $0 \leq p-p_o \leq N-N_o$.

De fato, ψ e φ coincidem para $N = N_o$ e $p \in \mathbb{Z}$. Então por indução finita, segue-se a coincidência para $N = N_o+1$, $N = N_o+2$, ..., e $p \in \mathbb{Z}$.

■

(5.2.9) **LEMA**

A aplicação ϕ (do teorema (5.2.6) e lema (5.2.7)), satisfaz à equação:

$$(5.2.9.1) \qquad \phi(N,p) = \sum_{p_o \in \mathbb{Z}} \phi(1,p_o) \binom{N-1}{p-p_o} .$$

Demonstração

Trata-se da decorrência imediata de (5.2.7) e (5.2.8), com ϕ em lugar de φ na equação (5.2.8.1) e tomando $N_o = 1$.

Para a demonstração de (5.2.7) ■

Para a demonstração de (5.2.7) $\Longrightarrow$ (5.2.6), consideremos a expressão, obtida de (5.2.9.1):

$$\phi(N,p) = \sum_{k \in \mathbb{Z}} \phi(1,p_o) \binom{N-1}{k} .$$

De uma parte, devemos ter $0 \leq k \leq N-1$; ademais, não pode ocorrer $k = p-p_o > p-1$, pois nesse caso teríamos $p_o < 1$, situação em que a expressão $\phi(1,p_o)$ fica destituída de sentido (portanto, sendo considerada nula).

Então, se $p \leq N$, obtem-se (5.2.6.1); se $p > N$, a soma dos índices se estende até N-1 e o valor da expressão fica igual a 2^N.

A tabela abaixo nos fornece os valores de $\phi(N,p)$ calculados através da fórmula de recorrência do lema (5.2.7) pelo computador DEC-10 da U.F.C.

(5.2.10) TABELA DOS VALORES DE $\phi(N,p)$

N	p=1	2	3	4	5	6	7	8	9	10
	2	22								
10	2	20	92	270	512	764	932	1004	1022	1024
	2	18	74	186	326	438	494	510	512	512
	2	16	58	128	198	240	254	256	256	256
	2	14	44	84	114	126	128	128	128	128
	2	12	32	52	62	64	64	64	64	64
5	2	10	22	30	32	32	32	32	32	32
	2	8	14	16	16	16	16	16	16	16
	2	6	8	8	8	8	8	8	8	8
	2	4	4	4	4	4	4	4	4	4
1	2	2	2	2	2	2	2	2	2	2

P

5.3 O TESTE DE SEPARABILIDADE

Nesta secção, passamos ao estudo propriamente dito do teste multidimensional não-paramétrico para o poder discriminante de um hiperplano, ou teste de separabilidade, nos termos anunciados na Introdução.

Agora, já dispomos de elementos para testar (não-paramétricamente) a hipótese de que duas classes se distribuem de forma idênticas, ou não, no $\mathcal{R}^p$; para esse fim, deve-se definir o que significa uma "dicotomia aleatória".

(5.3.1) DEFINIÇÃO

Considere-se a extração de N pontos no $\mathbb{R}^p$ segundo uma certa lei de probabilidade (sob a condição de que uns pontos se encontrem em situação geral afim, com probabilidade 1).

Por dicotomia aleatória entenden-se a distribuição de cada um desses pontos a uma ou outra dentre duas classes consideradas de maneira independente e equiprovável.

5.4 TEOREMA

A probabilidade de uma dicotomia aleatória de N pontos no $\mathbb{R}^p$ ser afinamente separável é dada por:

$$\text{(5.4.1)} \qquad \text{Prob}(N,p) = \left(\frac{1}{2}\right)^N \phi(N,p+1)$$

Demonstração

Sabemos que o número de partes B afinamente separáveis de A no $\mathbb{R}^p$, é igual a $\phi(N,p+1)$, representando o número de casos favoráveis. Por outro lado, sabemos que 2^N é o número de maneiras distintas segundo as quais podemos particionar um conjunto de N elementos em duas partes, representando o número de casos possíveis. Dessa forma, sendo a probabilidade de separabilidade, dada por:

$$\text{Prob}(N,p) = \frac{\text{número de casos favoráveis}}{\text{número de casos possíveis}} \quad ,$$

obtem-se:

$$\text{Prob}(N,p) = \frac{\phi(N,p+1)}{2^N} = \left(\frac{1}{2}\right)^N \phi(N,p+1) \quad .$$

