Documentos

ISSN 0101 - 9805
Dezembro, 2005
09

Análise de Agrupamento: Propriedades e Aplicações

Embrapa

Empresa Brasileira de Pesquisa Agropecuária
Centro de Pesquisa Agroflorestal de Roraima
Ministério da Agricultura, Pecuária e Abastecimento

ISSN 0101 – 9805
Dezembro, 2005

Documentos 09

Análise de Agrupamento: Propriedades e Aplicações

Moisés Mourão Jr.

Boa Vista, RR
2005

Exemplares desta publicação podem ser obtidos na:

Embrapa Roraima

Rod. BR-174 Km 08 - Distrito Industrial Boa Vista-RR

Caixa Postal 133.

69301-970 - Boa Vista - RR

Telefax: (095) 3626.7018

e-mail: sac@cpafrr.embrapa.br

www.cpafrr.embrapa.br

Normalização Bibliográfica: Maria José Borges Padilha

Editoração Eletrônica: Vera Lúcia Alvarenga Rosendo

1ª edição

1ª impressão (2005): 300

MOURÃO JÚNIOR, M. Análise de Agrupamentos: Propriedades e Aplicações. Boa Vista: Embrapa Roraima, 2005. 35.p (Embrapa Roraima. Documentos, 9)

1.Análise Multivariada. 2.Classificação. 3.Estatística Aplicada.

Autores

Moisés Mourão Jr.
Biólogo, M. Sc., pesquisador da Embrapa Roraima. BR 174, km 08. Distrito Industrial. Caixa Postal: 133. 69301-970. Boa Vista - Roraima, e-mail: mmourao@cpafrr.embrapa.br

SUMÁRIO

Análise de Agrupamento: Propriedades e Aplicações

Moisés Mourão Jr.

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 e 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; 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).

Este trabalho tem como objetivo apresentar a técnica multivariada análise de agrupamento [*cluster analysis*], assinalando sua dimensionalidade, propriedades, estimadores de similaridade entre objetos e os procedimentos de amalgamação destes objetos. A ausência de indicação de recursos computacionais, nesta publicação, justifica-se pela veiculação desta abordagem em outra publicação específica.

O 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 (Fig.1).

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 (Fig. 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).

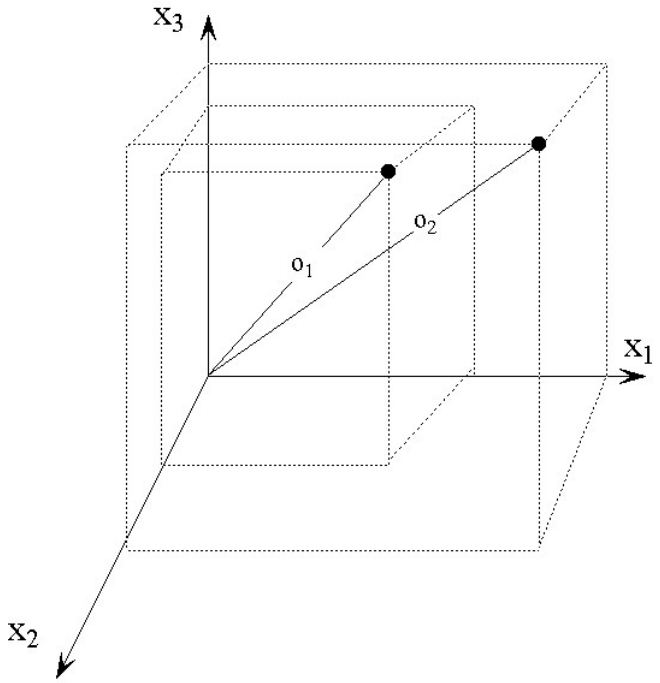

Fig.1 Realização de variáveis aleatórias em um espaço euclidiano R^3

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 (Fig. 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 (Fig. 3.b) (Pielou, 1984).

