# MISTURA DE MULTINORMAIS

COMO

# TÉCNICA DE ANÁLISE DE CONGLOMERADOS

EDINA SHISUE MIAZAKI

DISSERTAÇÃO APRESENTADA AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA A OBTENÇÃO DO GRAU DE MESTRE

EM

ESTATÍSTICA

ORIENTADOR:

*Prof. Dr. WILTON DE OLIVEIRA BUSSAB*

- SÃO PAULO, OUTUBRO DE 1979 -

*"O amor, o trabalho e o conhecimento*
*são as fontes da nossa vida.*
*Deveriam também governá-la"*

*(Wilhelm Reich)*

# A G R A D E C I M E N T O

Ao final deste trabalho gostaríamos de agradecer ao Prof. Wilton de Oliveira Bussab, orientador paciente, pelo muito que nos ensinou, e sobretudo por ter nos honrado com sua amizade.

Este trabalho foi realizado com auxílio parcial do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

# P R E F Á C I O

A Análise de Conglomerados trata de um problema comum em áreas de pesquisas observacionais, que é a descoberta automática da estrutura da população em estudo. Everitt (1974) descreve-a sucintamente como uma técnica de Análise Multivariada que tem por objetivo resolver o seguinte problema:

> — "Dado um conjunto de $n$ indivíduos ou objetos, os quais são medidos em $q$ variáveis, quer se determinar um esquema de classificação para agrupar os indivíduos em $g$ $(g < n)$ grupos. Também devem ser determinados o número de grupos e suas características".

Existe um número muito grande de propostas para a solução desse problema, tendo a maior parte delas surgido em áreas biológicas. Assim sendo, este trabalho não pretende fazer um estudo exaustivo das técnicas existentes, mas sim se restringir a apresentação e discussão de alguns métodos, considerados mais interessantes por terem algum suporte estatístico.

A leitura desta monografia pressupõe o conhecimento de conceitos básicos de estimação através do método da máxima verossimilhança.

# Í N D I C E

# CAPÍTULO I

## INTRODUÇÃO

Na área das ciências experimentais são comuns as situações em que se deseja estudar uma população na qual várias características são observáveis. Para tanto, toma-se uma amostra dessa população e fazem-se as várias medidas dessas características sobre cada unidade amostral.

Entretanto, em algumas dessas situações existem razões para se acreditar que a população considerada é heterogênea, dificultando a sua caracterização. Nestes casos a solução é descobrir as sub-populações homogêneas que compõem a população, permitindo ao pesquisador um estudo das propriedades subjacentes a cada uma delas isoladamente. Raramente o pesquisador dispõe de um critério teórico para definir essas sub-populações, tornando necessário o uso de técnicas numéricas para descobrir a estrutura dessa população. Como exemplo de aplicação considere um pesquisador que estava descontente com o método de classificação latos solos através do teor de Ferro e da coloração. Primeiro porque a variável coloração lhe parecia um tanto subjetiva, e segundo, porque as categorias assim obtidas pareciam ainda bastante heterogêneas. Assim, tomou um conjunto dos vários tipos de solo, estudou a composição química de cada um deles e, através de técnicas apropriadas, agrupou aqueles mais similares entre si. Dessa forma conseguiu criar uma classificação objetiva e mais refinada dos solos.

Em outras situações o pesquisador se vê com um número muito grande de observações para analisar e necessita de técnicas que resumam os dados para que estes se tornem

mais informativos. Ele conseguirá uma simplificação dos dados, perdendo um mínimo de informação, se agrupar esses dados em grupos internamente homogêneos. Poderá então reduzir seu conjunto de dados para um conjunto de valores típicos dos grupos. Como exemplo dessa técnica de redução de dados, suponha que um novo produto esteja para ser lançado no mercado. Para testar a sua aceitação junto ao público consumidor, a firma lançadora resolveu aplicar questionários nas cidades em que o produto deverá ser colocado à venda. Porém, como o número de cidades é grande, essa atitude iria onerar demasiadamente a pesquisa, e fornecer uma massa de dados muito grande e difícil de analisar. Assim, optou-se por aplicar questionários apenas nas cidades mais representativas do mercado consumidor. A seleção dessas cidades típicas pode ser feita através de uma técnica numérica. Com base em informações básicas de todas as cidades, pode-se agrupá-las em grupos internamente homogêneos, e escolher aquelas mais representativas dos seus respectivos grupos.

O método numérico que visa agrupar os $n$ elementos de uma amostra em grupos com alta homogeneidade interna é mais conhecido como Análise de Conglomerados (A.C.).

Existe uma confusão natural entre Classificação e Análise de Conglomerados. Aqui será feita uma distinção entre os dois tratamentos. O primeiro tem por objetivo alocar novos elementos em classes já conhecidas. A informação disponível sobre o número de classes e as características de cada uma delas é que possibilitam classificar cada um dos novos elementos em suas respectivas categorias. Já a Análise de Conglomerados tem por objetivo descobrir as classes e caracterizá-las. Pode-se dizer que a Análise de Conglomerados é o tratamento que antecede a Classificação.

Esta técnica se desenvolveu em áreas bastante di

versas como Biologia, Psicologia, Inteligência Artificial, etc., o que explica a variedade de nomes pelos quais ela é chamada, tais como Tipologia, Pré-Classificação, Taxonomia Numérica, Reconhecimento de Padrões, etc.. Neste trabalho serão usados indistintamente os termos Análise de Conglomerados e Análise de Agrupamento para a técnica em si, bem como grupos, categorias ou classes para os conglomerados.

A secção 1 apresenta resumidamente o desenvolvimento das técnicas de agrupamento, mostrando a transição da área da Biologia para a Estatística.

A secção 2 trata de alguns problemas mais comuns em Análise de Conglomerados.

Finalmente, a proposta deste trabalho está na secção 3.

## 1.1. DESENVOLVIMENTO DAS TÉCNICAS

As primeiras etapas de agrupamento surgiram na Taxonomia Animal e Vegetal, onde as plantas e os animais eram classificados por critérios subjetivos. A classificação das plantas e animais consistia mais em arte do que em um método científico. Talvez a imprecisão dos métodos até então utilizados tenha impelido os pesquisadores a procurar técnicas mais objetivas, e os métodos numéricos foram surgindo naturalmente.

A descrição inicial do que hoje é conhecido como Análise de Conglomerados foi formulada por Tryon em 1939.

Zubin e Thorndike (1953) tentaram usar os métodos de agrupamento em outras áreas além das Ciências Naturais mas não obtiveram sucesso pelo fato do trabalho operacional ser muito grande.

Somente a partir da última década, com o aparecimento de computadores eletrônicos, as técnicas numéricas

de classificação tiveram seu uso difundido em outras áreas.

Fisher (1968) discutiu métodos particulares relevante no problema de agregação em Economia. Tryon e Bailey (1970) lançaram um livro onde discutem um sistema para computador para a execução de Análise de Conglomerados e Análise Fatorial, do ponto de vista de um cientista social. Jardine e Sibson (1971) deram uma abordagem axiomática aos métodos relacionados com a taxonomia biológica, mas com aplicações em outras áreas.

Os computadores não só difundiram o uso das técnicas de Análise de Conglomerados como também propiciaram o desenvolvimento de inúmeros algorítmos para agrupar elementos semelhantes. A maior parte desses algorítmos são baseados em técnicas de otimização.

Uma forma de se buscar conglomerados com alta homogeneidade interna é buscar um agrupamento cuja dispersão dentro dos grupos seja pequena. Assim, com base na minimização da dispersão dentro dos conglomerados, surgiram algumas técnicas derivadas de resultados estatísticos. Estas técnicas são usualmente conhecidas como método da minimização da variância, e foram estudadas por Edwards & Cavalli-Sforza (1965) e Friedman & Rubin (1968) entre outros. Posteriormente foi desenvolvido um modelo estatístico (Scott & Symons (1972)) englobando os critérios citados.

Até então as técnicas de agrupamentos eram despidas de qualquer conceituação estatística. A partir de 1970, a Análise de Conglomerados vem recebendo uma abordagem mais rigorosa do ponto de vista da estatística teórica. Esta abordagem considera a distribuição de probabilidades subjacente aos dados, e conseqüentemente envolve conceitos de estimação de parâmetros, testes de significância,etc.. Sob este ponto de vista, um conglomerado é definido em termos das características da função densidade da população da

qual os dados foram amostrados, enquanto nos procedimentos heurísticos, as próprias observações é que sugerem os conglomerados. Já existem alguns modelos com suporte estatístico. Por exemplo, aquele baseado em mistura de distribuições, ou ainda, outro que se baseia na estimação das modas de uma distribuição multimodal. Outras propostas de modelo estão surgindo gradativamente, e a Análise de Conglomerados se afirma cada vez mais dentro da Estatística, principalmente como uma técnica descritiva de dados multivariados.

## 1.2. ALGUNS PROBLEMAS COMUNS

O conceito geral de que um conglomerado é a reunião de objetos similares e o fato da Análise de Conglomerados ser uma técnica aplicável a diferentes áreas de pesquisa, levaram ao aparecimento de um número muito grande de técnicas para agrupar dados. A maior parte delas consiste em definir medidas de similaridade e/ou dissimilaridade, para então, de acordo com essas medidas, agrupar os objetos mais similares entre si. Existe um número muito grande dessas medidas. Só Hartigan (1967) listou doze delas. O problema com essas técnicas, é a determinação da medida mais conveniente, uma vez que as técnicas baseadas em medidas de similaridade diferentes nem sempre levam aos mesmos resultados.

Um índice de similaridade largamente utilizado nos métodos de Análise de Conglomerados é a distância entre dois elementos. Elementos num mesmo conglomerado devem estar próximos entre si, enquanto que elementos de conglomerados diferentes devem estar distantes um do outro. As distâncias mais comuns nesses métodos são a distância Euclideana e a distância de Mahalanobis, denotadas por $d$ e

$D$ respectivamente, que estão definidas a seguir.

Sejam dois pontos $\underset{\sim}{x}$ e $\underset{\sim}{y}$ num espaço $p$-dimensional. definem-se

$$d(\underset{\sim}{x},\underset{\sim}{y}) = \left[(\underset{\sim}{x}-\underset{\sim}{y})' \ (\underset{\sim}{x}-\underset{\sim}{y})\right]^{\frac{1}{2}}$$

e

$$D(x,y) = \left[(\underset{\sim}{x}-\underset{\sim}{y})' \ \underset{\sim}{W}^{-1} \ (\underset{\sim}{x}-\underset{\sim}{y})\right]^{\frac{1}{2}}$$

onde $\underset{\sim}{W}$ é uma matriz simétrica positiva definida. Geralmente, toma-se como $\underset{\sim}{W}$ a matriz de dispersão dos dados.

Também nestes casos, as técnicas não são invariantes sob diferentes definições da distância entre os elementos.

Uma suposição implícita nas técnicas de agrupamento é que as variáveis medidas são aquelas relevantes na solução desejada. Porém, ainda assim é comum o número excessivamente grande de variáveis observadas. Este é um problema que além de dificultar a interpretação dos resultados, torna certas técnicas inaplicáveis, como será visto mais adiante. É possível solucionar esse problema de várias formas diferentes. A primeira delas é deixar a escolha das variáveis mais importantes por conta do pesquisador. Quando não for possível obter essa escala de importância das variáveis, recorre-se a técnicas estatísticas que buscam diminuir a dimensão do espaço de variáveis. Por exemplo, a Análise de Componentes Principais, ou a Análise Fatorial.

Em geral, as técnicas de Análise de Conglomerados constroem um agrupamento utilizando critérios que englobam todas as variáveis indistintamente. Assim, se as ordens de grandeza das variáveis observadas forem muito díspares entre si, o resultado da análise pode ficar seriamente comprometido. Isso porque, ao se considerar o conjunto de observações de uma maneira absoluta, algumas das variáveis

(as de magnitude menor) darão a falsa impressão de não serem discriminatórias, uma vez que as suas diferenças serão insignificantes quando comparadas com as diferenças nas demais variáveis. Qualquer critério global de agrupamento basear-se-á fortemente na variável cujas medidas tiverem maior ordem de grandeza, mesmo que ela não seja a mais discriminatória. Hartigan (1975) ilustra bem esse fato no exemplo que será transcrito aqui. Quer se agrupar um conjunto de alimentos (carnes, peixes e aves) segundo seus principais nutrientes (caloria, gordura, cálcio e ferro). Os dados foram obtidos tomando-se as unidades em três onças de peso, e estão na tabela 1.1. Como se vê as variáveis estão medidas em escalas diferentes, de forma que a diferença de uma unidade na variável caloria pode mascarar uma diferença de uma unidade na variável ferro, a qual seria mais significativa do que a primeira. Para perceber esse fato, é suficiente olhar a última linha da tabela 1.1, e verificar que os desvios padrões das variáveis são muito diferentes entre si. Necessita-se, portanto, de uma transformação que torne as variâncias mais homogêneas. Entretanto, a padronização, conforme é usual em Estatística, reduzindo todas as variáveis a uma mesma variância unitária, equivale a uma transformação em que todas as variáveis ficam com o mesmo grau de agrupabilidade. Isso não é desejável em Análise de Conglomerados porque é comum existirem variáveis que diferenciam os grupos mais do que outras, e esse caráter diferenciável é importante neste tipo de análise. O problema visa procurar por uma transformação que não tenha esses problemas, e permite às variáveis manter seu significado original. Assim, Hartigan optou naquele caso por uma padronização em que as unidades de medida dos nutrientes foram transformadas em porcentagem das necessidades diárias de cada um deles, segundo "Yearbook

of Agriculture" (1959). Dessa forma as variáveis transformadas têm a mesma magnitude, sem terem perdido suas importâncias individuais, conforme se vê na tabela 1.2. Para se ter uma idéia do ganho obtido com a transformação, basta comparar os desvios padrões das variáveis transformadas na tabela 1.2 com os das variáveis originais. Convém ressaltar que a escolha da transformação não é um problema estatístico, e ela deve ser sugerida pelo bom senso e pelo conhecimento da área em que se processa a pesquisa.

## 1.3. OBJETIVOS

Este trabalho pretende apresentar didaticamente as técnicas de Análise de Conglomerados que se desenvolvem baseadas nas propriedades da distribuição normal multivariada, de forma a fornecer um guia para o leitor, mesmo aquele pouco familiarizado com essa técnica, interessado na aplicação de uma análise de agrupamento.

Para isso, o trabalho será apresentado da seguinte forma.

No capítulo II, tem-se a descrição das técnicas que buscam minimizar a dispersão dentro dos conglomerados. O capítulo III trata de uma técnica um pouco mais refinada: a mistura de duas populações multinormais. No capítulo IV os resultados deduzidos no capítulo III são estendidos para o caso mais geral de mistura de várias populações multinormais. Em cada um desses capítulos está incluída uma descrição dos possíveis problemas que podem eventualmente surgir na aplicação prática dessa técnica. A mistura de distribuições multinormais é estudada em mais detalhe por ser ainda pouco explorada como técnica de agrupamento. Além disso é uma das técnicas de Análise de Conglomerados que mais se utiliza de resultados estatísticos.

Também por ser um problema pouco estudado, e constituir um sério problema em Análise de Conglomerados, os métodos de determinação do número de grupos estão no Capítulo V.

O capítulo VI apresenta exemplos de aplicação prática.

As conclusões estão no capítulo VII.

# TABELA 1.1.

| | | CALORIAS | PROTEÍNA | GORDURA | CÁLCIO | FERRO |
|---|---|---|---|---|---|---|
| BB | Bife grelhado | 340 | 20 | 28 | 9 | 2.6 |
| HR | Hamburger | 245 | 21 | 17 | 9 | 2.7 |
| BR | Rosbife | 420 | 15 | 39 | 7 | 2.0 |
| BS | Bife | 375 | 19 | 32 | 9 | 2.6 |
| BC | Carne enlatada | 180 | 22 | 10 | 17 | 3.7 |
| CB | Galinha cozida | 115 | 20 | 3 | 8 | 1.4 |
| CC | Galinha enlatada | 170 | 25 | 7 | 12 | 1.5 |
| BH | Coração de boi | 160 | 26 | 5 | 14 | 5.9 |
| LL | Pernil de carneiro assado | 265 | 20 | 20 | 9 | 2.6 |
| LS | Dianteiro de carneiro assado | 300 | 18 | 25 | 9 | 2.3 |
| HS | Presunto defumado | 340 | 20 | 28 | 9 | 2.5 |
| PR | Porco assado | 340 | 19 | 29 | 9 | 2.5 |
| PS | Porco ao bafo | 355 | 19 | 30 | 9 | 2.4 |
| BT | Língua de boi | 205 | 18 | 14 | 7 | 2.5 |
| VC | Costeleta de vitela | 185 | 23 | 9 | 9 | 2.7 |
| FI | Merlin assado | 135 | 22 | 4 | 25 | 0.6 |
| AR | Ostras cruas | 70 | 11 | 1 | 82 | 6.0 |
| AC | Ostras enlatadas | 45 | 7 | 1 | 74 | 5.4 |
| TC | Siri enlatado | 90 | 14 | 2 | 38 | 0.8 |
| HF | Haddock frito | 135 | 16 | 5 | 15 | 0.5 |
| MB | Cavala cozida | 200 | 19 | 13 | 5 | 1.0 |
| MC | Cavala enlatada | 155 | 16 | 9 | 157 | 1.8 |
| PF | Perca frita | 195 | 16 | 11 | 4 | 1.3 |
| SC | Salmão enlatado | 120 | 17 | 5 | 155 | 0.7 |
| DC | Sardinha enlatada | 180 | 22 | 9 | 367 | 2.5 |
| UC | Atum enlatado | 170 | 25 | 7 | 7 | 1.2 |
| RC | Camarão enlatado | 110 | 23 | 1 | 98 | 2.6 |
| | Média | 207,41 | 19,00 | 13,48 | 43,96 | 2,38 |
| | Desvio Padrão | 101,21 | 4,25 | 11,27 | 78,03 | 1,46 |

# TABELA 1.2.

| | ENERGIA (caloria) | PROTEÍNA (gramas) | GORDURA (gramas) | CÁLCIO (ng) | FERRO (ng) |
|---|---|---|---|---|---|
| Bife grelhado | 11 | 29 | 28 | 1 | 26 |
| Hamburger | 8 | 30 | 17 | 1 | 27 |
| Rosbife | 13 | 21 | 39 | 1 | 20 |
| Bife | 12 | 27 | 32 | 1 | 26 |
| Carne enlatada | 6 | 31 | 10 | 2 | 37 |
| Galinha cozida | 4 | 29 | 3 | 1 | 14 |
| Galinha enlatada | 5 | 36 | 7 | 2 | 15 |
| Coração de boi | 5 | 37 | 5 | 2 | 59 |
| Pernil de carneiro assado | 8 | 29 | 20 | 1 | 26 |
| Dianteiro de carneiro assado | 9 | 26 | 25 | 1 | 25 |
| Presunto defumado | 11 | 29 | 28 | 1 | 25 |
| Porco assado | 11 | 27 | 29 | 1 | 25 |
| Porco ao bafo | 11 | 27 | 30 | 1 | 25 |
| Língua de boi | 6 | 26 | 14 | 1 | 25 |
| Costeleta de vitela | 6 | 33 | 9 | 1 | 27 |
| Merlin assado | 4 | 31 | 4 | 3 | 6 |
| Ostras cruas | 2 | 16 | 1 | 10 | 60 |
| Ostras enlatadas | 1 | 10 | 1 | 9 | 54 |
| Siri enlatado | 3 | 20 | 2 | 5 | 8 |
| Haddock frito | 4 | 23 | 5 | 2 | 5 |
| Cavala cozida | 6 | 27 | 13 | 1 | 10 |
| Cavala enlatada | 5 | 23 | 9 | 20 | 18 |
| Perca frita | 6 | 23 | 11 | 2 | 13 |
| Salmão enlatado | 4 | 24 | 5 | 20 | 7 |
| Sardinha enlatada | 6 | 31 | 9 | 46 | 25 |
| Atum enlatado | 5 | 36 | 7 | 1 | 12 |
| Camarão enlatado | 3 | 33 | 1 | 12 | 26 |
| Média | 6,50 | 27,19 | 13,48 | 5,52 | 23,93 |
| Desvio Padrão | 3,25 | 6,08 | 11,26 | 9,78 | 14,62 |

# CAPÍTULO II

## MÉTODOS BASEADOS NA MATRIZ DE DISPERSÃO

Este capítulo consiste na apresentação de alguns métodos de A.C. desenvolvidos heuristicamente, e que mais tarde foram mostrados como sendo extensões de um modelo estatístico de classificação.

Na secção 1 são introduzidas algumas definições e conceitos básicos para a compreensão do texto.

A descrição dos métodos é feita na secção 2.

A secção 3 apresenta o modelo da máxima verossimilhança.

Alguns comentários sobre os métodos descritos são tecidos na secção 4.

### 2.1. DEFINIÇÕES

Seja uma coleção de observações $I = \{I_1,\ldots,I_n\}$, e considere associada a ela, um vetor observável de $p$ características $\underset{\sim}{X}' = (X_1,X_2,\ldots,X_p)$.