5.5 PRINCÍPIO DO TESTE E APLICAÇÕES

O princípio do teste consiste em, obtidas duas amostras de efetivos N_1 e N_2 (com $N_1 + N_2 = N$), rejeitar a hipótese de que elas estão sujeitas a uma mesma lei, desde que Prob(N,p) é suficientemente pequena (por exemplo inferior a 5% ou 1%); isto é, no que sugere a se encontrar um hiperplano que separa totalmente as duas classes.

De fato, desejamos testar a hipótese de que duas classes estão sujeitas à mesma lei ("hipótese nula" H_o), contra a alternativa que se encontram sujeitas a leis distintas (hipótese alternativa H).

Para a realização do teste necessitamos dispor de uma tabela, fornecendo para cada valor de p (dimensão do espaço) e para riscos respectivos α ($\alpha = 0{,}001$; $\alpha = 0{,}005$; $\alpha = 0{,}01$; $\alpha = 0{,}05$), os valores mínimos $N_\alpha(p)$ a partir dos quais se encontrarmos uma separabilidade total das duas amostras, rejeitaremos então a hipótese nula.

O valor $N_\alpha(p)$ por exemplo, é determinado por:

$$(5.5.1) \qquad N_\alpha(p) = \inf \left\{ N / \left(\frac{1}{2}\right)^N \phi(N, p+1) \leq \alpha \right\}$$

A tabela a seguir é construída a partir dos valores da tabela (5.2.10) o qual nos dá os valores de $\phi(N,p)$, e da fórma (5.4.1).

(5.5.2) TABELA (TESTE DE SEPARABILIDADE)

Valores mínimos de N , em função de p , tais que a separação seja significativa :

	N_α	1	2	3	4	5	6	7	8	9	10
risco	0,05	9	12	24	17	19	21	24	27	29	31
risco	0,01	12	15	18	20	23	26	28	31	34	37
risco	0,005	13	16	19	21	25	27	30	33	35	38
risco	0,001	15	19	22	25	28	31	34	36	39	42
											P

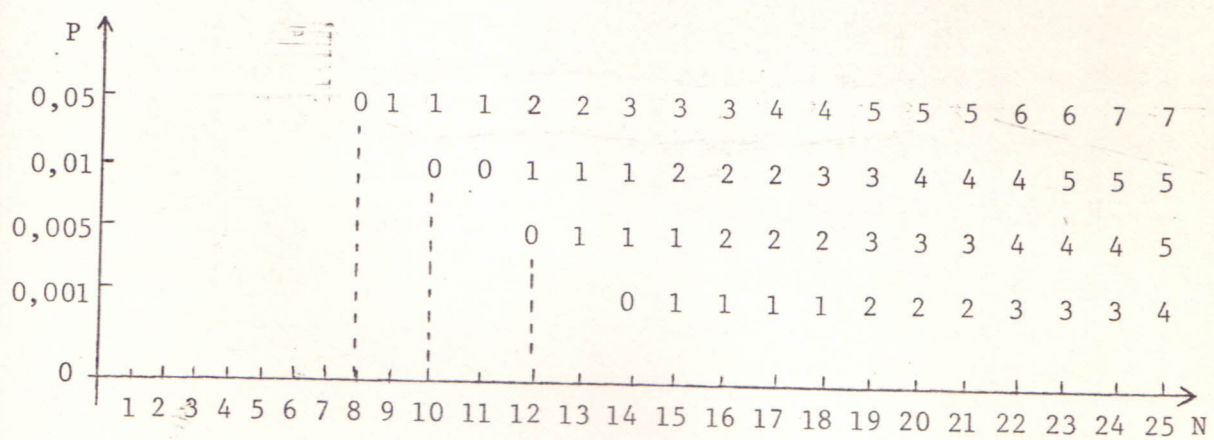
(5.6) EXEMPLOS DE APLICAÇÃO

I) Sejam sete indivíduos de uma classe, e dez indivíduos de outra classe. Suponhamos procurar discriminar entre estas duas classes com a ajuda de duas variáveis, e que saiamos bem sucedido em determinar um hiperplano, separando perfeitamente as duas classes. Uma vez que $17 > 16$, considera-se separação significativa; exceto no risco $\alpha = 0{,}001$. Vide tabela (5.5.2).

