UNIVERSIDADE FEDERAL DE LAVRAS

AVALIAÇÃO DE UM CRITÉRIO PROBABILÍSTICO EM ANÁLISE MULTIVARIADA DE AGRUPAMENTO (*CLUSTER ANALYSIS*), POR MEIO DE SIMULAÇÃO MONTE CARLO

MOISÉS MOURÃO JR.

2001

MOISÉS MOURÃO JR.

AVALIAÇÃO DE UM CRITÉRIO PROBABILÍSTICO EM ANÁLISE MULTIVARIADA DE AGRUPAMENTO (*CLUSTER ANALYSIS*), POR MEIO DE SIMULAÇÃO MONTE CARLO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Mestrado em Estatística e Experimentação Agropecuária, área de concentração em Estatística Aplicada em Genética e Melhoramento de Plantas, para obtenção do título de "Mestre".

Orientador

Prof. Dr. Daniel Furtado FERREIRA

LAVRAS
MINAS GERAIS - BRASIL
2001

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Mourão Junior, Moisés

Avaliação de um critério probabilístico em análise multivariada de agrupamento (*Cluster analysis*), por meio de simulação Monte Carlo / Moisés Mourão Junior.-- Lavras : UFLA, 2001.

79 p. : il.

Orientador: Daniel Furtado Ferreira.

Dissertação (Mestrado) – UFLA.

Bibliografia.

1. Análise de agrupamento. 2. T^2 de Hoteling. 3. D^2 de Mahalanobis. 4. Simulação Monte Carlo. 5 Método assintótico. I. Universidade Federal de Lavras. II. Título.

CDD-519.282

MOISÉS MOURÃO JR.

AVALIAÇÃO DE UM CRITÉRIO PROBABILÍSTICO EM ANÁLISE MULTIVARIADA DE AGRUPAMENTO (*CLUSTER ANALYSIS*), POR MEIO DE SIMULAÇÃO MONTE CARLO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Mestrado em Estatística e Experimentação Agropecuária, área de concentração em Estatística Aplicada em Genética e Melhoramento de Plantas, para obtenção do título de "Mestre".

APROVADA em 08 de fevereiro de 2001

Prof. Dr. João Bosco dos Santos UFLA

Prof. Dr. Marcelo Tavares UFU

Prof. Dr. Daniel Furtado Ferreira
UFLA (Orientador)

LAVRAS
MINAS GERAIS - BRASIL

"A criação é um tipo de expansão do Um.
(...)
O Um não deve estar só: se assim fosse, todas as coisas permaneceriam ocultas, não havendo forma no Um."

[Plotino *in* Eneidas II, i; III, x]

"Eu canto a mim mesmo e celebro a mim mesmo
E o que eu assumo também deves assumir
Por que todo átomo que a mim pertence, também pertence a ti.
(...)
Minha língua, cada átomo do meu sangue, formado desta terra, deste ar.
[Walt Whitman *in* Canto a Mim Mesmo; Folhas de Relva, Livro III]

À mulher e fêmeas da minha vida, respectivamente: Néa, Amélia, Passiflora, Lucrécia, Alice, Luna e Euterpe. Uma noite com vocês é inesquecível!

Ao meu pai, Moisés *Primus*, pela eterna confiança. À minha mãe, Graça, por sempre fazer jus ao nome que tem. À minha irmã, Giovana, por sempre ter algo a acrescentar. À minha outra irmã, Suzy, por ter me dado o sobrinho mais lindo do mundo, o Moisés *Tertius*. Ao Moisés *Tertius* por ser o que é.

Aos amigos, Lucíolo Camarão Brabo e Klaus Franz Alfred Geissler, por sempre estarem tão longe, tão perto ...

Á minha vó, Celina, que do alto de sua sábia ignorância achou que eu era "demais burro" por passar "dois anos estudando só Matemática!¿!?!".

Aos vivos, que lutam contra os cadáveres adiados que procriam.

À Lucinéa, que sempre foi meu consolo nos momentos mais difíceis. Este agradecimento é tão pequeno. Queria dizer mais, mas não sei como. Com certeza sabes o que devo estar sentindo, bem como o fato de seres o leito e o ascendente.

Ao 001001101101 11111001011100 001110011011000010 (Daniel Furtado Ferreira, em codificação binária), o algorítimico, pela humanidade e profissionalismo, que sem querer acabou me provando que orientador não é alguém com quem se compete. Além de agradecer, respeito seu trabalho.

Aos amigos do curso, da minha geração: Adriano, Aladir (*Horácio*) [o professor, no sentido mais estrito], Alex, Ana Rita, Andréa (Andreinha), Cristiane, Everton, Iara, Nagib e Teixeira; os da geração antiga: Ivani, Carlos, Mônica, Andréa, Marcelo e Letícia e aos da nova geração: Ceile, Douglas (*Coelhinho*), Ednaldo, Flávio (*Tipim*), José Marcelo (*Banana*), Lívia, Marcelo, Marcos, Paulo César, Paulo José e Sérgio; aos que virão: Denismar, Fabiano, Marielli e Roberta. Além dos honorários: Francisco (*Frank Zappa*) [o grande etáliecoisa], Ivana, João Luís, Glauber, Wladmir, Renato e Waldemar. A companhia de vocês foi fundamental.

Aos professores do curso: Augusto Ramalho por sua calma caraterística; Eduardo Bearzoti por ensinar a olhar as pedras do caminhos; Joel Muniz por cuidar de tudo e todos; Júlio Bueno pelos papos e derrotas no xadrez; Lucas Chaves pelo amor a Ecologia via Matemática; Luís Henrique de Aquino por ser sempre professor; Mário Javier-Vivanco pelas conversas; Ruben Delly pelo incentivo na profissão e na nobre carreira de "bikeiro"; Thelma Sáfadi porque ainda vai ser minha parceira de gamão.

Aos amigos da Ecologia e os da esquina com a Biologia Evolutiva: Marcelo, Júlio, Eduardo Van den Berg e Newton pelos papos, sempre, interessantes. *Darwinian Fields Forever*!

À Fundação de Amparo a Pesquisa de Minas Gerais «FAPEMIG», pelo válido incentivo sob a forma de bolsa. Mesmo que criticada auxiliou de maneira decisiva a independente Fundação Mourão Jr. para o Desenvolvimento da Ecologia na Amazônia, de parcos recursos e muitas expectativas.

Aos (in)confidentes: Karl Popper, Rudolf Carnap, Gaston Bachelard, Nélson Papavero, Abelardo, São Tomás de Aquino, Francis Bacon, George Berkeley, Paulinho da Viola e todos os grandes sambistas por não me fazerem esquecer que modelos são representações.

SUMÁRIO

RESUMO

MOURÃO JR., Moisés. **Avaliação de um critério probabilístico em análise multivariada de agrupamento (*cluster analysis*), por meio de simulação Monte Carlo**. Lavras: UFLA, 2001. 78p. (Dissertação - Mestrado em Estatística e Experimentação Agropecuária)[1]

Em pesquisa agropecuária, a utilização de um único critério nem sempre carateriza um julgamento adequado do fenômeno. Assim técnicas multivariadas acenam como uma alternativa a este problema. Estas técnicas consistem em simplificação, tanto de estruturas de variáveis, quanto de objetos. As técnicas de agrupamento são aplicadas na determinação de afinidade entre grupos de objetos. Entretanto sua representação e interpretação corrente são impregnadas de subjetivismo. Distâncias métricas estatísticas são a alternativa mais informativa dentre as outras, assim a distância generalizada de Mahalanobis (D^2)apresenta-se como uma alternativa praticável em análise de agrupamento. O trabalho apresenta um estudo de simulação para a obtenção de uma aproximação da distribuição empírica de D^2 de Mahalanobis pela T^2 de Hotelling via F e χ^2, sendo esta consistente e acurada. Sendo utilizadas configurações com número de objetos iguais a $p+1 \leq n \leq 250$ e número de variáveis $2 \leq p \leq 10$. A descrição da distribuição empírica apresentou sintonia com a de T^2 de Hotelling e sua aderência foi assinalada, pela regra empírica $n>80$. O número de objetos e a de variáveis apresentou elevada influência no ajuste das distribuições empírica e da aproximação. Métodos hierárquicos, como o ligação completa, UPGMA e de Ward apresentaram valores de classificação correta equivalente ou superiores aos considerados como teóricos, fornecidos pela análise de discriminantes. O método particional *k-means* não apresentou resultados satisfatórios. A estrutura dos dados regeu a discriminação dos métodos de agrupamento empregados, entretanto alguns como o de Ward de ligação completa mantiveram um limite aceitável, mesmo com a menor esfericidade, por extensão, maior heterogeneidade das populações avaliadas. Recomenda-se a utilização do critério através da aproximação via χ^2 por sua maior aderência, sob o limite empiricamente determinado, e por sua facilidade de implementação.

Palavras-chave: análise de agrupamento, *cluster analysis*, T^2 de Hoteling, D^2 de Mahalanobis, métodos assintóticos, simulação Monte Carlo

---

[1] Comitê orientador: Daniel Furtado FERREIRA, UFLA

ABSTRACT

MOURÃO JR., Moisés. **Evaluation of a probabilistic criterion in multivariate cluster analysis, by Monte Carlo simulation.** Lavras: UFLA, 2001. 78p. (Dissertation - Master in Statistics and Agronomical Experimentation)[2]

In agricultural research, the use of only one criterion not always characterizes an appropriate judgement of phenomenon. Thus, multivariate techniques present as an alternative to this problem. These techniques consist of simplification, both of variables and objetcs structures. The grouping techniques are applied in the similaritiness determination among groups of objects. However its current representation and interpretation are impregnated of subjetivism. Statistical metric distances are the most informative alternative among the other ones, like this the Mahalanobis' generalized distance (D^2) comes as a practical alternative in cluster analysis. The work presents a simulation study for the obtaining of an approach of the empiric distribution of Mahalanobis' D^2 based on Hotelling's T^2 through F and χ^2, being this consistent and acurated. Configurations were used with number of objects equal $p+1 \leq n \leq 250$ and number of variables $2 \leq p \leq 10$. The description of the empiric distribution presented syntony with the one of Hotelling's T^2 and its adherence was marked, for the empiric rule $n>80$. The number of objects and variables presented high influence in the adjustment of the empiric distributions and of the approach. Hierarchical methods, as the single linkage, UPGMA and Ward's, presented values of equivalent correct classification or superiors to the considered as theoretical, supplied by the discriminant analysis. The partitional method k-means didn't present satisfactory results. The structure of the data governed the discrimination of the employed grouping methods, however some like Ward's and complete linkage maintained an acceptable limit, even with the smallest spheracity, for extension, larger heterogeneity of the appraised populations. The use of the criterion is recommended through the χ^2 approach for its largest adherence, under the empirically determined limit, and for its implementation easiness.

Key-words: *cluster analysis*, Hotelling's T^2, Mahalanobis' D^2, assynthotic methods, Monte Carlo simulation

[2] Guidance Committee: Daniel Furtado FERREIRA, UFLA

1 INTRODUÇÃO

Os fenômenos naturais são estreitamente influenciados e associados a diversos efeitos. Deste modo, sua mensuração e expressão, devem ser concordantes com este paradigma. O enfoque multivariado surgiu como alternativa a esta questão, representando os fenômenos sob influência da realização de várias variáveis. As técnicas multivariadas disponíveis, de modo geral, habilitam o usuário a: (i) reduzir e simplificar dados, (ii) reunir e classificar grupos, (iii) investigar dependência entre variáveis, (iv) gerar modelos de predição e (v) testar hipóteses, sendo frequente o uso conjunto destas técnicas no decorrer da análise (Johnson Wichern, 1998).

Este enfoque muito mais preciso muitas vezes apresenta-se como um complicador, já que sua característica multidimensionalidade comumente não pode ser expressa em uma noção de espaço mais simplificada. Técnicas como o agrupamento (*cluster analysis*) apresentam a vantagem de reduzirem o espaço multidimensional a uma medida de distância entre os objetos, representando esta em um espaço bidimensional, muito mais simplificado do que o espaço multidimensional (Cormack, 1971; Mardia, Kent e Bibby, 1995). Esta capacidade de sumarização é o grande atrativo desta técnica multivariada, o que lhe confere grande aplicabilidade e difusão em diversos ramos da Ciência (Everitt, 1979; Manly, 1994). A definição de agrupamento adotada no trabalho, refere-se a arranjos entre objetos, dispostos em um espaço multidimensional, p-variado ou euclidiano, sem nenhuma definição de arranjo dos objetos *a priori* (Giri, 1996).

Como resultado da análise de agrupamento, tem-se o dendograma, que apresenta o arranjo entre os objetos em uma escala de distância. Este arranjo indica apenas afinidade entre os grupos, não definindo nenhuma ordenação entre estes. O caráter heurístico do resultado da análise de agrupamento é indicado

pelas inferências cabíveis: (i) esclarecimento de um dado fenômeno avaliado, (ii) geração de novas hipóteses, (iii) planejamento e organização de uma estrutura, baseada na disposição dos objetos e (iv) confecção de uma lista de categorias ou objetos afins (Cormack, 1971).

Sua interpretação é destituída de qualquer caráter probabilístico, já que sua escala é, comumente, definida como o somatório dos quadrados de diferenças entre pares de objetos, e de interpretação muitas vezes subjetiva, o que torna a técnica passível de críticas no que diz respeito à detectação de agrupamentos legítimos, estando muito mais sujeita a percepção do usuário. A técnica de agrupamento em si apresenta um apelo visual muito forte (prova disto são os recursos gráficos largamente utilizados no decorrer do trabalho); deste modo, a representação gráfica de similaridade ou dissimilaridade entre os objetos e mais especificamente de grupos de objetos afins mais polarizados, contribuem como forte critério de decisão (Kruskal e Landwehf, 1983; Lebart, Morineau e Warwick, 1984). Entretanto, como já citado anteriormente, sua escala não apresenta nenhuma propriedade probabilística, o que reitera a subjetividade da técnica (Forgy, 1965).

Mesmo com este caráter heurístico e profundamente subjetivo, três características, baseadas no procedimento fenético, são requeridos para a execução de uma análise de agrupamento efetiva e consistente (Sneath e Sokal, 1973), a saber: (i) objetividade, através da qual experimentador subsequente deve obter as mesmas conclusões quando comparadas as conclusões de um experimentador original, (ii) estabilidade, através da qual a análise subsequente deve refletir as mesmas conclusões ou padrões da análise original, dada a inclusão de uma nova variável ou caracter e (iii) preditibilidade, que promove inalteração do padrão ou conclusão iniciais, em uma análise subsequente, dada a inclusão de uma nova categoria. De modo geral, estas características são

cumpridas na íntegra, garantindo a determinação da estrutura latente de um fenômeno, ou seja, da organização, do padrão de comportamento deste, o que é a base do pensamento científico atual (Dolby, 1982).

A dissertação versa sobre as propriedades fundamentais da análise de agrupamento, avalia procedimentos anteriormente propostos e tem como objetivo principal propor um critério probabilístico para a determinação de agrupamentos legítimos, baseado na distribuição empírica da distância generalizada de Mahalanobis (D^2), bem como avaliar este critério em função de métodos consagrados na literatura, como os hierárquicos, particionais e discriminantes.

2 REFERENCIAL TEÓRICO

2.1 Espaço multidimensional

Um conjunto de dados, tanto de natureza univariada quanto multivariada, pode ser expresso via geometria vetorial (Bryant, 1984; Saville e Wood, 1986). A geometria de amostras multivariadas apresenta um espaço do tipo multidimensional, o que lhe confere maior complexidade. As realizações das variáveis medidas são vetores expressos em um plano ou espaço euclidiano (Figura 1).

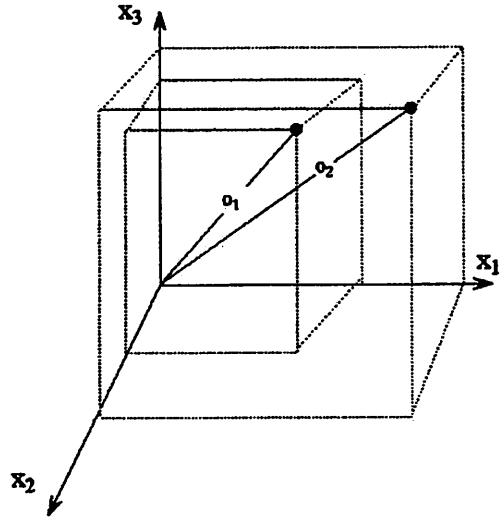

Figura 1 Realização de variáveis em um espaço euclidiano R^3

Deste modo, o arranjo entre populações ou objetos provém de sua disposição em um hiperespaço. A percepção neste caso é comprometida pela própria natureza dos dados, já que representações gráficas são perceptíveis em até três dimensões (Chatfield e Collins, 1986; Johnson e Wichern, 1998). Algumas alternativas a este problema têm sido propostas (Figura 2), como as

faces de Chernoff (Chernoff, 1973) ou ajustes do tipo Fourier (Andrews, 1972), consistindo de uma representação gráfica de cada observação através dos valores das várias variáveis mensuradas, entretanto estas não têm apresentado efetiva aplicação ou facilidade de interpretação (Manly, 1994).

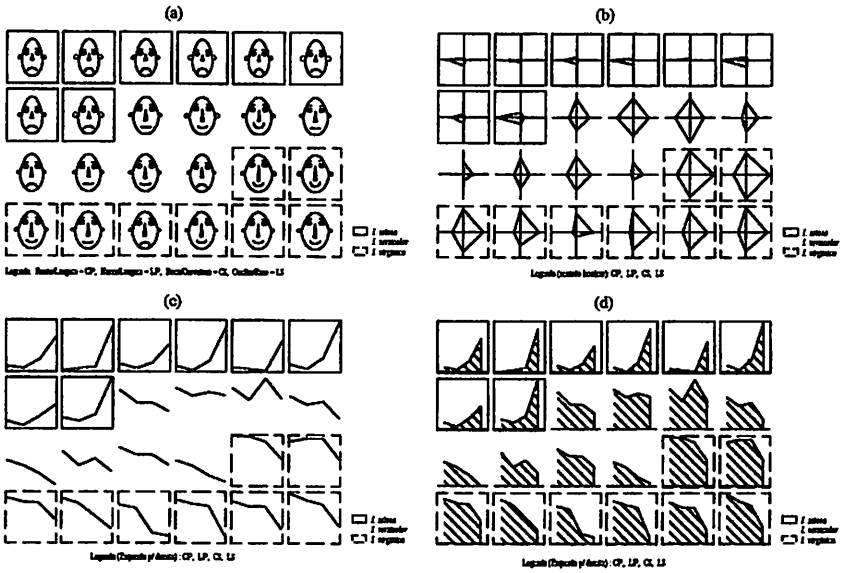

Figura 2 **Representações alternativas de processos multivariados sob a forma de iconográficos: (a) faces de Chernoff, (b) raio de sol, (c) linhas e (d) perfil de algumas observações do banco de dados *Iris***

A notação aplicada é definida pelo uso de uma matriz de dados com n valores de objetos como linhas e p valores de variáveis nas colunas (Figura 3.a). De modo geral, as técnicas multivariadas podem ser reduzidas a um princípio de simplificação, para o qual p variáveis e n observações ou objetos ou casos são reduzidos a grupos afins de variáveis, objetos ou observações. A escolha da técnica apropriada está intimamente vinculada à natureza dos dados e à proposição do usuário. Uma categorização inicial é empregada definindo: (i) técnicas variável-dependente, ditas técnicas-R, através das quais são avaliadas estruturas de covariância ou correlação entre as variáveis e (ii) técnicas

indivíduo-dependente, ditas técnicas-Q, para as quais distâncias entre indivíduos, objetos, listas em função das variáveis mensuradas são empregadas (Figura 3.b) (Pielou, 1984).