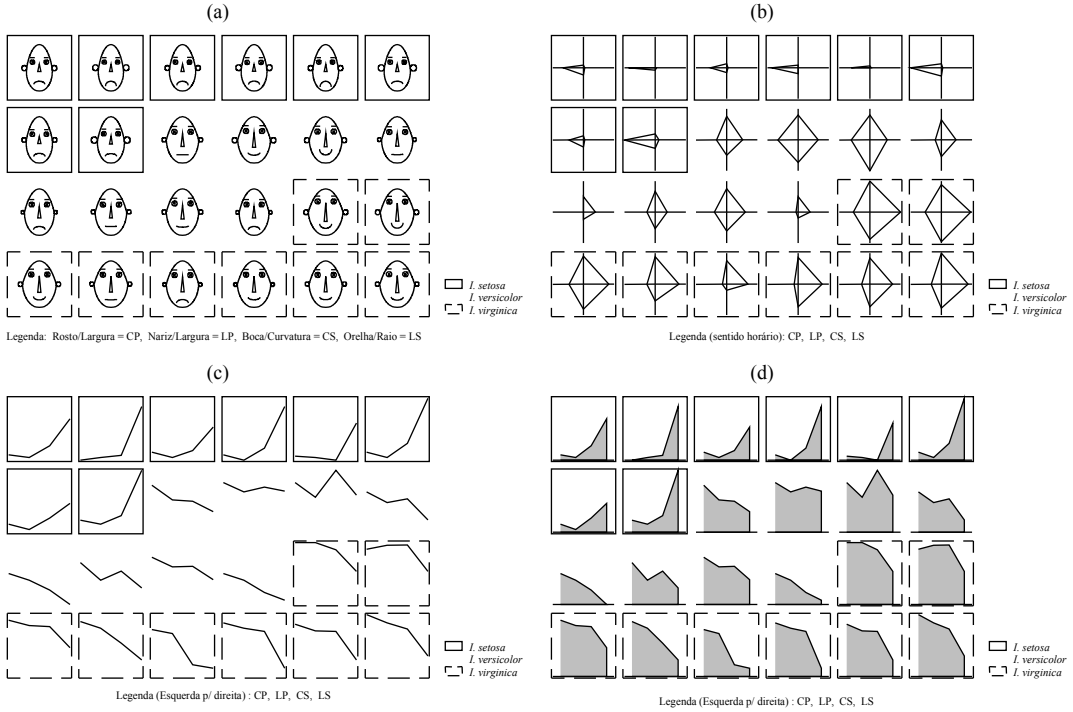

Fig. 2 Representações alternativas de processos multivariados sob a forma de iconográficos: (a) faces de Chernoff, (b) raio de sol, (c) linhas e (d) perfil de algumas observações do banco de dados *Iris.* (Fisher, 1936)

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).