II) Sejam dez indivíduos de uma classe e oito de uma outra, consegue-se separar estas duas classes no $\mathbb{R}^5$; pode-se dizer que esta separação não é significativa no risco $\alpha = 0{,}05$ pois $18 < 19$.

Pode-se da mesma maneira indicar, em função de N, o valor máximo de p, tal que chega-se a separar as duas classes significativamente . Para tal segue-se:

(5.6.1) TABELA (TESTE DE SEPARABILIDADE)



Observa-se que o valor 0 (zero), corresponde a N = 11 (respectivamente a N = 8), significa que é ilusório esperar discriminar, mesmo com uma única variável no risco $\alpha = 0{,}01$ (respectivamente a $\alpha = 0{,}05$) duas classes cuja soma dos efetivos seja inferior ou igual a 11, (respectivamente, igual a 8).

APÊNDICE I

A P Ê N D I C E I

OPERADORES E (VALOR ESPERADO) VAR(VARIÂNCIA) E COV(COVARIÂNCIA)

Na introdução deste trabalho tivemos a oportunidade de considerar, para cada classe C de efetivo N, indivíduos x_i (i = 1,2,...,N) pensados como p-uplas $(x_{i1},x_{i2},\ldots,x_{ip})$, onde cada componente x_{ij} (j = 1, 2,...,p) se interpreta como resultado de uma j-ésima medida efetuada sobre x_i; usando outra terminologia, trata-se do valor assumido pelo indivíduo relativamente a certa "j-ésima variável".

Ora, sendo cada indivíduo concebido como um p-vetor $x_i=(x_{ij})_{j=1}^{p}$; então no mesmo contexto, cada "j-ésima variável" (j = 1,2,...,p) concebe-se como um N-vetor $x^{(j)} = (x_{ij})_{j=1}^{N}$.

Note-se que a variável $x^{(j)}$, em vez de um vetor no R^N, pode também pensar-se como uma aplicação $x^{(j)}: \{1,2,\ldots,N\} \to \mathbb{R}$. Sob tal ponto de vista, tem sentido considerar a aplicação constante λ; bem como, expressões λu, $\lambda+u$, $u+v$, uv, u^2,...., no sentido do resultado de operações definidas "ponto a ponto", relativamente às aplicações u,v,..... sem dúvida, no caso de u,v,....., serem pensadas como vetores no $\mathbb{R}^N$, a interpretação correta corresponde ao fato de que as operações respectivas são agora definidas "coordenada a coordenada". Portanto, faz sentido considerar o "vetor λ", como a N-upla (...., λ,....); $u+v$ = (....., u_i+v_i,....); uv = (....,u_iv_i,.....); u^2 = (....., u_i^2,.....); etc.

(I.1) DEFINIÇÃO

O operador valor esperado **E**, é aplicação:

$$E: \mathbb{R}^N \longrightarrow \mathbb{R} \ ,$$

tal que:

$$E(u) = \frac{1}{N} \sum_{i=1}^{N} u_i \ ,$$

onde **u** é um vetor de componentes u_i.

Sem quaisquer dificuldades, podem ser evidenciadas as seguintes propriedades:

(I.1.1.1) $E(\lambda) = \lambda$

(I.1.1.2) $E(\lambda u) = \lambda E(u)$;

(I.1.1.3) $E(u+v) = E(u) + E(v)$;

quanto a esta última propriedade, pode generalizar-se, por indução finita.

(I.1.1.4) $E(x^{(1)}+x^{(2)} + \cdots + x^{(q)}) = E(x^{(1)}) + \cdots + E(x^{(q)})$.

(I.2) DEFINIÇÃO

O operador variância **VAR**, é aplicação:

$$\mathbf{VAR} : \mathbb{R}^N \longrightarrow \mathbb{R} \ ,$$

tal que:

(I.2.1) $$\mathbf{VAR}(u) = E(u^2) - [E(u)]^2$$

Para esse operador, valem as propriedades, cuja verificação não envolve quaisquer dificuldades;

(I.2.1.1) $\quad VAR(\lambda) = 0$;

(I.2.1.2) $\quad VAR(u+\lambda) = VAR(u)$;

(I.2.1.3) $\quad VAR(\lambda u) = \lambda^2 \, VAR(u)$;

(I.2.1.4) $\quad VAR(u) = E\,[u - E(u)]^2$.