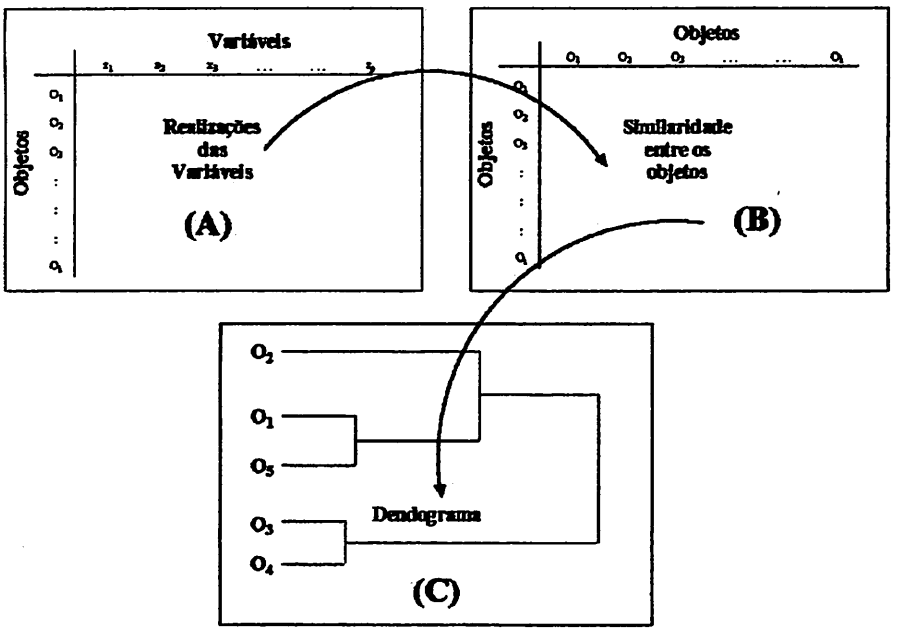

Figura 3 Passos da análise de agrupamento

2.2 Agrupamento, classificação e dissecação

A análise de agrupamento situa-se como uma técnica indivíduo-dependente, na qual valores de distâncias, sob a forma de matrizes, entre os objetos são arranjados. A estimação de parâmetro não é requerida, neste caso, o que lhe ratifica o caráter não-probabilístico (Chatfield e Collins, 1986). O fracionamento de um conjunto de dados, de unidades de observação ou casos em subconjuntos ou grupos homogêneos é o objetivo principal desta análise, definindo-se, assim, uma maior homogeneidade dentro do subconjunto e maior heterogeneidade em relação a outros subconjuntos (Fisher, 1958; Mardia, Kent e Bibby, 1995). Uma distinção cabível refere-se a conceitos tomados como

sinônimos, quais sejam: agrupamentos natural e legítimo, a adoção do conceito de naturalidade do agrupamento está associada a qualquer arranjo entre subconjuntos, separados por um critério objetivo ou não e legitimidade somente associada a subconjuntos definidos por critérios objetivos, como os probabilísticos (Chatfield e Collins, 1986).

Outra distinção refere-se à adoção dos termos classificação, agrupamento e dissecação. Define-se e adota-se classificação, como uma disposição ordenada entre objetos de maior ou menor afinidade em função de um ou mais atributos, sendo que esta ordenação reflete algum padrão entre os objetos e seus subconjuntos. Agrupamento é a disposição não necessariamente ordenada entre os objetos, os subconjuntos formados não tributam nenhuma informação a não ser sua afinidade latente. Já a dissecação refere-se exclusivamente à separação de objetos em subconjuntos, normalmente através da inspeção isolada de cada atributo. De modo geral, a classificação está associada à detecção de agrupamentos legítimos; o agrupamento é um dos meios de inferir sobre a existência destes e a dissecação está mais associada a agrupamentos naturais, em que a principal preocupação há a não ser a separação dos objetos (Cormack, 1971; Everitt, 1981).

Assim, os subconjuntos, dada sua legitimidade, apresentam zonas no hiperespaço com uma maior densidade de indivíduos, também assinalando-se zonas de menor densidade separando estes subconjuntos (Johnson e Wichern, 1998). Diferentes estruturas de agrupamento podem ser assinaladas, como grupos bem separados ou polarizados, formando agrupamentos esféricos (Figura 4.a) que são detectáveis até pela inspeção visual dos dados. Outras estruturas como agrupamentos pobremente separados (Figura 4.b) apresentam determinação mais difícil. Além destas, os agrupamentos elipsoidais ou elongados (Figura 4.c), que se apresentam sob grande influência da correlação entre as variáveis, os

agrupamentos com diferentes números de objetos ou variabilidade (Figura 5) e agrupamentos com estrutura não convencional (Figura 4.d), também assinalam dificuldade em sua determinação ou discriminação (Sarle, 1990).

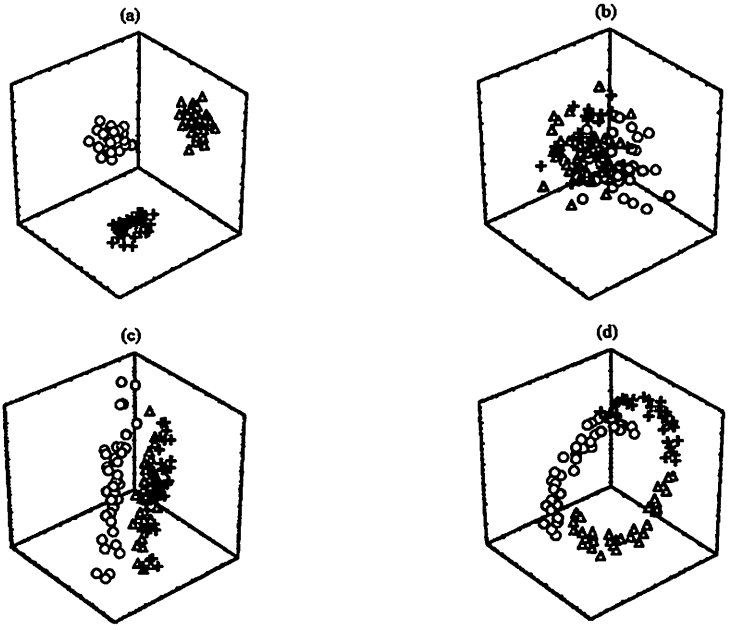

Figura 4 **Estruturas de agrupamento (a) esférico, (b) pobremente separados, (c) elipsoidal e (d) não convencional**

A implementação de extensões e particularizações é comumente observável. O desenvolvimento da técnica de análise de agrupamentos ocorre nos anos 60, com a disponibilização de ferramentas computacionais propícias e neste momento começam a surgir trabalhos de natureza aplicada e questionamentos sobre a validade das determinações fornecidas por esta. Críticas e alternativas vêm sendo apontadas desde então e atualmente algumas linhas de pesquisa têm apontado outros caminhos e formalizado alternativas para o manuseio de problemas de classificação.

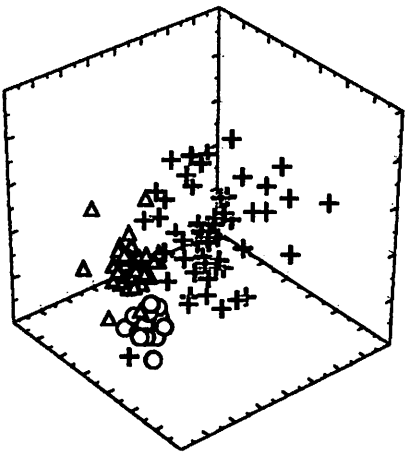

Figura 5 **Estrutura de agrupamento com diferente número de objetos e variabilidade dos dados**

A abordagem não paramétrica assinala alternativas definindo agrupamentos como modas (Silverman, 1992) e lança mão de procedimentos inspecionais para a determinação destes agrupamentos (Mardia, 1970; Hartingan e Hartingan, 1985), especialmente os com estrutura mais complexas ou que promovam ruptura da pressuposição de multinormalidade. Os estudos de estimação de densidade também apresentam-se como uma alternativa formal ao problema da determinações dos agrupamentos, e surgem como uma alternativa mais viável, especialmente nos problemas de ruptura da multinormalidade (Silverman, 1992).

Algumas alternativas, no caso univariado, já se apresentam bastante estruturadas, tais como os procedimentos de meta-análise, que consistem na análise conjunta de diferentes resultados citados na literatura, desde que estes guardem em comum ou variáveis classificatórias ou resposta. Este método de análise apresenta sensibilidade, apesar da baixa robustez na detectação de agrupamentos legítimos (Schwarzer, 1989; Mann, 1994; Manly, 1994; Friedman e Goldberg, 1996). Entretanto nenhum correspondente multivariado foi proposto.

Procedimentos baseados em redes neurais (*neural network*) apresentam a proposição fundamental na alocação de cada objeto em um determinado grupo, baseado em critérios de propagação da informação, no caso, similiaridade. A terminologia tem inspiração biológica, mas as sinapses, neste caso, apresentam-se associadas a probabilidades condicionais e não a potássio e cálcio. Diversos algoritmos têm aplicação no caso de classificação, entre eles as redes neurais lineares; redes neurais probabilísticas PNN; Multilayer Perceptrons (MLP); redes neurais do tipo função radial de base (RBF) e rede neural de Kohonen (StatSoft, 1996). Outra alternativa é a classificação automática, que não se relaciona diretamente com o conceito comum de automático, no sentido de imediato, mas sim de autômato. Estas técnicas têm sido empregadas largamente em estudos de inteligência artificial (IA), consistindo tanto de técnicas comuns e clássicas de classificação quanto de novas técnicas de classificação não supervisionada, como os ISODATA (*Iterative Self-Organization Data Analysis*) (Rower, Wynne-Jones e Wysotzki, 1994).

Os paradigmas atuais em análise de agrupamento referem-se especificamente ao poder de classificação, consistindo em determinar de maneira mais efetiva rupturas entre os subconjuntos, legitimando-os (Everitt, 1981). Algumas abordagens recentes, apresentam modelagem diferente de enfoque clássico de classificação, como a de *mixture-models*, em que a separação de grupos é feita através de médias de um modelo de mistura de distribuições, chamado modelo-mistura, e a *model-based*, consistindo de técnicas com objetivo de determinar a estrutura latente dos dados, no caso de classificação; A técnica pretende fornecer informações sem a definição de nenhuma *priori*; entretanto, tratando-se de um método iterativo, julga obter *posterioris* válidas, através do critério bayesiano de informação (BIC) (Fraley, 1998; Fraley e Raftery, 1998, 1999).

Além da análise de agrupamento, outras técnicas multivariadas, reunidas sob o rótulo de "ordenação", têm-se prestado como procedimentos classificatórios. Técnicas de redução de número de variáveis, como a análise fatorial e as variáveis canônicas, prestam-se na inspeção de agrupamentos e também em testes de hipóteses. Como estas também apresentam caráter intrínseco, em que não é necessária nenhuma pressuposição sobre os agrupamentos, a combinação destas técnicas é de uso recomendado na literatura (Lebart, Morineau e Warwick, 1984).

Já técnicas de caráter extrínseco, como a análise de discriminantes, não apresentam a mesma resposta. Esta análise consiste na determinação de quais variáveis resposta ou atributos discriminam de maneira efetiva um grupo de objetos, populações definidas *a priori*. Assim, o caráter heurístico da análise de agrupamento é totalmente desassociado na técnica de análise de discriminantes (Everitt, 1981; Manly, 1994). O uso desta técnica, na maioria das vezes, é efetuado após a condução de uma análise de agrupamentos, e o resultado desta inspeção ratifica a geração de hipóteses derivadas pela análise de agrupamentos. Assim, o uso de diferentes técnicas multivariadas presta-se de maneira efetiva nos procedimentos de ordenação ou classificação.

2.3 Propriedades das distribuições multivariadas

Como visto anteriormente, as realizações de um fenômeno de natureza multivariada são avaliadas de maneira conjunta. Deste modo, o tratamento estatístico consiste de uma extensão do caso univariado, na qual são consideradas as dependências entre as variáveis (Johnson e Wichern, 1998).

A abordagem paramétrica univariada centra-se na distribuição normal de probabilidade (2.1), com os parâmetros $N(\mu,\sigma)$ média e variância,

respectivamente. Esta centralização deve-se ao fato de esta distribuição ser completamente descrita com apenas os seus dois primeiros momentos, o que torna o cômputo muito mais simplificado, sendo que a utilização de momentos de ordem superior fornece informações adicionais como a forma e escala da distribuição (Johnson e Kotz, 1970a).

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-\left[\frac{(x-\mu)}{\sigma}\right]^2}{2}} \qquad -\infty<x<\infty \tag{2.1}$$

A distribuição apresenta o primeiro membro constante, sendo o segundo responsável pela conformação da distribuição ao longo das realizações da variável mensurada. Em uma abordagem matricial, todos os valores de (2.1) são definidos como escalares. Assim, podemos definir o componente estocástico no segundo membro como $(x-\mu)'(\sigma^2)^{-1}(x-\mu)$, esta uma distância quadrática (Anderson, 1984).

No caso multivariado, o correspondente engloba o número das *p* variáveis consideradas e os valores dos parâmetros, que agora correspondem a vetores e matrizes. A distribuição normal multivariada (2.2), representada na Figura 6, apresenta como parâmetros $N_p(\mu,\Sigma)$, onde μ é o vetor paramétrico de médias e Σ é a matriz de covariâncias

$$f(x)=\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp-\left\{\frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{2}\right\} \tag{2.2},$$

sendo *p* a dimensão no hiperespaço, $(x-\mu)'\Sigma^{-1}(x-\mu)$ uma distância quadrática generalizada e o $|\Sigma|$ representa uma variância generalizada (Johnson e Wichern, 1998).

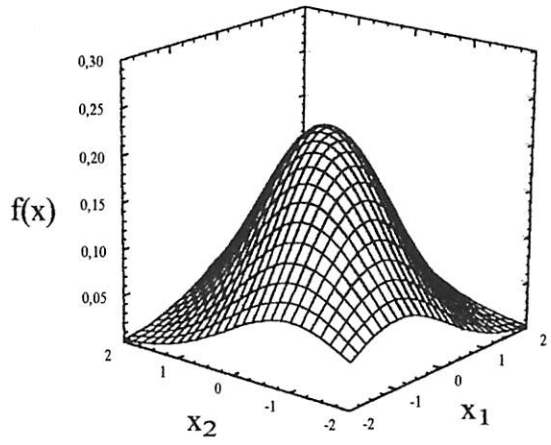

Figura 6 Função densidade de probabilidade da distribuição normal bivariada

Algumas propriedades de interesse da normal multivariada seguem: (i) combinações lineares dos componentes de x são normalmente distribuídos; (ii) todos os subconjuntos dos componentes de x têm uma distribuição normal multivariada; (iii) covariância nula implica em os componentes correspondentes serem independentemente distribuídos e (iv) as probabilidades condicionais dos componentes seguem uma normal multivariada ou univariada (Anderson, 1984; Mardia, Kent e Bibby, 1995; Johnson e Wichern, 1998).

Dado um conjunto de variáveis aleatórias normal univariada independentes, então tomando-se seu quadrado e soma, tem-se que a variável resultante segue uma distribuição de χ^2. Assim a variância, que é o produto de operações desta natureza, é representada por esta distribuição. A distribuição de χ^2 é dada por $\chi^2_{(v)}$ (2.3), em que v é o número de grau de liberdade, e ainda sendo considerada assimétrica à direita. A média desta distribuição é o número de graus de liberdade e sua variância equivale a duas vezes o número de graus de liberdade (Johnson e Kotz, 1970b). Uma propriedade da distribuição χ^2 é que uma distribuição com v_1 graus de liberdade pode ser adicionada a uma outra

distribuição com v_2 graus de liberdade, gerando uma nova distribuição de χ^2 com $v=v_1+v_2$ graus de liberdade

$$\chi^2 = \left\{ \frac{1}{\left[2v^{\frac{1}{2}} \Gamma\left(\frac{v}{2}\right) \right]} \right\} \left[x\left(v^{\frac{1}{2}}\right)^{-1} e^{\frac{-x}{2}} \right] \qquad v=1, 2, \ldots < x \qquad (2.3),$$

sendo v o número de graus de liberdade e Γ a função gama.

A variância no caso univariado é uma particularização da noção de covariância, em que a variância é tomada como a covariância de variável com ela mesma. Representações de variação no espaço multidimensional e tratamentos destas em modelos analíticos são mais difíceis. Assim, conceitos como variância generalizada são empregados a fim de solucionar o problema da multidimensionalidade.

A variância generalizada, geometricamente, pode ser representada pelo volume delimitado pelas variâncias marginais em um espaço multidimensional. Com fins algébricos, medidas como o determinante e traço de Σ podem ser empregadas como uma forma de representar a variação das variáveis de maneira conjunta (Johnson e Wichern, 1998).

A distribuição de χ^2 define a distribuição da variância, tendo na distribuição de Wishart seu correspondente multivariado, representando a distribuição das matrizes de covariância. Esta distribuição é denotada por $W_m(\cdot|\Sigma)$. Considerando uma matriz S, teríamos sua função densidade de probabilidade definida (2.4) em função da matriz considerada, do número de variáveis e das observações e da matriz de covariância.

$$W_{n-1}(S \mid \Sigma) = \frac{|S|^{\frac{(n-p-2)}{2}} e^{-tr\left[S\Sigma^{-1}\right]/2}}{2^{p(n-1)/2} \pi^{p(p-1)/4} |\Sigma|^{(n-1)/2} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n-i)\right)} \qquad (2.4),$$

sendo: S uma matriz qualquer positivo definida, p o número de variáveis, n o número de observações.

A aditividade prerrogada pela distribuição de χ^2 continua a ser válida para a distribuição de Wishart (Anderson, 1984; Giri, 1996).

O teste t fundamenta-se na diferença entre estimativas de parâmetros ponderados por sua variação (2.5), sendo o teste uma avaliação do grau de diferença entre as populações com base em uma combinação linear entre seus parâmetros. A esta diferença ponderada denominamos padronização.

$$t = \left(\frac{\bar{x}_i - \bar{x}_j}{\sqrt{s_p^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \right) \qquad (2.5) \qquad t^2 = \left(\frac{\left(\bar{x}_i - \bar{x}_j\right)^2}{\frac{s_p^2}{n_i + n_j}} \right) \qquad (2.6)$$

Esta padronização pode ser rescrita para uma aproximação de F, fundamentada no fato de que t^2=F, como (2.6). A estatística t, então, é definida como uma razão entre uma distribuição normal e uma $\sqrt{\chi^2}$, seguindo uma distribuição t de Student.

No caso multivariado, esta combinação linear é melhor expressa por uma função discriminante entre as populações, tendo nos centróides a representação do valor médio do caso univariado. Como diferença entre as funções discriminantes, tem-se uma distância generalizada de Mahalanobis (D^2).

Um teste formal para estas combinações lineares (2.7) é fornecido e apresenta distribuição $Z \simeq F_{(p, n_1+n_2-p-1)}$.

$$Z = \frac{n_1 n_2}{n_1 + n_2} \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} D^2 \qquad (2.7)$$

A partir da conhecida relação t^2=F, pode-se redefinir a expressão, para o caso multivariado (2.8), tendo, assim, um correspondente multivariado com distribuição $\approx F_{(p, n_1+n_2-p-1)}$.

$$T^2 = \frac{n_i n_j}{n_i + n_j} \left(\bar{x}_i - \bar{x}_j\right)' S^{-1} \left(\bar{x}_i - \bar{x}_j\right) \qquad (2.8)$$

Assim, correspondente multivariado à estatística t assinala-se a distribuição T^2 de Hotelling, de natureza generalizada e associada à normal multivariada. A extensão multivariada, define não mais escalares, como o caso univariado a média é representada por um vetor de parâmetros e a variância por uma matriz de covariância. Sua formalização, então, é dada pela combinação entre as distribuições correspondentes: normal multivariada e distribuição de Wishart (2.9).

$$T^2_{p,n-1} = N_p(0,\Sigma)' \left[\frac{1}{n-1} W_{p,n-1}(\Sigma)\right] N_p(0,\Sigma) \qquad (2.9)$$

No caso de um número igual de observações ou objetos nas populações, a aproximação de T^2 é dada por (2.10), sendo esta utilizada neste trabalho.

$$2\left[\frac{(2n-2)p}{(2n-p)}\right] F_{(\alpha; p, 2n-p)} \qquad (2.10)$$

Utilizou-se a aproximação via F por sua facilidade de implementação e por esta apresentar-se como bastante acurada, mesmo quando comparada com a exata (Hughes e Saw, 1971).