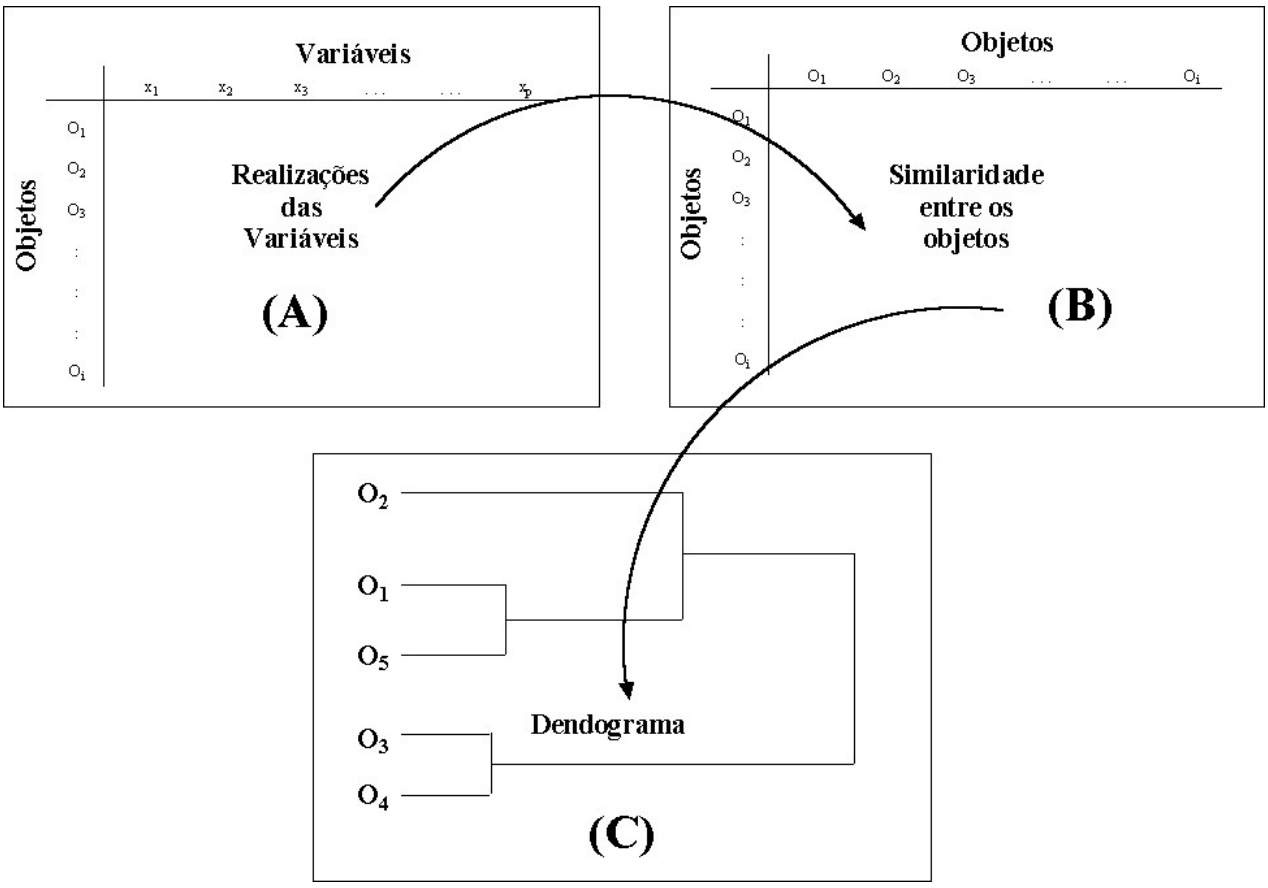

Fig. 3 Passos da análise de agrupamento

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).

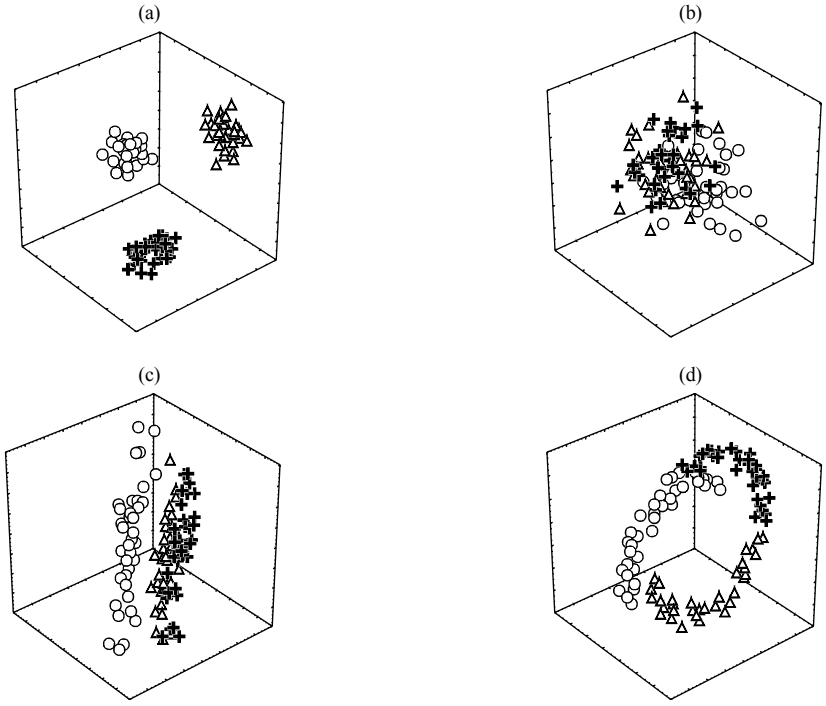

Fig. 4 Estruturas de agrupamento (a) esférico, (b) pobremente separados, (c) elipsoidal e (d) não convencional