(I.3) DEFINIÇÃO

O operador covariância **COV**, é a aplicação

$$\mathbf{COV} : \mathbb{R}^N \times \mathbb{R}^N \longrightarrow \mathbb{R} \ ,$$

tal que:

(I.3.1) $\quad \mathbf{COV}(u,v) = E[(u - E(u))(u - E(v))]$

A seguinte propriedade interrelaciona os operadores COV e VAR .

(I.3.2) $\quad VAR(u) = COV(u,u)$, de fato,

$$COV(u,u) = E[u - E(u)]^2 = VAR(u)$$

Usando os operadores VAR e COV, e notando $x^{(j)} = (x_{ij})$, observe-se que os σ_{jk} em (1.2.2.1), conforme Capítulo 1, são precisamente as $COV(x^{(j)},x^{(k)})$; se j = k, tem-se $\sigma_{jj} = (\sigma_j)^2 = VAR(x^{(j)})$. Ademais, com relação a (1.2.2), no referido capítulo, tem-se ;

(I.4) $$\sum = \begin{pmatrix} & \vdots & \\ \cdots\cdots & COV(x^{(j)},x^{(k)}) & \cdots\cdots \\ & \vdots & \end{pmatrix}$$

Esses resultados justificam a utilização das expressões "covariância" e "variância" para os σ_{jk} e $(\sigma_j)^2$; bem como, a designação de " matriz de variâncias-covariâncias" para $\sum$.

Por outro lado, considerada uma matriz $X = (x^{(1)},\ldots,x^{(r)})$ onde cada $x^{(j)}$; para $j = 1,2,\ldots,r$, é uma variável (ou N-vetor), definiremos:

(I.5) $$E(X) = (E(x^{(1)}),\ldots., E(x^{(r)})) .$$

Analogamente, dada a matriz $M = \begin{pmatrix} \vdots \\ \cdots\cdots\ x^{(jk)} \cdots\cdots \\ \vdots \end{pmatrix}$, onde cada $x^{(jk)}$, para $j = 1,\ldots,r$ e $k = 1,2,\ldots,s$, é uma variável (ou N-vetor), definiremos E(M) como a matriz cujas componentes são $E(x^{(jk)})$.

No contexto acima, faz sentido o seguinte resultado, onde se nota $E(X) = \overline{X}$; e bem como $X - \overline{X}$ tem componentes $x^{(j)} - E(x^{(j)})$.

(I.6) PROPOSIÇÃO :

Se $X = (x^{(1)},x^{(2)},\ldots,x^{(r)})$, então:

(I.6.1) $\sum = E[(X - \overline{X})(X - \overline{X})']$, onde $X - \overline{X}$ escreve-se como um vetor coluna e $(X - \overline{X})'$ é o vetor linha correspondente.

Demonstração

Uma vez que $X - \overline{X}$ possui componentes $x^{(j)} - E(x^{(j)})$, segue-se que $(X - \overline{X})(X - \overline{X})'$, possui, como termo geral

$$(x^{(j)} - E(x^{(j)}))(x^{(k)} - E(x^{(k)}) ,$$

donde $E[(X - \overline{X})(X - \overline{X})'$ possui, por termo geral:

$$E[(x^{(j)} - E(x^{(j)}))(x^{(k)} - E(x^{(k)}))] = Cov(x^{(j)},x^{(k)}))$$

Isso nos permite concluir, de fato pela validade de (I.6.1).

O outro resultado que se segue (Proposição I.8) vai permitir interrelacionar as matrizes de variâncias-covariâncias associadas as variáveis $x^{(j)}$, $j = 1,2,\ldots,p$, e as suas transformadas $y^{(j)} = A\,x^{(j)}$, onde A é uma matriz $N \times N$.