2.4 Distâncias

As distâncias são medidas utilizadas para a representação dos pontos na estrutura de similaridade. Esta medida representa o menor espaço entre dois pontos, sendo uma extensão do teorema de Pitágoras para o caso multidimensional. No caso bivariado, é definida por (2.11). A expressão desta expansão ao caso multivariado é apresentada na Tabela 3 ($d_{3(i,j)}$).

$$d_{i,j} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} \qquad (2.11)$$

Sendo que estas medidas de distância podem ser facilmente transformadas em medidas de associação ou similaridade, por meio de $\frac{1}{1+d_{ip}}$. Entretanto, a recíproca não é sempre verdadeira, devido ao fato das distâncias terem que, necessariamente satisfazer três condições:

(i) $d_{i,j} \geq 0$; $d_{i,j} = 0$ se i=j $\quad \forall i; \forall j$ positividade

(ii) $d_{i,j} = d_{j,i}$ $\quad \forall i; \forall j \quad i \neq j$ simetria

(iii) $d_{i,k} + d_{j,k} \geq d_{i,j}$ $\quad \forall i; \forall j; \forall k \quad i \neq j \neq k$ desigualdade do triângulo

Como afirmado anteriormente, os procedimentos em análise de agrupamento são influenciados pela natureza dos dados. Deste modo, desde a escolha ou determinação de uma distância deve-se considerar a natureza das variáveis agrupadoras ou dos atributos dos objetos. Problemas de escala podem ser previamente sanados, dada a escolha correta das distâncias, além de propiciar melhores resultados, já que os atributos podem apresentar-se como valores ou até como categorias (Gauch Jr., 1982; Beals, 1984; Pielou, 1984) (Tabela 2).

Dentre os níveis de mensuração das variáveis mais comumente utilizados, têm-se os nominais, para os quais o atributo é um caracter que pode ser

codificado de modo binário, assinalando (0) para a ausência e (1) para a presença.

Uma consideração deve ser feita, no caso das variáveis nominais, quanto a sua simetria, para a qual tem-se uma resposta exclusiva que associa os objetos (Tabela 1), já que a assimetria, não necessariamente os associa, como no caso de caracteres muito raros manifestos que acabam por associar objetos, onde estas não se manifestam, sendo esta dupla ausência representada por (-1) (Tabela 1) (Kuo, 1997).

Tabela 1 Codificação de variáveis nominais para os objetos *i* e *j*

		i (+)	*i* (-)	
j	(+)	*a*	*b*	*a+b*
	(-)	*c*	*d*	*c+d*
		a+c	*b+d*	*a+b+c+d*

Outros níveis de mensuração, como os ordinais, que assinalam uma quantificação entre os objetos em função de seus atributos em uma escala de ordem. Os níveis intervalares situam os atributos dos objetos em uma faixa de valores e também podem ser expressos como log-intervalares que assinalam a razão entre duas faixas intervalares.

As proporções são níveis de mensuração que assinalam a razão de atributos em função de um somatório e apresentam aplicação um pouco mais rara. Entretanto, os níveis absolutos, para os quais os atributos são variáveis aleatórias, discretas ou contínuas, apresentam a maior quantidade de informação dentre todos os citados. A utilização de distâncias comuns a vários níveis é possível, especialmente nos níveis não-nominais, mas algumas distâncias, como a de Gower, pode ser utilizada em qualquer um dos níveis citados acima (Gower, 1971; Everitt, 1981; Kuo, 1997).

Tabela 2 **Coeficientes para variáveis nominais simétricas e assimétricas utilizados em análise de agrupamento**

Coeficientes		Expressão	Amplitude	Variáveis
Hamming	{D}	$d_{20(i,j)} = b + c$	0 a p	{NS}
Coincidência simples	{S}	$s_{21(i,j)} = \frac{a+d}{a+b+c+d}$	0 a 1	{NS}
Coincidência quadrada	{D}	$d_{22(i,j)} = \frac{b+c}{a+b+c+d}$	0 a 1	{NS}
Hamann	{D}	$d_{23(i,j)} = \frac{[(a+d)-(b+c)]}{a+b+c+d}$	-1 a 1	{NS}
Roger & Tanimoto	{S}	$s_{24(i,j)} = \frac{a+d}{[(a+d)+2(b+c)]}$	0 a 1	{NS}
Sokal & Sneath 1	{S}	$s_{25(i,j)} = \frac{2(a+d)}{[2(a+d)+(b+c)]}$	0 a 1	{NS}
Sokal & Sneath 3	{S}	$s_{26(i,j)} = \frac{a+d}{b+c}$	0 a 1	{NS}
Jaccard	{S}	$s_{27(i,j)} = \frac{a}{(a+b+c)}$	0 a 1	{NA}
Sørensen[3]	{S}	$s_{28(i,j)} = \frac{2a}{(2a+b+c)}$	0 a 1	{NA}
Ochiai	{S}	$s_{29(i,j)} \frac{a}{\sqrt{[(a+b)(a+c)]}}$	0 a 1	{NA}
Baroni, Urbani & Buser	{S}	$s_{30(i,j)} = \frac{[a+(ad)]}{[a+b+c+(ad)]}$	0 a 1	{NS}

Onde: Codificações de presença e ausência expressas na tabela são sumarizadas na Tabela 1. {D} - coeficiente de dissimilaridade; {S} - coeficiente de similaridade; {NS} - nominal simétrica; {NA} - nominal assimétrica

De modo geral, as distâncias nominais recebem a denominação de coeficientes ou índices, já que estas não satisfazem a desigualdade triangular, o que não lhes confere a legitimidade de distâncias (Orloci, 1966). Já as não-

[3] Também denominado índice de Ney & Li

nominais satisfazem todas as três condições. Uma outra propriedade, mais rigorosa que as descritas acima, é a ultramétrica, definida por

$$d_{i,j} \leq \max\{d_{i,k}; d_{k,j}\} \qquad \forall i; \forall j; \forall k \qquad i \neq j \neq k$$

De onde obtém-se que toda ultramétrica é uma distância, mas nem toda distância é uma ultramétrica.

As distâncias aplicáveis em níveis de mensuração não-nominais ou quantitativos (Tabela 3) podem ser divididas em distâncias métricas, associadas exclusivamente com a medida vetorial das variáveis mensuradas, não incluindo nenhuma medida de variação e distâncias estatísticas, estas sim incluindo medidas de variação.

Tabela 3 Medidas de distância e similaridade utilizadas em análise de agrupamento

Medidas		Expressão	Limite	Variável
Gower	S	$s_{g(i,j)} = \dfrac{\sum_{v=1}^{p} w_v \delta_{i,j}^v d_{i,j}^v}{\sum_{v=1}^{p} w_v \delta_{i,j}^v}$	0 a 1	Todas
Gower transformada	D	$d_{2(i,j)} = 1 - S_{g(i,j)}$	0 a 1	Todas
Euclidiana	D	$d_{3(i,j)} = \sqrt{\sum_{v=1}^{p} w_v (x_{iv} - x_{jv})^2}$	≥0	{AIRO}
Size[4]	D	$d_{5(i,j)} = \dfrac{\left\lvert \sum_{v=1}^{p} w_v (x_{iv} - x_{jv}) \right\rvert}{\sqrt{\sum_{v=1}^{p} w_v}}$	≥0	{AIRO}
Shape[5]	D	$d_{6(i,j)} = \sqrt{\sum_{v=1}^{p} w_v \left[(x_{iv} - \bar{x}_v) - (x_{jv} - \bar{x}_v) \right]^2}$	≥0	{AIRO}

[4] Adotou-se a terminologia inglesa por esta ser mais conhecida do que o correspondente em língua portuguesa.
[5] Idem anterior.

Covariância	S	$s_{7(i,j)} = \frac{\sum_{v=1}^{p} w_v (x_{iv} - \bar{x}_v) - (x_{jv} - \bar{x}_v)}{v-1}$	≥0	{AIRO}
Correlação	S	$s_{8(i,j)} = \frac{\sum_{v=1}^{p} w_v (x_{iv} - \bar{x}_v) - (x_{jv} - \bar{x}_v)}{\sqrt{\sum_{v=1}^{p} w_v (x_{vi} - \bar{x}_v)^2 \sum_{v=1}^{p} w_v (x_{jv} - \bar{x}_v)^2}}$	-1 a 1	{AIRO}
Correlação transformada	D	$d_{9(i,j)} = \sqrt{1 - s_{8(i,j)}}$	0 a 2	{AIRO}
Minkowsky	D	$d_{10(i,j)} = \left[\sum_{v=1}^{p} w_v \lvert x_i - x_j \rvert^q \right]^{\frac{1}{q}}$	≥0	{AIRO}
Manhathann[6]	D	$d_{11(i,j)} = \sum_{v=1}^{p} w_v \lvert x_i - x_j \rvert$	≥0	{AIRO}
Chebychev	D	$d_{12(i,j)} = \max_{v=1}^{p} \left(w_v \lvert x_i - x_j \rvert \right)$	≥0	{AIRO}
Potência(q,r)[7]	D	$d_{13(i,j)} = \left[\sum_{v=1}^{p} w_v \lvert x_i - x_j \rvert^q \right]^{\frac{1}{r}}$	≥0	{AIRO}
Razão de Similaridade	S	$s_{14(i,j)} = \frac{\sum_{v=1}^{p} w_v (x_{iv} x_{jv})}{\sum_{v=1}^{p} w_v (x_{iv} x_{jv}) + \sum_{v=1}^{p} w_v (x_{iv} - x_{jv})^2}$	0 a 1	{R}
Canberra[8]	D	$d_{15(i,j)} = \sum_{v=1}^{p} \left(\frac{w_v \lvert x_{iv} - x_{ij} \rvert}{w_v (x_{iv} + x_{ij})} \right)$	0 a 1	{R}
Cosseno	S	$s_{16(i,j)} = \frac{\sum_{v=1}^{p} w_v (x_{iv} x_{jv})}{\sqrt{\sum_{v=1}^{p} w_v (x_{iv})^2 \sum_{v=1}^{p} w_v (x_{jv})^2}}$	0 a 1	{R}
Produto interno	S	$s_{17(i,j)} = \frac{\sum_{v=1}^{p} w_v (x_{iv} x_{jv})}{\sum_{v=1}^{p} w_v}$	≥0	{R}

[6] Sinonímia: city-block distance
[7] Sinonímia: distância euclidiana generalizada
[8] Sinonímia: coeficiente de Lance e Willians não-métrico (Kuo, 1997)

Sobreposição mínima[9]	S	$s_{18(i,j)} = \sum_{v=1}^{p} w_v[\min(x_{iv}, x_{jv})]$	≥0	{R}
Sobreposição	D	$d_{19(i,j)} = \max\left(\sum_{v=1}^{p} w_v(x_{iv}), \sum_{v=1}^{p} w_v(x_{jv})\right) - \sum_{v=1}^{p} w_v[\min(x_{iv}, x_{jv})]$	≥0	{R}
Cantell	D	$d_{20(i,j)} \frac{2\chi^2_{0,5[v]} - vd^2_{i,j}}{2\chi^2_{0,5[v]} + vd^2_{i,j}}$	≥0	{AIRO}

Onde: S - medida de similaridade; D - medida de dissimilaridade; {AIRO} - variáveis absolutas, intervalares, racionais e ordinais; {R} - racionais; *i, j* - objetos; *v* - variável mensurada; x - realização da variável *v* em um dado objeto; $\delta^v_{i,j}$ - presença ou ausência da variável nos objetos i e j; w_v - peso atribuído à variável; q, r - valores arbitrários atribuídos pelo usuário.

As distâncias métricas são definidas, assim, como a menor distância entre os objetos, que são representados pelas realizações das variáveis em um espaço multidimensional (Figura 7.b). A variabilidade dentro e entre as variáveis mensuradas é desconsiderada nesta caso (Figura 7.c,d), o que as torna de difícil manipulação dada a não inclusão de qualquer medida que possa formalizar um procedimento probabilístico. Entretanto, algumas destas distâncias podem apresentar caráter probabilístico, como no caso da distância de Canttell (Tabela 3 [$d_{20(i,j)}$]), também chamada de coeficiente de padrão de similaridade, em que as distâncias são relacionadas a um escore em função da distribuição χ^2 (Sneath Sokal, 1973).

A distribuição de probabilidade da distância euclidiana entre cada par de objetos foi estudada de forma empírica, sendo determinada por Goodall (1966) como tendendo a uma distribuição uniforme, o que nos sugere um procedimento não paramétrico (Purin & Sen, 1971).

[9] Sinonímia: porcentagem mínima ou índice de Renkönen (Pielou, 1984)

Já as distância estatísticas apresentam como diferença em relação às distâncias métricas a inclusão de uma medida de variabilidade. É o caso da distância de Penrose (2.12),

$$P_{jk} = \sum_{i=1}^{p} \frac{(x_{ij} - x_{ik})^2}{pV_i} \qquad (2.12),$$

em que V_i a variância amostral da i-ésima variável considerada (Manly, 1994).

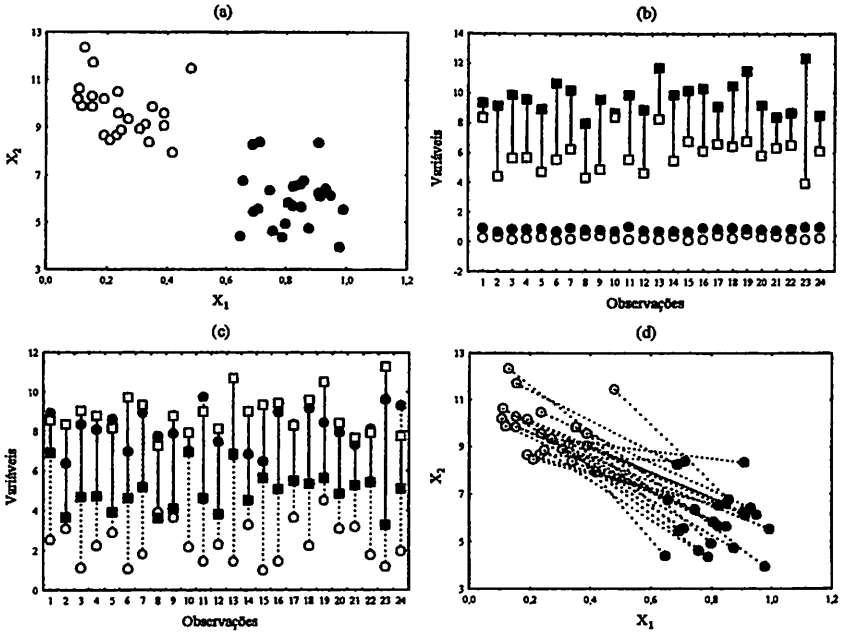

Figura 7 **(a) Valores de um agrupamento hipotético e propriedades de (b) distâncias métricas, (c) estatísticas, ponderadas pela variância e pela (d) covariância**

Adoção de distâncias com esta natureza trazem como resultado a redução de efeitos de escala, já que estas distâncias ponderam as diferenças entre objetos pelo efeito da variação no atributo, reduzindo consideravelmente os possíveis efeitos de escala (Figura 7.c), que tornam as distâncias métricas não indicadas em casos em que variáveis de diferentes unidades são manipuladas (Manly, 1994). Alternativas correntes são a adoção de variáveis padronizadas (Dillon e

Goldstein, 1984) ou de escores obtidos através de outras análises multivariadas, como variáveis canônicas ou análise fatorial (Lebart, Morineau e Warwick, 1984).

De modo geral, as distâncias métricas são utilizadas nos casos em que somente uma unidade é utilizada, como no caso dos estudos de composição florística e faunística, em que locais de amostragem são arranjados em função da abundância de espécies coletadas (Gauch Jr., 1982; Pielou, 1984).

Outra distância estatística também muito utilizada é a generalizada de Mahalanobis (D^2). Definida inicialmente em função da posição do centróide, que é o ponto médio de várias variáveis no hiperespaço, sendo a distância de cada observação ao centróide, é considerada uma D^2 de Mahalanobis em relação às variáveis independentes correlacionadas. Uma ressalva a ser feita é que se as variáveis independentes não apresentam correlação, D^2 é equivalente à distância euclidiana (Johnson e Wichern, 1998). Sua expressão neste contexto é (2.13)

$$D_i^2 = (x_i - \mu)' \Sigma^{-1} (x_i - \mu) \qquad (2.13),$$

em que x_i é um vetor de observações da população avaliada, μ_i é o centróide da população avaliada, Σ^{-1} é a inversa da matriz de covariância combinada (*pooled*), definida pela média das matrizes de covariâncias das populações avaliadas.

Definições associadas ao centróide e covariância viesada são pertinentes ao próprio espaço físico, no qual o centróide representa o centro de massa de um corpo (Figura 8.a), enquanto a variância combinada assinala o grau de inércia médio nos corpos (Figura 8.b).

Apesar de esta ter sido originalmente proposta para a mensuração de distância entre observações e seus centróides, uma generalização para qualquer par de objetos formalizada por Friedman e Rubin (1967) é atualmente aceita e bastante utilizada. Sua expressão, então, dada por (2.14)

$$D^2 = \left(x_i - x_j\right)' \Sigma^{-1} \left(x_i - x_j\right) \qquad (2.14),$$

assinala diferenças entre pares de objetos e não mais entre objetos de uma população e seu centróide, entretanto a matriz de covariância continua a ser a combinada (Everitt, 1981).

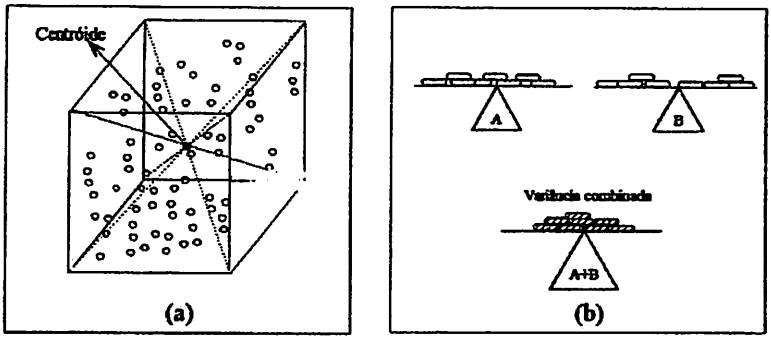

Figura 8 **Representação esquemática do (a) centróide e (b) variância combinada**

A inclusão de medidas com esta natureza em outros ramos da Estatística são observáveis, como no caso da distância de Cook (D_i); utlizada como diagnóstico em regressão, que descende desta idéia (Cook, 1977; Ramirez, 1998), em que esta medida provê a indicação, se uma observação pode ser considerada como um ponto mais distante, no caso de análise de agrupamento ou um *outlier* no caso de regressão. A única distinção entre D^2 e D_i é que Σ^{-1} é a inversa da matriz de covariância global. A distância generalizada de Mahalanobis apresenta vantagem sobre a de Penrose pela inclusão da matriz de covariância, o que lhe confere, além da medida de variação de uma dada variável, a relação desta com outras consideradas (Mardia, Kent e Bibby, 1995). Deste modo, agrupamentos com forte influência da estrutura de covariância podem ser analisados de maneira mais apropriada com esta distância, tornando-a a mais indicada em análise de agrupamento englobando variáveis quantitativas (Johnson e Wichern, 1998).

2.5 Procedimentos de classificação

Definem-se vários algoritmos para a análise de agrupamento, entretanto, definições acerca dos problemas relacionados à análise são necessárias (Figura 9). O caráter exclusivo em análise de agrupamento denota o fato de que um objeto pertence somente a um subconjunto dos dados, enquanto a não-exclusividade denota que um objeto pode situar-se em mais de um subconjunto; um exemplo deste caráter são palavras com diferentes sentidos semânticos que são alocadas em mais de um subconjunto (Henery, 1994).