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 (i) agrupamentos esféricos (Fig. 4.a) que são detectáveis até pela inspeção visual dos dados. Outras estruturas como (ii) agrupamentos pobremente separados (Fig. 4.b) apresentam determinação mais difícil. Além destas, os (iii) agrupamentos elipsoidais ou elongados (Fig. 4.c), que se apresentam sob grande influência da correlação entre as variáveis, os agrupamentos com diferentes números de objetos ou variabilidade (Fig. 5) e

(iv) agrupamentos com estrutura não convencional (Fig. 4.d), também assinalam dificuldade em sua determinação ou discriminação (Sarle, 1990).

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.

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).

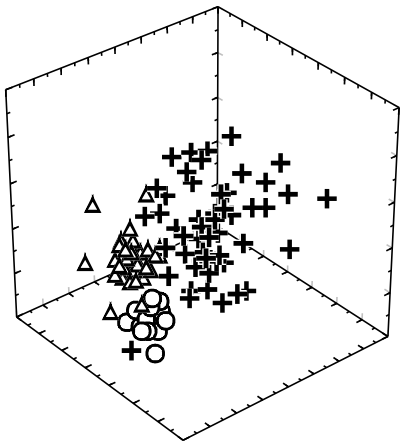

Fig. 5 Estrutura de agrupamento com diferente número de objetos e variabilidade dos dados

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.

Coeficientes e 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 . A expressão desta expansão ao caso multivariado é apresentada na Tabela 3 ($d_{3(i,j)}$).

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

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	$\forall i; \forall j$		positividade
(ii) $d_{i,j} = d_{j,i}$	$\forall i; \forall j$	$i \neq j$	simetria
(iii) $d_{i,k} + d_{j,k} \geq d_{i,j}$	$\forall i; \forall j; \forall k$	$i \neq j \neq k$	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		
		(+)	(-)	
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[1]	{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-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.