(I.7) LEMA

Se $M = (x^{(jk)})_{j,k=1}^{p}$, $A = (a_{ij})_{i,j=1}^{p}$ e $B = (b_{k\ell})_{k,\ell=1}^{p}$,

então:

(I.7.1) $E(AM) = A\,E(M)$;

(I.7.2) $E(MB) = E(M)B$;

(I.7.3) $E(AMB) = A\,E(M)\,B$.

Demonstração

Tem-se, para (I.7.1)

$$E(AM) = E[(a_{ij})(x^{(jk)})] =$$

$$= E[(\sum_{j} a_{ij}\, x^{(jk)})] =$$

$$= (a_{ij})\, E[(x^{(jk)})] = A\,E(M)$$

Para (I.7.2) é análogo, concluindo-se em seguida (I.7.3).

(I.8) PROPOSIÇÃO

Sejam Σ_y e Σ_x as matrizes de variâncias-covariâncias associadas às $x^{(j)}$ e $y^{(j)} = A\,x^{(j)}$, $j = 1,2,\ldots,p$, respectivamente,

onde A é uma matriz (constante) N x N. Então:

(I.8.1) $$\Sigma_Y = A \Sigma_X A' \ .$$

Demonstração:

Tem-se $$\Sigma_Y = E[\ (Y - \bar{Y})(Y - \bar{Y})'\ ] =$$

$$= E[\ (AX - A\bar{X})(AX - A\bar{X})'\ ] =$$

$$= E[\ A(X - \bar{X})(X - \bar{X})'A'\ ] =$$

$$= A\ E[\ (X - \bar{X})(X - \bar{X})'\ ]\ A' =$$

$$= A \Sigma_X A'$$

■

A P Ê N D I C E II

A P Ê N D I C E II

APLICAÇÕES DA ANÁLISE DISCRIMINANTE

Neste apêndice mostraremos a utilização de métodos de Análise Discriminante em duas situações distintas. A primeira delas refere-se a uma aplicação em gastroenterologia, envolvendo a discriminação de entidades mórbidas do antro do estômago. A outra, trata de um problema de discriminação em climatologia.

(II.1) UMA APLICAÇÃO EM GASTROENTEROLOGIA

Para esta aplicação lançamos mão de dados cedidos por de Amorim, W.P.D. (1984), oriundos de sua tese de Mestrado em Medicina, tendo em vista o estudo de câncer gástrico e de doenças gástricas benígnas, a partir das determinações de frações eletroforéticas de isoenzimas da dehidrogenase láctica (DHL).

Basicamente, dispomos de 65 (sessenta e cinco) indivíduos com diagnóstico de lesão do antro, classificados a priori nos seguintes grupos:

Grupo I = gastrite crônica quiescente

Grupo II = gastrite crônica ativa

Grupo III = úlcera péptica (benígna)

Grupo IV = câncer ;

onde os números de pacientes em cada grupo foram os seguintes: 13, 21, 17 e 14, respectivamente.

Para cada paciente dispomos de dados referentes à DHLT (dehidrogenase láctica total), ISODHL1, ISODHL2, ISODHL3, ISODHL4 e ISODHL5 (isoenzimas da DHL); bem como, os valores dos "monômeros H e M" contidos nas várias isoenzimas.

Na tabela (II.1.1) são exibidos os valores observados (NORD é o número de ordem de cada indivíduo, enquanto GPO indica o grupo a que pertence: I, II, III ou IV).