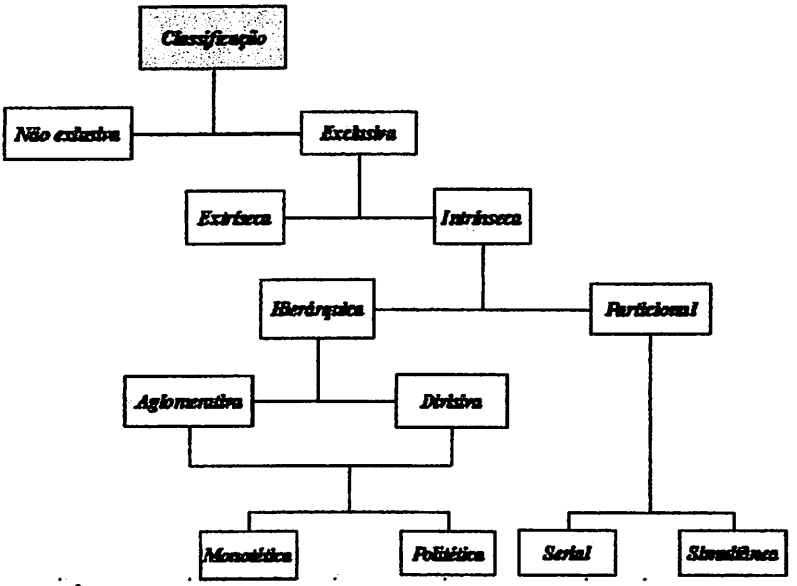

Figura 9 Árvore dos problemas de classificação

Dada a exclusividade, o caráter extrínseco refere-se a uma separação inicial das categorias de objetos, com o objetivo de determinar quais as afinidades e diferenças dos objetos previamente selecionados. Estudos epidemiológicos, utilizando a estrutura caso-controle, assinalam este caráter. Já o caráter intrínseco reafirma a proposição original da análise de agrupamento, que assume o desconhecimento *a priori* de qualquer organização entre os objetos, sendo as

informações contidas nos dados responsáveis pelo arranjo entre estes. Deste modo, tem-se no caráter intrínseco a essência da análise de agrupamento, o que pode explicar a razão de técnicas baseadas neste princípio apresentarem tanta aplicabilidade e discussão na literatura.

Os procedimentos hierárquicos apresentam como resultados séries de agrupamentos em uma escala de afinidade, partindo do pressuposto de que o conjunto de dados é um único supra-agrupamento e cada objeto forma ou formará um subconjunto próprio. Em contrapartida, os procedimentos particionais ou não-hierárquicos resultam em um arranjo dos objetos em um número de agrupamentos pré-definido. Estes procedimentos podem ser do tipo seriais, nos quais um objeto é alocado por vez, ou do tipo simultâneo, em que todos os objetos são alocados ao mesmo tempo.

Dentre os procedimentos hierárquicos, têm-se os aglomerativos, que descrevem a orientação do agrupamento partindo do princípio de que cada objeto é um agrupamento natural, posteriormente reunindo-se a outros de maior afinidade através de fusões de *n* objetos, que sucessivamente são reunidos até formar o supra-agrupamento, que é o conjunto de objetos como um todo. Do lado oposto, os procedimentos divisivos descrevem a orientação do agrupamento a partir de um supra-agrupamento, representado pelo conjunto de objetos, que é dividido em agrupamentos subsequentes de menor afinidade até o retorno ao objeto.

Em ambos os procedimentos pode-se ter um enfoque monotético, no qual apenas um atributo é mensurado; ou politético, em que vários atributos são mensurados. De modo geral, as aplicações em análise de agrupamento apresentam o enfoque politético, pela própria natureza multivariada dos fenômenos, e através de procedimentos hierárquicos, pelo próprio

desconhecimento da estrutura dos objetos. O objeto de estudo desta dissertação, centra-se neste ponto, pelos motivos já assinalados anteriormente.

Os métodos de agrupamento, ligação ou amalgamação para os procedimentos hierárquicos e particionais são apresentados em seguida. A subdivisão citada acima é representada por (a) métodos aglomerativos, atribuindo séries de fusões de *n* objetos em diferentes grupos e (b) métodos divisivos, determinando separações no conjunto de n objetos em subdivisões cada vez menores.

Dentre os métodos aglomerativos, podem ser citados:

(a.1) Ligação simples ou método do vizinho mais próximo (*Single linkage*; *Nearest-neighbor method*)

Este procedimento utiliza a distância mínima (Figura 10.b) entre dois objetos de um conjunto n, de grupos distintos como sendo a distância entre os grupos. O próximo grupo é representado pela menor distância entre o primeiro grupo determinado e o objeto mais próximo a este. Os passos seguem-se até o encadeamento de todos os objetos em um único agrupamento, este com diferentes arranjos de objetos em um dado nível da escala de distâncias.

(a.2) Ligação completa ou método do vizinho mais distante (*Complete linkage*; *Furthest-neighbor method*)

Este método é exatamente oposto ao da ligação simples (Figura 10.c), em que no primeiro passo considera-se a distância entre dois grupos como sendo a distância entre os objetos de maior distância, estes definindo grupos polarizados. Com a redução das distâncias entre os grupos e objetos, estes passam a formar agrupamentos com menor distância, encadeando-se.

(a.3) Ligação média (*Average linkage*)

Trata-se de uma variação dos procedimentos descritos anteriormente, sendo que neste, a distância entre dois grupos é representada pela média da

distância entre todos os pares de objetos pertencentes a cada grupo (Figura 10.d). Vários algoritmos são propostos para a condução deste procedimento (Dillon e Goldstein, 1984).

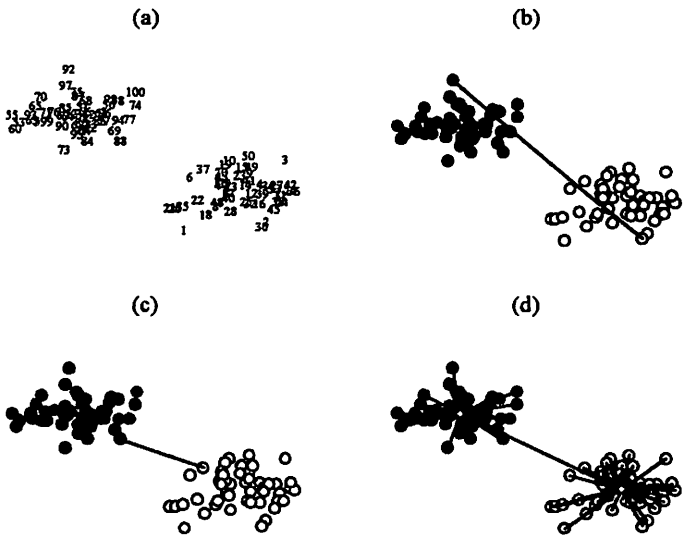

Figura 10 Métodos de ligação em análise hierárquica de agrupamento, (a) disposição das observações, (b) método de ligação simples, (c) método de ligação completa e (d) método de ligação média

Variações neste método podem ser encontradas na literatura. Destacam-se os procedimentos baseados diretamente na média entre as distâncias dos objetos, podendo estas serem ponderadas ou não. Neste caso, os correspondentes são, respectivamente, WPMGA e UPMGA, e baseados no centróide, valor central ou médio entre os objetos de um dado grupo, também com correspondentes ponderados ou não, respectivamente WPGMC e UPGMC.

(a.4) Método de Ward

Baseado na redução da informação resultante, dada a inclusão de um conjunto de objetos em um grupo. Esta redução de informação é determinada pela soma total do quadrado do erro de cada objeto, em função da média do grupo a

que este, supostamente, pertença (Figura 11). Esta regra de inclusão envolve todos os pares possíveis, sendo definidos como pertencente a um dado grupo o objeto que contribua o mínimo com o aumento da soma de quadrado do erro.

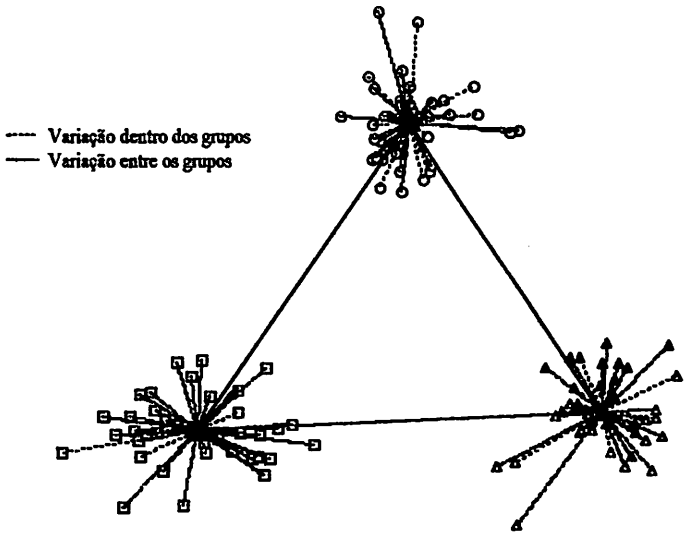

Figura 11 **Método hierárquico de Ward de redução da soma de quadrado do erro**

Informações mais detalhadas sobre os métodos hierárquicos podem ser obtidas em Manly (1994), Johnson e Wichern (1998).

Os (b) métodos divisivos baseiam-se na subdivisão do conjunto de objetos em dois grupos. Subdivisões posteriores são empregadas nos grupos inicialmente separados. Dentre os métodos divisivos, têm-se o método da distância média subdivisora, que consiste na divisão do conjunto de dados em dois grupos em um número de combinações 2^{n-1}-1. Define-se, então, qual dos pares apresenta a maior dissimilaridade. A este, o subdivisor, são adicionados seqüencialmente os objetos de maior dissimilaridade, até que os grupos possam ser polarizados em torno deste subdivisor. Outro método divisivo é a detectação automática de interação,

em que inicialmente são definidos subconjuntos de maior dissimilaridade, partições binárias são conduzidas dentro destes subconjuntos, com base em um enfoque monotético. Os subconjuntos que apresentarem redução na soma de quadrado em cada uma das variáveis são identificados como afins (Everitt, 1981). Em ambos os casos citados, a exigência computacional é intensiva, o que lhes confere uma maior dificuldade de implementação. Entretanto, no caso da detectação automática de interação, a escolha de um enfoque de decisão monotético reduz a complexidade do fenômeno e pode comprometer a decisão na análise.

Diferentes dos métodos de classificação hierárquica, os métodos de partição definem uma posição definitiva para os objetos no decorrer da sua condução, primando exclusivamente pelos critérios estabelecidos no início destas, no caso a determinação do número de grupos. Algumas técnicas que representam este método são baseadas em propriedades da matriz de soma de quadrados da análise de variância (ANAVA). A primeira, denominada *k-means*, é baseada no critério de maximização da soma de quadrado entre os subconjuntos e redução dentro dos subconjuntos definidos. Um número de subconjuntos é definido *a priori*, sendo então são aplicados os critérios assinalados de maneira iterativa (Dillon e Goldstein, 1984). Outras técnicas baseiam-se na matriz de soma de quadrados e produtos da análise de variância multivariada (MANAVA), composta de uma submatriz de efeito entre os tratamentos (B) e outra matriz de efeito dentro dos tratamentos (W). Os subconjuntos são considerados tratamentos, então critérios como minimização do traço de B ou do determinante de B atuam com o intuito de minimizar as diferenças dentro dos subconjuntos e maximizar aquelas entre os subconjuntos, já que o traço e o determinante destas matrizes são medidas de variância generalizada (Everitt, 1981; Johnson e Wichern, 1998).

2.6 O exemplo do estudo taxonômico do gênero *Iris* L.

O banco de dados *Iris*, composto pelas espécies *I. setosa*, *I. versicolor* e *I. virginica*, tornou-se notório com o célebre artigo de Sir Ronald A. Fisher, do ano de 1936, em que este definiu a análise de discriminantes lineares. O banco de dados consiste em medidas de comprimento das pétalas (CP), sépalas (SP) e da largura das mesmas (LP) e (LS), nas três espécies citadas (Figura 12).

A estrutura do agrupamento é mista, assinalando um componente esférico, em que é discriminada a espécie *I. setosa* das outras duas, por uma definida zona de menor densidade e um componente elipsoidal, provendo dificuldades na separação entre as outras espécies, com poucas zonas de menor densidade entre *I. versicolor* e *I. virginica* (Figura 12).

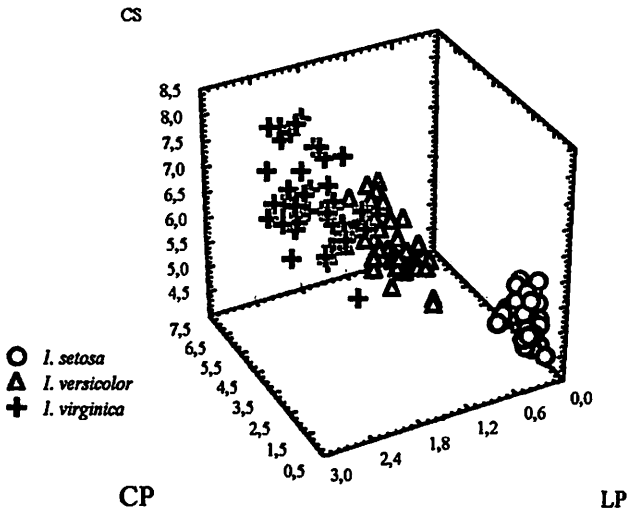

Figura 12 **Representação dos valores de comprimento das pétalas (CP), sépalas (CS) e largura da pétala (LP) no gênero *Iris* L.**

A validade dos caracteres medidos, no caso de morfologia floral, é dada pelo caráter conservativo destas, já preconizado em estudos taxonômicos

(Endress, 1996). Deste modo, as medidas tomadas servem como caracteres efetivamente discriminantes entre os taxa do gênero *Iris*.

Utilizou-se este banco de dados por já ser consagrado na literatura (Anderson, 1984; Chatfield e Collins, 1986; Mardia, Kent e Bibby, 1995; Johnson e Wichern, 1998; entre outros) e por ter sido submetido a diversas metodologias (Henery, 1994).

3 MATERIAL E MÉTODOS

3.1 Simulação

Foram efetuadas simulações Monte Carlo no programa SAS® System 6.12 (Anexo 1) (SAS Institute, 1990), consistindo de um procedimento gerador de populações com distribuição normal multivariada. Esta medida foi tomada a fim de se obterem os valores de quantis para a distribuição empírica da D^2 de Mahalanobis. Diferentes configurações, quanto à estrutura de covariância, ao número de objetos e ao número de variáveis, foram obtidas via simulação. Em todas as configurações foi utilizado um número de 5.000 simulações.

O número de objetos adotado foi de n=p+1, p+2, ..., 250, sendo com passos de uma unidade até o valor 30, com passos de cinco unidades até o valor 80, passos de dez unidades até o valor 120, passos de vinte unidades até o valor 200 e, finalmente, passo de cinquenta unidades até o valor 250. Esta estrutura foi respeitada considerando todas as configurações para o números de variáveis $2 \leq p \leq 10$.

As simulações consistem da geração de populações sob distribuição normal multivariada centrada, em que é definido um vetor μ de médias, sem perda de generalidade, e uma matriz Σ de covariâncias. O modelo adotado para a simulação é então representado por $\underset{\sim}{x}=F\underset{\sim}{Z}+\mu$, onde F é tal que $FF'=\Sigma$, podendo F ser obtido pela decomposição de Cholesky, e $\underset{\sim}{Z}$ é um vetor de variáveis aleatórias normais univariadas centrais independentes (Kennedy Jr. e Gentile, 1980; Khattree e Naik, 1995).

Outras formas de simulação podem ser utilizadas, consistindo da parametrização categorizada de cada população, adotando-se conceitos de

centralidade (vetor de médias) e dispersão ou escala (matriz de covariância), além de componentes latentes, como estruturas trigonométricas, por exemplo (Sarle, 1990). Este procedimento foi adotado para os outros tipos de agrupamentos avaliados na comparação da eficiência entre os métodos de agrupamento baseados no critério proposto. A estrutura de covariâncias (3.1) utilizou como valores de ρ_{ik}: 0,1; 0,3; 0,5 e 0,9. Após a verificação de não influência da estrutura de covariância, adotou-se o valor de ρ_{ij}=0,5 para os demais casos.

$$\Sigma = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & \rho_{2k} & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k} & \rho_{2k} & \cdots & \rho_{kk} \end{bmatrix} \qquad (3.1)$$

Os valores dos quantis da distribuição empírica da estatística D^2 de Mahalanobis, obtidos para cada configuração, foram de: 90; 92,5; 95; 97,5; 99 e 99,5%. Medidas da forma da distribuição, como a assimetria e a curtose, também foram tomadas.

Utilizou-se a aproximação de $2\chi^2_{(p)}$ com n≤15 e de T^2 de Hotelling aproximada por F para n>15, expressa por $2\left[\frac{(2n-2)p}{(2n-p)}\right]F_{(\alpha;p,2n-p)}$ (Johnson e Wichern, 1998). A restrição quanto ao número de objetos indicado na literatura não foi seguida, sendo as aproximações submetidas a todos os números de objetos simulados com o objetivo de determinar em qual número de objetos as aproximações apresentam maior concordância com a distribuição empírica. Os quantis das aproximações de χ^2 e F foram confrontados com o da distribuição empírica da D^2 de Mahalanobis .

3.2 Comparação entre os procedimentos de agrupamento alternativos

Três tipos de técnicas foram utilizadas, a saber: análise de discriminantes, esta com os métodos linear e quadrático; técnicas particionais representadas pelo métodos *k-means*, e técnicas hierárquicas, representadas pelos métodos de ligação simples, completa, UPGMA e de Ward. A seleção destes métodos foi regida pela larga utilização destes em análises estatísticas e pela disponibilidade destes em diversos programas de análise.

No caso dos métodos da análise de discriminantes e particional, foi tomado apenas o percentual de classificação correta. Para os métodos hierárquicos foi utilizado o ponto de corte fornecido pela aproximação da distribuição empírica da D^2 de Mahalanobis pela T^2 de Hotelling via F e χ^2, sendo tomados o percentual de classificação correta, o número de partições indicadas pelo ponto de corte, o número de subconjuntos em cada população amostrada ou simulada e a contiguidade dos agrupamentos.

O número dos k agrupamentos para o método particional *k-means* foi dado pela expressão (3.2) (Calinski e Harabasz, 1974, citado por Everitt, 1981)

$$C = \frac{tr(B)}{k-1} \Big/ \frac{tr(W)}{N-k} \qquad (3.2),$$

onde: *tr(B)* traço da matriz de hipóteses da matriz de soma de quadrados e produtos cruzados (SQPC-MANAVA), *tr(W)* traço da matriz de erro da matriz de soma de quadrados e produtos cruzados (SQPC-MANAVA), k número de agrupamentos, N o número de objetos.

Estes métodos foram utilizados para os dados de morfologia floral do gênero *Iris* e outras estruturas de agrupamento simuladas, à exceção do método *k-means*. O interesse maior desta comparação situou-se sobre os métodos hierárquicos, que são os mais comumente utilizados em pesquisa agropecuária,

procurando destacar quais foram suas limitações ou incompatibilidades com o critério de aproximação da distribuição empírica de D^2 de Mahalanobis sob diferentes estruturas de agrupamentos.

Em todos os casos, os dados foram analisados: (i) sob sua condição natural, considerando-os brutos [B] e (ii) através de amostragem aleatória sem reposição, considerando-os sob subamostragem [S]. O objetivo desta medida é inferir sobre o caráter conservativo dos resultados indicados pelas aproximações.