Dessa maneira, obtém-se os $n$ vetores:

$$\underset{\sim}{X}_i' = (x_{i1},\ldots,x_{i2},\ldots,x_{ip}) \quad i = 1,\ldots,n$$

onde

$x_{ij}$ é a medida observada na *j-ésima* variável do *i-ésimo* elemento.

Se cada vetor de observação for pensado como um ponto num espaço *p-dimensional*, cujos eixos são representados pelas variáveis em questão, os conglomerados podem ser descritos como regiões desse espaço contendo uma concentração relativamente grande de pontos, separados uma das outras por regiões cuja densidade de pontos é relativamente

baixa. Com esta definição de conglomerado, a dispersão dos pontos no espaço parece ser um bom critério para as técnicas de agrupamento. Para melhor entendimento das técnicas a serem descritas, seguem algumas definições.

*Definição 1:*

A matriz simétrica $\underset{\sim}{SS}$ $(p \times p)$

$$\underset{\sim}{SS} = \sum_{i=1}^{n} (\underset{\sim}{X}_i - \overline{\underset{\sim}{X}})\ (\underset{\sim}{X}_i - \overline{\underset{\sim}{X}})' \quad \text{onde} \quad \overline{\underset{\sim}{X}} = \frac{\sum_{i=1}^{n} \underset{\sim}{X}_i}{n} \tag{2.1}$$

é chamada de matriz de dispersão, ou matriz soma de quadrados para vetor $\underset{\sim}{X}$.

$$\underset{\sim}{SS} = [ss_{ij}], \quad i = 1,\ldots,p; \quad j = 1,\ldots,p$$

$$ss_{ij} = \begin{cases} \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)^2 , & i = j \\ \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)\ (x_{kj} - \overline{x}_j) , & i \neq j \end{cases}$$

onde

$$\overline{x}_i = \sum_{k=1}^{n} \frac{x_{ki}}{n}$$

Os elementos da diagonal de $\underset{\sim}{SS}$ fornecem as dispersões de cada uma das variáveis observadas, enquanto os elementos fora da diagonal fornecem as somas dos produtos cruzados das variáveis duas a duas.

$\underset{\sim}{SS}$ é uma medida de dispersão multivariada.

*Definição 2:*

A matriz de variâncias - covariâncias do vetor $\underset{\sim}{X}$, é definida como:

$$\underset{\sim}{S} = \frac{1}{n-1}\ \underset{\sim}{SS} \tag{2.2}$$

$\underset{\sim}{S}$ é a medida multivariada análoga à variância $\sigma^2$ em distribuições multivariadas.

*Definição 3:*

Define-se como variância generalizada do vetor $\underset{\sim}{X}$ ao determinante da matriz $\underset{\sim}{S}$.

Variância generalizada de $\underset{\sim}{X} = |\underset{\sim}{S}|$

A interpretação geométrica da variância generalizada pode ser feita em termos dos $n$ pontos num espaço $p$-*dimensional*.

Sejam $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_p$ - pontos num espaço $p$-*dimensional*. Ligando cada um desses pontos à origem, tem-se $p$ vetores, também denotados por $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_p$.

Seja $A$ uma matriz $p \times p$ dada por:

$$a_{ij} = \underset{\sim}{y}_i' \underset{\sim}{y}_j = ||\underset{\sim}{y}_i|| \; ||\underset{\sim}{y}_j|| \cos\theta, i=1,\ldots,p, j=1,\ldots,p$$

onde $\theta$ é o ângulo formado pelos dois vetores $\underset{\sim}{y}_i$ e $\underset{\sim}{y}_j$.

Se o espaço for bi-dimensional, os vetores $\underset{\sim}{y}_1$ e $\underset{\sim}{y}_2$ definem um paralelogramo. Basta traçar uma linha paralela ao vetor $\underset{\sim}{y}_2$ e que passe pelo ponto $\underset{\sim}{y}_1$; e outra, passando pelo ponto $\underset{\sim}{y}_2$ e paralela ao vetor $\underset{\sim}{y}_1$. Veja a figura 1 baseada nesses vetores. A área desse paralelogramo é dada por $||\underset{\sim}{y}_1|| \; ||\underset{\sim}{y}_2|| \; sen\,\theta$, onde $||\underset{\sim}{y}_i||$ é o comprimento do vetor $\underset{\sim}{y}_i$, $i = 1,2$ e $\theta$ é o ângulo formado pelos vetores $\underset{\sim}{y}_1$ e $\underset{\sim}{y}_2$. (Anderson, (1958)).

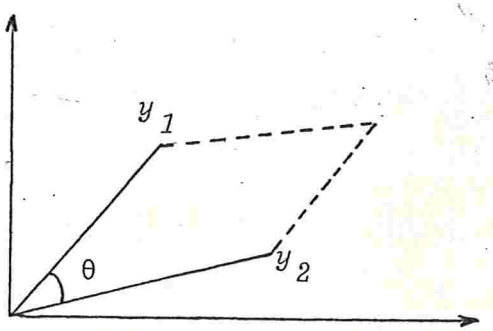



Figura 1

Os vetores $\underset{\sim}{y}_1$ e $\underset{\sim}{y}_2$ são chamados de margens principais do paralelogramo definido por eles.

$$(Área)^2 = \underset{\sim}{y}_1' \underset{\sim}{y}_1 \; \underset{\sim}{y}_2' \underset{\sim}{y}_2 \; sen^2 \theta = .$$

$$= \underset{\sim}{y}_1' \underset{\sim}{y}_1 \; \underset{\sim}{y}_2' \underset{\sim}{y}_2 - \underset{\sim}{y}_1' \underset{\sim}{y}_1 \; \underset{\sim}{y}_2' \underset{\sim}{y}_2 \; cos^2 \theta =$$

$$= y_1' y_1 \; y_2' y_2 - y_1' y_2 \, | y_2' y_1 = |\underset{\sim}{A}|$$

Se a dimensão do espaço for 3, a figura definida por $\underset{\sim}{y}_1$, $\underset{\sim}{y}_2$, $\underset{\sim}{y}_3$ será um paralelepípedo, cujo volume ao quadrado será igual ao $|\underset{\sim}{A}|$.

$$(Volume)^2 = |\underset{\sim}{A}|$$

Para o caso geral de dimensão $p$, a figura formada será um paralelepípedo *p-dimensional*, que é o sólido gerado por $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_p$. O volume ao quadrado desta figura é dado por $|\underset{\sim}{A}|$.

Nos casos em que se tem $n$ $(n > p)$ vetores *p-dimensionais*, $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n$, $|\underset{\sim}{A}| = \sum_i |A_i|$, onde a somatória é sobre todos os conjuntos possíveis de $p$ vetores em $n$. Ou seja, $|\underset{\sim}{A}|$ é a soma dos volumes ao quadrado de todos os sólidos definidos por $p$ vetores em $n$. (Anderson, (1958)).

Assim, o $|\underset{\sim}{A}|$ dá uma idéia da dispersão dos pontos no espaço. Quanto menor o valor do $|\underset{\sim}{A}|$, tanto maior será a concentração deles. Parece então explicado porque um dos critérios mais utilizados em técnicas de agrupamento está fundamentado na minimização do $|\underset{\sim}{A}|$. Ver-se-á essa técnica bem detalhada mais adiante.

Fazendo $\underset{\sim}{y}_i = \underset{\sim}{x}_i - \overline{\underset{\sim}{x}}$, tem-se o seguinte teorema: (Anderson (1958))

*Teorema 1:* "Seja $\underset{\sim}{S}$, definida por (2.2), onde $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ são $n$ pontos de uma amostra. Então $|\underset{\sim}{S}|$ é proporcional a soma dos volumes ao quadrado de todos os paralelepípedos de dimensão 1 formados usando como margens principais $p$ veto

res dentre um conjunto de $n$ vetores definidos em uma extremidade por $\bar{\underset{\sim}{x}}$ e em outra por $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$. O fator proporcionalidade é dado por $(n-1)^{-p}$".

Se $|\underset{\sim}{S}|$ assumir um valor numérico alto, é sinal de que as observações estão muito espalhadas, e portanto a variância de pelo menos uma das variáveis é grande. Ao contrário, $|\underset{\sim}{S}|$ igual a zero indica que o volume de todos os sólidos formados por conjuntos de $p$ vetores $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_p$ são nulos, ou seja, um dos vetores é combinação linear dos demais, produzindo uma forte correlação entre as variáveis.

*Definição 3:*

O traço da matriz $\underset{\sim}{S}\underset{\sim}{S}$, é conhecido como a dispersão estatística do vetor $\underset{\sim}{X}$.

$$
\begin{aligned}
tr\,(\underset{\sim}{S}\underset{\sim}{S}) &= tr\Big[\sum_{i=1}^{n} (\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})\,(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})'\Big] = \\
&= \sum_{i=1}^{n} tr\,(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})\,(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})' = \\
&= \sum_{i=1}^{n}\sum_{k=1}^{p} (x_{ik} - \bar{x}_k)^2 = \\
&= \sum_{i=1}^{n} (\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})'\,(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}}).
\end{aligned}
\qquad (2.3)
$$

Vê-se que o $tr\,(\underset{\sim}{S}\underset{\sim}{S})$ também fornece uma medida de dispersão conjunta de $\underset{\sim}{X}$, uma vez que é simplesmente a soma das dispersões de cada variável. É uma medida que não considera a estrutura de correlação entre as variáveis observadas.

O $tr\,(\underset{\sim}{S}\underset{\sim}{S})$ também é chamado de soma de quadrados devido ao erro, ou, soma de quadrados da população $I$.

Nas aplicações da A.C., além da homogeneidade interna dos grupos, há interesse em maximizar a disparida

de entre os grupos. Convém verificar como se relacionam essas duas medidas e como é definida uma medida de disparidade entre conglomerados.

Suponha que o conjunto $I$ é formado por dois conglomerados $J = \{J_1, \ldots, J_m\}$ e $K = \{K_1, K_2, \ldots, K_n\}$ $(I = J \cup K)$ mutuamente exclusivos, que geram dois conjuntos de observações $\underset{\sim}{X} = \{\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_m\}$ e $\underset{\sim}{Y} = \{\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, \ldots, \underset{\sim}{Y}_n\}$ associados a $J$ e a $K$ respectivamente. Seja

$$\underset{\sim\sim}{SS}_I = \sum_{i=1}^{m} (\underset{\sim}{X}_i - \underset{\sim}{M})(\underset{\sim}{X}_i - \underset{\sim}{M})' + \sum_{i=1}^{n} (\underset{\sim}{Y}_i - \underset{\sim}{M})(\underset{\sim}{Y}_i - \underset{\sim}{M})',$$

onde

$$\underset{\sim}{M} = \frac{\sum_{i=1}^{m} \underset{\sim}{X}_i + \sum_{i=1}^{n} \underset{\sim}{Y}_i}{m + n}$$

Então

$$\underset{\sim\sim}{SS}_I = m(\overline{\underset{\sim}{X}} - \underset{\sim}{M})(\overline{\underset{\sim}{X}} - \underset{\sim}{M})' + n(\overline{\underset{\sim}{Y}} - \underset{\sim}{M})(\overline{\underset{\sim}{Y}} - \underset{\sim}{M})' + \underset{\sim\sim}{SS}_J + \underset{\sim\sim}{SS}_K.$$

Chamando $\underset{\sim}{S}_I$ de soma de quadrados total, e denotando por $T$, e $(\underset{\sim}{S}_J + \underset{\sim}{S}_K)$ de soma de quadrados dentro dos grupos $J$ e $K$, e denotando por $\underset{\sim}{W}$, pode-se escrever

$$\underset{\sim}{T} = \underset{\sim}{B} + \underset{\sim}{W} \qquad (2.4)$$

onde

$$\underset{\sim}{B} = m(\overline{\underset{\sim}{X}} - \underset{\sim}{M})(\overline{\underset{\sim}{X}} - \underset{\sim}{M})' + n(\overline{\underset{\sim}{Y}} - \underset{\sim}{M})(\overline{\underset{\sim}{Y}} - \underset{\sim}{M})'$$

é a matriz soma de quadrados entre os grupos $J$ e $K$, e representa a "economia" de variabilidade que se faz numa população quando ela é particionada em duas.

*Definição 5:*

O $tr(\underset{\sim}{B})$,

$$tr(\underset{\sim}{B}) = \frac{n_1 n_2}{n_1 + n_2} (\overline{\underset{\sim}{X}} - \overline{\underset{\sim}{Y}})'(\overline{\underset{\sim}{X}} - \overline{\underset{\sim}{Y}}) \qquad (2.5)$$

é chamado de soma de quadrados entre os grupos.

O $tr(\underset{\sim}{B})$ representa a contribuição na soma de quadrados devido às diferenças entre os grupos $J$ e $K$, fornecendo, portanto, uma medida de heterogeneidade entre os grupos $J$ e $K$.

Os conceitos introduzidos formam a base da Análise de Variância, uma técnica estatística cujo objetivo é buscar evidências de que vários grupos conhecidos diferem entre si.

Muito embora existam diferenças básicas nos objetivos de Análise de Variância e Análise de Conglomerados, os dois métodos procuram explicar a variação dos dados através de uma estrutura de grupos. Daí, vários critérios de agrupamento foram desenvolvidos explorando-se a mesma conceituação da Análise de Variância.

## 2.2. DESCRIÇÃO DOS MÉTODOS

A expressão (2.4) dá uma idéia clara da dispersão das observações dentro e entre grupos. Conhecendo esta relação, é intuitivo que a criação de grupos homogêneos seja baseada na minimização de uma medida da dispersão dentro dos grupos. Ou, equivalentemente, maximização da diferenciação entre grupos.

### *2.2.1. Critério Invariante sob Transformação Ortogonal*

Maximização do $tr(\underset{\sim}{B})$

De (2.4) Edwards & Cavalli - Sforza (1965) derivaram a seguinte expressão:

$$tr(\underset{\sim}{T}) = tr(\underset{\sim}{B} + \underset{\sim}{W}) = tr(\underset{\sim}{B}) + tr(\underset{\sim}{W}).$$

Considerando que o $tr(\underset{\sim}{T})$ é constante para um conjunto fixado de observações, a proposta é particionar este conjunto de tal forma a miximizar o $tr(\underset{\sim}{B})$. Porém, computacionalmente é mais simples minimizar o $tr(\underset{\sim}{W})$.

$$tr\ (\underset{\sim}{W}) = tr\ (\underset{\sim}{W}_1) + tr\ (\underset{\sim}{W}_2)$$

$$tr\ (\underset{\sim}{W}_k) = \frac{1}{n_k} \sum\sum_{i<j} d^2\ (\underset{\sim}{x}_{ik}, \underset{\sim}{x}_{jk}),\ k = 1,2$$

onde $d^2(\underset{\sim}{x}_{ik}, \underset{\sim}{x}_{jk})$ é a distância Euclideana entre os pontos $\underset{\sim}{x}_{ik}$ e $\underset{\sim}{x}_{jk}$.

Toda a informação numérica necessária para este procedimento pode ser colocada na matriz triangular das distâncias entre os elementos, o que simplifica enormemente os cálculos. A aplicação deste procedimento evita o cálculo da matriz $\underset{\sim}{W}$ a cada partição, tornando este método muito atrativo.

Porém, embora o $tr\ (\underset{\sim}{W})$ seja uma medida invariante sob transformações ortogonais, não o é sob qualquer transformação linear não singular nos dados originais. Isso torna esta técnica indesejavelmente sensível, porque pode produzir partições diferentes para um mesmo conjunto de observações quando aplicada a dados diretos e a dados transformados.

### 2.2.2. *Critérios Invariantes sob Transformações Lineares Não Singulares*

Para contornar o problema da Invariância, Friedman & Rubin (1967) propuseram dois critérios extraídos da mesma igualdade básica (2.4).

Adicionando-se a suposição de que as $p$ variáveis medidas são linearmente independentes e, $n \geq p + 1$, tem-se que $\underset{\sim}{W}$ além de ser uma matriz simétrica, é também positiva definida, garantindo a existência da inversa $\underset{\sim}{W}^{-1}$.

Pós-multiplicando (2.4) por $\underset{\sim}{W}^{-1}$, obtém-se

$$\underset{\sim}{T}\underset{\sim}{W}^{-1} = \underset{\sim}{B}\ \underset{\sim}{W}^{-1} + \underset{\sim}{I} \qquad (2.6)$$

Desta nova expressão foram derivados mais 2 critér

rios a saber:

*i) Maximização do* $|\underset{\sim}{T}| \; / \; |\underset{\sim}{W}|$

De (2.6):

$$|\underset{\sim}{T}\ \underset{\sim}{W}^{-1}| = |\underset{\sim}{B}\ \underset{\sim}{W}^{-1} + I|$$

$$|\underset{\sim}{T}| \; / \; |\underset{\sim}{W}| = |\underset{\sim}{B}\ \underset{\sim}{W}^{-1} + I|$$

A maximização do $|\underset{\sim}{T}| \; / \; |\underset{\sim}{W}|$ se resume a minimizar $|\underset{\sim}{W}|$.

Este método exige uma complexidade de cálculo maior do que o anterior.

*ii) Maximização do* $tr\ (\underset{\sim}{B}\ \underset{\sim}{W}^{-1})$

Ainda de (2.6):

$$tr\ (\underset{\sim}{T}\ \underset{\sim}{W}^{-1}) = tr\ (\underset{\sim}{B}\ \underset{\sim}{W}^{-1}) + p$$

Como $\underset{\sim}{B}$ é uma matriz simétrica e $\underset{\sim}{W}$ é uma matriz simétrica positiva definida, (Singer (1977))

$$\frac{|\underset{\sim}{B} + \underset{\sim}{W}|}{|\underset{\sim}{W}|} = |\underset{\sim}{B}\ \underset{\sim}{W}^{-1} + I| = \prod_{i=1}^{p} (\lambda_i + 1)$$

$$tr(\underset{\sim}{B}\ \underset{\sim}{W}^{-1}) = \sum_{i=1}^{p} \lambda_i$$

onde os $\lambda_i$'s são as raízes obtidas resolvendo-se a equação dada por:

$$|\underset{\sim}{B} - \lambda\ \underset{\sim}{W}| = 0.$$

Pode-se demonstrar que as raízes desta equação são invariantes sob transformações lineares não singulares na matriz de dados originais (Anderson, (1958)). Portanto, embora os dois últimos critérios exijam cálculos mais complicados são mais abrangentes do que o primeiro.

*2.2.3. Procedimentos*

Quando o objetivo é particionar a população em apenas duas sub-populações, basta aplicar o critério escolhido, e determinar a melhor partição. Como existem $(2^{n-1} - 1)$ maneiras diferentes de se particionar o conjunto inicial, é importante dispor de métodos eficientes de efetuação de cálculos.

Em geral, o pesquisador interessado em criar uma tipologia espera encontrar mais que duas caracterizações para seus dados. Existem dois procedimentos diferentes quando se pretende particionar os dados em mais de dois grupos. Esta diferença, permite uma classificação grosseira das técnicas de A.C. em técnicas hierárquicas e técnicas de partição.

As técnicas hierárquicas divisivas consistem em particionar o conjunto inicial de valores em partições cada vez mais finas.

Isto é, o conjunto $I$ de observações é particionado segundo um dos critérios em dois conjuntos $I_1$ e $I_2$. Em seguida, cada um desses conjuntos é particionado em dois pelo mesmo critério. E, assim sucessivamente, até se obter o número de grupos desejados.

As técnicas de partição pressupõem um número $k$ de grupos, e maximizam a função escolhida sobre todas as possíveis partições dos $n$ elementos em $k$ grupos.

As técnicas de partição diferem basicamente das hierárquicas porque, obtida uma partição, admitem a reclassificação dos elementos, permitindo que uma partição inicialmente pobre seja corrigida num estágio posterior.

## 2.3. MODELO DE MÁXIMA VEROSSIMILHANÇA

Os métodos apresentados foram sugeridos para agru

par uma população, e não consideram a aleatoriedade contida em uma amostra. Não obstante, pode se mostrar que são casos particulares de um modelo de agrupamento com base na teoria estatística de verossimilhança.

Iniciando com um modelo geral de classificação e particularizando para o caso de não haver qualquer informação a priori sobre as categorias ou classes, Scott & Symons (1971) construíram um modelo estatístico de agrupamento.

Sejam $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ observações independentes de um vetor aleatório $\underset{\sim}{X}$ e provenientes de uma de $k$ populações normais $p$ - variadas, com médias $\underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_k$ e matriz de variâncias e covariâncias $\underset{\sim}{\Sigma}_1, \underset{\sim}{\Sigma}_2, \ldots, \underset{\sim}{\Sigma}_k$. Seja $\underset{\sim}{\gamma}$ o parâmetro de classificação, $\underset{\sim}{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)$, onde $\gamma_i = g$ quando a *i-ésima* observação pertencer a *g-ésima* sub-população.

A distribuição de probabilidades de $\underset{\sim}{X}$ ficará completamente especificada ao se conhecer o parâmetro $\underset{\sim}{\theta} = (\underset{\sim}{\gamma}, \underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_k, \underset{\sim}{\Sigma}_1, \ldots, \underset{\sim}{\Sigma}_k)$. O parâmetro de maior interesse é $\underset{\sim}{\gamma}$, uma vez que é ele quem aloca os elementos aos conglomerados $C_1, C_2, \ldots, C_k$. Portanto, o objetivo é estimar $\underset{\sim}{\gamma}$. Um método utilizado é o da máxima verossimilhança.