(II.1.1) TABELA

NORD	DHLT	ISODHL1	ISODHL2	ISODHL3	ISODHL4	ISODHL5	H	M	GPO
1	0.87	10.10	12.60	23.10	26.10	28.10	62.30	37.70	1
2	0.66	3.00	8.30	30.70	33.10	25.40	67.40	32.60	1
3	0.23	2.80	17.10	26.50	14.90	38.70	67.40	32.60	1
4	0.40	1.20	3.60	21.10	31.30	42.80	77.60	22.40	1
5	1.57	10.60	11.90	24.30	26.60	26.60	61.60	38.40	1
6	0.50	2.90	10.40	28.50	32.10	26.10	67.00	33.00	1
7	1.04	5.20	7.60	23.00	30.60	34.60	70.90	29.10	1
8	0.61	2.40	8.40	25.70	34.10	29.40	69.90	30.10	1
9	0.94	6.50	11.40	20.30	23.60	38.20	68.80	31.20	1
10	0.63	5.50	13.60	31.60	31.60	17.40	60.40	39.60	1
11	0.36	2.60	9.10	26.40	33.20	28.70	69.10	30.90	1
12	0.63	3.20	6.90	23.40	33.10	29.40	70.60	29.40	1
13	0.32	3.20	11.80	24.70	25.80	34.50	69.00	31.00	1
14	1.07	2.90	5.80	28.60	36.20	26.50	69.30	30.70	2
15	0.64	7.30	14.40	26.00	27.90	24.20	61.70	38.30	2
16	1.02	6.60	11.30	31.60	30.20	20.30	61.50	38.50	2
17	0.86	5.50	16.70	31.90	29.10	16.80	58.80	41.20	2
18	0.46	5.90	9.00	31.40	27.90	25.80	64.60	35.40	2
19	0.79	3.80	10.80	30.80	32.30	22.30	64.60	35.40	2
20	0.37	6.80	10.20	27.10	28.20	27.70	64.80	35.20	2
21	0.57	3.50	13.30	26.90	31.80	24.50	65.00	35.00	2
22	0.64	4.80	11.80	27.90	33.10	22.40	64.10	35.90	2
23	1.15	3.70	14.60	26.80	35.10	19.60	62.80	37.20	2
24	0.53	3.60	6.70	27.80	32.10	29.80	69.50	30.50	2
25	0.80	4.50	9.30	23.50	34.80	28.10	68.10	31.90	2
26	0.80	10.40	19.10	23.10	24.20	23.00	57.60	42.40	2
27	0.65	4.30	9.80	27.50	35.60	22.80	65.60	34.40	2
28	0.77	6.20	17.90	30.90	29.40	15.60	57.50	42.50	2
29	0.57	4.90	11.50	28.60	30.00	23.00	62.70	37.30	2
30	0.96	7.90	19.90	30.10	27.70	14.50	55.30	44.70	2
31	0.68	2.90	8.70	29.20	34.70	24.60	67.40	32.60	2
32	0.47	5.00	7.00	28.00	29.40	35.60	73.30	26.70	2
33	0.46	4.80	10.80	28.00	33.30	23.10	64.80	35.20	2
34	0.48	1.10	4.50	27.10	33.90	29.40	70.40	29.60	2
35	0.80	1.40	10.40	34.00	33.00	21.20	65.50	34.50	3
36	0.78	1.40	12.70	37.50	30.50	17.90	62.70	37.30	3
37	0.67	3.50	9.40	30.60	34.10	22.00	65.30	34.70	3
38	1.37	7.30	12.70	32.00	32.30	15.70	59.10	40.90	3
39	0.69	10.20	15.10	24.90	25.60	24.20	59.60	40.40	3
40	0.56	12.10	9.90	26.80	27.60	23.60	60.20	39.80	3
41	0.60	7.10	16.70	23.10	40.00	13.10	58.80	41.20	3
42	0.97	9.10	14.60	28.70	23.80	24.00	59.70	40.30	3
43	0.91	7.90	14.10	26.90	30.60	20.50	60.40	39.60	3
44	0.66	7.20	10.90	37.40	29.60	19.90	63.50	36.50	3
45	0.63	5.10	7.40	28.60	37.10	21.80	65.70	34.30	3
46	0.40	5.00	5.70	46.40	23.60	19.30	61.60	38.40	3
47	0.49	2.90	15.60	38.70	33.00	9.80	57.70	42.30	3
48	0.75	4.80	11.70	29.70	31.60	23.50	64.20	35.80	3
49	0.61	5.70	12.20	29.10	33.90	19.60	62.60	37.40	3
50	0.78	6.60	17.20	26.60	26.90	12.70	55.50	44.50	3
51	1.07	5.40	12.10	22.40	28.20	31.70	67.10	32.90	3
52	1.06	2.60	13.20	29.40	32.00	22.80	64.80	35.20	4
53	0.68	1.20	4.80	15.60	21.70	56.70	82.20	18.00	4
54	0.54	5.60	20.50	21.70	16.80	35.40	63.90	36.10	4
55	0.62	3.30	10.10	27.40	31.00	28.20	67.80	32.20	4
56	0.62	4.90	13.20	24.80	19.50	37.60	67.90	32.10	4
57	0.83	7.70	13.50	25.10	28.20	25.50	62.50	37.50	4
58	0.78	4.70	13.10	22.40	27.60	32.20	67.40	32.60	4
59	0.80	10.90	16.90	23.30	28.50	20.40	57.70	42.30	4
60	0.79	5.00	11.20	20.90	26.30	36.60	69.50	30.50	4
61	0.74	7.30	20.60	32.60	23.10	16.40	55.10	44.90	4
62	0.50	4.30	14.60	23.30	21.30	34.50	65.80	34.20	4
63	0.93	7.60	15.30	21.80	24.40	30.90	63.90	36.10	4
64	0.95	9.80	11.90	24.60	26.20	27.50	62.40	37.60	4
65	1.21	3.90	13.70	26.30	28.60	27.50	65.40	34.60	4