4 RESULTADOS

4.1 Aproximações

4.1.1 Caso bivariado

A distribuição empírica da D^2 de Mahalanobis apresentou-se nos menores valores de número de objetos, levemente platicúrtica com assimetria à esquerda (Figura 13). A partir de um valor de 2p+1, pode-se observar uma reversão de assimetria à direita (Figura 14.a), e quanto à sua forma adotando uma forma leptocúrtica (Figura 14.b), esta tendência acentuou-se com o aumento de número de objetos.

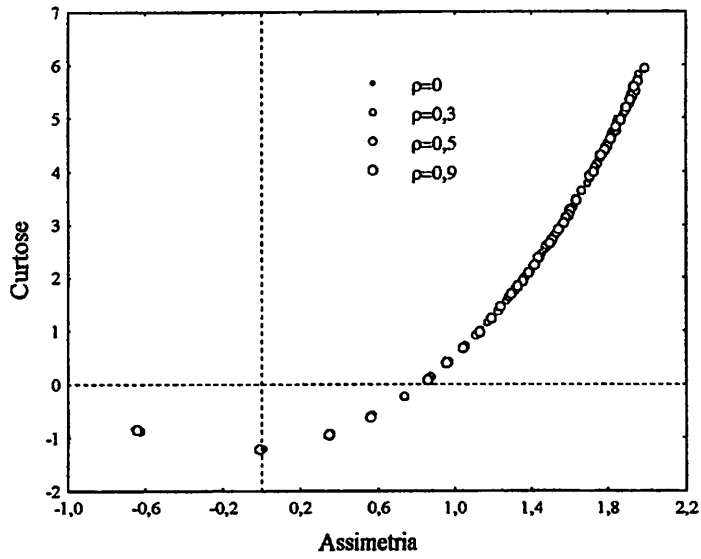

Figura 13 **Diagrama de razão de momentos (assimetria e curtose) em função das estruturas de covariância no caso bivariado**

As semelhanças preconizadas com a distribuição F são um prévio indicador de que aproximações podem ser satisfatórias. Também pode-se

observar que a estrutura de covariância não influenciou as medidas de assimetria e curtose (Figura 13), entretanto o número de objetos influiu de maneira efetiva na forma da distribuição empírica (Figura 14.a,b).

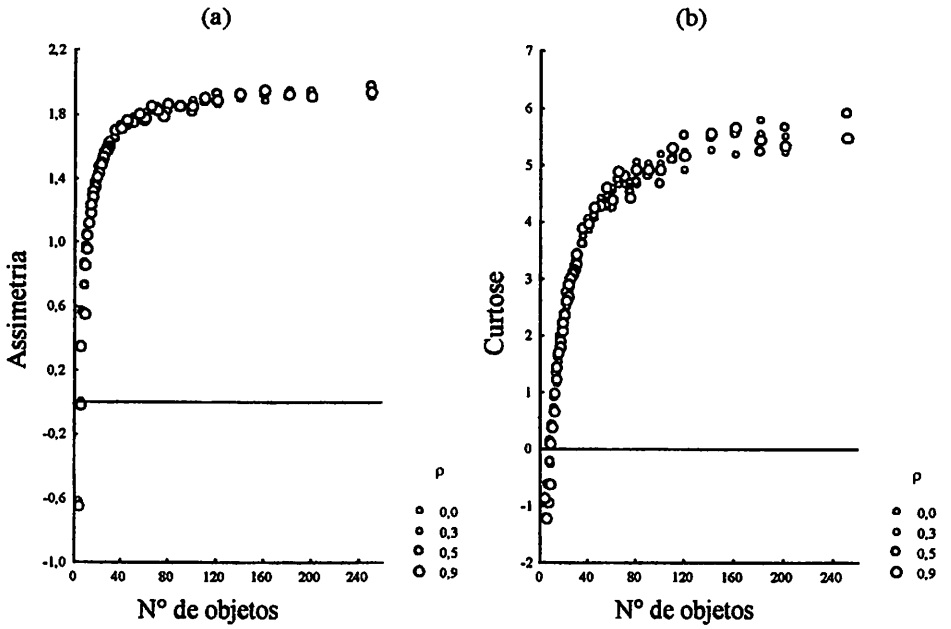

Figura 14 **Relação entre o número de objetos e (a) coeficiente de assimetria e (b) curtose em função das estruturas de covariância, no caso bivariado**

As aproximações conduzidas também indicaram independência da estrutura de covariância (Figura 15.a). A aproximação através de χ^2 apresentou-se como menos discrepante em relação à empírica de D^2 de Mahalanobis, quando comparada à aproximação via F, especialmente nos casos de menores números de objetos (Figura 15.b). Esta aproximação apresentou subestimação nos menores valores de número de objetos, enquanto a aproximação via F apresentou superestimação na mesma condição.

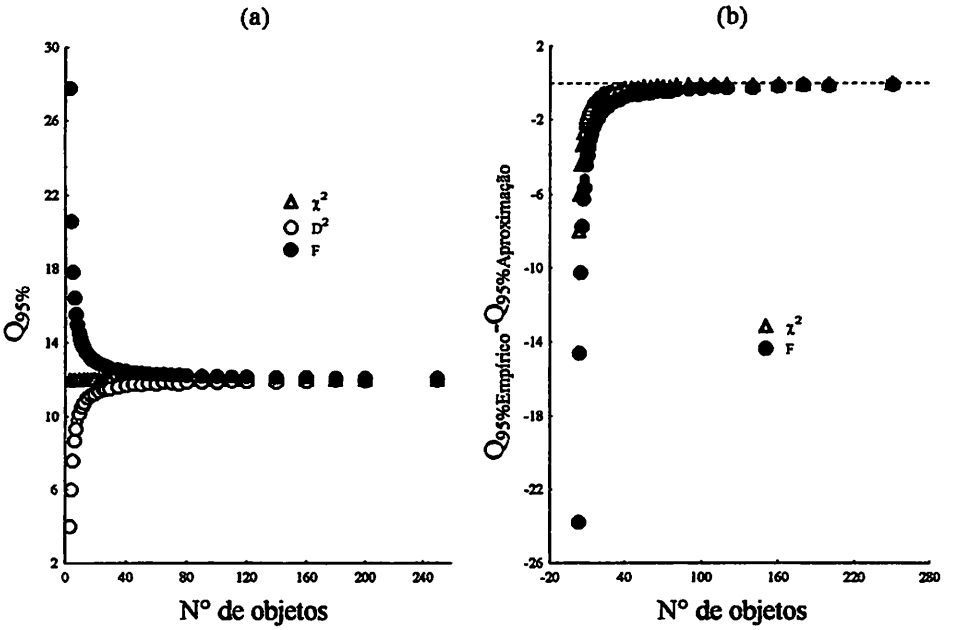

Figura 15 **(a) Valores do quantil 95% para as aproximações F, χ^2 e distribuição empírica de D^2 de Mahalanobis; (b) diferença entre valores obtidos da distribuição empírica de D^2 de Mahalanobis e aproximações F, χ^2 no caso bivariado**

A aproximação via F tende a concordar com a distribuição empírica da D^2 de Mahalanobis. Em ambos os casos, assinala-se a concordância assintótica com a aproximação via χ^2 (Figura 15.a).

Este padrão foi observado em todos os outros coeficientes de confiança, ressaltando-se que em coeficientes de confiança mais elevados, a utilização de um maior número de objetos assegura uma aproximação mais fidedigna.

4.1.2 Extensão para casos p-variados

Dada a independência da estrutura de covariância, assinalada no caso bivariado e em outras configurações p-variadas, adotou-se uma estrutura única de

covariância. Os padrões de resposta assinalados no caso bivariado são concordantes com sua extensão nos casos p-variados.

A forma da distribuição apresentou o mesmo padrão, indicado anteriormente (Figura 16), em que a distribuição empírica de D^2 de Mahalanobis com menores valores de números de objetos apresenta-se platicúrtica com assimetria à esquerda, e com o aumento do número de objetos, comporta-se como leptocúrtica com assimetria à direita.

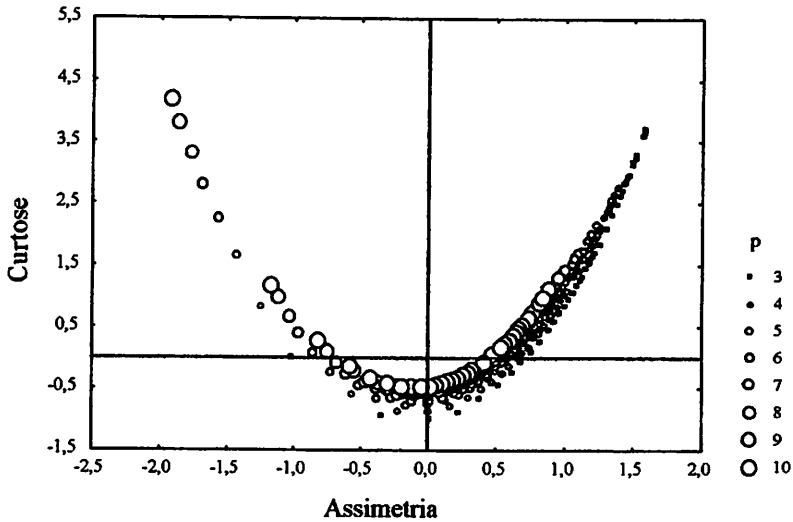

Figura 16 **Diagrama de razão de momentos (assimetria e curtose) em função das estruturas de covariância nos casos p-variados**

Entretanto, com o aumento do número de variáveis a alteração da forma platicúrtica para leptocúrtica deu-se somente em valores de número de objeto iguais a 2p+1. Este limite também indica a reversão da assimetria à esquerda para à direita (Figura 17a,b).

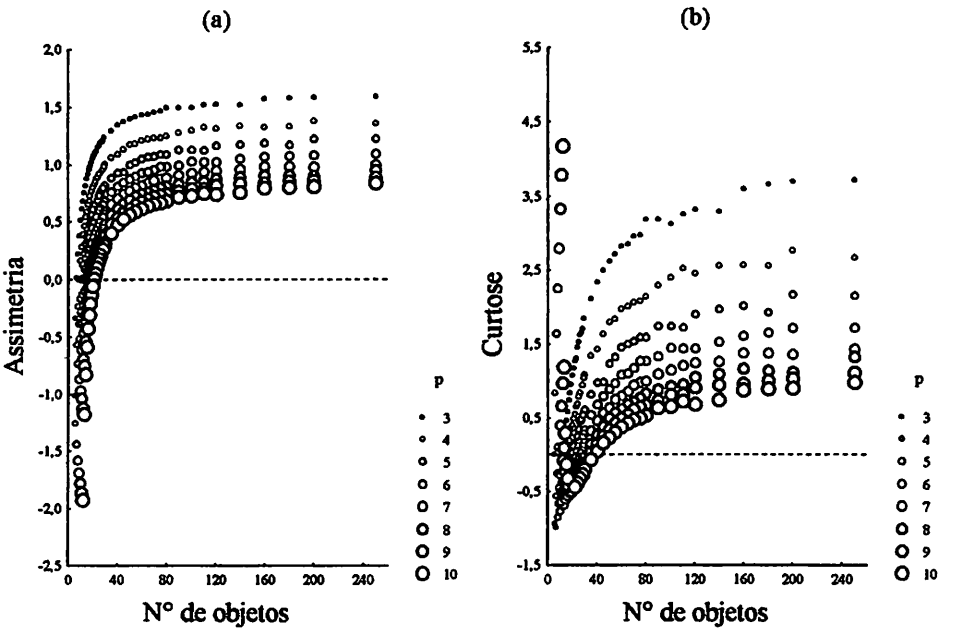

Figura 17 **Relação entre o número de objetos e (a) coeficiente de assimetria e (b) curtose em função das estruturas de covariância, no caso p-variado**

As aproximações via χ^2 e F apresentaram o mesmo padrão descrito para o caso bivariado, a subestimação no caso da aproximação de χ^2 e a superestimação no caso da aproximação via F. O ajustes apresentaram-se comprometidos com o aumento do número de variáveis, sendo que a aderência, no caso da aproximação via F, deu-se em números maiores de objetos (Figura 18).

A aproximação via χ^2, também no caso p-variado, apresentou maior concordância com a distribuição empírica da D^2 de Mahalanobis (Figura 19), mesmo que em um número pequeno de objetos.

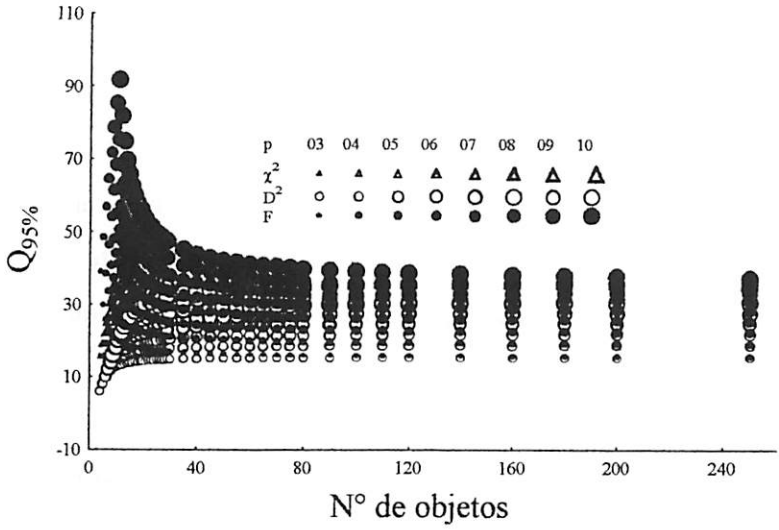

Figura 18 **Valores do quantil 95% para as aproximações F, χ^2 e distribuição empírica de D^2 de Mahalanobis, no caso p-variado**

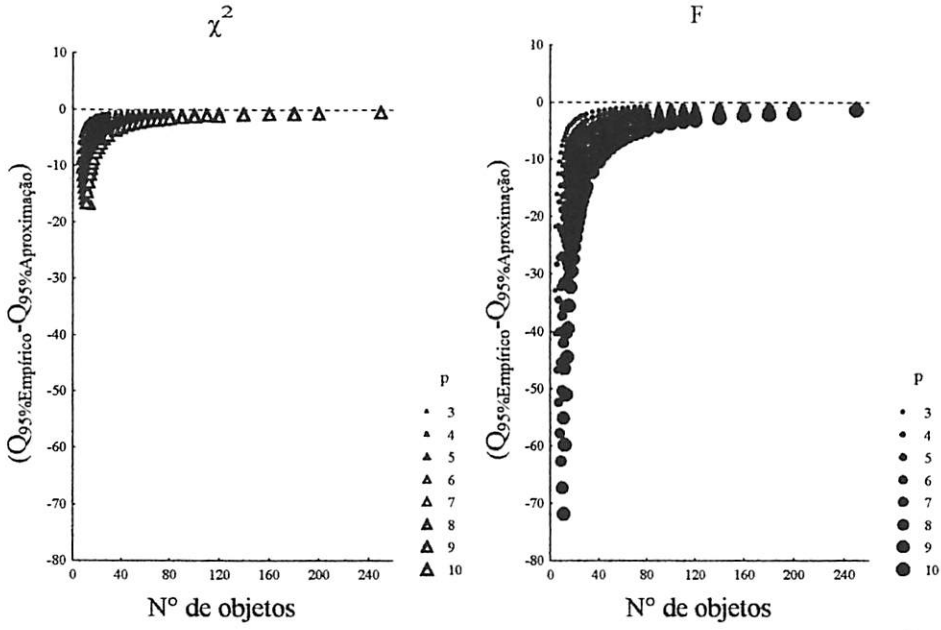

Figura 19 **Diferença entre valores obtidos da distribuição empírica de D^2 de Mahalanobis e aproximações F, χ^2 no caso p-variado**

Uma avaliação das influências do número de variáveis e objetos sobre o grau de aderência das aproximações à distribuição empírica de D^2 de Mahalanobis foi conduzida. O modelo testado consistia dos efeitos lineares e quadráticos do número de variáveis e objetos e conduziu-se uma análise de regressão múltipla por passos. Os efeitos lineares de ambas e o produto cruzados das variáveis foi significativo; entretanto, somente o efeito quadrático do número de objetos foi significativo.

Assim, determina-se que ajustes mais fidedignos à distribuição empírica de D^2 de Mahalanobis são obtidos com um aumento do número de objetos (Figura 20), respeitando-se também o número de variáveis tomadas. Empiricamente, sugere-se a utilização um número de objetos superior a 80.

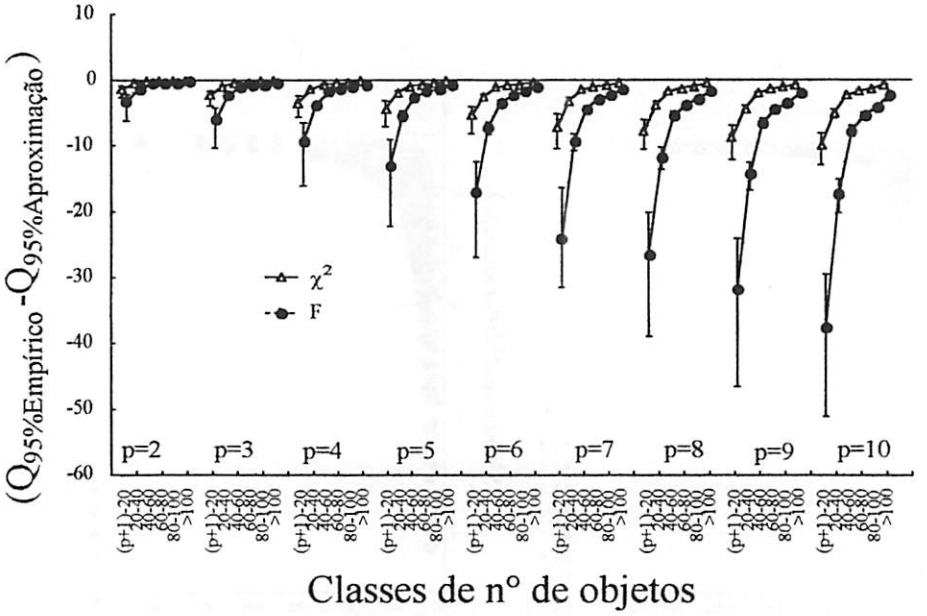

Figura 20 **Valores de mediana e quartis inferior e superior das aproximações via F e χ^2 em função das classes de número de objetos e variáveis**

A diferença entre a aderência das aproximações via F e χ^2 em relação a D^2 de Mahalanobis foi marcante em todos as de números de objetos nos diferentes

números de variáveis. O teste da Mann-Whitney detectou diferença em todas as classes de número de objetos, sendo evidente a superestimação da aproximação via F em números pequenos de objetos (Figura 20).

4.2 Comparação entre os métodos de agrupamento

4.2.1 Exemplo taxonômico do gênero *Iris* L.

A separação entre *I. setosa* e as demais espécies do gênero, foi indicada em todos os métodos utilizados (Figura 21 e Tabela 4). Esta separação é dada pela própria esfericidade dos dados, em que *I. setosa* apresentou um valores de $tr(\Sigma)$ e $|\Sigma|$ baixos, o que indica uma homogeneidade. A análise de discriminantes, tanto linear, quanto quadrática proporcionou uma alta taxa de classificação correta. Este padrão também foi observado no caso de subamostragem dos dados, situando-se sempre acima de 92% (Tabela 4). Os métodos de discriminantes quadráticos e lineares apresentaram resultados equivalentes neste caso, ressaltando-se que nos casos em que se utilizou subamostragem, houve uma leve redução na detectação de agrupamentos corretos (Tabela 4).

Os valores da distribuição empírica de D^2 de Mahalanobis e das aproximações via F e χ^2 não forneceram agrupamentos diferentes, sendo estes concordantes na legitimação dos agrupamentos.