O logaritmo da função de verossimilhança da amostra é dado por:

$$l(\theta) = \frac{-1}{2} \sum_{g=1}^{k} \left[ \sum_{C_g} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_g)' \, \underset{\sim}{\Sigma}_g^{-1} \, (\underset{\sim}{x}_i - \mu_g) + n_g \log |\underset{\sim}{\Sigma}_g| \right] - \frac{n\,p}{2} \log 2\pi \qquad (2.7)$$

onde

$C_g$ é o conjunto dos $\underset{\sim}{x}$'s alocados ao *g-ésimo* grupo através de $\underset{\sim}{\gamma}$.

$n_g$ é o número de observações em $C_g$.

A estimativa de $\underset{\sim}{\theta}$ é obtida de forma a maximizar o valor da função $l(\underset{\sim}{\theta})$.

Para cada partição em $k$ grupos é possível a obtenção de estimativas de máxima verossimilhança (M.V.) de $\underset{\sim}{\mu}_g$ e $\underset{\sim}{\Sigma}_g$, $g = 1,2,\ldots,k$.

Para $\gamma$ fixado, tem-se:

$$\underset{\sim}{\mu}_g(\gamma) = \overline{\underset{\sim}{x}}_g$$

$$\hat{\underset{\sim}{\Sigma}}_g(\gamma) = \frac{1}{n_g} \sum_{i \in C_g} (\underset{\sim}{x}_i - \overline{\underset{\sim}{x}}_g)(\underset{\sim}{x}_i - \overline{\underset{\sim}{x}}_g)' = \frac{1}{n_g} \underset{\sim}{W}_g$$

$$l(\underset{\sim}{\theta}) = \frac{-1}{2} \sum_{g=1}^{k} \left[ tr\ n_g \underset{\sim}{I} + \log \left| \underset{\sim}{W}_g \right|^{n_g} \right] - \frac{np}{2} \log 2\pi$$

Assim, a estimativa de $\gamma$ é dada pela partição que minimiza $\prod_{g=1}^{k} \left| \underset{\sim}{W}_g \right|^{n_g}$.

Este é o caso mais geral, e exige um número muito grande de observações para que estas estimativas façam sentido.

São necessárias pelo menos $(p+1)$ observações em cada grupo para garantir a não singularidade da matriz de variâncias e covariâncias, e consequentemente, evitar pontos de descontinuidade na função de verossimilhança. É possível diminuir o número de parâmetros a serem estimados fazendo algumas suposições adicionais. Alguns casos particulares merecem destaque.

*i) Matrizes de variâncias e covariâncias iguais*

a) $\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \ldots = \underset{\sim}{\Sigma}_k = \underset{\sim}{\Sigma}$ ($\underset{\sim}{\Sigma}$ *desconhecida*)

Neste caso, para uma dada partição, $\hat{\underset{\sim}{\Sigma}} = \frac{1}{n} \underset{\sim}{W}$, onde

$$\underset{\sim}{W} = \sum_{g=1}^{k} \underset{\sim}{W}_g .$$

Substituindo $\underset{\sim}{\Sigma}$ por sua estimativa em (2.7):

$$l(\underset{\sim}{\theta}) = \frac{-1}{2}\left[np + log\left|\frac{1}{n}\ \underset{\sim}{W}\right|^{n}\right] - \frac{np}{2}\ log\ 2\pi,$$

e $\hat{\underset{\sim}{\gamma}}$ será dado pela partição que minimizar $|\underset{\sim}{W}|$, ou a variância generalizada dentro dos grupos.

Este critério coincide com aquele sugerido por Friedman & Rubin mostrado anteriormente na secção 2.2.

Nas situações em que a suposição sobre as matrizes de variâncias e covariâncias dos grupos é violada, Marriott (1971) mostrou, através de aplicações práticas, que esse critério tende a formar conglomerados similares em dispersão.

*Matrizes de variâncias e covariâncias iguais*

b) $\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \ldots = \underset{\sim}{\Sigma}_k = \underset{\sim}{\Sigma}$ ($\underset{\sim}{\Sigma}$ *conhecida*)

Para $\underset{\sim}{\gamma}$ fixado:

$$l(\hat{\underset{\sim}{\theta}}) = \frac{-1}{2}\left[tr(\underset{\sim}{W}\ \underset{\sim}{\Sigma}^{-1}) + log\left|\underset{\sim}{\Sigma}\right|^{n}\right] - \frac{np}{2}\ log\ 2\pi.$$

E, portanto, $\hat{\underset{\sim}{\gamma}}$ minimiza $tr(\underset{\sim}{W}\ \underset{\sim}{\Sigma}^{-1})$. Ou, equivalentemente, $\hat{\underset{\sim}{\gamma}}$ maximiza $tr(\underset{\sim}{B}\ \underset{\sim}{\Sigma}^{-1})$, onde

$$\underset{\sim}{B} = \sum_{g=1}^{k} n_g\ (\bar{\underset{\sim}{x}}_g - \bar{\underset{\sim}{x}})\ (\bar{\underset{\sim}{x}}_g - \bar{\underset{\sim}{x}})'$$

Portanto o critério da maximização do $tr\ (\underset{\sim}{B}\ \underset{\sim}{W}^{-1})$ apresentado por Friedman & Rubin na secção 2.2 se aplica nos casos em que as matrizes de variâncias e covariâncias dos grupos são iguais e conhecidas.

*Matrizes de variâncias e covariâncias iguais a* $\underset{\sim}{I}$

c) $\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \ldots = \underset{\sim}{\Sigma}_k = \underset{\sim}{I}$

Para uma partição fixada,

$$l(\hat{\underset{\sim}{\theta}}) = \frac{-1}{2}\ tr(\underset{\sim}{W}) - \frac{np}{2}\ log\ 2\pi$$

e $\hat{\gamma}$ será dado pela partição que minimizar o $tr(\underset{\sim}{W})$.

Como dantes, este critério já havia sido apresentado anteriormente por Edwards & Cavalli - Sforza (1965), na secção 2.1.

A tendência apresentada por esse método em formar conglomerados esféricos (Everitt, (1974)) fica explicada pela estrutura da matriz de variâncias e covariâncias dos conglomerados.

*ii) Matrizes de variâncias e covariâncias desiguais e conhecidas*

Esta é uma situação mais rara, e portante de pouco interesse.

Fixando-se $\underset{\sim}{\gamma}$,

$$l(\hat{\theta}) = \frac{-1}{2} \sum_{g=1}^{k} \left[ tr(\underset{\sim}{W}_g \underset{\sim}{\Sigma}_g^{-1}) + n_g \log|\underset{\sim}{\Sigma}_g| \right] - \frac{np}{2} \log 2\pi$$

e portanto $\hat{\gamma}$ deve minimizar

$$\sum_{g=1}^{k} \left[ tr\ \underset{\sim}{W}_g \underset{\sim}{\Sigma}_g^{-1} + n_g \log|\underset{\sim}{\Sigma}_g| \right]$$

O método baseado na razão de verossimilhança, além de unificar vários procedimentos num único modelo, elucida as situações em que cada uma das técnicas é aplicável, permitindo um resumo dos resultados encontrados nesse capítulo no seguinte roteiro:

i) Se as matrizes de variâncias e covariâncias dos grupos forem iguais, deve-se procurar a partição do conjunto de observações que minimiza o $|\underset{\sim}{W}|$.

ii) Se as matrizes de variâncias e covariâncias dos grupos forem iguais, e, além disso, conhecidas, então o critério a ser adotado deverá ser a minimização do $tr(\underset{\sim}{B}\ \underset{\sim}{W}^{-1})$.

iii) Se as variáveis forem não correlacionadas entre si, então o agrupamento deverá ser obtido pela minimização do $tr(\underset{\sim}{W})$. Este critério é largamente utilizado, principalmente em uma de suas versões, desenvolvida por Mac Queen (1967), num procedimento conhecido como "$k$ médias".

iv) Se as matrizes de variâncias e covariâncias forem diferentes entre si, e desconhecidas, tem-se o caso mais geral, e portanto, mais complicado. Os conglomerados devem ser obtidos minimizando-se o $\prod_{g=1}^{k} |\underset{\sim}{W}_g|^{n_g}$.

## 2.4. COMENTÁRIOS

As técnicas apresentadas consistem em determinar uma partição que maximiza (ou minimiza) uma função. Teoricamente, este é um problema de fácil solução. Entretanto, a quantidade de cálculo exigida torna estas técnicas impraticáveis.

Existem várias sugestões de procedimentos não exaustivos que, embora não produzam a melhor solução, produzem resultados satisfatórios. O mais comum deles consiste em fixar uma partição inicial em $k$ grupos. Em seguida, toma-se um elemento qualquer (em geral, aquele que estiver mais afastado do centro de gravidade do seu grupo) e transfere-se-o para todos os outros conglomerados. A cada transferência, o valor da função critério é calculado. Se nenhuma das novas partições mostrar uma otimização na função critério, o elemento permanecerá no conglomerado em que estava. Caso contrário, ele deverá ser transferido de forma a se obter o maior acréscimo (ou decréscimo) no valor da função critério. Usando a nova partição o segundo elemen

to é processado, e depois o terceiro, etc., até se alcançar um ponto em que nenhuma transferência otimize a função critério. Este é o ponto de máximo (ou mínimo) local da função.

A vantagem de se transferir um único elemento a cada passo é evitar que a matriz $\underset{\sim}{W}$ seja recalculada por completo. É possível calcular somente a alteração devida a essa única mudança.

Para se ter maior confiança nos resultados, recomenda-se repetir o procedimento com diferentes partições iniciais, e selecionando a que apresentar maior (ou menor) valor da função critério.

Evéritt (1974) fornece outros procedimentos adequados para a determinação de máximos (ou mínimos) locais.

# C A P Í T U L O  III

## MISTURA DE DUAS POPULAÇÕES MULTINORMAIS

Será apresentado aqui o caso mais simples de separar duas populações. O objetivo é familiarizar o leitor com a técnica de mistura de distribuições, e dar-lhe conhecimento dos problemas que surgem na sua aplicação.

Na secção 1 é colocado, de maneira geral, o conceito da mistura de distribuições como uma técnica na Análise de Conglomerados.

O modelo matemático, e a derivação dos estimadores dos parâmetros envolvidos são apresentados na secção 2.

Um teste de significância para os grupos encontrados é discutido na secção 3.

A secção 4 faz alguns comentários sobre esta técnica.

### 3.1. INTRODUÇÃO

Numa tentativa de dar à Análise de Conglomerados uma abordagem estatística mais rigorosa, Wolfe (1970) sugeriu um modelo baseado em mistura de distribuições. Porém, anteriormente, Day (1969) havia feito um estudo em mistura de duas populações multinormais, cujo trabalho será resumido aqui.

No modelo proposto, cada elemento da amostra sobre a qual será feita a análise, é admitida como selecionada aleatoriamente de uma entre duas populações multinormais. Dessa forma a amostra pode ser considerada como pro

veniente de uma população constituída ao acaso por duas populações diferentes, e portanto caracterizada por uma função distribuição que é a composta de duas diferentes distribuições de probabilidade. Criar dois grupos a partir dessa amostra é equivalente a separar uma mistura de distribuições caracterizando cada uma de suas componentes. Assim, a identificação dos conglomerados se transforma num problema de estimação dos parâmetros envolvidos numa mistura de distribuições.

O método de estimação utilizado é o da máxima verossimilhança porque os seus estimadores apresentam um comportamento superior aqueles derivados por outros métodos usuais de estimação, tais como, método dos momentos, Bayes, $\chi^2$ mínimo. O método dos momentos apresenta propriedades amostrais muito pobres em relação a variância dos seus estimadores, enquanto que os outros dois métodos mencionados acima apresentam cálculos muito complexos na sua aplicação (Day (1969)).

## 3.2. MODELO

### *2.1. Estimação*

Seja $\{\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n\}$ uma amostra aleatória retirada de uma população, *p-dimensional*, composta por duas populações multinormais e, cuja função densidade de probabilidade (f.d.p.) é dada por:

$$f(\underset{\sim}{x}) = (2\pi)^{-p/2} |\underset{\sim}{\Sigma}|^{p/2} \{\lambda \exp[-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)' \overset{-1}{\underset{\sim}{\Sigma}} (\underset{\sim}{x} - \underset{\sim}{\mu}_1)] +$$

$$+ (1 - \lambda) \exp[-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)' \overset{-1}{\underset{\sim}{\Sigma}} (\underset{\sim}{x} - \underset{\sim}{\mu}_2)]\} =$$

$$= \lambda g_1(\underset{\sim}{x}) + (1 - \lambda) g_2(\underset{\sim}{x})$$

onde:

$\lambda$ - é a proporção da primeira componente na mistura

$g_i$ - é a f.d.p. da população $\pi_i$, $i = 1,2$.

$\underset{\sim}{\mu}_i$ - é a média da $i$-ésima componente, $i = 1,2$.

$\underset{\sim}{\Sigma}$ - é a matriz de variâncias e covariâncias comum às duas componentes.

Define-se ainda a probabilidade de pertinência de um vetor a primeira componente da população como:

$$P[\pi_1 \mid \underset{\sim}{x}] = \frac{P[\pi_1]\ P[\underset{\sim}{x} \mid \pi_1]}{P[\underset{\sim}{x}]} = \frac{\lambda\ g_1(\underset{\sim}{x})}{f(\underset{\sim}{x})}$$

Analogamente para a segunda componente:

$$P[\pi_2 \mid \underset{\sim}{x}] = \frac{P[\pi_2]\ P[\underset{\sim}{x} \mid \pi_2]}{P[\underset{\sim}{x}]} = \frac{(1-\lambda)\ g_2(\underset{\sim}{x})}{f(\underset{\sim}{x})}$$

Na técnica de agrupamento esta medida desempenha um papel muito importante, uma vez que é ela quem sugere a distribuição dos elementos em um ou outro grupo. Para determinar esta medida, é necessário antes, conhecer as estimativas dos parâmetros populacionais.

A função de verossimilhança da amostra é dada por:

$$L(\underset{\sim}{\theta}) = (2\pi)^{-\frac{np}{2}} |\underset{\sim}{\Sigma}|^{\frac{np}{2}} \prod_{i=1}^{n} \{\lambda\ exp[-\frac{1}{2}(\underset{\sim}{x}_i - \underset{\sim}{\mu}_1)' \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_1)] +$$

$$+ (1-\lambda)\ exp[-\frac{1}{2}(\underset{\sim}{x}_i - \underset{\sim}{\mu}_2)' \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_2)]\}$$

onde $\underset{\sim}{\theta} = (\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \underset{\sim}{\Sigma}, \lambda)$.

O máximo de $L(\underset{\sim}{\theta})$ será atingido quando o sistema de equações abaixo for satisfeito:

$$\frac{\partial\ log\ L(\underset{\sim}{\theta})}{\partial\ \lambda} = \sum_{i=1}^{n} \frac{e_{1i} - e_{2i}}{\lambda\ e_{1i} + (1-\lambda)\ e_{2i}} = 0 \quad (3.1)$$

$$\frac{\partial\ log\ L(\underset{\sim}{\theta})}{\partial\ \underset{\sim}{\mu}_1} = \sum_{i=1}^{n} \frac{(x_i - \hat{\underset{\sim}{\mu}}_1)\ \lambda}{\lambda\ e_{1i} + (1-\lambda)\ e_{2i}} = \underset{\sim}{0} \quad (3.2)$$

$$\frac{\partial \log L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\mu}_2} = \sum_{i=1}^{n} \frac{(\underset{\sim}{x}_i - \underset{\sim}{\mu}_2)(1-\lambda)}{\lambda e_{1i} + (1-\lambda) e_{2i}} = \underset{\sim}{0} \quad (3.3)$$

$$- n \hat{\underset{\sim}{\Sigma}} + \sum_{i=1}^{n} \left[ (\underset{\sim}{x}_i - \underset{\sim}{\mu}_1)(\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_1)' \lambda e_{1i} + \right.$$

$$\left. + (\underset{\sim}{x}_i - \underset{\sim}{\mu}_2)(\underset{\sim}{x}_i - \underset{\sim}{\mu}_2)' (1-\lambda) e_{2i} \right]$$

$$\left[\lambda e_{1i} + (1-\lambda) e_{2i}\right]^{-1} = \underset{\sim}{0},$$

onde

$$e_{ji} = \exp\left[-\frac{1}{2}(\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_j)' \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_j)\right], i=1,\ldots,n, j=1,2$$

$$\hat{P}[\pi_1 | \underset{\sim}{x}_j] = \hat{\lambda} e_{1j} / \left[\hat{\lambda} e_{1j} + (1-\hat{\lambda}) e_{2j}\right]$$

$$\hat{P}[\pi_2 | \underset{\sim}{x}_j] = 1 - \hat{P}[1 | \underset{\sim}{x}_j].$$

Uma forma mais simples para o sistema acima é dada por:

$$\hat{\lambda} = \frac{1}{n} \sum_{j=1}^{n} \hat{P}[\pi_1 | \underset{\sim}{x}_j]$$

$$\hat{\underset{\sim}{\mu}}_1 = \frac{1}{n\hat{\lambda}} \sum_{j=1}^{n} \hat{P}[\pi_1 | \underset{\sim}{x}_j] \cdot \underset{\sim}{x}_j$$

$$\hat{\underset{\sim}{\mu}}_2 = \frac{1}{n(1-\lambda)} \sum_{j=1}^{n} \hat{P}[\pi_2 | \underset{\sim}{x}_j] \cdot \underset{\sim}{x}_j$$

$$\hat{\underset{\sim}{\Sigma}} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{2} (\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)' \hat{P}[\pi_i | \underset{\sim}{x}_j]$$

Outro parâmetro de interesse nesta técnica de agrupamento é a distância ponderada que separa as distribuições componentes da mistura. Ou seja, é a distância de Mahalanobis entre as médias das distribuições componentes,

dada por:

$$\Delta = \left[(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \, \underset{\sim}{\Sigma}^{-1} \, (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\right]^{\frac{1}{2}}$$

Pelo princípio de invariância que rege os estimadores de máxima verossimilhança, $\Delta$ pode ser estimado pelo método da máxima verossimilhança por:

$$\hat{\Delta} = \left|(\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)' \, \hat{\underset{\sim}{\Sigma}}^{-1} \, (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)\right|^{\frac{1}{2}}.$$

$\Delta$ é uma medida indicadora da heterogeneidade entre os grupos. Quanto maior o valor numérico de $\Delta$ mais diferenciadas são as duas populações de interesse.

### *3.2.2. Solução Numérica*

A solução das equações de M. V. obtidas de (3.1) a (3.3) é obtida através de processos iterativos, ou seja, processos que determinam uma solução a partir de um conjunto arbitrário de valores iniciais.

Entretanto, o sistema de equações pode ter mais de uma solução, e, neste caso, o processo converge para a solução mais próxima do conjunto inicial de valores, que, no caso em estudo, é um ponto de máximo local da função de verossimilhança. Numa tentativa de se atingir o máximo global, recomenda-se repetir o processo com diferentes estimativas iniciais, e, dentre essa multiplicidade de soluções, escolher a que der maior valor para $L(\underset{\sim}{\theta})$.

A convergência do processo também é afetada pela distância $\Delta$ definida acima. Parece intuitivo que o processo converge tão mais rapidamente quanto mais distanciadas estiverem os dois conglomerados.

$\Delta$, além de ser uma medida importante na solução numérica do problema, fornece também uma estatística importante para futuros testes de significância.

## 3.3. SIGNIFICÂNCIA DOS GRUPOS

Tendo identificado as duas componentes da mistura, e alocado os elementos em seus respectivos grupos, resta ainda saber se eles (os conglomerados) diferem significativamente. Isso pode ser colocado na forma de um teste de hipótese:

$H_0$ : existe uma única distribuição multinormal

$H_1$ : existe uma mistura de duas distribuições multinormais

Ou, equivalentemente

$H_0 : \Delta = 0$

$H_1 : \Delta \neq 0$

Ou ainda,

$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$

$H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$

No caso de estar se trabalhando com grupos definidos a priori, e sob a suposição de normalidade e matriz de variâncias e covariâncias constante, a estatística $T^2$ de Hotelling, dada por

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$

fornece um teste estatístico poderoso para as hipóteses formuladas acima, pois sob $H_0$,

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \sim F(p, n_1 + n_2 - p - 1).$$

Entretanto, na Análise de Conglomerados, os grupos são obtidos artificialmente, de forma a maximizar a função de verossimilhança. Portanto, a estatística não se

distribui mais segundo uma distribuição $F$, invalidando formalmente o teste.

Porém, essa estatística ainda pode fornecer um critério estatístico para se tomar uma decisão. Os valores observados de $T^2$ tenderão a ser maiores do que aqueles baseados em grupos pré-definidos. Se o valor observado da estatística for não significante, aceita-se $H_0$. Caso contrário, um estudo qualitativo mais cuidadoso deverá acompanhar a análise. Quão maior deverá ser o valor observado em relação ao crítico, para se aceitar que realmente existem duas populações, é um problema ainda a ser estudado.

## 3.4. COMENTÁRIOS

O conhecimento da distribuição de $\hat{\Delta}$ é de grande valia nas aplicações práticas. As conclusões obtidas por Day (1969) sobre o comportamento do estimador $\hat{\Delta}$ estão resumidas a seguir.