[1] Também denominado índice de Ney & Li

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

Medidas		Expressão	Limite	Variável
Gower	S	$s_{g(i,j)} = \frac{\sum_{v=1}^{p} w_v \delta_{i,j}^{v} d_{i,j}^{v}}{\sum_{v=1}^{p} w_v \delta_{i,j}^{v}}$	0 a 1	Todas
Gower transformada	D	$d_{2(i,j)} = 1 - S_{g(i,j)}$	0 a 1	Todas
Euclidiana	D	$d_{3(i,j)} = \sqrt{\sum_{v=1}^{p} w_v \left(x_{iv} - x_{jv}\right)^2}$	≥0	{AIRO}
Size[2]	D	$d_{5(i,j)} = \frac{\left\lvert \sum_{v=1}^{p} w_v \left(x_{iv} - x_{jv}\right) \right\rvert}{\sqrt{\sum_{v=1}^{p} w_v}}$	≥0	{AIRO}
Shape[3]	D	$d_{6(i,j)} = \sqrt{\sum_{v=1}^{p} w_v \left[\left(x_{iv} - \bar{x}_v\right) - \left(x_{jv} - \bar{x}_v\right)\right]^2}$	≥0	{AIRO}
Covariância	S	$s_{7(i,j)} = \frac{\sum_{v-1}^{p} w_v \left(x_{iv} - \bar{x}_v\right) - \left(x_{jv} - \bar{x}_v\right)}{v - 1}$	≥0	{AIRO}
Correlação	S	$s_{8(i,j)} = \frac{\sum_{v-1}^{p} w_v \left(x_{iv} - \bar{x}_v\right) - \left(x_{jv} - \bar{x}_v\right)}{\sqrt{\sum_{v=1}^{p} w_v \left(x_{vi} - \bar{x}_v\right)^2 \sum_{v=1}^{p} w_v \left(x_{jv} - \bar{x}_v\right)^2}}$	-1 a 1	{AIRO}
Correlação transformada	D	$d_{9(i,j)} = \sqrt{1 - s_{8(i,j)}}$	0 a 2	{AIRO}
Minkowsky	D	$d_{10(i,j)} = \left[\sum_{v=1}^{p} w_v \left\lvert x_i - x_j \right\rvert^q\right]^{\frac{1}{q}}$	≥0	{AIRO}
Manhathann[4]	D	$d_{11(i,j)} = \sum_{v=1}^{p} w_v \left\lvert x_i - x_j \right\rvert$	≥0	{AIRO}
Chebychev	D	$d_{12(i,j)} = \max_{v=1}^{p} \left(w_v \left\lvert x_i - x_j \right\rvert\right)$	≥0	{AIRO}
Potência(q,r)[5]	D	$d_{13(i,j)} = \left[\sum_{v=1}^{p} w_v \left\lvert x_i - x_j \right\rvert^q\right]^{\frac{1}{r}}$	≥0	{AIRO}
Razão de Similaridade	S	$s_{14(i,j)} = \frac{\sum_{v=1}^{p} w_v \left(x_{iv} x_{jv}\right)}{\sum_{v=1}^{p} w_v \left(x_{iv} x_{jv}\right) + \sum_{v=1}^{p} w_v \left(x_{iv} - x_{jv}\right)^2}$	0 a 1	{R}
Canberra[6]	D	$d_{15(i,j)} = \sum_{v=1}^{p} \left(\frac{w_v \left\lvert x_{iv} - x_{ij} \right\rvert}{w_v \left(x_{iv} + x_{ij}\right)} \right)$	0 a 1	{R}
Cosseno	S	$s_{16(i,j)} = \frac{\sum_{v=1}^{p} w_v \left(x_{iv} x_{jv}\right)}{\sqrt{\sum_{v=1}^{p} w_v \left(x_{iv}\right)^2 \sum_{v=1}^{p} w_v \left(x_{jv}\right)^2}}$	0 a 1	{R}

continua...

[2] Adotou-se a terminologia inglesa por esta ser mais conhecida do que o correspondente em língua portuguesa.
[3] Idem anterior.
[4] Sinonímia: city-block distance
[5] Sinonímia: distância euclidiana generalizada
[6] Sinonímia: coeficiente de Lance e Willians não-métrico (Kuo, 1997)

Medidas		Expressão	Limite	Variável
Produto interno	S	$s_{17(i,j)} = \dfrac{\sum_{v=1}^{p} w_v\left(x_{iv}x_{jv}\right)}{\sum_{v=1}^{p} w_v}$	≥0	{R}
Sobreposiçãomínima[7]	S	$s_{18(i,j)} = \sum_{v=1}^{p} w_v\left[\min\left(x_{iv}, x_{jv}\right)\right]$	≥0	{R}
Sobreposição	D	$d_{19(i,j)} = \max\left(\sum_{v=1}^{p} w_v\left(x_{iv}\right), \sum_{v=1}^{p} w_v\left(x_{jv}\right)\right) - \sum_{v=1}^{p} w_v\left[\min\left(x_{iv}, x_{jv}\right)\right]$	≥0	{R}
Cantell	D	$d_{20(i,j)} \dfrac{2\chi^2_{0,5[v]} - \mathrm{vd}^2_{i,j}}{2\chi^2_{0,5[v]} + \mathrm{vd}^2_{i,j}}$	≥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 (Fig. 6.b). A variabilidade dentro e entre as variáveis mensuradas é desconsiderada nesta caso (Fig. 6.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 e 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).

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,

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

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

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

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 (Fig. 6.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).