Também no caso das técnicas hierárquicas, a distinção entre o componente esférico, representado por *I. setosa*, foi detectada em todos os métodos, sob as situações de subamostragem ou não (Tabela 4). Entretanto, no tratamento do componente elipsoidal, o método de ligação simples apresentou-se como ineficiente na separação das espécies *I. versicolor* e *I. virginica*, baseando-se no critério de aproximação da D^2 de Mahalanobis via F (Tabela 4). Este método somente apresentou resposta positiva no caso de subamostragem,

inferindo-se sobre o papel centralizador deste procedimento, o que reduz a chance de classificação errada, já que este método, também conhecido como o do "vizinho mais próximo", pressupõe maiores semelhanças entre os objetos (Sarle, 1990).

Todos os outros métodos hierárquicos apresentaram uma taxa de classificação correta semelhante aos métodos da técnica de análise de discriminantes (Tabela 4). Em todos estes, exceto ligação simples, a taxa mínima foi de 94%, sendo que nos casos em que foi utilizada subamostragem, esta taxa foi de 100% de classificações corretas.

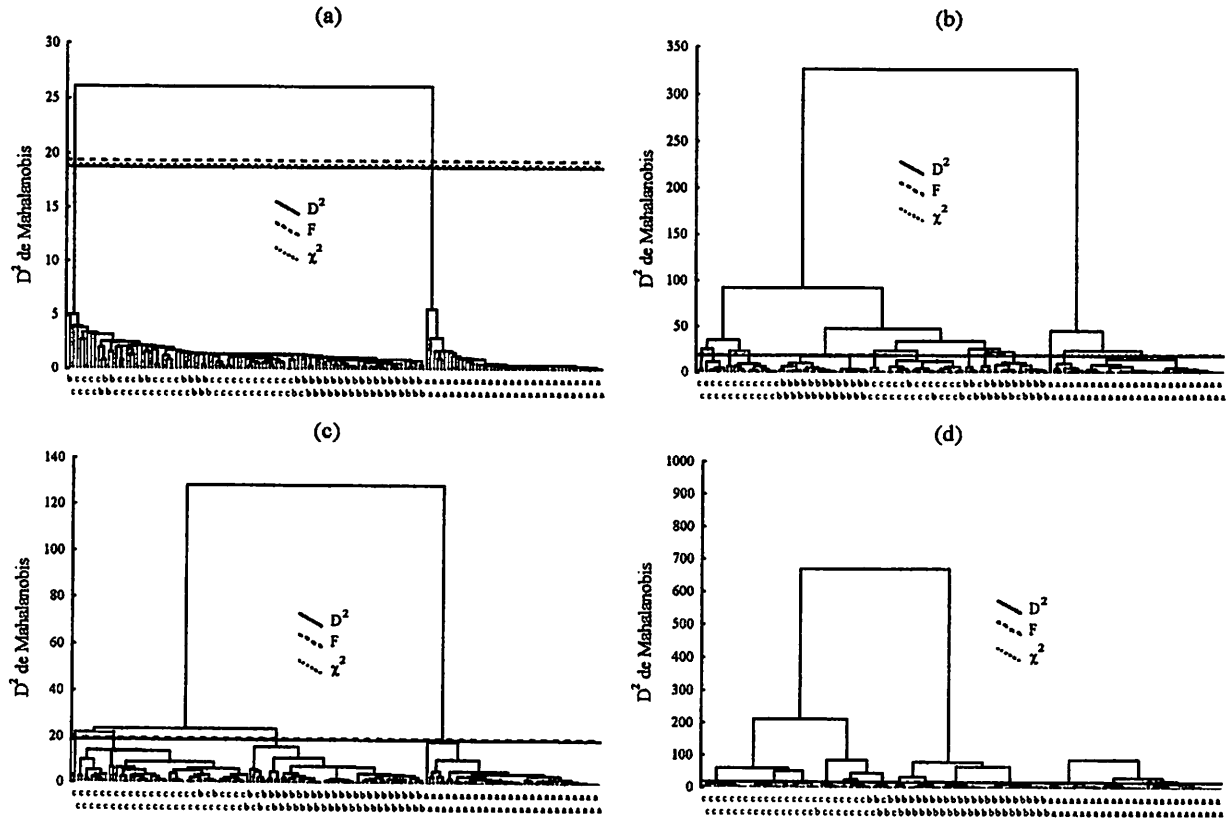

Figura 21 **Dendogramas de dissimilaridade para os dados do gênero *Iris* , segundo os métodos hierárquicos de (a) ligação simples, (b) ligação completa, (c) UPGMA e (d) de Ward**

O número de partições impostas pelos métodos hierárquicos foi sempre concordante com o número de populações avaliadas, exceto no caso do método de ligação simples. Entretanto, em todos os casos, a contigüidade entre os arranjos propostos foi evidenciada (Tabela 4).

Tabela 4 **Percentuais de classificação correta nas espécies do gênero *Iris* em função dos diferentes métodos nas técnicas análise de discriminantes {AD} e técnicas hierárquicas {HIER}**

		set		*vers*		*virg*		*Geral*	
Técnicas	**Métodos**	**B**	**S**	**B**	**S**	**B**	**S**	**B**	**S**
{AD}	Linear	100,0	100,0	96,0	94,0	96,0	96,0	97,3	96,7
	Quadrático	100,0	100,0	94,0	94,0	96,0	96,0	96,7	96,7
	Geral	**100,0**	**100,0**	**93,5**	**92,5**	**96,0**	**96,0**	**96,5**	**96,2**
{HIER}	Simples	100,0	100,0	0,0	100,0	0,0	100,0	33,3	100,0
	Completa	100,0	100,0	94,0	100,0	94,0	100,0	96,0	100,0
	UPGMA	100,0	100,0	94,0	100,0	94,0	100,0	96,0	100,0
	Ward	100,0	100,0	96,0	100,0	94,0	100,0	96,7	100,0
	Geral	**100,0**	**100,0**	**71,0**	**100,0**	**70,5**	**100,0**	**80,5**	**100,0**

Onde: set *I. setosa*; ver *I. versicolor*; virg *I. virginica*; B Dados brutos; S Dados sob subamostragem

O método *k-means*, de caráter divisivo, apresentou um valor de C positivo e decrescente, o que reforça a sua natureza hierárquica (Figura 22). O número de partições indicada foi de 50, segundo o critério de Beale, já que a contribuição de um maior não contribuiria para uma melhoria na definição de agrupamentos (Everitt, 1981). Valores de número de partições entre o mínimo e o valor indicado foram conduzidos a fim de determinar a consistência do critério aplicado.

Apenas a espécie *I. setosa* apresentou taxa de classificação correta semelhante aos demais critérios avaliados (Figura 23). Mesmo com o aumento do número de partições ao nível tido como ótimo, a taxa de classificação correta para *I. virginica* e *I. versicolor* foi inferior à determinada pelos outros métodos, sendo, respectivamente, 85,7 e 88% para um número de partições igual a 50

(Figura 23). Os métodos hierárquicos que apresentaram um número de partições mínimo obtiveram uma taxa de classificação correta muito mais elevada.

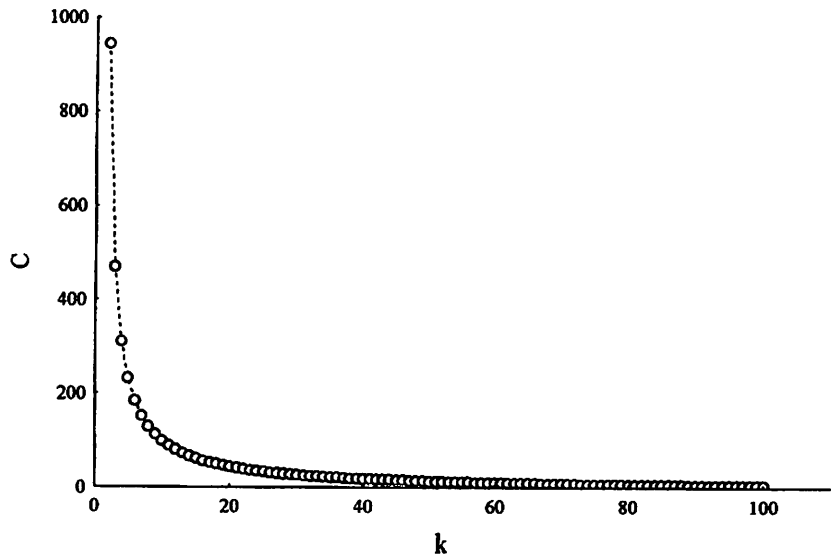

Figura 22 Valores do coeficiente de determinação da natureza e número dos agrupamentos em função do número de partições imposta

A limitação imposta pelos dados de *Iris* restringe-se apenas ao componente elipsoidal, que engloba as espécies *I. versicolor* e *I. virginica*. Mesmo com esta limitação, as técnicas de análise de discriminantes e hierárquicas, com a inclusão do ponto de corte dado pela aproximação de D^2 via F, apresentaram resultados válidos. O método diviso *k-means* não apresentou resultado desejável, especialmente quando comparado ao poder de discriminação de técnicas em um mesmo número de subconjuntos. Dentre os métodos avaliados, o de ligação completa, UPGMA e de Ward apresentaram-se como válidos, incluindo também a análise de discriminantes.

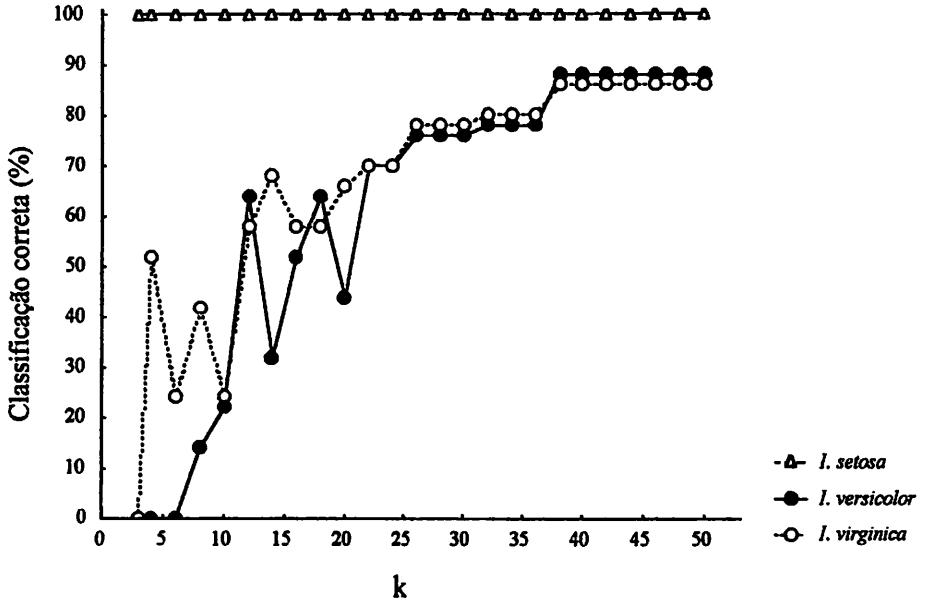

Figura 23 Percentual de classificação correta para as espécies do gênero Iris em função do número de partições imposta

4.2.2 Outros exemplos de agrupamentos

4.2.2.1 Esféricos

Em todos os métodos avaliados, a discriminação entre os grupos simulados em uma estrutura de agrupamento esférica (Figura 4.a) foi efetiva para os critérios da distribuição empírica de D^2 de Mahalanobis e as aproximações via F e χ^2. A taxa de classificação correta foi de 100% para quaisquer métodos, utilizando-se subamostragem ou não.

4.2.2.2 Elipsoidais

Os métodos de análise de discriminantes, também apresentaram resposta positiva à taxa de classificação correta entre os grupos simulados sob uma estrutura elipsoidal (Figura 4.c). O método quadrático apresentou um

desempenho levemente superior ao linear, entretanto estes podem ser considerados como idênticos. O uso ou não de subamsotragem também não indicou diferença entre os métodos empregados na técnica de análise de discriminantes (Tabela 5).

A concordância entre os critérios de distribuição empírica de D^2 de Mahalanobis e aproximações via F e χ^2 também foi assinalada nesta estrutura de agrupamento.

Tabela 5 **Percentuais de classificação correta para a simulação de um agrupamento elipsoidal em função dos diferentes métodos nas técnicas análise de discriminantes {AD} e técnicas hierárquicas {HIER}**

		a		*b*		*c*		*Geral*	
Técnicas	**Métodos**	**B**	**S**	**B**	**S**	**B**	**S**	**B**	**S**
	Linear	100,0	100,0	100,0	100,0	92,5	92,5	97,5	97,5
{AD}	Quadrático	100,0	100,0	97,5	100,0	95,0	94,0	97,5	98,0
	Geral	**100,0**	**100,0**	**98,8**	**100,0**	**93,8**	**93,3**	**97,5**	**97,8**
	Simples	0,0	100,0	0,0	0,0	0,0	0,0	0,0	33,3
	Completa	100,0	100,0	95,0	100,0	92,5	100,0	95,8	100,0
{HIER}	UPGMA	100,0	100,0	97,5	100,0	82,5	100,0	93,3	100,0
	Ward	100,0	100,0	100,0	100,0	85,0	100,0	95,0	100,0
	Geral	**75,0**	**100,0**	**73,1**	**75,0**	**65,0**	**75,0**	**71,0**	**83,3**

Onde: B - Dados brutos; S - Dados sob subamostragem; a, b e c - populações simuladas para os outros agrupamentos

Os métodos hierárquicos apresentaram resultados semelhantes aos da análise de discriminantes, à exceção do método de ligação simples, que apresentou elevadas taxas de classificação errônea, tanto nos casos sob subamostragem ou não. Este método apresentou-se efetivo somente na determinação da população que apresentava maior esfericidade, dada esta sob subamostragem (Tabela 5).

Todos os outros métodos hierárquicos apresentaram valores superiores a 85%, no caso dos dados brutos. Dada a subamostragem a taxa de classificação para estes foi de 100%. Em todos os métodos hierárquicos, sem exceção

assinalou-se contiguidade entre os subconjuntos evidenciados pela aproximação da D^2 via F e χ^2. Os métodos de ligação completa e de Ward apresentaram respostas mais concordantes com o padrão simulado, sendo indicados para agrupamentos desta natureza.

4.2.2.3 Pobremente separados

No caso de dados simulados com uma estrutura pobremente discriminada (Figura 4.b), a redução na taxa de classificação correta foi drástica no caso dos métodos hierárquicos (Tabela 6), especialmente os métodos de ligação simples e UPGMA, que apresentaram, em média, esta taxa abaixo de 5% (Tabela 6). Os métodos de ligação completa e de Ward apresentaram maiores taxas de classificação correta, respectivamente 48 e 65%, nos casos em que não foi utilizada subamostragem.

Mesmo com a redução do poder de discriminação, não foram indicados agrupamentos diferentes pelos critérios empírico e aproximações.

Tabela 6 **Percentuais de classificação correta para a simulação de um agrupamento em estrutura pobremente separada em função dos diferentes métodos nas técnicas análise de discriminantes {AD} e técnicas hierárquicas {HIER}**

		a		*b*		*c*		*Geral*	
Técnicas	**Métodos**	**B**	**S**	**B**	**S**	**B**	**S**	**B**	**S**
{AD}	Linear	67,5	72,5	55,0	45,0	70,0	72,5	64,0	63,3
	Quadrático	67,5	72,5	55,0	45,0	75,0	72,5	66,0	63,3
	Geral	**67,5**	**72,5**	**55,0**	**45,0**	**73,0**	**72,5**	**65,0**	**63,3**
{HIER}	Simples	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	Completa	50,0	0,0	57,5	0,0	35,0	0,0	48,0	0,0
	UPGMA	12,5	0,0	0,0	0,0	0,0	0,0	4,2	0,0
	Ward	67,5	100,0	55,0	100	73,0	100,0	65,0	100,0
	Geral	**32,5**	**25,0**	**28,1**	**25,0**	**27,0**	**25,0**	**29,0**	**25,0**

Onde: B Dados brutos; S Dados sob subamostragem; a, b e c populações simuladas para os outros agrupamentos

Os métodos em análises de discriminantes apresentaram valores sempre inferiores ou iguais aos obtidos através do método de Ward (Tabela 6). Somente o método de Ward apresentou taxa de classificação correta elevada no caso de subamostragem, tendo este obtido uma taxa igual a 100% em todos os grupos simulados.

A contigüidade foi assinalada somente nas situações em que se utilizou subamostragem, sendo que todos os métodos hierárquicos a assinalaram, apesar da baixa discriminação destes.

O método de Ward apresenta-se como a alternativa mais viável para a condução de análises desta natureza, já que seu poder de discriminação foi aceitável em todas as populações avaliadas. Este método apresenta como principal empecilho a monotonicidade. Entretanto, esta característica não é exclusividade deste método, já que todos os métodos hierárquicos também são monotônicos (Everitt, 1981). A sua limitação consiste na criação de agrupamentos com uma nova escala de distâncias muito maior, entretanto equivalente à original da matriz de distâncias, o que dificulta a inspeção visual dos resultados.

4.2.2.4 Número de objetos e Σ desiguais

As populações com estrutura simulada de número de objeto e matriz de covariância desiguais assinalaram um componente esférico e outro representando esta estrutura proposta (Figura 5). A taxa de classificação correta nos métodos da técnica de análise de discriminantes foi comparável ao caso pobremente discriminado, no qual os valores de classificação correta para a simulação foram de 63%, havendo redução do poder de discriminação nos casos em que foi utilizada subamostragem (Tabela 7).

Os critérios empírico e aproximações apresentaram concordância na determinação dos agrupamentos nesta estrutura de agrupamento.

Os métodos hierárquicos também apresentaram resposta semelhante ao caso do agrupamento pobremente separado, no qual ligação simples e UPGMA tiveram redução no poder de discriminação, especialmente para o primeiro, que apresentou taxa nula para as populações com menor esfericidade.

Os métodos de ligação completa e de Ward apresentaram maior discriminação e legitimaram os agrupamentos. Um adendo refere-se ao fato de que, mais uma vez, somente o método de Ward assinalou taxa de classificação correta elevada nos casos em que se utilizou subamostragem (Tabela 7).

Tabela 7 **Percentuais de classificação correta para a simulação de um agrupamento com estrutura de número de objetos e Σ desiguais em função dos diferentes métodos nas técnicas análise de discriminantes {AD} e técnicas hierárquicas {HIER}**

		a		*b*		*c*		*Geral*	
Técnicas	**Métodos**	**B**	**S**	**B**	**S**	**B**	**S**	**B**	**S**
	Linear	100,0	100,0	100,0	42,5	88,3	48,3	96,1	63,6
{AD}	Quadrático	100,0	100,0	98,0	42,5	95,0	46,5	97,7	63,0
	Geral	**100,0**	**100,0**	**99,0**	**42,5**	**91,7**	**47,4**	**96,9**	**63,3**
	Simples	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	Completa	100,0	0,0	97,5	0,0	93,3	0,0	96,9	0,0
{HIER}	UPGMA	100,0	0,0	90,0	0,0	56,7	0,0	82,2	0,0
	Ward	100,0	100,0	97,5	100,0	90,0	100,0	95,8	100,0
	Geral	**75,0**	**25,0**	**71,3**	**25,0**	**60,0**	**25,0**	**68,8**	**25,0**

Onde: B Dados brutos; S Dados sob subamostragem; a, b e c populações simuladas para os outros agrupamentos

Também foi assinalado contiguidade em todos os métodos hierárquicos, apesar da baixa discriminação propiciada por estes.

4.2.2.5 Arranjos não convencionais

Este tipo de agrupamento (Figura 4.d) apresentou apesar da estrutura um pouco mais complexa, elevada taxa de classificação correta nos métodos em análise de discriminantes, perfazendo um valor médio global acima de 99% (Tabela 8).

Também nesta estrutura, os critérios empírico e aproximações foram concordantes na determinação dos agrupamentos. Os métodos hierárquicos, com exceção do de ligação simples, apresentaram também elevada taxa de classificação correta. O método de ligação simples apresentou taxa de classificação correta igual a 100% em todos os casos em que foi utilizada subamostragem, bem como todos os outros métodos hierárquicos (Tabela 8).