Para valores de $\Delta$ suficientemente grandes, a probabilidade de cada observação pertencer a uma das classes é próxima de *1*, e as estimativas dos parâmetros de cada componente são próximas das que seriam encontradas ao se trabalhar com amostras separadas de duas populações.

Além disso, sabe-se que os estimadores de máxima verossimilhança apresentam boas propriedades assintóticas. Entretanto, para valores de $\Delta$ próximos de *zero*, a amostra tem que ser muito grande para que a teoria assintótica seja aplicável.

Assim, para pequenas amostras e/ou pequenos valores de $\Delta$, o estimador $\hat{\Delta}$ pode não apresentar as características do que é usualmente um bom estimador.

Conjugando alguns resultados assintóticos com outros obtidos por simulação, Day (1969) chegou a alguns re

sultados interessantes. Abaixo mostrar-se-á o procedimento utilizado por ele, e algumas de suas conclusões.

Inicialmente, supondo $n \to \infty$, tem-se que a variância de $\hat{\Delta}$ é dada por $r(\Delta) / n$, onde $r(\Delta)$ é definida como:

$$r(\Delta) = \frac{1}{E_{\Delta}\{\left[\frac{\partial \log f(x)}{\partial \Delta}\right]^2\}}, \quad \Delta > 0 \qquad (4.1)$$

Day aproximou esta função por outras mais simples em dois casos distintos:

*i)* para valores altos de $\Delta$ $(\Delta > 1)$

$$r(\Delta) \simeq \frac{1}{\lambda(1-\lambda)} + \frac{\lambda(1-\lambda)\,\Delta^4}{1 + 2\lambda(1-\lambda)\,\Delta^2} \qquad (4.2)$$

*ii)* para valores relativamente baixos de $\Delta$ $(\Delta < 0,5$ e $\lambda \neq 0,1$ ou $0,5)$

$$[r(\Delta)]^{-1} \simeq \frac{3}{2}\,\lambda^2(1-\lambda)^2(1-2\lambda)^2\,\Delta^4 + o(\Delta^6) \qquad (4.3)$$

Em seguida, determinou um limite inferior para a tendenciosidade de $\hat{\Delta}$, $b(\Delta)$, através da desigualdade de Cramér-Rao:

$$Var(\hat{\Delta}) \geq r(\Delta)\left[1 + \frac{\partial b(\Delta)}{\partial \Delta}\right]^2 / n$$

$$\frac{\partial b(\Delta)}{\partial \Delta} \leq -1 + \{n[r(\Delta)]^{-1}\,Var(\hat{\Delta})\}^{\frac{1}{2}} \qquad (4.4)$$

Usando as aproximações (4.2) e (4.3) na relação (4.4) em amostras geradas artificialmente, Day mostrou que para pequenos valores de $\Delta$, $\hat{\Delta}$ é um estimador altamente tendencioso. Por exemplo, no caso univariado para $\Delta = 0$, $n = 500$ e $\lambda = 0,3$, a tendenciosidade é $1,43 \pm 0,15$.

Nos casos multivariados a situação é mais comple

xa. Em particular, para $n$ fixado e $\Delta = 0$, tem-se que a tendenciosidade cresce com a dimensão do espaço amostral. A tabela 3.1 mostra claramente esse fato. .

TABELA 3.1. Estimativas de Máxima Verossimilhança de $\Delta$ de uma população com $\Delta = 0$

| *Tamanho da Amostra* | *Dimensão do Espaço Amostral* | | | | |
|---|---|---|---|---|---|
| | *1* | *2* | *3* | *5* | *10* |
| 50 | 50<br>0.94–3.56<br>2.317 | 67<br>1.92–3.86<br>2.754 | 36<br>2.03–3.72<br>3.042 | 20<br>3.10–5.37<br>4.239 | 3<br>4.75–5.76<br>5.36 |
| 100 | 27<br>1.08–2.76<br>1.866 | — | — | 4<br>2.88–3.48<br>3.16 | 6<br>3.95–4.20<br>4.06 |
| 200 | — | 28<br>0.79–2.58<br>1.945 | — | — | — |
| 500 | 14<br>0.95–1.91<br>1.432 | 37<br>1.00–1.99<br>1.667 | 6<br>1.65–1.93<br>1.758 | — | — |

(Retirado de Biometrika (1969), 56, página 472)

O corpo da tabela fornece o número de amostras utilizadas, a amplitude dos valores de $\Delta$, e o valor médio de $\hat{\Delta}$.

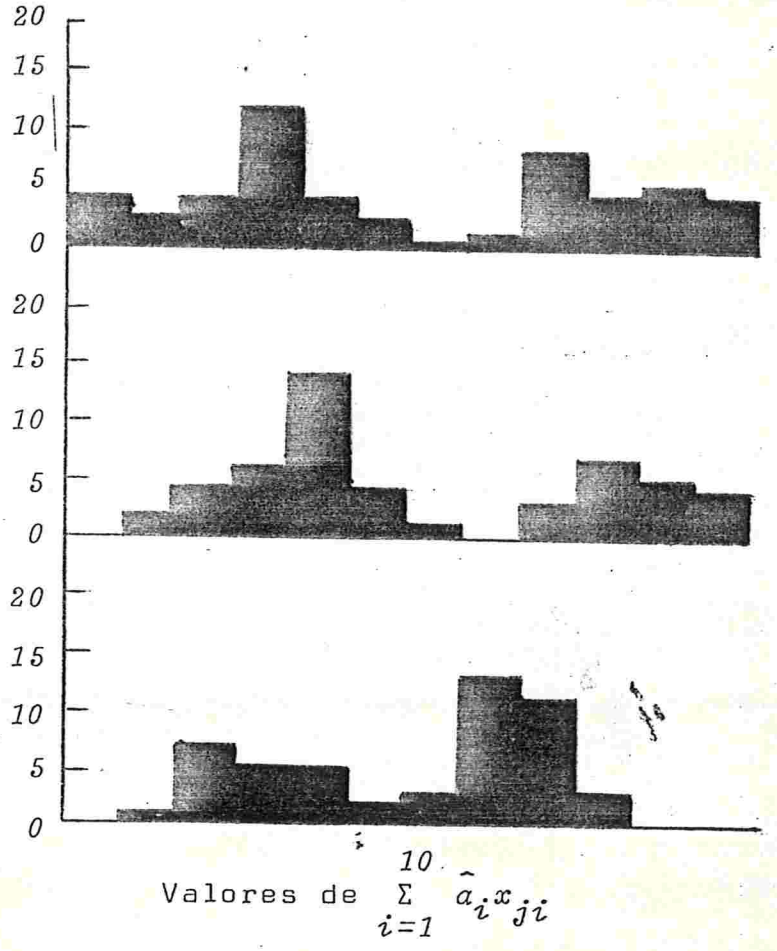
Esta tabela pode ser utilizada para se ter uma idéia do tamanho de amostra necessário, dada a dimensão do espaço amostral, para se resguardar contra um possível indício falso de agrupamento.

Os valores altos de $\hat{\Delta}$, quando $\Delta = 0$, apenas evidenciam uma bimodalidade casual na amostra, conforme está

mostrado nos gráficos abaixo. Isso indica que valores altos de $\hat{\Delta}$ não devem ser utilizados como prova conclusiva da existência de duas populações distintas.

Os gráficos são as projeções dos pontos amostrais $\underset{\sim}{x}_j$ $(j = 1,\ldots,50)$ para as três amostras da tabela 3.1, onde $k = 10$ e o valor médio de $\hat{\Delta}$ é $5.36$, sobre o eixo que liga as duas médias estimadas.

FIGURA 3.1. Histograma de $\sum_{i=1}^{10} \hat{a}_i x_{ji}$ de três amostras de 50 observações de uma distribuição normal multivariada de dimensão $p = 10$.




(Retirado de Biometrika (1969), 56, página 470)

O estudo da distribuição de $\hat{\Delta}$ mostra que este método é aplicável para as situações em que as populações estiverem bem separadas, ou em que a amostra seja muito grande. Caso contrário, as partições obtidas podem não ter qualquer significado.

# CAPÍTULO IV

## ALGORITMO PARA MISTURA DE MULTINORMAIS

Este capítulo consiste na generalização da técnica apresentada no capítulo anterior. Aqui serão estudadas as situações em que a população misturada é composta por mais de duas sub-populações, cada qual com uma estrutura de covariância.

As estimativas dos parâmetros envolvidos na mistura de distribuições estão na secção 1.

A secção 2 apresenta a descrição de um algoritmo para a solução das equações de máxima verossimilhança.

Os comentários sobre o algoritmo estão na secção 3.

### 4.1. MODELO

Sejam $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ observações do vetor aleatório $\underset{\sim}{X}$, de dimensão $p$, selecionados independentemente de uma população composta $\pi = \sum_{i=1}^{k} \lambda_i \pi_i$, com $\sum_{i=1}^{k} \lambda_i = 1$ e $0 < \lambda_i < 1$ $(i = 1, \ldots, k)$. Cada $\lambda_i$ indica a proporção da população do tipo $\pi_i$ na mistura. As populações $\pi_i$ têm distribuição normal *p-variada*, especificada pelo parâmetro $\underset{\sim}{\theta}_i = (\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma}_i)$. Assim, pode-se escrever:

$$g_i(\underset{\sim}{x} \mid \underset{\sim}{\theta}_i) = \frac{1}{(2\pi)^{p/2} \mid \underset{\sim}{\Sigma}_i \mid^{\frac{1}{2}}} \exp\left[\frac{-1}{2} (\underset{\sim}{x} - \underset{\sim}{\mu}_i)' \underset{\sim}{\Sigma}_i^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_i)\right] \quad (4.1)$$

e, consequentemente, a função densidade de probabilidade

do vetor $\underset{\sim}{X}$ será:

$$f(\underset{\sim}{x}, |\underset{\sim}{\theta}) = \sum_{i=1}^{k} \lambda_i \, g_i \, (\underset{\sim}{x}, |\underset{\sim}{\theta}_i) \qquad (4.2)$$

onde

$$\underset{\sim}{\theta} = (\lambda, \underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2, \ldots, \underset{\sim}{\theta}_k) \quad \text{e}$$

$$\lambda' = (\lambda_1, \lambda_2, \ldots, \lambda_k).$$

Analogamente ao caso de duas componentes, define-se a probabilidade de pertinência de um vetor $\underset{\sim}{x}$ a população do tipo $\pi_i$ como sendo

$$P[\pi_i \mid |\underset{\sim}{x}] = \frac{P[\pi_i] \cdot P[\underset{\sim}{x} \mid |\pi_i]}{P[\underset{\sim}{x}]} = \frac{\lambda_i \, g_i \, (\underset{\sim}{x}, |\underset{\sim}{\theta}_i)}{\Sigma \, \lambda_i \, g_i \, (\underset{\sim}{x}, \underset{\sim}{\theta}_i)} \qquad (4.3)$$

Além de alocar os elementos em seus respectivos conglomerados, esta medida tem uma função importante na solução numérica das equações de máxima verossimilhança, como será visto a seguir.

A função de verossimilhança é dada por:

$$L(\underset{\sim}{\theta}) = \prod_{j=1}^{n} \left[ \sum_{i=1}^{k} \lambda_i \, g_i \, (\underset{\sim}{x}_j, |\underset{\sim}{\theta}_i) \right] \qquad (4.4)$$

As estimativas de máxima verossimilhança são dadas pelos valores de $\lambda_i, \underset{k}{\theta}_i$ $(i = 1, \ldots, k)$ que maximizam $L(\underset{\sim}{\theta})$, com a restrição $\sum_{i=1}^{k} \lambda_i = 1$. Pelos procedimentos usuais, chegam-se aos seguintes resultados:

$$\hat{\lambda}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{P}[\pi_i \mid \underset{\sim}{x}_j] \qquad (4.5)$$

$$\hat{\underset{\sim}{\mu}}_i = \frac{1}{n \, \hat{\lambda}_i} \sum_{j=1}^{n} \hat{P}[\pi_i \mid \underset{\sim}{x}_j] \cdot \underset{\sim}{x}_j \qquad (4.6)$$

$$\hat{\underset{\sim}{\Sigma}}_i = \frac{1}{n \, \hat{\lambda}_i} \sum_{j=1}^{n} \hat{P}[\pi_i \mid \underset{\sim}{x}_j] \, (\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)' \qquad (4.7)$$

$$\hat{P}\left[\pi_i \mid \underset{\sim}{x}_j\right] = \hat{\lambda}_i \left|\hat{\underset{\sim}{\Sigma}}_i\right|^{-1/2} \gamma_j \exp\left[\frac{-1}{2}\left(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i\right)' \underset{\sim}{\Sigma}_i^{-1}\left(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i\right)\right] \qquad (4.8)$$

onde os $\gamma_j$ são determinados de tal forma que

$$\sum_{i=1}^{k} \hat{P}\left[\pi_i \mid \underset{\sim}{x}_j\right] = 1.$$

Este modelo tem um número muito grande de parâmetros a serem estimados, e que cresce com a dimensão $p$ do espaço amostral. É possível diminuir este número fazendo algumas restrições sobre a matriz de variâncias e covariâncias do mesmo modo já apresentado na secção 2.3. Alguns modelos reduzidos merecem destaque.

*i)* Matrizes de variâncias e covariâncias iguais

$$\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \ldots = \underset{\sim}{\Sigma}_k = \underset{\sim}{\Sigma} \qquad \text{(desconhecida)}$$

As equações de máxima verossimilhança fornecem as mesmas estimativas (4.5) e (4.6) para as proporções e médias cada componente da mistura. Porém, a matriz de variâncias e covariâncias passa a ser estimada por:

$$\hat{\underset{\sim}{\Sigma}} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} \left(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i\right)\left(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i\right)' P\left[\pi_i \mid \underset{\sim}{x}_j\right] \qquad (4.9)$$

*ii)* Matrizes de variâncias e covariâncias iguais e diagonais

$$\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \ldots = \underset{\sim}{\Sigma}_k = \underset{\sim}{\Delta}$$

Neste caso, as estimativas são idênticas às da situação *i)*, com a única restrição que os elementos fora da diagonal são postos iguais a zero.

*iii)* Matrizes de variâncias e covariâncias iguais a $\sigma^2 \underset{\sim}{I}$

$$\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \ldots = \underset{\sim}{\Sigma}_k = \sigma^2 \underset{\sim}{I}$$

Quando as variáveis observadas são independentes duas a duas, e tem a mesma variância, as estimativas dos parâmetros são dadas por:

$$\hat{\underset{\sim}{\mu}}_i = \frac{1}{n\,\hat{\lambda}_i} \sum_{j=1}^{n} \hat{P}\left[\pi_i \mid \underset{\sim}{x}_j\right] \underset{\sim}{x}_j$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)'(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)\, \hat{P}\left[\pi_i \mid \underset{\sim}{x}_j\right], \quad \hat{\underset{\sim}{\Sigma}} = \hat{\sigma}^2 \underset{\sim}{I}$$

$$\hat{\lambda}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{P}\left[\pi_i \mid \underset{\sim}{x}_j\right]$$

e, finalmente

$$\hat{P}\left[\pi_i \mid \underset{\sim}{x}_j\right] = \hat{\lambda}_i\, \gamma_j\, exp\left[-\frac{(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)'(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)}{2\sigma^2}\right]$$

onde os $\gamma_j$'s são determinados de forma a satisfazer

$$\sum_{i=1}^{k} \hat{P}\left[\pi_i \mid \underset{\sim}{x}_j\right] = 1$$

As estimativas das probabilidades de pertinência é que sugerem os agrupamentos. Cada elemento é alocado no grupo cuja probabilidade de pertinência é maior. Porém essas estimativas dependem das estimativas de $\underset{\sim}{\theta}$. Precisa-se, portanto, de um algorítmo para resolver as equações de máxima verossimilhança iterativamente.

## 4.2. DESCRIÇÃO DO ALGORITMO

Para determinar as estimativas dos parâmetros de

uma mistura de multinormais, Hartigan (1975) propôs um procedimento iterativo, que será descrito aqui. O procedimento é inicializado com uma partição do conjunto de observações em $k$ grupos aproximadamente do mesmo tamanho. Em seguida, obtém-se as estimativas das médias, das matrizes de variâncias e covariâncias e das proporções na mistura através das equações (4.5) a (4.7). Calcula-se o valor do logaritmo da função de verossimilhança. Finalmente são calculadas as estimativas das probabilidades de pertinência pela fórmula (4.8), e reestimados os parâmetros. Este ciclo é repetido até que o processo convirja. A regra de parada sugerida é terminar o procedimento quando, de um ciclo para outro, o logaritmo da função de verossimilhança tiver um acréscimo menor do que $0,01$.

Este algoritmo permite a introdução das seguintes estruturas para as matrizes de variâncias e covariâncias das populações componentes:

*i)* Matrizes de variâncias e covariâncias desiguais e arbitrárias;

*ii)* Matrizes de variâncias e covariâncias iguais;

*iii)* Matrizes de variâncias e covariâncias iguais e diagonais;

*iv)* Matrizes de variâncias e covariâncias iguais e do tipo $\sigma^2 \underset{\sim}{I}$.

Os passos do algoritmo são dados por:

1) Inicializar as probabilidades de pertinência $P[\pi_i \mid \underset{\sim}{x}_j]$ da seguinte maneira:

$$P[\pi_1 \mid \underset{\sim}{x}_1] = P[\pi_1 \mid \underset{\sim}{x}_2] = \ldots = P[\pi_1 \mid \underset{\sim}{x}_{j(1)}] = 1,$$

$$P[\pi_2 \mid \underset{\sim}{x}_{j(1)+1}] = P[\pi_2 \mid \underset{\sim}{x}_{j(1)+2}] = \ldots = P[\pi_2 \mid \underset{\sim}{x}_{j(2)}] = 1$$

$$\vdots$$

$$P[\pi_k \mid \underset{\sim}{x}_{j(k-1)+1}] = P[\pi_k \mid \underset{\sim}{x}_{j(k-1)+2}] = \ldots = P[\pi_k \mid \underset{\sim}{x}_{j(k)}] = 1,$$

onde $j(i)$ é o maior inteiro contido em $i*n/k$.

2) Atualizar as proporções na mistura

$$\hat{\lambda}_i = \sum_{j=1}^{n} \hat{P}[\pi_i \mid \underset{\sim}{x}_j]$$

3) Atualizar as médias por

$$\hat{\mu}_{(i)} = \sum_{j=1}^{n} \frac{\underset{\sim}{x}_j \hat{P}[\pi_i \mid \underset{\sim}{x}_j]}{n\hat{\lambda}_i}$$

4) Atualizar as matrizes de variâncias e covariâncias por

$$\sigma_r(i) = \sum_{j=1}^{n} \frac{(x_{jr} - \hat{\mu}_r(i))(x_{js} - \hat{\mu}_s(i))\, \hat{P}[\pi_i \mid \underset{\sim}{x}_j]}{\hat{\lambda}_i}$$

onde $x_{jr}$ - representa o valor da *r-ésima* variável no *j-ésimo* elemento amostral.

$u_r(i)$- representa a média da *r-ésima* variável no *i-ésimo* grupo.

O procedimento adotado a seguir depende das estruturas de covariâncias das populações componentes, conforme foi dito anteriormente. Assim,

sob opção i) passar para o passo 5.

sob opções ii) - iv), para $i$, $1 \leq i \leq k$

$$\hat{\sigma}_{rs}(i) = \frac{1}{n} \sum_{i=1}^{k} \hat{\sigma}_{rs}(i) \cdot \hat{\lambda}_i$$

sob opção iii),

$$\tilde{\sigma}_{rs}(i) = 0 \quad \text{para} \quad r \neq s, \quad 1 \leq i \leq k$$

sob opção iv),

$$\hat{\sigma}_{rr}(i) = \frac{1}{p} \sum_{r=1}^{p} \hat{\sigma}_{rr}(1), \quad \text{para} \quad 1 \leq r \leq p, \quad 1 \leq i \leq k$$

5) Calcular os determinantes e as inversas das matrizes de variâncias e covariâncias para cada grupo; os valores das funções densidades de probabilidade das populações componentes para cada observação; o valor da função densidade da mistura para cada observação e o logaritmo da função de verossimilhança.

6) Atualizar as probabilidades de pertinência através dos valores obtidos em 5).

7) Se o valor do logaritmo da função de verossimilhança não exceder seu valor prévio por mais de $0,01$, parar o procedimento. Caso contrário, retornar a 2).

## 4.3. COMENTÁRIOS

Este modelo, como será visto a seguir, guarda uma grande semelhança com o proposto por Scott & Symons (1971). Naquele, a função de verossimilhança $L(\underset{\sim}{\theta})$, pode ser escrita como:

$$L(\underset{\sim}{\theta}) = (2\pi)^{\frac{-np}{2}} \prod_{i=1}^{n} \left\{ \sum_{i=1}^{k} \delta_{ig} \left[\underset{\sim}{\Sigma}_{g}\right]^{n_i} exp\left[-\frac{1}{2}(\underset{\sim}{x}_i - \underset{\sim}{\mu}_g)' \underset{\sim}{\Sigma}_{i}^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_g)\right]\right\} \qquad (4.10)$$

onde

$$\delta_{ig} = \begin{cases} 1 & \text{se} \quad \gamma_i = g \\ 0 & \text{caso contrário} \end{cases}$$

Portanto, basta definir o vetor $\underset{\sim}{\delta}_i$, $i = 1,\ldots,n$, de $0's$ e $1$, da seguinte forma:

$\underset{\sim}{\delta}_i' = (0,0,0,\ldots,1,0,\ldots,0)$, onde o $1$ está na $g$-$ésima$ posição.