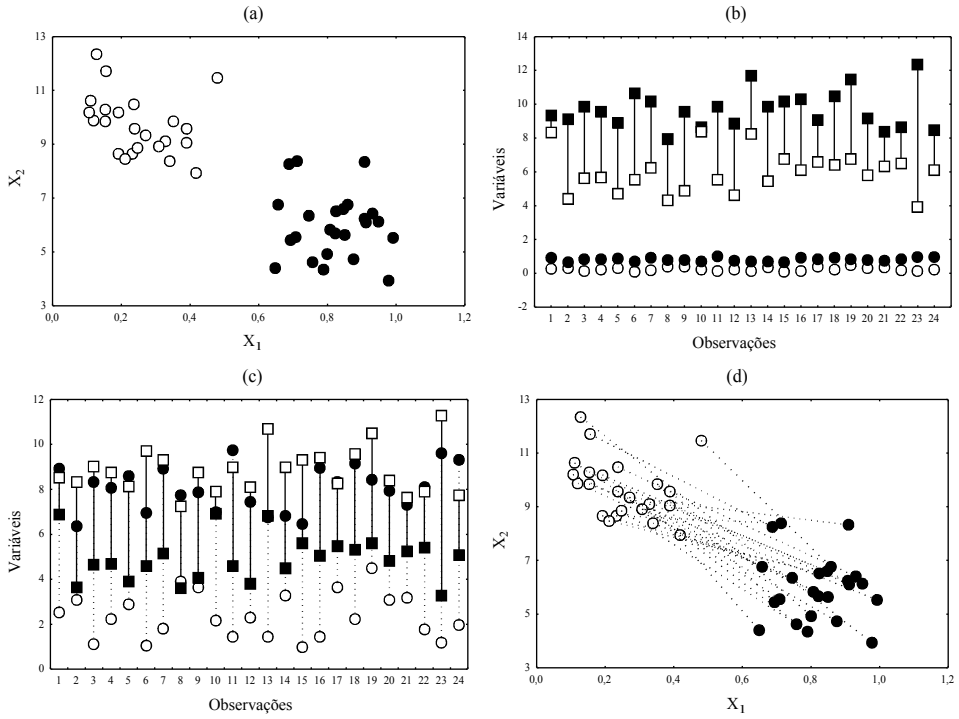

Fig. 6 (a) Valores de um agrupamento hipotético e propriedades de (b) distâncias métricas, (c) estatísticas, ponderadas pela variância e pela (d) covariância

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

$$D_i^2 = \left(x_i - \mu \right)' \Sigma^{-1} \left(x_i - \mu \right) \quad ,$$

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 (Fig. 7.a), enquanto a variância combinada assinala o grau de inércia médio nos corpos (Fig. 7.b).

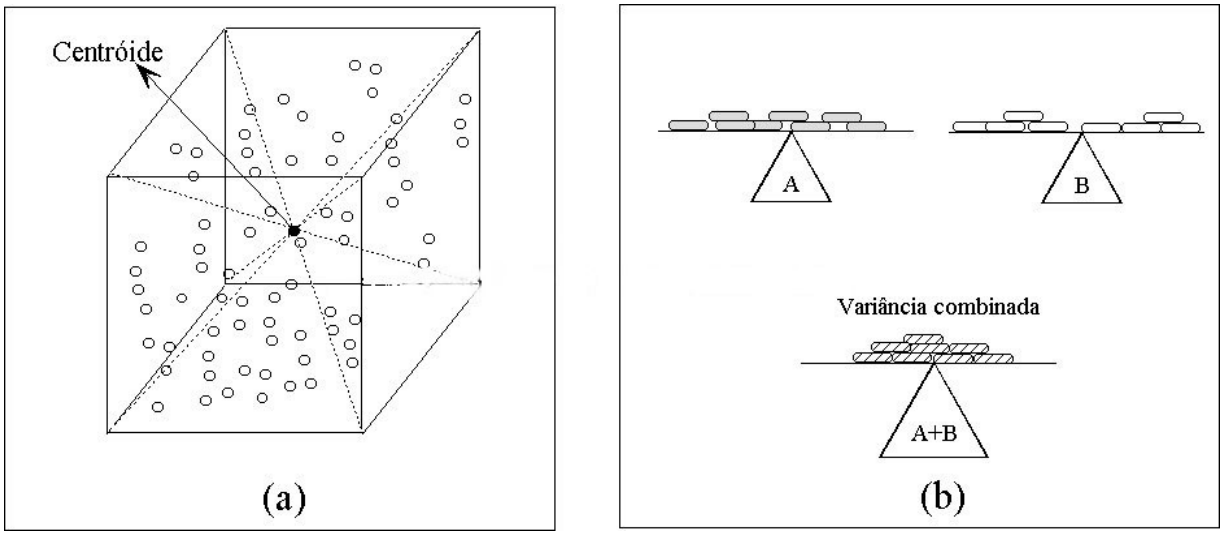

Fig. 7 Representação esquemática do (a) centróide e (b) variância combinada

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 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).