Mesmo em casos de maior redução da taxa de classificação correta os valores neste exemplo não foram inferiores a 80%. A contiguidade dos subconjuntos foi assinalada em todos os métodos hierárquicos.

Tabela 8 **Percentuais de classificação correta para a simulação de um agrupamento com estrutura não convencional em função dos diferentes métodos nas técnicas análise de discriminantes {AD} e técnicas hierárquicas {HIER}**

		a		*b*		*c*		*Geral*	
Técnicas	**Métodos**	**B**	**S**	**B**	**S**	**B**	**S**	**B**	**S**
{AD}	Linear	100	100	100	100	98	97,5	99	99,2
	Quadrático	100	100	100	100	98	97,5	99	99,2
	Geral	**100**	**100**	**100**	**100**	**98**	**97,5**	**99**	**99,2**
{HIER}	Simples	0,0	100	0,0	100	0,0	100	0,0	100
	Completa	100	100	100	100	95	100	98	100
	UPGMA	100	100	100	100	80	100	93	100
	Ward	100	100	100	100	90	100	97,0	100
	Geral	**75**	**100**	**75**	**100**	**66**	**100**	**72**	**100**

Onde: B Dados brutos; S Dados sob subamostragem; a, b e c populações simuladas para os outros agrupamentos

4.2.3 Ordenação entre os métodos hierárquicos

Considerando a taxa de classificação correta, o número de partições e de subconjuntos em cada população, bem como a contiguidade dos subconjuntos, tem-se uma definição do padrão de resposta dos métodos utilizados em função destes critérios de avaliação do poder de discriminação dos métodos.

De modo geral, dada uma maior variação da população, o poder de discriminação dos métodos decai abruptamente. Todos os métodos apresentaram correlação negativa entre a taxa de classificação correta com o $tr(\Sigma)$ e o $|\Sigma|$, sendo significativa somente com $|\Sigma|$ ($r_s$min=-0,52) (Figura 24).

Pouca discriminação, também indicada pelo número de partições obtidas, assinala um decréscimo, já que a relação entre estas foi positiva e significante ($r_s$min=0,42). Entretanto, um elevado número de partições relaciona-se com uma baixa na contiguidade destes ($r_s$min=-0,45), valendo ressaltar que a contiguidade *per si* não assegura discriminação entre os objetos.

Tomando-se a relação categorizada entre $|\Sigma|$ e a taxa de classificação correta, obtém-se sob subamostragem somente nos casos em que há menores valores do determinante da matriz de covariância amostral, e por extensão, maior esfericidade. Tem-se uma elevada taxa de classificação correta, à exceção do método de Ward que a apresenta elevada em qualquer estrutura de covariância (Figura 24).

Os métodos de ligação simples e UPGMA apresentam-se como estreitamente dependentes da estrutura de variação para uma maior discriminação dos agrupamentos. Nestes casos, a taxa de classificação correta reduziu-se rapidamente em função de uma menor homogeneidade nas populações.

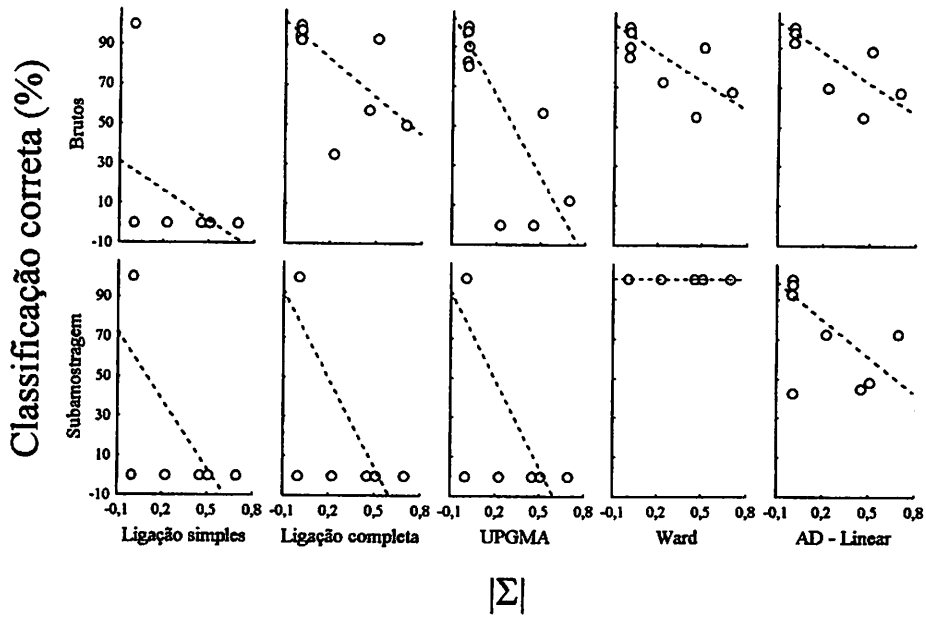

Figura 24 **Relação entre |Σ| e taxa de classificação correta nos métodos hierárquicos em função do tipo de amostragem utilizada**

Já os métodos de ligação completa e de Ward preservaram um nível de discriminação aceitável, mesmo em condições de menor homogeneidade. Vale ressaltar que sob subamostragem, apenas o método de Ward apresentou taxa de classificação correta elevada, enquanto, no caso da não utilização de subamostragem, os dois apresentaram taxas de classificação semelhantes (Figura 24).

De modo geral, o uso de subamostragem comprometeu a taxa de classificação correta, à exceção do métodos de ligação simples e de Ward.

Dentre todos os métodos, o de Ward e o de ligação completa, dada a não utilização de subamostragem, foram os que tiveram melhor desempenho quanto à taxa de classificação correta. Nestes métodos também foram assinalados os maiores números de partições. Entretanto, estes não apresentaram necessariamente contiguidade entre os subconjuntos das populações.

Os métodos hierárquicos foram os mais influenciados pela estrutura de variação dos dados. À exceção das populações com maior homogeneidade, leia-se esfericidade, a taxa apresentou-se reduzida. Métodos como o de Ward, sob qualquer situação de amostragem, e o de ligação completa e UPGMA, sem subamostragem, foram os que apresentaram maior conservação da taxa de classificação correta, apresentando menor dependência da esfericidade nas populações.

Assim, determina-se que o ponto de corte indicado pela distribuição empírica de D^2 de Mahalanobis e sua aproximação através de uma T^2 de Hotelling via χ^2 apresentam concordância e consistência. Métodos mais segregativos, como o de ligação completa e Ward, apresentaram melhores resultados, mesmo com uma maior heterogeneidade nas populações avaliadas. Estas técnicas apresentaram resultados semelhantes e muitas vezes superiores aos obtidos pelos métodos linear e quadrático de análise de discriminantes.

5 DISCUSSÃO

O comportamento da distribuição empírica da estatística D^2 de Mahalanobis, reforçou as prerrogativas de sensibilidade da T^2 de Hotelling às alterações de assimetria e curtose, especialmente a primeira (Mardia, 1970). A partir da estabilização destas razões, a distribuição empírica também direcionou-se a um valor assintótico. Mesmo tratando-se de uma aproximação, o critério avaliado assinala a sensibilidade da T^2 de Hotelling, o que reforça a consistência da aproximação. O limiar indicado neste caso foi acima de 80 observações.

Assim, diagramas de razão de momentos têm um papel importante na determinação de similaridade entre distribuições e/ou aproximações, com base no fornecido pela distribuição empírica, podem-se observar semelhanças com a distribuição F e creditar confiança à aproximação avaliada (Stuart e Ord, 1994). Outra informação importante fornecida por este foi a independência da matriz de covariância, que não afetou o comportamentos das razões de momentos, mesmo sem a inclusão de medidas de variação generalizada na aproximação, o que não implica necessariamente em independência destas medidas, especialmente em casos de não-centralidade (Pearson e Tiku, 1970), nas quais esta não se apresentou afetada.

A aproximação de T^2 de Hotelling via F, mesmo quando comparada com a distribuição exata, apresenta elevada acuidade (Hughes e Saw, 1971). Assim, a validade da aproximação indicada para a distribuição empírica de D^2 de Mahalanobis é a partir de um maior número de objetos. Nos casos em que o $n \leq 80$ objetos, a aproximação via χ^2 apresentou-se como muito mais eficiente. Assim, pela maior concordância com a distribuição empírica da D^2 de Mahalanobis e

pela facilidade de determinação, a aproximação via χ^2 é indicada como critério na determinação de agrupamentos legítimos.

Assim, a aplicação do critério de corte para determinação dos agrupamentos legitima-se dado um número de objetos, dados empiricamente por $n<80$. Regras desta natureza, empregadas especialmente em teoria assintótica, são fartas na literatura aplicada (Sneath e Sokal, 1973; Manly, 1994; Mardia, Kent e Bibby, 1995).

O rigor exigido para a determinação de um número de objetos ótimo ainda não pode ser cumprido, especialmente no que se refere a problemas de ruptura da normalidade multivariada, situação sob a qual a distribuição empírica foi obtida. Mesmo com todas convenções definidas, não foi possível determinar qual seria este valor exato, especialmente por tratar-se de uma aproximação, eleita por sua facilidade de implementação. Entretanto, todos os deméritos indicados podem ser reduzidos a problemas decorrentes da própria natureza assintótica do trabalho.

Exemplos citados na literatura, restritos à distribuição da estatística generalizada T^2 de Hotelling (Ito, 1960; Constantine, 1966; Davis, 1970), provam que resultados desta natureza podem não representar a solução definitiva, mas auxiliam muito no esclarecimento dos processos sob avaliação.

A natureza dos agrupamentos foi a grande responsável pelo melhor desempenho dos métodos avaliados, sendo que os agrupamentos que apresentaram maior esfericidade em suas populações tiveram uma maior taxa de classificação correta.

Os métodos de análises de discriminantes apresentaram resultados bastante satisfatórios em todas as situações, porém, os métodos linear e quadrático não apresentaram diferença sensível no exemplo real ou nos simulados. Nesta técnica, a subamostragem apresentou-se como não aconselhável, devendo-se ao fato de que a D^2 de Mahalanobis calculada utilizava-

se da matriz de covariância da população, e estas, por sua vez, eram combinadas. Como a subamostragem foi sem reposição, os resultados expressos por esta abordagem assinalam um resultado singular. A utilização de técnicas de reamostragem como *jackkniffe* ou *bootstrap*, podem vir a sanar este contratempo (Manly, 1994).

Um adendo aos uso mais restrito da técnica de análise de discriminantes, apesar do bom desempenho promovido por esta, é a sua extrinsicidade, o que limita o caráter heurístico em análises exploratórias e é o grande atrativo da análise de agrupamento (Henery, 1994)

O método particional *k-means* apresentou limitações inerentes a sua implementação, especialmente ao número de partições a serem definidos *a priori*, o que compromete de maneira decisiva sua utilização em análises exploratórias (Everitt, 1981). Mesmo utilizando-se o número de partições indicado ou valores superiores, a taxa de classificação correta foi muito baixa, quando comparada aos outros métodos utilizados. No exemplo real com o banco de dados do gênero *Iris*, sua taxa foi inferior a 90%, sendo que o número de partições utilizadas foi pelo menos vinte vezes maior que os obtidos nos métodos hierárquicos, que apresentaram, em média, 98% de classificação correta. Deste modo, estes métodos não são recomendados, já que todos os outros apresentaram melhor desempenho e maior facilidade de implementação.

Os métodos hierárquicos, baseados no ponto de corte indicado pela distribuição empírica e pelas aproximações da T^2 de Hotelling via F e χ^2, apresentaram resultados consistentes. Em todas as situações, as aproximações apresentaram o mesmo ponto de divergência entre os agrupamentos que a distribuição empírica.

Dentre os métodos hierárquicos, o de ligação simples apresentou os piores resultados, à exceção dos casos em que se utilizou subamostragem. Sua taxa de

classificação correta somente elevava-se nos casos em que as populações apresentavam maior esfericidade. Assim, este método apresenta desempenho satisfatório somente nos casos em que os agrupamentos têm elevada densidade (Hartingan, 1981), o que não é comumente assinalado em pesquisa agropecuária.

Já os outros apresentaram melhores taxas; os métodos UPGMA e de ligação completa somente em condições em que não foi utilizada subamostragem, e o de Ward em qualquer tipo de amostragem.

Sobre o papel da subamostragem, infere-se mais uma vez sobre a não reposição da amostragem e da não extensividade. As alternativas são as mesmas citadas anteriormente para as técnicas de análise de discriminantes, consistindo de técnicas de computação intensiva, como *jackkniffe* e *bootstrap*. Estudos posteriores poderão validar esta hipótese, agora em aberto.

Deste modo, propõe-se que técnica hierárquica deva ser utilizada, já que alguns de seus métodos apresentaram valores iguais ou superiores aos obtidos através da técnica de análise de discriminante, servindo como melhor resultado teórico de discriminação, além de conservarem o caráter intrínseco requerido para a análise de agrupamentos. Os métodos mais segregativos apresentaram melhores resultados na taxa de classificação correta (Kuiper e Fisher, 1975).

6 CONCLUSÃO

A observação da razão entre momentos da distribuição empírica auxilia bastante na descrição de seu comportamento, servindo também como um indicador de aderência. A sensibilidade da distribuição T^2 de Hotelling às alterações na assimetria e curtose, especialmente a primeira, também foram assinaladas na distribuição empírica de T^2 de Hotelling. Com o aumento do número de objetos, as distribuições tendem a uma estabilidade.

As aproximações da distribuição empírica de D^2 de Mahalanobis pela T^2 de Hotelling via aproximação F e χ^2 apresentaram independência da estrutura de covariância. O número de objetos e variáveis influi muito no ajuste das distribuições, sendo que, em ambos os casos, a relação foi negativa, sendo o número de objetos o maior responsável pela qualidade dos ajustes e os melhores ajustes obtidos com $n>80$.

As técnicas hierárquicas, mesmo quando comparadas à de análise de discriminantes, tida como o melhor valor teórico de discriminação, apresentaram taxas de classificação correta equivalentes ou superiores.

A natureza dos dados teve uma influência determinante sobre a taxa de classificação correta nos métodos empregados. Os métodos mais segregativos, como o de Ward, ligação simples e UPGMA, apresentaram resultados satisfatórios, especialmente o de Ward e ligação completa, que se apresentaram menos influenciados pela menor homogeneidade das populações.

O critério avaliado deve ser utilizado, desde que o número de objetos empregado seja superior ao limite indicado empiricamente.

Estudos baseados em simulação apresentam respostas consistentes, mesmo que não rigorosas. Recomenda-se a utilização de metodologia afim em outros problemas de natureza multivariada.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, T.W. **An introduction to multivariate statistical analysis**. 2.ed. John Willey & Sons, 1984. 675p.

ANDREWS, D.F. Plots of high-dimensional data. **Biometrics**, Washington, v.28, n,1, p.125-136, Mar. 1972.

BEALS, E.W. Bray-Curtis Ordination: An effective strategy for analysis of multivariate ecological data. **Advances in Ecological Research**, London, v.14, p.1-55, 1984.

BRYANT, P. Geometry, Statistics, Probability: Variations on a common theme. **The American Statistician**, Washington, v.38, n.1, p.38-48, Feb. 1984.

CHATFIELD, C.; COLLINS, A.J. **Introduction to multivariate analysis**. London: Chapman & Hall, 1986. 246p.

CHERNOFF, H. Using faces to represent points in k-dimensional space graphically. **Journal of the American Statistical Association**, Washington, v.68, n.342, p.361-368, June 1973.

CONSTANTINE, A.G. The Distribution of Hotelling's generalized T_0^2. **Annals of Mathematical Statistics**, Baltimore, v.37, n.3, p.215-225, June 1966.

COOK, R. Detection of influential observations in linear regression models. **Technometrics**, Washington, v.19, n.1, p.15-18, Feb. 1977

CORMACK, R.M. A Review of classification. **Journal of Royal Statistical Society**, Serie A, London, v.134, n.3, p.321-367, Nov. 1971.

DAVIS, A.W. Exact Distribution of Hotelling's T_0^2. **Biometrika**, London, v.57, n.1, p.187-191, Apr. 1970.

DILLON, W.R.; GOLDSTEIN, M. **Multivariate analysis:** methods and applications. New York: John Willey & Sons, 1984. 575p.

DOLBY, G.R. The role of statistics in methodology in life science. **Biometrics**, Washington, v.38, n.4, p.1069-1083, Dec. 1982.

ENDRESS, P.K. **Diversity and evolutionary biology of tropical flowers**. London: Cambridge University Press, 1996. 511p.

EVERITT, B.S. **Cluster analysis**. 2.ed. London: Social Science Research Council/ Halsted Press, 1981. 136p.

EVERITT, B.S. Unresolved problems in cluster analysis. **Biometrics,** Washington, v.35, n.1, p.169-181, Mar. 1979.

FISHER, R.A. The use of multiple measurement in taxonomic problems. **Annals of Eugenetics,** New York, v.7, p.179-188, 1936.

FISHER, W.D. On grouping for maximum homogeneity. **Journal of the American Statistical Association,** Washington, v.53, n.3, p.789-798, Oct. 1958.

FORGY, E.W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. **Biometrics,** Washington, v.21, n.3, p.768-769, Sept. 1965

FRALEY, C. Algorithms for model-based Gaussain hierarchical clustering. SIAM **Journal on Scientific Computing,** New York, v.20, n.1, p.270-281, 1998

FRALEY, C.; RAFTERY, A.E. How many clusters? Which clustering method ? Answers via model-based cluster analysis. **The Computer Journal,** Cambridge, v.41, n.8, p.578-588, 1998.

FRALEY, C.; RAFTERY, A.E. MCLUST: Software for model-based cluster and discriminant analysis: user's guide. 1999.

FRIEDMAN, H.P.; GOLDBERG, J.D. Meta-analysis: An Introduction and Point of View. **Hepatology,** v.23, n.4, p.917-928, 1996.

FRIEDMAN, H.P.; RUBIN, J. On some invariant criteria for grouping data. **Journal of Amerian Statistical Association,** Washington, v.62, n.320, p.1159-1178, Dec. 1967.

GAUCH JR., H.G. **Multivariate analysis in community ecology**. New York: Cambridge University Press, 1982. 384p.

GIRI, N.C. **Multivariate statistical analysis**. New York: Marcel Dekker, 1996. 378p.

GOODALL, D.W. Hypotesis testing in classification. **Nature,** London, v.11, n.5045, p.329-330, July 1966.

GOWER, J.C. A general coefficient of similarity and some one of its properties. **Biometrics,** Washington, v.27, n.4, p.857-872, Dec. 1971.

HARTINGAN, J.A. Consistensy of single linkage for high-density clusters. **Journal of the American Statistical Association,** Washington, v.76, n.2, p.388-394, June 1981

HARTINGAN, J.A.; HARTINGAN, P.M. The dip test of unimodality. **Annals of Statistics,** Baltimore, v.13, n.1, p.80-84, Jan. 1985.

HENERY, R.J. Classification. In: MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C.C. (eds). **Machine learning, neural and statistical classification**. 1994. 290p.

HUGHES, D.T.; SAW, J.G. Aproximating the percentage points of Hotelling's generalized T_0^2 statistics. **Biometrics,** Washington, v.24, n.1, p.224-226, Mar. 1971.

ITO, K. Asymptotic formulae for the distribution of Hotelling's generalized T_0^2 statistics, II. **Annals of Mathematical Statistics,** Washinton, v.31, n.2, p.1148-1153, June 1960

JOHNSON, N.L.; KOTZ, S. **Distributions in statistics:** continuous multivariate distributions. New York: John Willey & Sons, 1972. 333p.

JOHNSON, N.L.; KOTZ, S. **Distributions in statistics:** continuous univariate distributions 1. New York: John Willey & Sons, 1970a. 300p.

JOHNSON, N.L.; KOTZ, S. **Distributions in statistics:** continuous univariate distributions 2. New York: John Willey & Sons, 1970b. 306p.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate Statistical analysis**. 4.ed. New Jersey: Prentice Hall, 1998. 815p.