Se agora, adicionar-se a suposição de que os $\delta_{ig}'s$ são variáveis aleatórias independentes e identicamente distribuídas sobre as $n$ observações com

$$P[\delta_{ig} = 1] = \lambda_g \quad \text{e} \quad \sum_{g=1}^{k} \lambda_g = 1$$

tem-se que os $\delta_i's$, $i = 1,\ldots,n$, são $n$ ensaios independentes de uma distribuição multinomial. $\underset{\sim}{\gamma}$ se transforma, então, em uma variável aleatória não observável, cujos componentes são os resultados de ensaios independentes de uma distribuição multinomial.

Com estas suposições adicionais, a função de verossimilhança $L(\underset{\sim}{\theta})$ dada por (4.10) pode ser escrita como:

$$L(\underset{\sim}{\theta}) = (2\pi)^{\frac{-np}{2}} \left\{ \prod_{j=1}^{n} \sum_{g=1}^{k} \lambda_g |\underset{\sim}{\Sigma}_g|^{-n_g} \exp\left[\frac{-1}{2} (\underset{\sim}{x}_j - \underset{\sim}{\mu}_g)' \underset{\sim}{\Sigma}_g^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}_g)\right]\right\}$$

que é exatamente a função de verossimilhança dada pelo modelo de mistura de multinormais.

A distribuição condicional de $\delta_{ig}$, $g = 1,\ldots,k$, dada a observação $\underset{\sim}{x}_i$, é que fornecerá a probabilidade de pertinência, $P[\pi_g \mid \underset{\sim}{x}_i]$, tão importante na solução iterativa descrita na secção anterior.

No algoritmo descrito, a partição inicial do conjunto de observações é feita através das probabilidades de pertinência. Porém, diferentes inicializações podem produzir diferentes soluções. A menos que os conglomerados es

tejam muito bem definidos, a solução iterativa das equações de máxima verossimilhança é sensível às partições iniciais dos dados. Para contornar esse problema, recomenda-se obter várias soluções. O procedimento proposto por Hartigan permite variações nas partições iniciais, bastando para isso alterar a ordem de entrada das observações. De posse dessas múltiplas soluções, esta técnica permite ainda um critério estatístico para escolher a melhor delas. Como o método é baseado na maximização da função de verossimilhança, basta optar pela solução que fornecer maior valor para $L(\underset{\sim}{\theta})$.

Com sorte, pode-se obter uma solução em que as probabilidades de pertinência dos elementos em um dos conglomerados é bastante alta em relação aos demais, e assim, o agrupamento fica bem definido. Se, no entanto, as probabilidades de pertinência de um elemento em vários conglomerados forem próximas entre si, a Análise de Conglomerados deve ser seguida de uma análise qualitativa mais cuidadosa. Isto é, antes de simplesmente alocar um elemento ao grupo cuja probabilidade de pertinência for a mais alta, é recomendável analisar os vários grupos prováveis de conter essa observação, e só então tomar uma decisão.

Esta técnica requer um número grande de observações devido ao número de parâmetros que estima. No caso mais geral são $\frac{1}{2}(p+1)(p+2)(K-1)$ parâmetros, e Hartigan (1975), então, cautelosamente recomenda que,

$$n \geq \frac{1}{2}(p+1)(p+2)(K-1). \qquad (4.11)$$

Além disso, o número mínimo de elementos por grupo deve ser $(p+1)$, de forma a garantir que as matrizes de variâncias e covariâncias sejam não singulares (com probabilidade 1). Quando o número de observações não é suficientemente grande, uma estratégia é impor algumas restri

ções sobre as matrizes de variâncias e covariâncias. Isso diminui o número de parâmetros a estimar, e consequentemente, o limite imposto por (4.11) também é diminuído. É bom lembrar, porém, que os modelos simplificados devem ser adotados quando existirem razões para justificar esse procedimento.

# CAPÍTULO V

## ESCOLHA DO NÚMERO DE GRUPOS

Todos os métodos vistos até agora procuram a melhor partição de um conjunto de observações em $k$ grupos. Porém, são raros os casos em que o valor de $k$ é conhecido. Este fato gera um dos mais graves problemas em Análise de Conglomerados que é a determinação do número de grupos. Neste capítulo tratar-se-á desse problema.

A secção 1 faz uma rápida introdução do problema.

Na secção 2 são descritos métodos da determinação do número de grupos para as técnicas baseadas na minimização da variância.

Os métodos ligados à mistura de distribuições multinormais estão na secção 3.

### 5.1. INTRODUÇÃO

Nos raros casos em que se aplica a Análise de Conglomerados e se tem a priori a informação do número aproximado de grupos, a análise fica bastante simplificada. A configuração final do agrupamento é obtida através da aplicação de uma das técnicas apresentadas. Se a técnica for aplicada hierarquicamente, basta interromper o processo divisivo ao se atingir os $k$ grupos desejados. Se a técnica for aplicada por partição obtém-se diretamente os $k$ grupos.

Nas situações mais comuns em que se desconhece por completo a estrutura dos dados, além de escolher um critério de agrupamento e uma técnica de aplicação, resta ainda a determinação do número ideal de grupos. Esta ques

tão é colocada da seguinte forma:

i) em que estágio se deve interromper o processo divisivo (se a técnica for hierárquica).

ii) qual é o número $k$ de grupos em que deve ser particionado o conjunto de observações (se o agrupamento for obtido por uma técnica de partição).

Nestes casos é necessário definir alguns critérios que auxiliem a tomada de decisão. Existem vários deles, e na sua maioria são fundamentados em princípios heurísticos. Com base nesses critérios surgiram vários métodos de determinação do número de grupos, que costumam produzir resultados fortemente baseados na estrutura dos dados, porém são válidos na medida em que está se fazendo uma análise exploratória e não se dispõe de recursos maiores.

São poucos os procedimentos que contam com a teoria de Inferência Estatística para ajudar na questão. A quase inexistência de métodos estatísticos que tratem desse assunto talvez se deva ao fato do assunto exigir o estudo de distribuições bastante complexas, como se verá nas secções 2 e 3.

Apesar disso, alguns testes estatísticos de comparação de várias médias são utilizados de forma a fornecer algumas indicações sobre a significância dos grupos. As hipóteses são testadas de maneira usual, porém as interpretações devem ser feitas de maneira mais cautelosa. Na secção seguinte esta situação será estudada mais detalhadamente.

Os mesmos critérios usados para agrupar os dados costumam ser utilizados na determinação do número ideal de

conglomerados. Por essa razão as descrições dos procedimentos referentes aos métodos da minimização da variância estão na secção 2, e aqueles que se relacionam com mistura de distribuições multinormais estão na secção 3.

Antes de passar à descrição dos procedimentos, será definido um método para a escolha do número $k$ de grupos. Chama-se de procedimento hierárquico para determinação do número de grupos, ao procedimento em que a escolha do número $k$ é feita analisando, em sequência, as partições em $1,2,\ldots,g$ grupos. Note que esse método pode ser aplicado independentemente da forma como se obtém as partições.

## 5.2. REGRAS PARA O CAPÍTULO II

### *5.2.1. Procedimentos Baseados na ANOVA*

Na Análise de Variância, supondo que as observações são independentes e identicamente distribuídas, com matriz de variâncias e covariâncias $\Sigma$, a significância dos grupos pode ser verificada através das seguintes hipóteses:

$$H_0 : \textit{existe uma única população}$$
$$H_1 : \textit{existe mais de uma população} \qquad (5.1)$$

Wilks, conforme descreve Morrison (1976), desenvolveu o teste da razão de verossimilhança para as hipóteses acima. Sob $H_0$, a teoria assintótica da razão de verossimilhança implica que a estatística

$$-\left[n - k - \frac{1}{2}(p - k + 2)\right] \ln \Lambda, \qquad (5.2)$$

onde $\Lambda = |\underset{\sim}{W}| \; / \; |\underset{\sim}{T}|$, converge para uma distribuição de $\chi^2$ com $p(k-1)$ graus de liberdade.

No modelo apresentado por Scott & Symons, sob a restrição de que as matrizes de variâncias e covariâncias

dentro dos grupos são iguais, as hipóteses (5.1) podem ser colocadas na forma abaixo:

$$H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_n$$
$$H_1 : \gamma_i \neq \gamma_j, \text{ algum } i \neq j, \; \begin{matrix} i = 1,\ldots,n \\ j = 1,\ldots,n \end{matrix} \qquad (5.3)$$

Entretanto, em Análise de Conglomerados, os grupos são construídos de forma a maximizar o valor de $|\underset{\sim}{T}| \; / \; |\underset{\sim}{W}|$ sobre todas as possíveis partições de $n$ elementos em $k$ grupos. Portanto, o valor obtido para $|\underset{\sim}{T}| \; / \; |\underset{\sim}{W}|$ será, em geral, maior do que aquele observado para grupos conhecidos. Para se estudar a significância dos grupos encontrados pela Análise de Conglomerados, deveria então se fazer um estudo da distribuição do $max \; |\underset{\sim}{T}| \; / \; |\underset{\sim}{W}|$. Apesar de não poder ser utilizado como esse teste estatístico, o valor do $|\underset{\sim}{T}| \; / \; |\underset{\sim}{W}|$, sob $H_0$, ainda pode ser usado como um indicador da existência de grupos. Para isso procede-se como se fosse fazer o teste das hipóteses (5.1). Se o valor observado da estatística for menor do que o valor crítico dado por uma tabela de $\chi^2$ é indicação clara de que os dados não devem ser agrupados.

Porém, para a hipótese $H_0$ ser rejeitada, é intuitivo que o valor observado da estatística não somente deva ser maior do que o valor crítico fornecido pela distribuição de $\chi^2$, mas sim algumas vezes maior. Entretanto, é desconhecida a existência de trabalhos que investiguem acerca de quantas vezes maior deve ser o valor observado da estatística em relação ao $\chi^2$ crítico para que se possa aceitar a existência de uma estrutura de grupos nos dados. Assim, para os casos em que a estatística assume um valor muito alto em relação ao valor crítico do $\chi^2$, $H_0$ deve ser rejeitada. A situação se torna mais delicada quando, embora o valor observado da estatística seja maior do que o $\chi^2$ crítico, ele não é relativamente tão alto, e neste caso não

se dispõe então de qualquer indicação.

Um procedimento análogo pode ser utilizado empregando-se a estatística $T_0^2 = tr\ \underset{\sim}{B}\ \underset{\sim}{W}^{-1}$, chamada de traço de Lawley-Hotelling. Sob a hipótese de que os grupos provêm de uma única distribuição multinormal, tem-se que:

$$n\ T_0^2 \xrightarrow[n \to \infty]{} \chi^2$$

com $p(k-1)$ graus de liberdade (Morrison (1976)). As mesmas considerações tecidas para $|\underset{\sim}{T}| \ / \ |\underset{\sim}{W}|$ se aplicam aqui.

Para o caso particular em que $p=1$ e $k=2$, as hipóteses (5.1) podem ser colocadas da seguinte forma:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

onde $\mu_i$ é a média do *i-ésimo* grupo, $i = 1,2$.

Se os grupos forem construídos de modo a minimizar a variância dentro deles, Engelman & Hartigan, através de simulação, construiram uma tabela da distribuição do *max SSB/SSW* para testar as hipóteses acima.

A estatística para esse teste é dada por *SSB/SSW* onde

$$SSB = \frac{(x_1 - \bar{x}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$SSW = \sum_{C_g} (x_{ig} - \bar{x}_g)^2$$

onde:

$\bar{x}_g$ - é a média do *g-ésimo* grupo, $g = 1,2$

$n_g$ - é o número de elementos do *g-ésimo* grupo $g = 1,2$

$C_g$ - conjunto dos $x_i$'s alocados ao *g-ésimo* grupo, $g = 1,2$.

Os dois métodos acima somente dão uma idéia da existência dos grupos, mas não do número deles. Hartigan (1975) propõe uma regra, baseada em uma única variável, que auxilia a decisão de se escolher entre $k$ e $(k+1)$ grupos.

Seja a partição $P_k$ dos $n$ elementos em $k$ conglomerados, com matriz de dispersão dentro dos conglomerados dada por

$$\underset{\sim}{W}(k) = \left[w_{ij}(k)\right] = \sum_{g=1}^{k} \sum_{C_g} (\underset{\sim}{x}_{ig} - \overline{\underset{\sim}{x}}_g)(\underset{\sim}{x}_{ig} - \overline{\underset{\sim}{x}}_g)' \quad (5.4)$$

onde

$C_g$ é o conjunto dos $\underset{\sim}{x}_i$'s alocados ao *g-ésimo* grupo.

A dispersão da *j-ésima* variável é dada por:

$$w_{jj}(k) = \sum_{g=1}^{k} \sum_{C_g} (x_{ig} - \overline{x}_g)^2$$

$x_{ig}$ - é a medida da *j-ésima* variável no *i-ésimo* elemento do *g-ésimo* grupo.

$\overline{x}_g$ - é a média da *j-ésima* variável no *g-ésimo* grupo.

Sob as suposições de normalidade e matrizes de variâncias e covariâncias iguais, $\sigma_j^{-2}\, w_{jj}(k)$ tem distribuição de $\chi^2$ com $(n-k)$ graus de liberdade.

Similarmente para a partição $P_{k+1}$ dos $n$ elementos em $(k+1)$ conglomerados, $\sigma_j^{-2}\, w_{jj}(k+1)$ tem distribuição de $\chi^2$ com $(n-k-1)$ graus de liberdade.

Assim, pode-se fazer o teste:

$$H_0(k) : \text{existem } k \text{ populações}$$
$$H_1(k+1) : \text{existem } (k+1) \text{ populações} \qquad (5.5)$$

utilizando-se a estatística $R$,

$$R = \left( \frac{w_{jj}(k)}{w_{jj}(k+1)} - 1 \right) (n-k-1) \qquad (5.6)$$

Se $P_{k+1}$ for obtido particionando-se um dos conglomerados de $P_k$ em dois, então $R$ se distribui segundo uma $F$ com $1$ e $(n-k-1)$ graus de liberdade. A razão $R$ fornece uma medida da redução da variância dentro dos grupos ao se passar de $k$ para $(k+1)$ grupos, quando se considera tão somente a *j-ésima* variável.

Também aqui não se dispõe de um teste estatístico. Além dos grupos serem criados aritificialmente maximizando a disparidade entre os grupos, este método não considera a influência das outras variáveis existentes. Ainda assim, para valores bastante altos de $R$ é justificável o aumento do número de grupos para $(k+1)$. (Hartigan sugere $R > 10$). Uma vez que este indicador considera uma única variável, um procedimento a ser utilizado é o seguinte: aplica-se o teste sobre cada uma das variáveis observadas, e, aceita-se a partição com $(k+1)$ grupos, quando todas as razões $R_i$, $i = 1,2,\ldots,p$ forem bastante altas.

Mas este é um procedimento um tanto rígido. É comum, ao se passar de $k$ para $(k+1)$ grupos, que uma das razões $R_i$, $i = 1,2,\ldots,p$, seja muito alta, contrapondo-se a outra bastante baixa. Para contornar esse problema, procurou-se por um critério que leve em conta as $p$ variáveis conjuntamente.

Assim, se além da normalidade for possível supor que as matrizes de variâncias e covariâncias dentro dos grupos são iguais a $\sigma^2 \underset{\sim}{I}$, isto é, as variáveis observadas

são não correlacionadas, pode-se deduzir um "teste" que considera todas as variáveis simultaneamente.

Seja:

$$\underset{\sim}{W}(k) = [w_{ij}(k)] = \sum_{g=1}^{k} (\underset{\sim}{x}_{ig} - \overline{\underset{\sim}{x}})(\underset{\sim}{x}_{ig} - \overline{\underset{\sim}{x}}_{g})' \qquad (5.7)$$

e consequentemente:

$$tr(\underset{\sim}{W}(k)) = \sum_{j=1}^{p} w_{jj}(k).$$

Como as variáveis são independentes, $tr(\underset{\sim}{W}(k))$ tem distribuição proporcional ao $\chi^2$ com $p$ graus de liberdade.

Analogamente para $(k+1)$ grupos, $tr(\underset{\sim}{W}(k+1))$ tem distribuição proporcional ao $\chi^2$ com $p$ graus de liberdade.

Assim,

$$R = \left( \frac{tr(\underset{\sim}{W}(k))}{tr(\underset{\sim}{W}(k+1))} - 1 \right) (n-k-1) \qquad (5.8)$$

tem distribuição $F$ com $p$ e $p(n-k-1)$ graus de liberdade.

Da mesma forma que o teste precedente, valores bastante altos de $R$ justificam o aumento do número de grupos. (Hartigan novamente sugere $R > 10$.)

Apesar disso, convém ainda analisar as variáveis individualmente.

### 5.2.2. *Análise Gráfica*

Um método bastante comum e eficiente na escolha do número de grupos é a análise gráfica.

Geralmente os agrupamentos são obtidos de forma a maximizar uma função, então a idéia é verificar o ganho dessa função ao se passar de $k$ para $(k+1)$ grupos. Assim, obtém-se as configurações para $1, 2, \ldots, g$ grupos, por qualquer tipo de técnica, e os correspondentes valores do critério usado. Em seguida constrói-se o gráfico do número

de grupos contra a função critério. Dessa forma tem-se uma visualização do ganho (ou perda) da função que está sendo otimizada, conforme se aumenta o número de grupos. Decide-se por $k (k \leq g)$ grupos quando o ganho (ou perda) ao se passar de $k$ para $(k+1)$ grupos for pequeno em relação aos demais.

Friedman & Rubin (1967) sugeriram usar o logaritmo do máximo $|T| / |W|$ em função do número de grupos. Mas a utilização de gráficos na escolha do número $k$ é um método geral, e permite que se utilize qualquer função usada no agrupamento.

Este método é bastante difundido por ser simples, rápido e aplicável nos casos mais gerais.

## 5.3. REGRAS PARA O CAPÍTULO IV

### *5.3.1. Limite Máximo para o Número de Grupos*

Como foi visto no capítulo anterior, esta técnica requer que se estime um número muito grande de parâmetros. E que, embora para a existência desses estimadores, necessite-se de uma relação diferente, Hartigan (1975) sugere que:

$$n \geq \left[ \frac{(p+1)(p+2)}{2} \right] k - 1$$

Dessa forma, para uma amostra de tamanho $n$ tem-se que

$$k \leq \frac{2(n+1)}{(p+1)(p+2)} \quad . \qquad (5.9)$$

É fácil ver que esse limite é menor quanto maior for o número de variáveis $p$. A fim de se obter um limite confortável para $k$ é necessário ou um número excessivo de observações, ou um número menor de parâmetros a estimar.

Como nem sempre é possível obter amostras maiores, a maneira de resolver esse problema é impor restrições sobre as matrizes de variâncias e covariâncias dos grupos, reduzindo o número de parâmetros a serem estimados. Porém, como já foi dito, esta é uma estratégia que deve ser usada somente quando existem motivos para isso.

### 5.3.2. *Análise Gráfica*

Também neste caso se recomenda o uso de um gráfico para ajudar na escolha do número de grupos. Aqui, a função critério é dada pela função de verossimilhança, e o gráfico pode ser feito usando-se o seu logaritmo.

Por este método é fácil detectar pontos onde o ganho da função de máxima verossimilhança é muito pequeno para justificar o acréscimo do número de grupos.

### 5.3.3. *Teste de Significância*

Wolfe (1970) construiu o teste da razão de verossimilhança para testar a hipótese nula

$$H_0(k): \text{o número de grupos é igual a } k \tag{5.10}$$

contra a hipótese alternativa

$$H_1(k): \text{o número de grupos é igual a } k' \ (k' > k)$$

e sugeriu utilizar a estatística

$$T_n = \frac{-2}{n}\left(n-1-p-\frac{k}{2}\right) \log Q \tag{5.11}$$

onde $Q$ é a razão de verossimilhança.

Ele mostrou (Everitt (1974)) através de simulação, que a estatística modificada $T_n$, converge mais rapida

mente para a distribuição assintótica de $X^2$ com $2p(k'-k)$ graus de liberdade, do que a tradicional $-2\ log\ Q$.

No caso particular de mistura de distribuições multinormais,

$$Q = \prod_{j=1}^{n} \frac{\sum_{i=1}^{k} \hat{\lambda}_i |\hat{\Sigma}_i|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(x_j - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} (x_j - \hat{\mu}_i)\right]}{\sum_{i=1}^{k} \hat{\lambda}_i |\hat{\Sigma}_i|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(x_j - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} (x_j - \hat{\mu}_i)\right]} \quad (5.12)$$

Este teste pode ser usado hierarquicamente para auxiliar na escolha do número de grupos.

Inicia-se o procedimento testando-se $H_0(1)$ versus $H_1(2)$, caso $H_0$ seja rejeitada, passa-se a $H_0(2)$ versus $H_1(3)$, e assim sucessivamente até encontrar-se uma hipótese não significante.