Utilizou-se o método de Análise Discriminante Passo a Passo, com o emprego dos programas MAHAL 2 e MAHAL 3 descritos na obra de ROMEDER ( ).

A Análise Discriminante com base nos 4 grupos originais, atingiu uma porcentagem de indivíduos bem classificados da ordem de 50,77%, no passo de ordem número 7, o que é considerado um resultado sofrível do

ponto de vista da discriminação.

Na tabela (II.1.2), tem-se a distribuição de indivíduos, em termos dos grupos de origem e dos grupos de afetação.

(II.1.2) TABELA

GRUPO DE AFETAÇÃO

		1	2	3	4
GRUPO	1	6	2	1	4
DE	2	6	8	6	1
ORIGEM	3	0	5	11	1
	4	2	3	1	8

VARIÁVEIS: 6, 3, 5, 1, 7, 8, 2

Em seguida, consideram-se os resultados, caso sejam reunidos os grupos I e II numa única classe gastrite crônica ativa ou quiescente). Nesta situação, no passo de número 8 é atingido um prcentual de 67,69% de pacientes bem classificados; vide Tabela (II.1.3).

(II.1.3) TABELA

GRUPO DE AFETAÇÃO

		1	2	3
GRUPO	1	23	6	5
DE	2	4	11	2
ORIGEM	3	2	2	10

VARIÁVEIS: 4, 5, 1, 2, 6, 3, 7, 8

Por outro lado, levando em conta que os indivíduos do grupo I (gastrite crônica quiescente) constituem aqueles pacientes que são menos levados a consultar o médico, experimentou-se realizar a discriminação

com base, unicamente, na presença dos grupos II, III e IV, sendo os grupos (gastrite crônica ativa ou úlcera péptica) por sua vez reunidos numa única classe. Neste caso, é atingido um porcentual de indivíduos bem classificados da ordem de 82,69%, no passo número 8, vide tabela II.1.4 abaixo.

(II.1.4) TABELA

GRUPO DE AFETAÇÃO

GRUPO DE ORIGEM		1	2
	1	30	8
	2	1	13

VARIÁVEIS: 5, 4, 2, 1, 6, 8, 3, 7

Note-se que todos os indivíduos foram utilizados para constituir nossa "amostra-de-base", tendo em vista que seu número total era pequeno. Assim, não foi possível considerar uma amostra de base. Por outro lado, também se dispunha de um número reduzido de variáveis a serem empregadas com finalidades discriminatórias.

(II.2) UMA APLICAÇÃO EM CLIMATOLOGIA

Para esta aplicação os dados foram obtidos pela Profa. Teresinha de M^{a} B.S. Xavier, conforme Girardi A. Teixeira (1979), que trabalharam com Análise Harmônica, foi possível extrair da série de totais pluviométricos de Fortaleza (no intervalo compreendendo os anos hidrológicos 1848/49 - 1977/78), duas componentes cíclicas dominantes, com períodos de 13 a 26 anos, aproximadamente. Conforme esses autores concluiram, a concordância entre os mínimos e máximos das curvas sinussóides correspondentes aos dois ciclos mencionados, corresponderiam a uma probabilidada máxima para

a ocorrência respectivamente, de anos secos e de anos excepcionalmente chuvosos. Por sua vez, Xavier A, Xavier (1981) reexaminaram a questão, tendo concluído por um reduzido poder explicativo dos ciclos, quanto à previsão de mínimos e de máximos para a pluviometria anual.