KHATTREE, R.; NAIK, D.N. **Applied multivariate statistics with SAS® software**. Cary: SAS Institute, 1995. 396p.

KENNEDY JR., W.J.; GENTILE, J.E. **Statistical computing**. New York: Marcel Dekker, 1980. 591p.

KRUSKAL, J.B.; LANDWEHF, J.M. Icicle plots: Better displays for hierarchial clustering. **The American Statistician,** Washington, v.37, n.2, p.162-168, May 1983.

KUIPER, F.K.; FISHER, L. A Monte Carlo comparsion of six clustering procedures. **Biometrics,** Washington, v.31, n.3, p.777-783, Sept. 1975.

KUO, A. **The macro distance**: technical report. Cary, NC.: SAS Institute, 1997. 33p.

LEBART, L.; MORINEAU, A.; WARWICK, K.M. **Multivariate descriptive statistical analysis**: correspondence analysis and related techniques for large matrices. New York: John Willey & Sons, 1984. 231p.

LING, R.F. A probability theory of cluster analysis. **Journal of the American Statistical Association,** Washington, v.68, n.341, p.159-164, Mar. 1973.

MANLY, B.F.J. **Multivariate statistical methods:** a primer. 2.ed. London: Chapman & Hall, 1994. 215p.

MANN, C. Meta-analysis in breech. *Science*: Research News. August, 3th. 1990. p.476-480.

MARDIA, K.V. Measures of multivariate skewness and kurtosis with applications. **Biometrika,** London, v.57, n.3, p.519-530, Dec. 1970.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. **Multivariate analysis.** London: Academic Press, 1995. 518p.

ORLOCI, L. Geometric models in ecology I. The theory and applications of some ordination methods. **Journal of Ecology,** Oxford, v.54, p.193-215, 1966.

PEARSON, E.S.; TIKU, M.L. Some notes on the relationship between the distributions of central and non-central F. **Biometrika,** London, v.57, n.1, p.175-179, Apr. 1970

PIELOU, E.C. **The interpretation of ecological data.** New York: John & Wiley Sons, 1984. 263p.

PURI, M.L.; SEN, P.K. **Nonparametric methods in multivariate analysis.** New York: John Willey & Sons, 1971. 440p.

RAMIREZ, D.E. The generalized F distribution. **Journal of Computational Statistics,** v.3, 1998.15p.

ROHWER, R.; WYNNE-JONES, M.; WYSOTZKI, F. Neural Network In: MICHIE, D.; SPIEGELHALTER, D.J.; TAYLOR, C.C. (eds). **Machine learning, neural and statistical classification.** 1994. 290p.

SARLE, W.W. Introduction to clustering procedures. In: SAS INSTITUTE. **SAS/STAT user's guide, version 6.** 4.ed. Cary, NC, 1990. v.1, 889p.

SAS INSTITUTE. **SAS/STAT user's guide, version 6.** 4.ed. Cary, NC, 1990. v.1, 889p.

SAVILLE, D.J.; WOOD, G.R. A method for teaching Statistics using N-dimensional Geometry. **The American Statistician,** Washington, v.40, n.3, p.205-214, Aug. 1986.

SCHWARZER, R. **Meta-analysis user guide.** Berlin: Institut für Psychologie. Freie Universität, 1989. 48p.

SILVERMAN, B.W. **Density Estimation.** New York: Chapman and Hall, 1992. 175p.

SNEATH, P.H.A.; SOKAL, R.R. **Numerical taxonomy:** The principles and practice of numerical classification. San Francisco: W. H. Freeman and Company, 1973. 573p.

STATSOFT. **STATISTICA for Windows** [Computer program manual]. 1996.

STUART, A.; ORD, J.K. **Kendall's Advanced Theory of Statistics**. New York: John Willey & Sons, 1994. 676p.

Índice de anexos

ANEXOS

ANEXO 1 Macro para simulação da distribuição empírica da D^2 de Mahalanobis, sob condição de normalidade multivariada central

```
options nodate nonumber ps=1000 ls=76;
proc iml;
  create moises var {x1, x2, x3, x4, x5, x6, x7, x8, x9, x10};
  exp=5000;                          /* número de simulações */
  sig={1.0 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5,
       0.5 1.0 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5,
       0.5 0.5 1.0 0.5 0.5 0.5 0.5 0.5 0.5 0.5,
       0.5 0.5 0.5 1.0 0.5 0.5 0.5 0.5 0.5 0.5,
       0.5 0.5 0.5 0.5 1.0 0.5 0.5 0.5 0.5 0.5,
       0.5 0.5 0.5 0.5 0.5 1.0 0.5 0.5 0.5 0.5,
       0.5 0.5 0.5 0.5 0.5 0.5 1.0 0.5 0.5 0.5,
       0.5 0.5 0.5 0.5 0.5 0.5 0.5 1.0 0.5 0.5,
       0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 1.0 0.5,
       0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 1.0};
p=nrow(sig);n=250;                    /* número de objetos */
  mu={0,0,0,0,0,0,0,0,0,0};           /* vetor de médias */
  s=root(sig);
  do ii=1 to exp;
    do i=1 to n;
      x1=normal(0); x2=normal(0); x3=normal(0); x4=normal(0);
x5=normal(0); x6=normal(0); x7=normal(0); x8=normal(0);
x9=normal(0); x10=normal(0);
      y=x1||x2||x3||x4||x5||x6||x7||x8||x9||x10;
      z=y*s;
      x1=z[1]; x2=z[2]; x3=z[3]; x4=z[4]; x5=z[5]; x6=z[6];
x7=z[7]; x8=z[8]; x9=z[9]; x10=z[10];
      append var {x1, x2, x3, x4, x5, x6, x7, x8, x9, x10};
    end;
  end;
quit;

proc IML;
  create t2ps var {d2};
  use moises;
  exp=5000;                           /* número de simulações */
  n=250;                              /* número de objetos */
  do i=1 to exp;
    read next 250 into X;             /* número de objetos */
    n=nrow(X);p=ncol(X);
    q=i(n) - (1/n)*j(n,n,1);          /* criando q=I-1/nJ, auxiliar */
    S=(1/(n-1))*x`*q*x;               /* matriz de covariancias não-
viesada */
```

```
    S_inv=inv(S);                          /* inversa de S  */
    do ii=1 to n-1;
     do jj=ii+1 to n;
       x1=x[ii,];
       x2=x[jj,];
       d2=(x1-x2)*s_inv*(x1-x2)`;
       append var {d2};
     end;
    end;
  end;
quit;
proc univariate data=t2ps plots;
 var d2;
run;
proc univariate normal data=t2ps PCTLDEF=1;
 var d2;
 output out=saida p95=d2p95 p99=d2p99
        PCTLPRE=d2 PCTLPTS=90    PCTLNAME=p90
        PCTLPRE=d2 PCTLPTS=92.5  PCTLNAME=p925
        PCTLPRE=d2 PCTLPTS=95    PCTLNAME=p95
        PCTLPRE=d2 PCTLPTS=97.5  PCTLNAME=p975
        PCTLPRE=d2 PCTLPTS=99    PCTLNAME=p99
        PCTLPRE=d2 PCTLPTS=99.5  PCTLNAME=p995;
run;
proc print data=saida;
       run;quit;
```

ANEXO 2 Valores tabelados da distribuição empírica de D^2 de Mahalanobis em função no número de objeto para $2 \leq p \leq 10$, significância 95%

	Número de variáveis								
n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	4,000								
4	5,982	6,000							
5	7,593	7,989	8,000						
6	8,703	9,673	9,990	10,000					
7	9,312	10,853	11,691	11,990	12,000				
8	9,783	11,644	12,907	13,707	13,990	14,000			
9	10,165	12,209	13,824	15,009	15,736	15,990	16,000		
10	10,402	12,677	14,525	15,964	17,045	17,734	17,991	18,000	
11	10,546	13,089	14,967	16,709	18,048	19,076	19,745	19,991	20,000
12	10,644	13,252	15,425	17,306	18,832	20,126	21,106	21,748	21,991
13	10,924	13,488	15,744	17,763	19,487	20,923	22,178	23,120	23,746
14	10,950	13,767	16,072	18,144	19,966	21,601	22,993	24,221	25,148
15	10,950	13,879	16,276	18,437	20,396	22,146	23,720	25,065	26,224
16	11,051	13,994	16,544	18,741	20,760	22,616	24,293	25,804	27,119
17	11,165	14,023	16,621	19,021	21,153	23,019	24,762	26,380	27,857
18	11,162	14,157	16,791	19,021	21,363	23,376	25,247	26,912	28,534
19	11,216	14,276	16,949	19,336	21,635	23,686	25,574	27,408	29,024
20	11,302	14,347	17,100	19,546	21,773	23,928	25,913	27,794	29,555
21	11,295	14,478	17,146	19,645	21,980	24,173	26,174	28,124	29,902
22	11,432	14,483	17,236	19,735	22,139	24,393	26,496	28,430	30,295
23	11,407	14,538	17,367	19,856	22,370	24,582	26,662	28,676	30,641
24	11,388	14,612	17,454	19,875	22,473	24,717	26,874	28,948	30,940
25	11,407	14,674	17,513	20,067	22,570	24,888	27,127	29,202	31,183
26	11,454	14,696	17,517	20,251	22,659	24,997	27,239	29,359	31,477
27	11,469	14,716	17,601	20,250	22,794	25,149	27,457	29,623	31,630
28	11,541	14,712	17,645	20,347	22,883	25,248	27,545	29,841	31,830
29	11,548	14,797	17,668	20,428	22,992	25,391	27,771	29,972	32,064
30	11,514	14,859	17,820	20,492	23,035	25,516	27,784	30,110	32,262
35	11,580	14,964	17,943	20,728	23,429	25,909	28,344	30,672	32,945
40	11,659	15,073	18,039	20,940	23,639	26,237	28,689	31,106	33,462
45	11,695	15,107	18,208	20,958	23,813	26,433	29,001	31,412	33,859
50	11,744	15,171	18,310	21,185	23,969	26,642	29,191	31,625	34,118
55	11,746	15,240	18,334	21,257	24,045	26,756	29,361	31,882	34,379
60	11,722	15,240	18,404	21,352	24,196	26,847	29,522	32,044	34,567

(Continua ...)

(Continuação)

65	11,774	15,265	18,441	21,452	24,267	26,966	29,599	32,269	34,731
70	11,841	15,314	18,476	21,463	24,319	27,117	29,761	32,402	34,864
75	11,779	15,367	18,492	21,541	24,465	27,148	29,838	32,490	34,973
80	11,815	15,360	18,540	21,548	24,453	27,251	29,925	32,540	35,060
90	11,826	15,360	18,613	21,609	24,486	27,311	30,039	32,645	35,282
100	11,765	15,400	18,608	21,692	24,662	27,455	30,164	32,805	35,428
110	11,880	15,449	18,708	21,704	24,613	27,460	30,222	32,857	35,589
120	11,896	15,466	18,732	21,857	24,642	27,520	30,321	33,042	35,544
140	11,915	15,466	18,776	21,808	24,768	27,580	30,379	33,161	35,709
160	11,848	15,514	18,691	21,905	24,807	27,697	30,433	33,160	35,918
180	11,951	15,509	18,751	21,876	24,781	27,711	30,515	33,283	35,907
200	11,928	15,542	18,909	22,037	24,807	27,673	30,476	33,328	35,997
250	11,975	15,567	18,806	22,054	24,917	27,753	30,692	33,475	36,229

ANEXO 3 Valores tabelados da distribuição empírica de D^2 de Mahalanobis em função no número de objeto para $2 \leq p \leq 10$, significância 90%

	Número de variáveis								
n	2	3	4	5	6	7	8	9	10
3	4,000								
4	5,933	6,000							
5	7,203	7,953	8,000						
6	7,841	9,333	9,956	10,000					
7	8,201	10,170	11,378	11,957	12,000				
8	8,455	10,610	12,246	13,412	13,961	14,000			
9	8,565	10,935	12,835	14,393	15,445	15,960	16,000		
10	8,662	11,170	13,293	15,016	16,445	17,456	17,965	18,000	
11	8,732	11,401	13,573	15,540	17,156	18,488	19,485	19,962	20,000
12	8,801	11,473	13,809	15,922	17,692	19,238	20,541	21,489	21,964
13	8,906	11,606	14,005	16,142	18,086	19,787	21,315	22,566	23,487
14	8,881	11,691	14,184	16,384	18,412	20,218	21,880	23,383	24,593
15	8,944	11,775	14,274	16,558	18,632	20,593	22,403	23,985	25,403
16	8,928	11,839	14,402	16,738	18,893	20,911	22,743	24,441	26,056
17	8,981	11,855	14,472	16,888	19,094	21,142	23,085	24,878	26,528
18	8,984	11,901	14,549	16,888	19,222	21,384	23,394	25,241	27,017
19	8,947	11,977	14,619	17,061	19,433	21,527	23,597	25,555	27,327
20	9,018	12,003	14,716	17,146	19,490	21,724	23,780	25,782	27,697
21	9,007	12,045	14,740	17,232	19,605	21,849	23,954	25,998	27,935
22	9,087	12,077	14,778	17,283	19,681	21,998	24,159	26,214	28,160
23	9,033	12,109	14,824	17,360	19,845	22,108	24,261	26,362	28,417
24	9,042	12,090	14,877	17,340	19,877	22,170	24,394	26,519	28,599
25	9,063	12,145	14,909	17,446	19,961	22,307	24,525	26,684	28,752
26	9,068	12,156	14,912	17,545	19,952	22,306	24,617	26,782	28,924
27	9,093	12,177	14,952	17,590	20,047	22,414	24,703	26,950	29,035
28	9,083	12,149	14,975	17,606	20,089	22,481	24,789	27,070	29,159
29	9,086	12,181	14,968	17,655	20,148	22,542	24,895	27,144	29,324
30	9,091	12,196	15,057	17,685	20,189	22,600	24,946	27,241	29,446
35	9,107	12,249	15,105	17,771	20,378	22,850	25,249	27,576	29,853
40	9,133	12,289	15,165	17,918	20,497	23,033	25,457	27,829	30,163
45	9,141	12,319	15,224	17,908	20,585	23,147	25,637	28,000	30,383
50	9,152	12,322	15,256	18,009	20,688	23,255	25,727	28,115	30,542
55	9,152	12,363	15,279	18,050	20,725	23,291	25,845	28,279	30,719
60	9,151	12,366	15,313	18,088	20,799	23,370	25,921	28,380	30,808

(Continua ...)

(Continuação)

65	9,147	12,369	15,322	18,157	20,833	23,433	25,941	28,499	30,910
70	9,184	12,385	15,340	18,157	20,860	23,517	26,049	28,564	30,983
75	9,151	12,412	15,350	18,182	20,947	23,512	26,087	28,644	31,046
80	9,164	12,395	15,378	18,201	20,931	23,566	26,134	28,652	31,117
90	9,166	12,395	15,400	18,226	20,951	23,608	26,198	28,713	31,218
100	9,165	12,421	15,408	18,268	21,027	23,678	26,280	28,794	31,333
110	9,187	12,433	15,415	18,272	21,015	23,692	26,296	28,850	31,392
120	9,184	12,452	15,454	18,350	21,009	23,720	26,349	28,932	31,362
140	9,218	12,452	15,483	18,316	21,075	23,755	26,378	28,977	31,487
160	9,164	12,455	15,433	18,374	21,123	23,789	26,400	28,994	31,583
180	9,215	12,447	15,467	18,363	21,071	23,814	26,460	29,070	31,570
200	9,196	12,475	15,541	18,425	21,095	23,794	26,440	29,110	31,649
250	9,216	12,480	15,482	18,444	21,164	23,850	26,552	29,178	31,772

ANEXO 4 Valores tabelados da distribuição empírica de D^2 de Mahalanobis em função no número de objeto para $2 \leq p \leq 10$, significância 99%

	Número de variáveis								
n	2	3	4	5	6	7	8	9	10
3	4,000								
4	5,999	6,000							
5	7,915	7,999	8,000						
6	9,523	9,934	10,000	10,000					
7	10,743	11,634	11,941	12,000	12,000				
8	11,784	12,895	13,653	13,944	14,000	14,000			
9	12,618	14,043	15,062	15,671	15,940	16,000	16,000		
10	13,111	15,042	16,246	17,099	17,682	17,945	18,000	18,000	
11	13,741	15,659	17,178	18,370	19,168	19,688	19,950	20,000	20,000
12	14,034	16,188	18,057	19,401	20,416	21,191	21,702	21,949	22,000
13	14,662	16,716	18,594	20,227	21,486	22,450	23,241	23,689	23,950
14	14,706	17,216	19,331	20,959	22,320	23,568	24,483	25,252	25,723
15	15,028	17,590	19,691	21,618	23,165	24,464	25,672	26,515	27,224
16	15,198	17,930	20,246	22,079	23,853	25,287	26,605	27,691	28,583
17	15,406	18,253	20,546	22,654	24,467	26,025	27,362	28,668	29,732
18	15,536	18,509	20,867	22,654	24,951	26,641	28,144	29,524	30,743
19	15,692	18,683	21,251	23,420	25,472	27,258	28,833	30,363	31,608
20	15,882	18,950	21,510	23,798	25,875	27,781	29,518	31,013	32,501
21	16,002	19,196	21,716	24,165	26,282	28,223	30,030	31,648	33,094
22	16,335	19,324	21,964	24,278	26,509	28,601	30,483	32,223	33,789
23	16,310	19,320	22,254	24,783	26,879	28,992	30,871	32,698	34,400
24	16,309	19,649	22,393	24,635	27,172	29,325	31,271	33,214	34,886
25	16,427	19,710	22,519	24,978	27,481	29,672	31,782	33,580	35,377
26	16,571	19,763	22,635	25,389	27,687	29,885	32,055	34,004	35,860
27	16,581	20,040	22,871	25,556	27,966	30,183	32,358	34,432	36,249
28	16,595	19,977	22,912	25,653	28,066	30,449	32,566	34,814	36,561
29	16,725	20,088	23,088	25,893	28,353	30,655	32,984	35,074	37,041
30	16,733	20,274	23,338	25,928	28,503	30,960	33,098	35,198	37,375
35	16,912	20,533	23,717	26,472	29,327	31,802	34,220	36,528	38,753
40	17,249	20,938	24,021	27,057	29,824	32,535	34,979	37,350	39,682
45	17,261	21,045	24,387	27,179	30,261	33,019	35,553	37,995	40,524
50	17,404	21,234	24,691	27,758	30,643	33,330	36,079	38,484	41,041
55	17,513	21,389	24,755	27,899	30,875	33,659	36,406	39,032	41,532
60	17,340	21,468	24,988	28,126	31,152	33,911	36,739	39,377	41,890

(Continua ...)

(Continuação)

65	17,676	21,492	25,093	28,366	31,341	34,200	36,899	39,761	42,385
70	17,904	21,720	25,170	28,442	31,465	34,490	37,311	40,053	42,622
75	17,544	21,689	25,183	28,565	31,724	34,554	37,394	40,312	42,822
80	17,674	21,794	25,329	28,600	31,800	34,716	37,644	40,389	43,017
90	17,727	21,794	25,493	28,758	31,850	34,999	37,924	40,705	43,525
100	17,632	21,890	25,607	29,040	32,226	35,247	38,203	41,019	43,766
110	17,951	22,021	25,818	28,990	32,218	35,371	38,258	41,141	44,098
120	18,092	22,107	25,826	29,255	32,228	35,473	38,581	41,527	43,960
140	18,101	22,107	25,957	29,390	32,577	35,615	38,674	41,729	44,337
160	18,017	22,329	25,770	29,525	32,706	35,912	38,841	41,770	44,831
180	18,179	22,309	25,901	29,369	32,699	35,910	38,979	42,001	44,844
200	18,144	22,348	26,330	29,820	32,769	35,819	38,934	42,134	45,055
250	18,365	22,472	26,085	29,880	32,988	36,007	39,477	42,444	45,502