Como não se conhece de antemão, o número de testes a serem realizados, e a fim de garantir um nível de significância geral não superior a um nível pré-fixado $\alpha$, recomenda-se um procedimento semelhante àquele utilizado em comparações múltiplas, e baseado na desigualdade de Bonferroni (Morrison (1976)). A cada hipótese testada corrige-se o nível de significância $\alpha$ para $\alpha/r$, onde $r$ é o número de hipóteses testadas $+1$.

Além disso, retestam-se todas as hipóteses anteriores ao nível de significância corrigido. Se qualquer uma das hipóteses for não significante, no estágio do teste $H_0(k)$ contra $H_1(k+1)$, pára-se o processo e adota-se $k$ como número final de grupos. Caso contrário, continua-se o procedimento até encontrar um valor não significante.

## 5.4. COMENTÁRIOS

Os procedimentos aqui descritos não exaurem a ga

ma de sugestões para a escolha do valor de $k$, número de grupos. Entretanto, foram escolhidos por serem derivados de métodos estatísticos.

O método devido a Hartigan, descrito na secção 2.1, pode também ser enquadrado como um método hierárquico para a determinação do número de grupos, se for aplicado sucessivamente para $g = 1,2,\ldots,k$ grupos.

Em aplicações práticas, às vezes é conveniente combinar dois ou mais métodos. Por exemplo, no caso de mistura de multinormais, usar o limite máximo sugerido por Hartigan, combinado com o teste de significância de Wolfe. Assim, adotar-se-ia $k$ grupos quando ou $k$ for o número sugerido pelo teste e menor do que o limite máximo; ou $k$ já tiver atingido esse limite.

# C A P Í T U L O VI

## APLICAÇÕES

Serão apresentadas aqui duas aplicações da técnica de mistura de multinormais como forma de se detectar uma estrutura de agrupamento nos dados.

O algorítmo proposto por Hartigan foi utilizado com todos os cuidados sugeridos anteriormente. Várias execuções foram realizadas com diferentes estimativas iniciais. Nos casos em que se obteve soluções diferentes, somente a que forneceu maior valor para a função de verossimilhança foi utilizada nas análises subsequentes.

Este algorítmo apresenta um problema que é a velocidade de convergência do processo iterativo. Como a convergência é muito vagarosa, Hartigan propõe a seguinte técnica de aceleração. Depois de um certo número de iterações, atribui-se às probabilidades de pertinência os valores *1* ou *0*, conforme as probabilidades estimadas até esse passo estejam próximas desses valores. Neste trabalho também se usou este artifício.

A primeira aplicação foi feita sobre dados já bastante explorados em técnicas de Análise Multivariada: o conjunto de dados da Iris. E a segunda trata de uma pesquisa de identificação de solos cujos dados foram cedidos pelo Setor de Estatística Aplicada do IME-USP.

### 6.1. ANÁLISE DOS DADOS DA IRIS

#### *6.1.1. Aplicações*

A eficiência desse algoritmo, foi verificada testando-o com o clássico conjunto de dados da Iris, utiliza

dos por Fisher em 1936 quando estudava a técnica de Análise Discriminatória (Kendall & Stuart (1961)). Trata-se de 150 plantas, classificadas em três categorias conhecidas: Iris Virginica (50), Iris Setosa (50) e Iris Versicolor (50). As observações são as quatro medidas: comprimento da pétala, largura da pétala, comprimento da sépala e largura da sépala. Supondo a inexistência de qualquer informação a priori sobre a classificação desse conjunto de plantas, tentou-se descobrir e caracterizar cada um dos tipos de íris.

Aplicou-se o algorítmo sob duas suposições diferentes. Inicialmente supondo o caso mais geral de matrizes de variâncias e covariâncias diferentes dentro dos grupos. Em seguida, impondo a restrição de que as matrizes de variâncias e covariâncias dos grupos deveriam ser iguais. No segundo caso se obteve resultados mais condizentes com a classificação real.

Com uma partição inicial próxima da verdadeira, chegou-se ao ponto de máximo global do logaritmo da função de verossimilhança, que é dado pelo valor *295,00907*, obtido com três alocações erradas. A tabela 6.1 mostra estas plantas e as respectivas estimativas das probabilidades de pertinência em cada um dos grupos. O asterisco indica as probabilidades estimadas de se alocar corretamente essas plantas. As probabilidades de pertinência ao terceiro grupo são iguais a zero, e portanto não constam da tabela.

TABELA 6.1

| *Planta* | $P[\hat{\pi}_1 \mid \underset{\sim}{x}]$ | $P[\hat{\pi}_2 \mid \underset{\sim}{x}]$ |
|---|---|---|
| *Iris Versicolor 21* | *0,865529* | *0,134421** |
| *Iris Versicolor 34* | *0,739051* | *0,260949** |
| *Iris Virginica 34* | *0,126620** | *0,873380* |

Estas três plantas mal alocadas coincidem com as

três classificações erradas utilizando-se a técnica de maximizar $|\underset{\sim}{T}| \; / \; |\underset{\sim}{W}|$. (Friedman & Rubin (1967).) A partição verdadeira fornece um valor para $log\ L(\underset{\sim}{\theta})$ igual a *294,71694*.

Iniciando o processo iterativo com partições iniciais bastante arbitrárias, obteve-se uma partição em três grupos com 17 plantas alocadas erroneamente, que fornecem um valor para $log\ L(\underset{\sim}{\theta})$ igual a *287,88925*. O valor menor de $log\ L(\theta)$ sugere que o processo convergiu para um máximo local, fato já discutido anteriormente.

A Tabela 6.2 das plantas mal classificadas segue as mesmas especificações da Tabela 6.1.

TABELA 6.2

| *Planta* | $\hat{P}[\pi_1 \mid \underset{\sim}{x}]$ | $\hat{P}[\pi_2 \mid \underset{\sim}{x}]$ |
|---|---|---|
| *Iris Versicolor 21* | *0,644660* | *0,355340** |
| *Iris Virginica 04* | *0,045600** | *0,954000* |
| *Iris Virginica 06* | *0,171832** | *0,828168* |
| *Iris Virginica 08* | *0,000307** | *0,999693* |
| *Iris Virginica 09* | *0,031536** | *0,968464* |
| *Iris Virginica 17* | *0,040123** | *0,959877* |
| *Iris Virginica 18* | *0,144829** | *0,855171* |
| *Iris Virginica 20* | *0,002020** | *0,997980* |
| *Iris Virginica 23* | *0,010592** | *0,989408* |
| *Iris Virginica 26* | *0,000636** | *0,999364* |
| *Iris Virginica 30* | *0,000009** | *0,999991* |
| *Iris Virginica 31* | *0,433838** | *0,566162* |
| *Iris Virginica 32* | *0,001519** | *0,998481* |
| *Iris Virginica 34* | *0,000131** | *0,999869* |
| *Iris Virginica 35* | *0,000002** | *0,999998* |
| *Iris Virginica 38* | *0,035019** | *0,964981* |
| *Iris Virginica 50* | *0,433838** | *0,566162* |

A Iris Setosa sempre se separou num grupo a parte,

numa indicação de que se os grupos estiverem bem separados, este procedimento deve fornecer bons resultados.

*6.1.2. Determinação do Número de Grupos*

Na determinação do número de grupos, segundo o capítulo anterior, poder-se-ia optar por três formas diferentes. A primeira delas seria o limite máximo de grupos, que, neste caso, supondo igualdade das matrizes de variâncias e covariâncias, é dado por:

$$k \leq \frac{n+1}{p+1} - \frac{p}{2}$$

e portanto,

$$k \leq 28.$$

A segunda alternativa é dada pelo teste de significância descrito em 3.3, onde se verifica hierarquicamente a significância da existência de um, dois, ..., etc grupos. A tabela 6.3 apresenta os valores observados do teste.

TABELA 6.3

| $k$ \| $k'$ | $\log L_k(\underset{\sim}{\theta})$ | $T_n$ | $\chi^2_{8,\alpha/k}$ |
|---|---|---|---|
| 1 / 2 | 171, 44849 | 21, 446762 | 15,507 |
| 2 / 3 | 182, 58003 | 202, 19369 | 17,534 |
| 3 / 4 | 287, 88925 | 77, 338596 | 19,456 |
| 4 / 5 | 328, 31012 | 15, 182919 | 19,610 |
| 5 / 6 | 336, 27319 | 5, 938042 | 20,090 |
| 6 / 7 | 339, 40948 | | |

Aceita-se então a hipótese de que existem quatro grupos.

Finalmente, a análise gráfica está na figura 6.1.

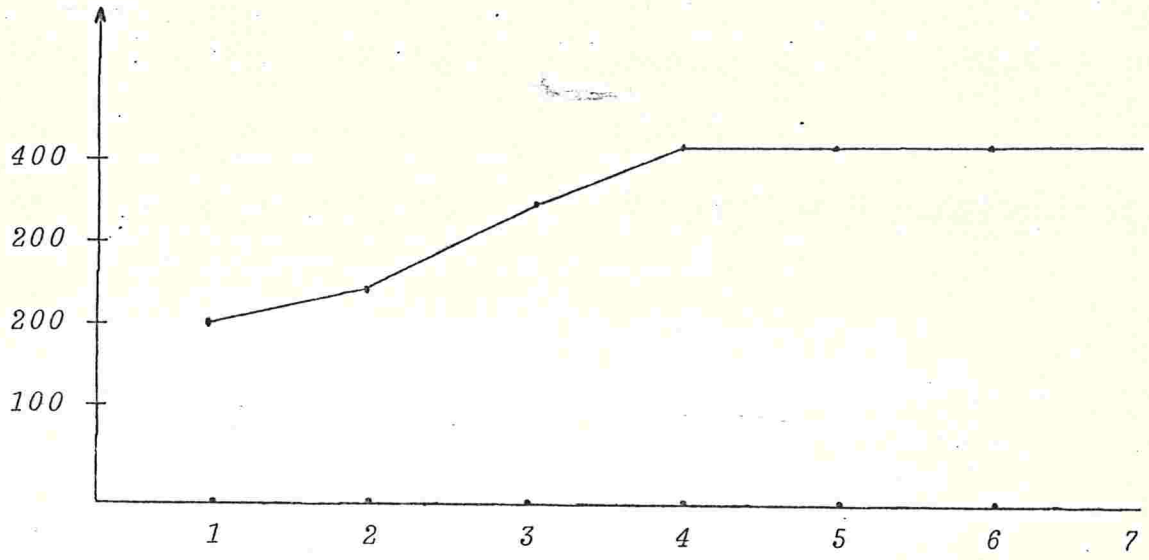



FIGURA 6.1

Pelo gráfico percebe-se que depois de quatro grupos, os ganhos na função de verossimilhança, ao se aumentar o número de grupos, são muito discretos, não justificando essas partições. O ponto razoável de parada é quatro.

A partição obtida com três grupos já foi analisada anteriormente, assim, agora a análise vai se ater à partição com quatro grupos. O agrupamento ficou assim configurado:

Grupo 1: *48 Iris Versicolor, 1 Iris Virginica (VG34)*

Grupo 2: *50 Iris Setosa*

Grupo 3: *15 Iris Virginica, 1 Iris Versicolor (VS34)*

Grupo 4: *34 Iris Virginica, 1 Iris Versicolor (VS21)*

A análise de pertinência dos elementos aos grupos, fornece as seguintes conclusões:

a) a fusão dos grupos 3 e 4 fornece a partição em três grupos que maximiza $log\ L(\theta)$, e com ape

nas três classificações erradas;

b) a partição "ótima", com 17 classificações erradas, é obtida pela fusão dos grupos 1 e 3.

Esses fatos parecem mostrar uma grande dissimilaridade entre os 17 elementos e os grupos aos quais eles deveriam pertencer. Este fato pode ser avaliado quando se observa as distâncias de Mahalanobis entre os centros dos grupos 1, 3 e 4, conforme aparece na Tabela 6.4.

TABELA 6.4.

| | *1* | *3* | *4* |
|---|---|---|---|
| *1* | *0,0000* | *13,8699* | *21,0832* |
| *2* | *13,8699* | *0,0000* | *14,0690* |
| *4* | *21,0832* | *14,0690* | *0,0000* |

O centro de massa do grupo 3 está "mais próximo" do centro de massa do grupo 1 do que do centro do grupo 4, explicando a composição da partição em 3 grupos.

Embora sabendo da não validade do método nesta situação, mas com o objetivo único de explorar os dados, foi feita uma análise de variância multivariada sobre o agrupamento obtido. A MANOVA rejeitou a hipótese de igualdade dos grupos, e o emprego posterior de técnicas de comparações múltiplas confirmou a significância de cada grupo isoladamente. Assim, apesar da classificação biológica da Iris Virginica em um único grupo, os dados revelam certa heterogeneidade interna nesta classe.

### *6.1.3. Sobre Diferentes Proporções*

Para verificar a tendência de alguns processos de

Análise de Conglomerados, em dividir os dados em grupos do mesmo tamanho (Everitt (1974)), alguns testes foram feitos variando-se as proporções na mistura. O trabalho foi feito sempre com duas populações apenas.

Inicialmente tomou-se as 50 plantas de uma das espécies combinadas com 10 de qualquer das outras espécies. Nas combinações em que as Iris Setosa estiveram presentes, sempre se obteve bons resultados. Já nas combinações das Iris Versicolor com as Iris Virginica, o algorítmo não convergiu. Isto é, os valores de $log\ L(\theta)$ foram diminuindo, até que todos os elementos ficassem num único grupo.

Em seguida, restringindo os testes às misturas de Iris Versicolor com Iris Virginica, e aumentando para 15 o número de plantas da espécie minoritária na mistura, o algorítmo ainda não pareceu eficiente, apresentando resultados semelhantes aos do caso anterior.

Finalmente, para misturas 20:50, conseguiu-se partições em 2 grupos, mas ainda assim com várias plantas mal alocadas.

Assim, parece que se os grupos forem bastante diferenciados, o algoritmo consegue identificá-los mesmo que suas proporções na mistura sejam bem diferentes. Se, no entanto, os conglomerados estiverem bem próximos um do outro, os resultados obtidos podem ser de nenhuma valia.

## 6.2. ANÁLISE DE UMA PESQUISA DE IDENTIFICAÇÃO DE SOLOS

A classificação de latossolos em roxo (LR), vermelho-escuro (LVE) e vermelho-amarelo (LVA) é usualmente feita através da cor que os caracteriza, e do teor de $Fe_2O_3$ (trióxido de ferro) contido no solo. Através desse critério, espera-se uma superposição dos elementos classificados como latossolo roxo e latossolo vermelho-escuro; bem

como entre latossolo vermelho-escuro e latossolo vermelho-amarelo.

Num trabalho apresentado ao Setor de Estatística Aplicada do IME-USP, um agrônomo, suspeitando da consistência desse método de classificação, mostrou interesse na utilização de métodos estatísticos na diferenciação de latossolos. Para tal, tomou uma amostra aleatória de 79 latossolos em vários estados do país, e observou onze variáveis que considerou de importância no processo. A saber: Silte, Argila, PH, % de matéria orgânica (C), saturação com bases (S), total de bases (T), $Al_2O_3$ (óxido de Alumínio), $Fe_2O_3$ (trióxido de Ferro), $TiO_2$ (óxido de Titânio), $SiO_2$ (óxido de Silício), V.

Os dados, para conferência, estão no apêndice 1.

Considerando que o número de variáveis iniciais é grande e muito correlacionadas umas com as outras, os dados foram inicialmente submetidos a uma Análise Fatorial, extraindo-se daí três fatores, que explicam, conjuntamente, 77% da variação original dos dados. O Fator 1 está mais correlacionado com as variáveis Silte, Argila, PH, $Al_2O_3$, C, $FeO_3$ e $TiO_2$. O Fator 2 está alta e negativamente correlacionado com as variáveis T e $SiO_2$. E, finalmente, o Fator 3 está mais correlacionado com as variáveis V e S. Os demais resultados dessa análise estão no apêndice 1. Ainda nesta fase preliminar foram retiradas algumas observações que pareceram espúrias dentro da amostra, reduzindo-se assim a amostra a 75 observações.

Destarte, ao invés da técnica de agrupamento ser aplicada diretamente sobre os dados originais, ela foi aplicada sobre os escores produzidos pela Análise Fatorial, que parecem se enquadrar num modelo de mistura de multinormais.

A teoria de Análise Fatorial garante que os escores fatoriais ainda não agrupados têm a mesma estrutura de

variâncias e covariâncias. Entretanto, ao se agrupar os dados, essa estrutura é quebrada, e trabalhos aplicados anteriores revelaram que os dados agrupados segundo algum critério, já não suportam a suposição de matrizes de variâncias e covariâncias idênticas. Em raras situações essas matrizes resistiram a um teste de igualdade de matrizes de variâncias e covariâncias.

Dessa forma, o algorítmo proposto por Hartigan (1975) foi utilizado na sua forma genérica. Os resultados estão reportados a seguir.

Pela fórmula (5.9), o número máximo de grupos é $k = 7$. E, aplicando-se o teste (5.11), tem-se os resultados na tabela 6.5.

TABELA 6.5

| $k$ | $\log L_k(\theta)$ | $T_n$ | $\chi^2_{6,\alpha/k}$, $\alpha = 0,05$ |
|---|---|---|---|
| 1 | -112,50 | 64,8788 | 12,592 |
| 2 | -77,99 | 19,9547 | 14,449 |
| 3 | -67,37 | 40,0320 | 15,567 |
| 4 | -45,70 | 21,8960 | 16,278 |
| 5 | -33,80 | 3,6716 | 16,812 |
| 6 | -31,79 | 18,7499 | |
| 7 | -21,45 | | |

Por outro lado, o gráfico da função de verossimilhança também sugere um ponto de parada seja para $k = 5$ grupos (Figura 6.2).

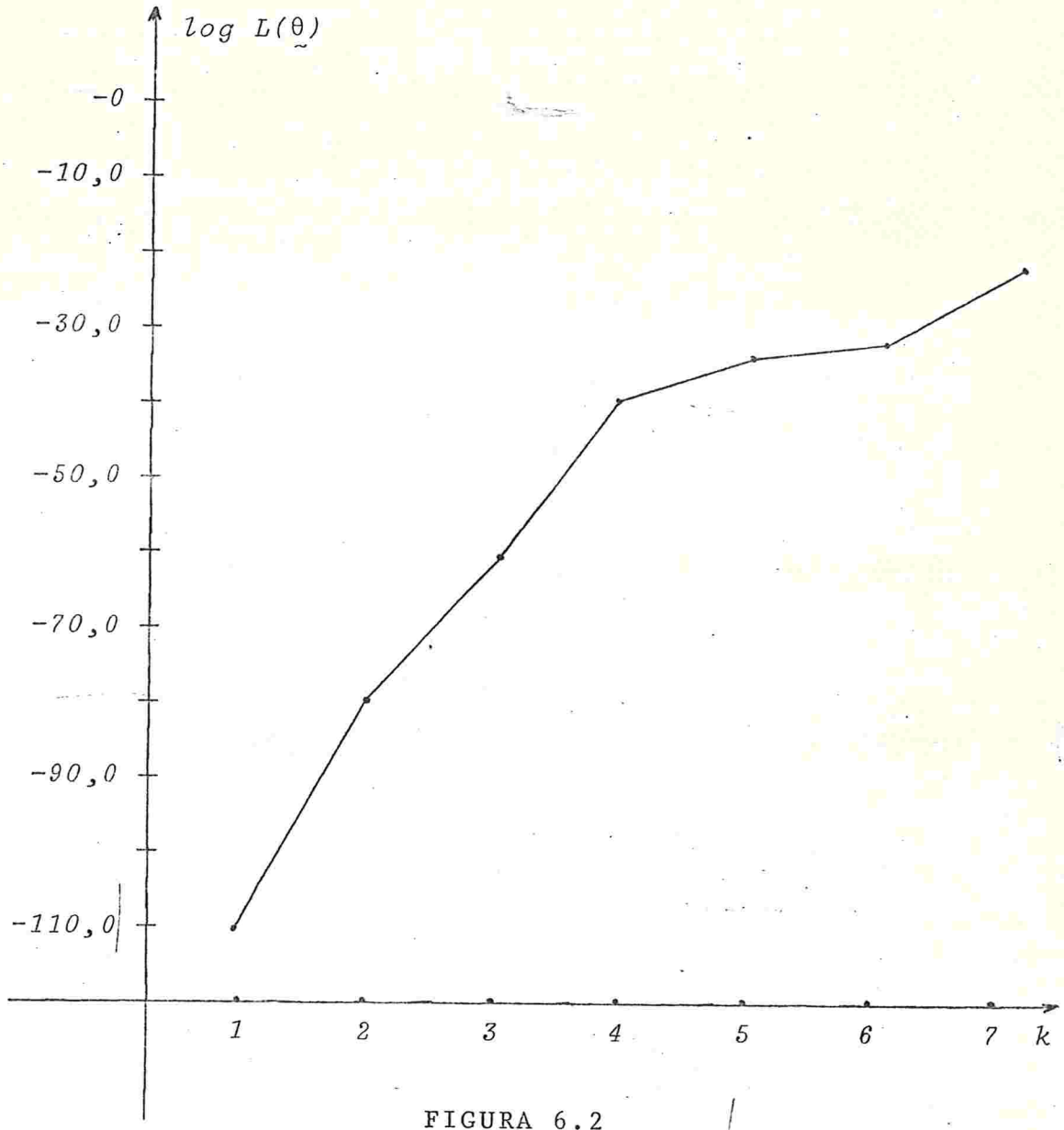



FIGURA 6.2

Como a Análise de Conglomerados é essencialmente uma análise exploratória, não obstante as indicações para cinco grupos, outros agrupamentos serão examinados com algum detalhe.