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

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).

Procedimentos de classificação

Definem-se vários algoritmos para a análise de agrupamento, entretanto, definições acerca dos problemas relacionados à análise são necessárias (Fig. 8). 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).

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.

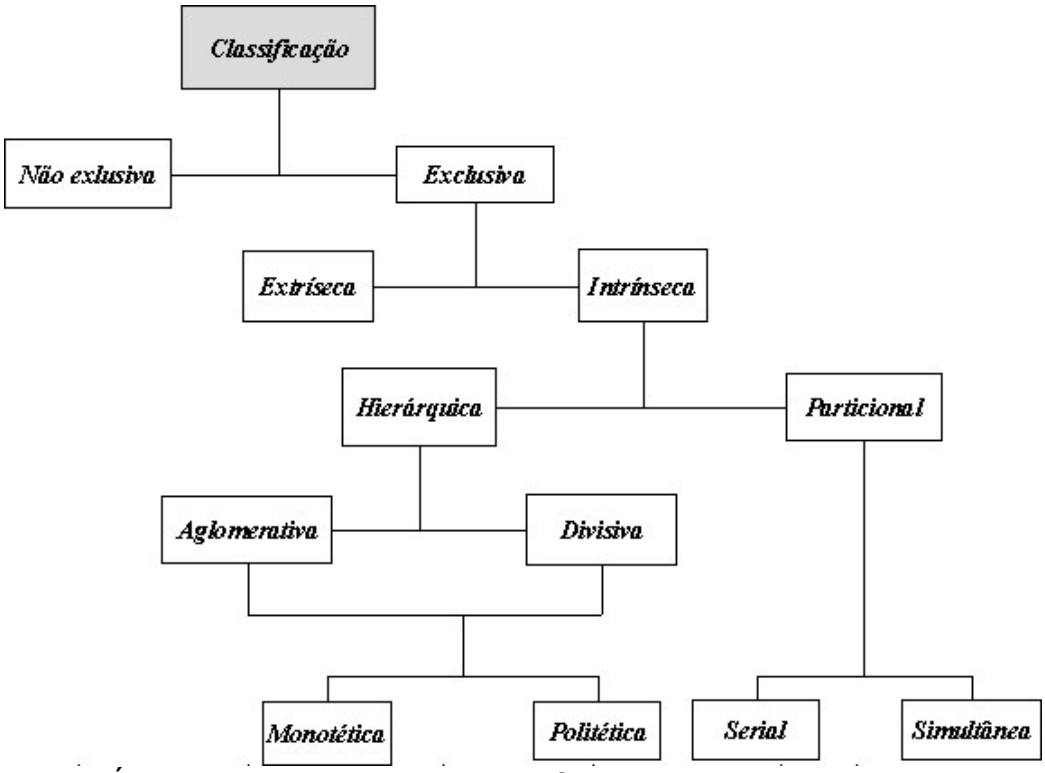

Fig. 8 Árvore dos problemas de classificação

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 (Fig. 9.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 (Fig. 9.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.