Neste exemplo, consideremos as variáveis X(1),X(2),...,X(12) que são os totais pluviométricos mensais, tomados em ordem decrescente das alturas pluviométricas em cada ano hidrológico. Esses, foram separados em três grupos A, M e B, correspondentes aos anos com previsão de precipitação alta, média e baixa, conforme Teixeira A. Girardi (op. cit). Constituímos uma amostra-de-base (compreendendo 30, 40 e 30 anos, nos grupos B, A e M respectivamente); e uma amostra-teste (com 10 para cada um dos grupos).

Utilizamos o programa MAHAL 3 FOR, o qual, no passo número 4 nos deu um porcentual de 54% para anos hidrológicos bem classificados, na amostra-de-base; na amostra-teste o porcentual cai para 36.67%.

No passo número 10, o porcentual de bem classificado na amostra-de-base sobe apenas para 56% enquanto na amostra-teste permanece invariável.

Conclui-se, pois, que as previsões antes mencionadas, em termos da ocorrência de anos secos, normais ou excepcionalmente chuvosos, não resistem convenientemente, a uma análise discriminante, quando se considera o conjunto dos totais pluviométricos nos diversos anos hidrológicos.

REFERÊNCIAS BIBLIOGRÁFICAS

AMORIM, W.P.D. de, Isoenzimas da LDH (Dehidrogenase Láctica) em Mucosa Gástrica de Pacientes com Câncer Gástrico e Doenças Gástricas Benígnas, Tese de Mestrado em Medicina , Instituto de Estudos e Pesquisas em Gastroenterologia, São Paulo, 1984.

ANDERSON, T.W. , An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, 1958.

FISHER, R.A. , Statistical Methods for Research Workers, New York, Hafner, 1970.

FISZ, M. , Probability Theory and Mathematical Statistica , John Wiley & Sons Inc, New York, London, 1963.

GIRARDI, L.A. & TEIXEIRA, L. , Prognóstico de Tempo a Longo Prazo (Prognóstico de Período de Seca para o Nordeste Brasileiro), CTA/IAE, Relatório Técnico ECA nº 06/72, 1978.

GOLDMAN, M. , Introduction to Probability and Statistics , Harcourt, Brace & World Inc, New York, 1970.

GRAYBILL, A.A. , An Introduction to Linear Statistical Models, Volume I, McGraw Hill, New York, 1961.

HOFFMAN, K. & KUNZE, R. , Álgebra Linear, LTC, Rio de Janeiro, 1979.

JOHNSON, N.L. , Distribuitions in Statistics:Continuous Multivariate Distribuitions, John Wiley, New York, 1972

KAPLAN, W. , Advanced Calculus, Reading Addison Wesley, 1969.

KING, L. J. , Statistical Analysis in Geography, Prentice Hall, Englewood Cliffs, 1969.

MILLER, K. S. , Multidimentional Gaussian Distribuitions, John Wiley, New York, 1964.

MURDOCH, D. C. , Álgebra Linear, LTC, Rio de Janeiro, 1972.

NOBLE, B. , Applied Linear Algebra, Prentice Hall, Englewood Cliffs, 1969.

NAKACHE, J. P. , Méthodes de Discrimination sur Variables de Nature Qualconque, Thêse de Doctorat d'État, Univer - sité Pierre et Marie Curie, Paris, 1980.

RAO, C. R. , Advanced Statistical Methods in Biometric Research, John Wiley, 1952.

ROMEDER, J.-H. , Méthodes et Programes d'Analyse Discriminante, Dunod, Paris, 1972.

SEBESTYEN, G. S. , Decision Making Processes in Pattern Recognition, Macmillan Co. , 1962.

VOLLE, M. , Analyse des Donneês, Economica, Paris, 1981.

XAVIER, T. Mº. B.S. & XAVIER, A. F. S. , Periodicidades nas Séries Pluviométricas de Fortaleza e Quixeramobim (Ceará) e de Mossoró (Rio Grande do Norte), Anais do IV Simpósio Brasileiro de Hidrologia e Recursos Hídricos , Fortaleza - 15 a 19 nov/81, Volume I, p- 423-440, 1981.