Tomando inicialmente a partição com três grupos, e comparando-a com a partição dada pela classificação usual de latossolos, obtém-se a tabela 6.6.

TABELA 6.6

| usual \ conglomerados | 1 | 2 | 3 | TOTAL |
|---|---|---|---|---|
| LR | 5 | 14 | 0 | 19 |
| LVE | 17 | 4 | 14 | 35 |
| LVA | 8 | 1 | 12 | 21 |
| TOTAL | 30 | 19 | 26 | 75 |

Olhando a tabela 6.6, verifica-se que apesar da discordância entre as duas classificações, existe um núcleo comum aos dois agrupamentos. O conglomerado 1, por exemplo, tem uma concentração alta de amostras de latossolo vermelho-escuro, enquanto no conglomerado 2 a concentração maior é de latossolos roxos. Já no conglomerado 3, há uma quantidade quase igual de latossolos vermelho-escuros e vermelho-amarelos, mas em compensação, não há nenhuma amostra de latossolo roxo.

Segundo esta nova classificação, além das suposições já citadas, parece existir também uma superposição entre latossolo roxo e latossolo vermelho-amarelo.

Considerando agora a partição para $k = 4$.

A tabela 6.7 mostra o número de elementos classificados segundo as partições em 3 e 4 conglomerados.

TABELA 6.7

| | 1 | 2 | 3 | 4 | TOTAL |
|---|---|---|---|---|---|
| 1 | 16 | 0 | 0 | 14 | 30 |
| 2 | 0 | 10 | 9 | 0 | 19 |
| 3 | 11 | 0 | 5 | 10 | 26 |
| TOTAL | 27 | 10 | 14 | 24 | 75 |

Pela tabela 6.7, parece que existia, embutido no grupo 2 da partição anterior (em 3 conglomerados), um grupo mais distante dos demais, que é composto por dez observações, e será denotado por 4(2).

Usar-se-á a notação $i(j)$ para denotar o *j-ésimo* conglomerado na partição em $i$ grupos.

A tabela 6.8 mostra a composição dos 4 conglomerados, segundo a classificação tradicional.

TABELA 6.8

| | *1* | *2* | *3* | *4* | *TOTAL* |
|---|---|---|---|---|---|
| *LR* | *5* | *10* | *4* | *0* | *19* |
| *LVE* | *12* | *0* | *7* | *16* | *35* |
| *LVA* | *10* | *0* | *3* | *8* | *21* |
| *TOTAL* | *27* | *10* | *14* | *24* | *75* |

Dos 12 LVE de 4(1), sete deles estavam em 3(1), e os outros cinco em 3(3), dos 10 LVA de 4(1), quatro deles estavam em 3(1) e os demais estavam em 3(3). Assim, embora um dos grupos atuais estivesse inteiramente contido em um dos conglomerados da partição anterior, houve ainda modificações substanciais na composição dos demais conglomerados.

Finalmente, examine-se a partição com 5 conglomerados. A relação entre a partição atual com a anterior em 4 conglomerados está exposta na tabela 6.9.

A tabela 6.9 mostra uma certa estabilidade dos conglomerados 5(1), 5(3) e 5(5), uma vez que são quase idênticos aos conglomerados 4(1), 4(2) e 4(4), respectivamente. Os conglomerados 5(2) e 5(4) foram obtidos a partir de uma partição no grupo 4(3).

TABELA 6.9

| | 1 | 2 | 3 | 4 | 5 | TOTAL |
|---|---|---|---|---|---|---|
| 1 | 27 | 0 | 0 | 0 | 0 | 27 |
| 2 | 0 | 0 | 10 | 0 | 0 | 10 |
| 3 | 1 | 5 | 0 | 8 | 0 | 14 |
| 4 | 0 | 1 | 0 | 0 | 23 | 24 |
| TOTAL | 28 | 6 | 10 | 8 | 23 | 75 |

A constituição dos conglomerados conforme a classificação tradicional está na tabela 6.10.

TABELA 6.10

| | 1 | 2 | 3 | 4 | 5 | TOTAL |
|---|---|---|---|---|---|---|
| LR | 5 | 2 | 10 | 2 | 0 | 19 |
| LVE | 12 | 4 | 0 | 4 | 15 | 35 |
| LVA | 11 | 0 | 0 | 2 | 8 | 21 |
| TOTAL | 28 | 6 | 10 | 8 | 23 | 75 |

Assim, mesmo segundo esse novo conjunto de variáveis, e esta nova classificação; existe um grupo formado apenas por latossolos roxos, o que indica uma acentuada diferenciação em relação aos demais.

A seguir está a composição e caracterização de alguns parâmetros (vetor de médias, matrizes de variâncias, covariâncias e proporções na mistura) de cada um dos cinco grupos. Ao lado de cada observação, está transcrita a sua classificação pelo método usual, fornecida pelo pesquisador.

CONGLOMERADO 1

| | | |
|---|---|---|
| GO14 - LR | P282 - LVE | P47S - LVA |
| P36S - LR | P405 - LVE | P48S - LVA |
| GO16 - LR | P22N - LVE | GOO5 - LVA |
| GOO1 - LR | P73N - LVE | P49S - LVA |
| P33S - LR | RS04 - LVE | GOO2 - LVA |
| GO17 - LVE | P33N - LVE | P1PB - LVA |
| GOO6 - LVE | P16N - LVE | P5PB - LVA |
| GO19 - LVE | PO8N - LVE | GOO8 - LVA |
| P39S - LVE | GO23 - LVA | GO13 - LVA |
| GO25 - LVE | GO24 - LVA | |

$$\hat{\underset{\sim}{\mu}}_1 = \begin{bmatrix} 0,2504 \\ -0,4131 \\ -0,1546 \end{bmatrix}, \quad \hat{\underset{\sim}{\Sigma}} = \begin{bmatrix} 0,9226 & -0,4784 & 0,0268 \\ & 0,4780 & -0,0007 \\ & & 0,1287 \end{bmatrix}$$

$$\hat{\lambda}_1 = 0,3974$$

CONGLOMERADO 2

| | | |
|---|---|---|
| RS02 - LR | P42S - LVE | P41S - LVE |
| P90N - LR | P102 - LVE | P46S - LVE |

$$\hat{\underset{\sim}{\mu}}_2 = \begin{bmatrix} 0,1008 \\ 0,7566 \\ 0,9584 \end{bmatrix}, \quad \hat{\underset{\sim}{\Sigma}}_2 = \begin{bmatrix} 0,8037 & 0,5150 & -0,0352 \\ & 0,3786 & -0,0238 \\ & & 0,0100 \end{bmatrix}$$

$$\hat{\lambda}_2 = 0,0742$$

*CONGLOMERADO 3*

| | | |
|---|---|---|
| P11N - LR | P38N - LR | RS01 - LR |
| P12N - LR | P15N LR | P24N - LR |
| P23N - LR | P20N LR | P35S - LR |
| P34N - LR | | |

$$\hat{\underset{\sim}{\mu}} = \begin{bmatrix} 1,1595 \\ 1,0436 \\ -0,8803 \end{bmatrix} \quad , \quad \hat{\underset{\sim}{\Sigma}}_3 = \begin{bmatrix} 0,0913 & -0,0449 & -0,0361 \\ & 0,2136 & -0,0470 \\ & & 0,0881 \end{bmatrix}$$

$$\hat{\lambda}_3 = 0,1310$$

*CONGLOMERADO 4*

| | | |
|---|---|---|
| P35N - LR | P47N - LVE | P4PB - LVA |
| P27N - LR | R503 - LVE | BR1R - LVA |
| P113N - LVE | P51N - LVE | |

$$\hat{\underset{\sim}{\mu}}_4 = \begin{bmatrix} -0,0731 \\ 1,3037 \\ 1,0368 \end{bmatrix} \quad , \quad \hat{\underset{\sim}{\Sigma}}_4 = \begin{bmatrix} 0,1225 & 0,0401 & 0,3456 \\ & 0,5514 & -0,1050 \\ & & 4,1994 \end{bmatrix}$$

$$\hat{\lambda}_4 = 0,1124$$

*CONGLOMERADO 5*

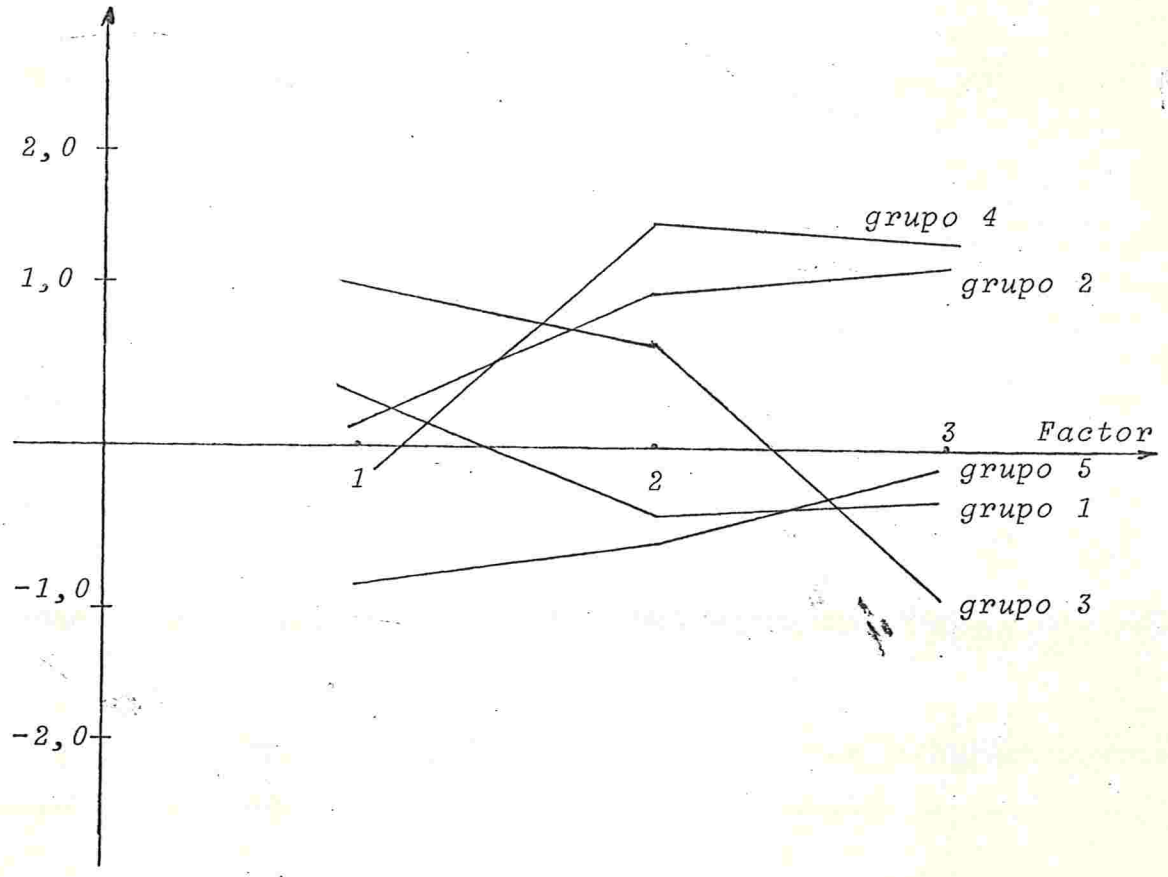
| | | |
|---|---|---|
| P45N - LVE | PO3N - LVE | GO20 - LVA |
| P53N - LVE | PO4N - LVE | P52S - LVA |
| P41N - LVE | PO7N - LVE | P53S - LVA |
| P43S - LVE | PO9N - LVE | P54S - LVA |
| P18N - LVE | P17N - LVE | P2PB - LVA |
| P44S - LVE | P28N - LVE | P3PB - LVA |
| P19N - LVE | P30N - LVE | GO18 - LVA |
| P45S - LVE | GO21 - LVA | |

$$\hat{\underset{\sim}{\mu}}_5 = \begin{bmatrix} -0,8794 \\ -0,6144 \\ -0,0381 \end{bmatrix} \quad , \quad \hat{\underset{\sim}{\Sigma}}_5 = \begin{bmatrix} 0,3931 & -0,3964 & 0,3463 \\ & 0,4466 & -0,3419 \\ & & 0,5302 \end{bmatrix}$$

$$\hat{\lambda}_5 = 0,2850$$

Os conglomerados encontrados parecem ter pouco em comum com a classificação usual dos latossolos.

Analisando o gráfico dos perfis dos grupos, nota-se que o grupo 2 se caracteriza por ter todas as médias positivas, em contraposição ao grupo 5, cujas médias são todas negativas. Já o grupo 1 tem as três médias relativamente próximas de zero. E, finalmente os grupos 3 e 4 contrastam fortemente para os fatores 1 e 3.



Embora esses resultados sejam tão somente o produto de uma análise exploratória de dados, eles revelam a possibilidade de se encontrar uma diferenciação mais refinada dos solos, e mais importante, que talvez valha a pena repensar essa classificação a partir de novos parâmetros.

# C A P Í T U L O VII

## CONCLUSÕES

Os resultados apresentados neste trabalho se restringem a distribuições normais multivariadas, e estão baseados nos trabalhos desenvolvidos principalmente por Scott & Symons (1971) e Wolfe (1970).

Deve-se lembrar que os métodos vistos se aplicam a amostras aleatórias, que nem sempre são as que aparecem nos trabalhos de pré-classificação.

As técnicas aqui estudadas apresentam um problema comum em Análise de Conglomerados. A quantidade de cálculo envolvida nas soluções numéricas é excessiva, tornando impraticável a determinação da melhor partição de $n$ elementos em $k$ grupos. Dessa forma, o resultado é dado por partições localmente ótimas. Como o agrupamento final é o produto de uma partição inicial arbitrária, talvez alguns estudos devessem ser feitos a respeito da escolha do agrupamento inicial. Ou ainda, sobre métodos computacionais mais simples e eficientes de se resolver o problema. Enquanto não se dispõe desses recursos, a sugestão ainda é utilizar diferentes partições iniciais, e escolher aquela que estiver mais próxima de satisfazer o critério adotado.

Aplicações práticas dos testes para a determinação do número de grupos através da estatística $T_n$ revelaram uma tendência em apresentar níveis de significância descritivos muito baixos, sugerindo assim uma partição com um número de grupos maior do que o necessário para se explicar a variabilidade dos dados. Uma explicação possível para esse fenômeno é o fato da amostra ser muito pequena

para estimar tantos parâmetros, e daí, a distribuição da estatística $T_n$ não é bem aproximada pela distribuição assintótica de qui-quadrado. Binder (1978) mostrou que em algumas situações a distribuição dessa estatística realmente não converge. Porém, como Everitt (1979) ponderou: " ..., interpretabilidade e simplicidade são importantes na análise de dados, e, qualquer inferência rígida do número ótimo de grupos, à luz de valores observados de um índice numérico de bondade de ajustamento, pode ser improdutivo, ...".

Então dois pontos fundamentais parecem ser merecedores de investigações futuras:

i) as distribuições das estatísticas utilizadas na determinação do número de grupos;

ii) a criação de algoritmos eficientes para a solução numérica do problema.

O segundo item é sumamente importante, uma vez que sem bons métodos numéricos e computacionais, a aplicação de Análise de Conglomerados se torna impraticável.

Na aplicação da técnica de mistura de multinormais convém ter em mente dois fatores que já foram discutidos no capítulo 3. Primeiro, que ela produz bons resultados quando os grupos estão bem separados, e segundo, que ela apresenta uma tendência em fornecer uma indicação falsa da existência de grupos.

Não obstante essas limitações, os modelos estudados se afirmam como técnicas estatísticas descritivas de dados multidimensionais. Eles têm por objetivo reduzir o conjunto excessivamente grande de observações em estudo, para $k$ conjuntos menores, de forma a ter um mínimo de informações perdidas. A Análise de Conglomerados trata então de descrever o comportamento de dados multivariados a

fim de facilitar a compreensão e a interpretação das observações. Ainda, como toda análise exploratória, não deve simplesmente ter um fim com os resultados produzidos pela aplicação da técnica em si. O bom emprego do método de agrupamento requer a aliação da técnica numérica com o conhecimento teórico e experimental do pesquisador.

Existem outras abordagens estatísticas ao problema de conglomerados. Uma delas consiste em se imaginar que a distribuição da população em estudo é multimodal, e, assim a cada moda da distribuição corresponde um conglomerado. Se a forma da distribuição for conhecida, o problema de conglomerados se transforma num de estimação de modas. Existem alguns trabalhos a esse respeito, tais como o de Dalenius (1965) e o de Robertson & Cryer (1974).

Para os casos em que não se dispõe de qualquer informação a respeito da população em estudo, Bryan (1971) propôs um método para estimação de funções densidade multivariadas através de modelos de núcleos e daí, derivou um método empírico para detectar as modas da distribuição estimada.

Numa abordagem mais recente, utilizando o modelo de mistura de distribuições, mas com a suposição adicional de que as probabilidades de pertinência aos grupos são variáveis aleatórias com distribuições conhecidas, Binder (1978) enfocou a Análise de Conglomerados de uma forma Bayesiana.

Assim, parece que os métodos estatísticos nessa área estão se refinando mais e mais, possibilitando, num futuro breve, a construção de uma teoria apoiada em conceitos estatísticos de estimação e testes de hipótese. De qualquer forma há um largo e interessante campo de estudos a frente.

# A P Ê N D I C E 1

Os resultados da Análise Fatorial foram obtidos utilizando-se o computador IBM/370 mod. 155 do IPEN (Instituto de Pesquisas Energéticas e Nucleares), e através da rotina FACTOR do SAS (Statistical Analysis System), de propriedade do ICMSC-USP.

A tabela 1 apresenta os dados iniciais.

As comunalidades estimadas estão na tabela 2.

Finalmente, a tabela 3 contém os escores fatoriais, sobre os quais foi aplicada a técnica de mistura de multinormais.

# T A B E L A  1

## DADOS ORIGINAIS

| OBS | SILTE | ARGILA | PH | C | S | T | V | $SiO_2$ | $Al_2O_3$ | $Fe_2O_3$ | $TiO_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 14.0 | 59.0 | 6.2 | 0.65 | 0.10 | 0.90 | 11.0 | 4.90 | 27.10 | 28.90 | 2.91 |
| 2 | 10.0 | 51.0 | 5.5 | 0.56 | 0.40 | 2.00 | 20.0 | 7.50 | 24.40 | 11.50 | 1.07 |
| 3 | 12.0 | 56.0 | 5.7 | 0.34 | 0.10 | 1.00 | 10.0 | 10.20 | 24.70 | 15.20 | 1.79 |
| 4 | 26.0 | 23.0 | 5.6 | 0.34 | 0.20 | 1.40 | 14.0 | 7.30 | 7.70 | 10.50 | 0.32 |
| 5 | 12.0 | 44.0 | 4.8 | 0.37 | 0.40 | 2.50 | 16.0 | 9.80 | 19.70 | 7.30 | 0.87 |
| 6 | 15.0 | 41.0 | 5.9 | 0.31 | 0.50 | 2.60 | 19.0 | 12.10 | 16.70 | 6.60 | 0.73 |
| 7 | 19.0 | 40.0 | 5.6 | 0.36 | 0.20 | 2.00 | 10.0 | 11.20 | 17.70 | 13.50 | 0.36 |
| 8 | 16.0 | 82.0 | 5.5 | 0.35 | 0.50 | 7.50 | 6.0 | 28.00 | 22.70 | 22.30 | 4.18 |
| 9 | 17.0 | 74.0 | 5.4 | 0.51 | 1.70 | 6.30 | 26.0 | 26.90 | 23.70 | 22.70 | 3.84 |
| 10 | 11.0 | 55.0 | 4.8 | 0.30 | 0.50 | 6.40 | 8.0 | 20.50 | 17.70 | 10.30 | 1.52 |
| 11 | 8.0 | 32.0 | 5.0 | 0.32 | 1.00 | 3.50 | 18.0 | 12.10 | 10.30 | 4.70 | 0.83 |
| 12 | 11.0 | 67.0 | 5.7 | 0.54 | 0.30 | 2.40 | 13.0 | 12.20 | 27.10 | 20.70 | 2.50 |
| 13 | 10.0 | 57.0 | 5.6 | 0.57 | 0.20 | 2.40 | 8.0 | 8.60 | 25.50 | 17.00 | 2.79 |
| 14 | 8.0 | 43.0 | 5.8 | 0.43 | 0.20 | 1.20 | 17.0 | 5.20 | 20.10 | 9.90 | 1.03 |
| 15 | 19.0 | 41.0 | 5.9 | 0.34 | 0.10 | 0.20 | 50.0 | 6.50 | 19.30 | 5.80 | 0.84 |
| 16 | 10.0 | 52.0 | 5.2 | 0.36 | 0.10 | 1.00 | 10.0 | 12.20 | 22.30 | 14.10 | 1.36 |
| 17 | 5.0 | 55.0 | 5.6 | 0.49 | 0.20 | 2.00 | 10.0 | 15.30 | 23.00 | 8.10 | 0.87 |
| 18 | 11.0 | 59.0 | 5.8 | 0.38 | 0.40 | 1.80 | 22.0 | 16.80 | 29.00 | 8.20 | 0.63 |

(Continuação da Tabela 1)

| OBS | SILTE | ARGILA | PH | C | S | T | V | $SiO_2$ | $Al_2O_3$ | $Fe_2O_3$ | $TiO_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 14.0 | 24.0 | 5.8 | 0.29 | 0.30 | 1.70 | 18.0 | 5.30 | 8.10 | 6.60 | 0.71 |
| 20 | 14.0 | 18.0 | 5.3 | 0.27 | 0.30 | 1.40 | 21.0 | 5.40 | 8.30 | 1.90 | 0.26 |
| 21 | 14.0 | 49.0 | 5.2 | 0.35 | 1.30 | 4.50 | 29.0 | 20.70 | 17.20 | 7.10 | 0.66 |
| 22 | 12.0 | 70.0 | 5.2 | 0.45 | 1.80 | 6.30 | 29.0 | 26.20 | 24.80 | 8.70 | 0.67 |
| 23 | 6.0 | 81.0 | 5.5 | 0.47 | 1.00 | 7.20 | 14.0 | 29.20 | 26.70 | 9.70 | 0.71 |
| 24 | 6.0 | 39.0 | 5.3 | 0.16 | 0.30 | 2.60 | 12.0 | 14.70 | 13.20 | 5.90 | 1.01 |
| 25 | 7.0 | 42.0 | 5.2 | 0.18 | 0.60 | 3.70 | 16.0 | 16.20 | 14.30 | 5.20 | 0.95 |
| 26 | 9.0 | 39.0 | 5.2 | 0.31 | 0.40 | 3.40 | 12.0 | 13.70 | 13.10 | 5.60 | 0.91 |
| 27 | 7.0 | 45.0 | 5.2 | 0.33 | 0.30 | 4.20 | 7.0 | 14.80 | 14.40 | 11.30 | 3.54 |
| 28 | 10.0 | 34.0 | 5.4 | 0.21 | 0.50 | 3.30 | 15.0 | 12.20 | 11.00 | 5.30 | 0.77 |
| 29 | 4.0 | 18.0 | 5.3 | 0.08 | 0.20 | 1.40 | 14.0 | 6.80 | 5.60 | 3.30 | 0.61 |
| 30 | 4.0 | 20.0 | 4.8 | 0.13 | 0.30 | 2.10 | 14.0 | 7.20 | 6.10 | 3.10 | 0.36 |
| 31 | 4.0 | 17.0 | 4.6 | 0.13 | 0.30 | 1.80 | 17.0 | 7.90 | 6.50 | 3.70 | 0.67 |
| 32 | 7.0 | 27.0 | 5.0 | 0.22 | 0.40 | 3.00 | 13.0 | 9.90 | 8.60 | 5.40 | 0.80 |
| 33 | 6.0 | 20.0 | 4.2 | 0.12 | 0.30 | 2.00 | 15.0 | 7.10 | 6.00 | 3.50 | 0.50 |
| 34 | 5.0 | 17.0 | 5.0 | 0.12 | 0.10 | 2.00 | 5.0 | 7.00 | 6.00 | 3.40 | 0.50 |
| 35 | 10.0 | 14.0 | 3.7 | 0.20 | 0.30 | 3.00 | 10.0 | 5.50 | 5.10 | 3.30 | 0.68 |
| 36 | 8.0 | 16.0 | 4.0 | 0.13 | 0.50 | 2.20 | 23.0 | 6.10 | 5.90 | 3.50 | 0.59 |
| 37 | 7.0 | 20.0 | 5.3 | 0.18 | 0.20 | 2.20 | 9.0 | 7.50 | 6.00 | 3.30 | 0.69 |
| 38 | 8.0 | 22.0 | 4.6 | 0.24 | 0.70 | 3.40 | 21.0 | 8.30 | 7.10 | 2.10 | 0.35 |

(Continuação da Tabela 1)

| OBS | SILTE | ARGILA | PH | C | S | T | V | $SiO_2$ | $Al_2O_3$ | $Fe_2O_3$ | $TiO_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | 7.0 | 15.0 | 5.0 | 0.18 | 0.20 | 2.10 | 10.0 | 6.70 | 5.60 | 2.30 | 0.71 |
| 40 | 5.0 | 16.0 | 5.3 | 0.11 | 0.20 | 1.50 | 13.0 | 6.10 | 5.20 | 4.60 | 0.84 |
| 41 | 8.0 | 23.0 | 5.8 | 0.14 | 0.20 | 2.00 | 10.0 | 9.80 | 8.90 | 4.20 | 0.63 |
| 42 | 7.0 | 19.0 | 5.5 | 0.13 | 0.20 | 1.60 | 13.0 | 6.60 | 5.90 | 3.40 | 0.53 |
| 43 | 13.0 | 77.0 | 5.1 | 0.35 | 0.80 | 5.90 | 14.0 | 26.20 | 24.10 | 27.10 | 3.92 |
| 44 | 10.0 | 81.0 | 5.0 | 0.76 | 0.30 | 6.10 | 5.0 | 25.10 | 23.80 | 25.00 | 2.81 |
| 45 | 14.0 | 76.0 | 5.7 | 0.20 | 0.50 | 5.40 | 9.0 | 26.10 | 25.00 | 25.80 | 3.85 |
| 46 | 12.0 | 71.0 | 5.6 | 0.40 | 0.70 | 5.70 | 12.0 | 24.10 | 21.20 | 26.00 | 4.37 |
| 47 | 9.0 | 61.0 | 5.5 | 0.38 | 0.40 | 3.80 | 11.0 | 19.30 | 18.50 | 25.40 | 4.20 |
| 48 | 21.0 | 68.0 | 5.5 | 0.40 | 0.40 | 5.00 | 8.0 | 22.30 | 22.70 | 31.40 | 5.38 |
| 49 | 17.0 | 71.0 | 5.0 | 0.39 | 0.40 | 4.40 | 9.0 | 21.90 | 23.20 | 26.60 | 3.47 |
| 50 | 23.0 | 61.0 | 5.4 | 0.45 | 0.50 | 4.90 | 10.0 | 22.90 | 21.70 | 28.90 | 6.32 |
| 51 | 14.0 | 75.0 | 4.2 | 0.31 | 3.20 | 7.20 | 44.0 | 24.70 | 23.90 | 27.00 | 4.24 |
| 52 | 15.0 | 68.0 | 5.5 | 0.36 | 1.80 | 6.00 | 30.0 | 24.10 | 20.50 | 28.90 | 5.55 |
| 53 | 16.0 | 51.4 | 5.8 | 0.51 | 0.52 | 2.00 | 26.0 | 8.45 | 26.40 | 33.80 | 7.16 |
| 54 | 13.5 | 48.2 | 5.0 | 0.81 | 0.84 | 4.35 | 15.7 | 13.61 | 18.03 | 21.21 | 4.47 |
| 55 | 17.7 | 60.4 | 5.8 | 0.44 | 0.58 | 2.96 | 19.59 | 14.36 | 27.60 | 26.25 | 4.74 |
| 56 | 9.1 | 69.2 | 5.9 | 0.22 | 0.67 | 3.45 | 19.40 | 21.83 | 28.37 | 10.23 | 0.46 |
| 57 | 6.6 | 60.0 | 5.3 | 0.43 | 0.64 | 2.96 | 21.60 | 18.30 | 23.19 | 8.85 | 0.81 |
| 58 | 9.4 | 56.0 | 4.7 | 0.29 | 1.23 | 3.85 | 33.20 | 16.44 | 19.42 | 11.92 | 0.73 |

(Continuação da Tabela 1)

| OBS | SILTE | ARGILA | PH | C | S | T | V | $SiO_2$ | $Al_2O_3$ | $Fe_2O_3$ | $TiO_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 59 | 11.1 | 78.6 | 5.3 | 0.24 | 1.31 | 5.33 | 24.60 | 27.20 | 27.37 | 13.16 | 1.19 |
| 60 | 2.6 | 24.7 | 5.1 | 0.28 | 0.60 | 2.22 | 27.00 | 9.19 | 8.41 | 5.76 | 1.46 |
| 61 | 1.0 | 18.8 | 4.8 | 0.24 | 0.65 | 2.88 | 22.60 | 6.59 | 6.11 | 3.53 | 0.75 |
| 62 | 3.1 | 26.4 | 5.6 | 0.18 | 0.56 | 1.73 | 32.40 | 6.40 | 9.60 | 9.44 | 2.46 |
| 63 | 0.6 | 17.9 | 4.5 | 0.29 | 0.61 | 1.62 | 37.70 | 6.40 | 6.42 | 4.61 | 1.00 |
| 64 | 7.2 | 52.1 | 5.4 | 0.42 | 0.47 | 3.11 | 13.10 | 16.42 | 23.84 | 7.62 | 1.06 |
| 65 | 5.5 | 52.7 | 4.4 | 0.31 | 0.74 | 4.65 | 15.90 | 18.73 | 21.42 | 6.06 | 0.80 |
| 66 | 5.6 | 53.1 | 4.9 | 0.15 | 0.54 | 3.01 | 17.90 | 20.87 | 21.24 | 4.93 | 0.59 |
| 67 | 1.2 | 24.7 | 5.2 | 0.16 | 0.44 | 1.60 | 27.50 | 9.01 | 10.03 | 3.01 | 0.65 |
| 68 | 5.0 | 35.4 | 4.9 | 0.45 | 0.44 | 1.59 | 38.30 | 4.23 | 17.11 | 8.70 | 1.38 |
| 69 | 2.3 | 23.9 | 5.4 | 0.15 | 0.45 | 1.92 | 23.40 | 8.79 | 13.53 | 5.78 | 1.08 |
| 70 | 5.0 | 28.0 | 5.1 | 0.49 | 0.60 | 4.00 | 15.00 | 12.00 | 11.70 | 3.20 | 0.89 |
| 71 | 4.0 | 25.0 | 4.9 | 0.20 | 0.30 | 2.20 | 14.00 | 11.00 | 11.00 | 3.30 | 0.33 |
| 72 | 2.0 | 18.0 | 5.2 | 0.18 | 0.30 | 2.10 | 14.00 | 5.00 | 5.90 | 2.90 | 0.56 |
| 73 | 8.0 | 59.0 | 5.4 | 0.31 | 2.50 | 3.50 | 71.00 | 26.10 | 20.90 | 6.30 | 1.02 |
| 74 | 8.0 | 30.0 | 5.0 | 0.34 | 1.50 | 3.50 | 4.60 | 13.20 | 12.00 | 3.20 | 0.48 |
| 75 | 10.0 | 40.0 | 4.4 | 0.57 | 0.20 | 5.20 | 4.00 | 12.80 | 17.20 | 7.80 | 0.92 |
| 76 | 2.0 | 64.0 | 5.0 | 0.40 | 0.90 | 4.60 | 20.00 | 27.20 | 24.80 | 1.00 | . |

T A B E L A 2

| FATORES ROTACIONADOS | | | |
|---|---|---|---|
| | FATOR 1 | FATOR 2 | FATOR 3 |
| SILTE | 0.71091 | -0.08676 | -0.03625 |
| ARGILA | 0.75794 | -0.55107 | 0.12614 |
| PH | 0.68445 | 0.45560 | 0.07699 |
| C | 0.70385 | -0.12958 | -0.03809 |
| S | 0.04177 | -0.61985 | 0.70570 |
| T | 0.24054 | -0.93223 | -0.01050 |
| V | -0.04707 | 0.04256 | 0.95914 |
| $SiO_2$ | 0.42848 | -0.79987 | 0.15990 |
| $Al_2O_3$ | 0.81623 | -0.29117 | 0.17273 |
| $Fe_2O_3$ | 0.84768 | -0.32686 | -0.08699 |
| $TiO_2$ | 0.71562 | -0.32353 | -0.08992 |

# TABELA 3

| ESCORES FATORIAIS | | | |
|---|---|---|---|
| OBS | FATOR 1 | FATOR 2 | FATOR 3 |
| GO14 | 2.2855 | 1.8608 | -0.3993 |
| GO17 | 0.8063 | 0.9829 | 0.3045 |
| GO06 | 0.9956 | 1.1244 | -0.4258 |
| GO19 | 0.5372 | 1.4756 | -0.3629 |
| GO25 | -0.1213 | 0.1468 | -0.2123 |
| GO05 | 0.4654 | 0.9287 | 0.3615 |
| GO02 | 0.6872 | 0.9413 | -0.5153 |
| RS01 | 1.0547 | -1.7577 | -1.2002 |
| RS02 | 1.1319 | -1.4811 | 1.0003 |
| RS03 | -0.3848 | -1.5623 | -1.0329 |
| RS04 | -0.7463 | -0.2762 | 0.2359 |
| GO16 | 1.4253 | 0.6469 | -0.2798 |
| GO01 | 1.2086 | 0.7232 | -0.7165 |
| GO08 | 0.6216 | 1.4850 | 0.1049 |
| GO18 | 1.0068 | 2.4540 | 2.2169 |
| GO13 | 0.4708 | 0.6538 | -0.5878 |
| GO23 | 0.4741 | 0.5376 | -0.3802 |
| GO24 | 0.8679 | 0.8535 | 0.6965 |
| GO21 | 0.0886 | 1.4460 | 0.0595 |
| GO20 | -0.4057 | 1.2210 | 0.1779 |
| P102 | -0.1107 | -0.6235 | 1.2068 |
| P113N | 0.1251 | -1.5590 | 1.5343 |
| P47N | 0.2755 | -1.7580 | 0.0674 |

(Continuação da Tabela 3)

| OBS | FATOR 1 | FATOR 2 | FATOR 3 |
|---|---|---|---|
| P16N | -0.5347 | 0.1057 | -0.4276 |
| P22N | -0.6071 | -0.3407 | -0.0585 |
| P73N | -0.3723 | -0.0562 | -0.4774 |
| P51N | -0.0208 | -0.4763 | -1.1008 |
| P33N | -0.4407 | 0.1956 | -0.1574 |
| P03N | -1.0426 | 0.8338 | -0.3459 |
| P04N | -1.3526 | 0.2344 | -0.4736 |
| P07N | -1.4203 | 0.1786 | -0.3465 |
| P08N | -0.8684 | 0.0423 | -0.4973 |
| P09N | -1.6374 | -0.1669 | -0.6115 |
| P17N | -1.1908 | 0.3838 | -1.1265 |
| P28N | -1.8164 | -0.6952 | -1.2197 |
| P30N | -1.7125 | -0.2273 | -0.0654 |
| P45N | -0.8312 | 0.6061 | -0.7444 |
| P28N | -1.2883 | -0.2258 | 0.0831 |
| P53N | -1.1010 | 0.4807 | -0.7579 |
| P41N | -0.9468 | 0.8509 | -0.4569 |
| P18N | -0.4076 | 0.9624 | -0.4625 |
| P19N | -0.7436 | 0.9987 | -0.3678 |
| P11N | 0.8473 | -1.6007 | -0.4795 |
| P12N | 1.0899 | -1.5146 | -1.4315 |
| P23N | 1.1170 | -1.0006 | -0.7324 |
| P34N | 1.1136 | -1.0530 | -0.5690 |
| P39N | 0.9051 | -0.3725 | -0.7487 |
| P15N | 1.7103 | -0.7070 | -1.1267 |
| P20N | 1.0127 | -0.9366 | -0.9996 |
| P24N | 1.7457 | -0.7229 | -1.0158 |
| P27N | -0.0659 | -2.9699 | 2.7557 |
| P90N | 1.1230 | -1.3202 | 1.2040 |

(Continuação da Tabela 3)

| OBS | FATOR 1 | FATOR 2 | FATOR 3 |
|---|---|---|---|
| P33S | 2.2875 | 1.0946 | 0.3388 |
| P35S | 1.0681 | -0.4373 | -0.4122 |
| P36S | 1.8663 | 0.5297 | 0.1418 |
| P39S | 0.5827 | 0.0227 | 0.6514 |
| P40S | 0.2237 | -0.0680 | 0.5006 |
| P41S | -0.3939 | -0.7281 | 1.3238 |
| P42S | 0.2767 | -1.3301 | 1.0218 |
| P43S | -0.7945 | 0.4308 | 0.5781 |
| P44S | -1.3644 | 0.0602 | 0.2191 |
| P45S | -0.3142 | 1.0825 | 1.0283 |
| P46S | -1.3368 | 0.4009 | 1.1128 |
| P47S | 0.2574 | 0.0318 | -0.1397 |
| P48S | -0.8216 | -1.3574 | -0.1996 |
| P49S | -0.6245 | -0.6008 | 0.1206 |
| P52S | -0.9698 | 0.7006 | 0.6893 |
| P53S | -0.1586 | 0.9426 | 1.2306 |
| P54S | -0.6548 | 0.7364 | 0.4747 |
| P1PB | -0.5801 | -0.1711 | -0.2528 |
| P2PB | -1.0497 | 0.1528 | -0.3858 |
| P2PB | -1.1150 | 0.6306 | -0.3626 |
| P4PB | -0.1031 | -0.3961 | 4.9364 |
| P5PB | -0.8635 | -0.6622 | -0.1721 |
| BR1R | -0.4411 | -1.0157 | -1.5391 |
| BR21 | . | . | . |

# R E F E R Ê N C I A S

ANDERSON, T.W. - *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, Inc., New York, (1958).

BINDER, D.A. - *Bayesian Cluster Analysis.* Biometrika, vol. 65, (1978), 31-38.

BRYAN, J. - *Classification and Clustering Using Density Estimation.* Tese de Doutoramento, University of Missouri, (1971).

CORMACK, R.M. - *A Review of Classification.* Journal of the Royal Statistical Society, Series A, 134, (1971), 321-367.

DALENIUS, T. - *The Mode - A Neglected Statistical Parameter.* Journal the Royal Statistical Society, Serie A, 128, 110-117.

DAY, N.E. - *Estimating the Componentes of a Mixture of Normal Distributions.* Biometrika, vol. 56, (1969), 463-474.

DURAN, B.S. and ODELL, P.L. - *Cluster Analysis, A Survey.* Springer Verlag, New York, (1974).

EDWARDS, A.W.F. and CAVALLI-SFORZA, L.L. - *A Method for Cluster Analysis.* Biometrics, vol. 21, nº 2, (1965), 362-375.

ENGELMAN, L. and HARTIGAN, J.A. - *Percentage Points of a Test for Clusters.* Journal of American Statistical Association, vol. 64, (1969), 1947-1948.

EVERITT, B.S. - *Cluster Analysis.* Heinemann, London, (1974).

EVERITT, B.S. - *Unresolved Problems in Cluster Analysis.* Biometrics, vol. 35, (1979), 169-181.

FISHER, W.D. - *Clustering and Aggregation in Economics.* The John Hopkins Press, Baltimore, Maryland, (1968).

FRIEDMAN, H.P. and RUBIN, J. - *On Some Invariant Criteria for Grouping Data.* Journal of American Statistical Association, vol. 62, (1967), 1159-1178.

HARTIGAN, J.A. - *Representation of Similarity Matrices by Trees.* Journal of American Statistical Association, vol. 62, (1967), 1140-1158.

HARTIGAN, J.A. - *Clustering Algorithms.* John Wiley and Sons, Inc., New York, (1975).

JARDINE, N. and SIBSON, R. - *Mathematical Taxonomy.* John Wiley and Sons, New York, (1971).

KENDALL, M.G. and STUART, A. - *The Advanced Theory of Statistics.* vol. III, Charles Griffin & Co., Ltd., London, (1961).

MARRIOT, F.H.C. - *Practical Problems in a Method of Cluster Analysis.* Biometrics, vol. 27, (1971), 501-514.

MAC QUEEN, J.B. - *Some Methods for Classification and Analysis of Multivariate Observations.* Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, (1967), 281-297.

MORRISON, D.F. - *Multivariate Statistical Methods*. McGraw-Hill Book Company, New York, (1976).

ROBERTSON, T. and CRYER, J. - *An Iterative Procedure for Estimating the Mode*. Journal of American Statistical Association, 6p, 1012-1016.

SCOTT, A.J. and SYMONS, M.J. - *Clustering Methods Based on Likelihood Ratio Criteria*. Biometrics, vol. 27, nº 2, (1971), 387-398.

SINGER, J.M. - *Análise de Curvas de Crescimento*. Tese de Mestrado, USP, (1977).

THORNDIKE, R.L. - *Who Belongs in a Family*. Psychometrika, 18, (1953), 267-296.

TRYON, R.C. - *Cluster Analysis*. Ann Arbor, Edwards Bros., (1939).

TRYON, R.C. and BAYLEY, D.E. - *Cluster Analysis*. McGraw-Hill Book Company, New York, (1970).

WOLFE, J.H. - *Pattern Clustering by Multivariate Mixture Analysis*. Multivariate Behavioral Research, vol. 5, nº 3, (1970), 329-350.

ZUBIN, J. - *A Technique for Measuring Likemindedness*, J. Abnormal Soc. Psychology, 33, (1938), 508-516.