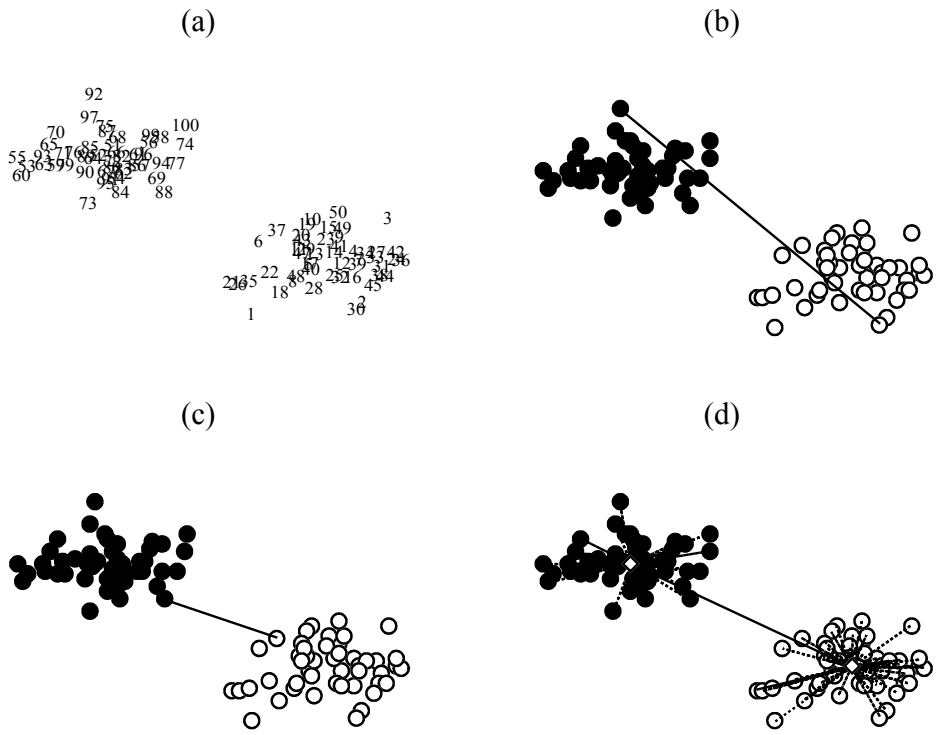

Fig. 9 Métodos de ligação em análise hierárquica de agrupamento, (a) disposição das observações, (b) método de ligação simples, (c) método de ligação completa e (d) método de ligação média

(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 (Fig. 9.d). Vários algoritmos são propostos para a condução deste procedimento (Dillon e Goldstein, 1984).

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 (Fig. 10). 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.

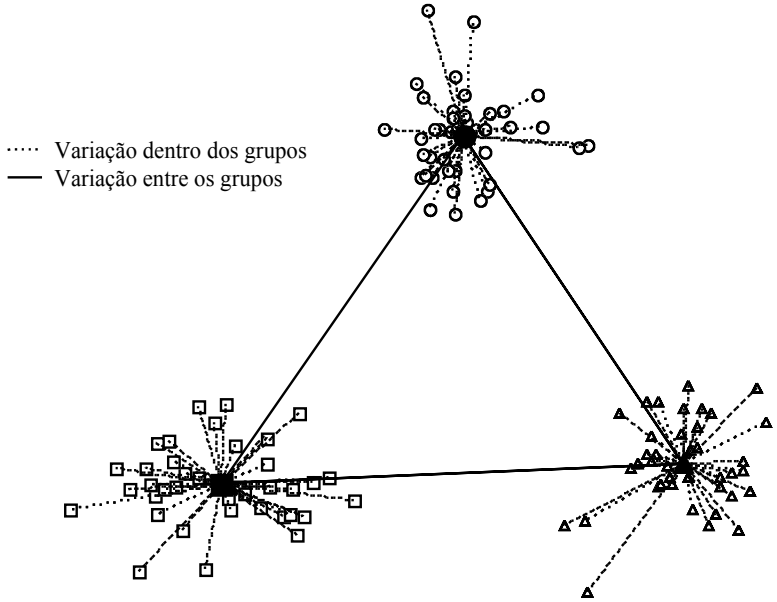

Fig. 10 Método hierárquico de Ward de redução da soma de quadrado do erro

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 (ANOVA). 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 (MANOVA), 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).

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.

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.

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.

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 American 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.; 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.

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.

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.

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.

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.

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. (Ed.). **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.

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.

Apêndice I:

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 , com os parâmetros N(μ,σ) 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).

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

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 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 , representada na Fig. 11, apresenta como parâmetros $N_p(\mu,\Sigma)$, onde μ é o vetor paramétrico de médias e Σ é a matriz de covariâncias

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

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

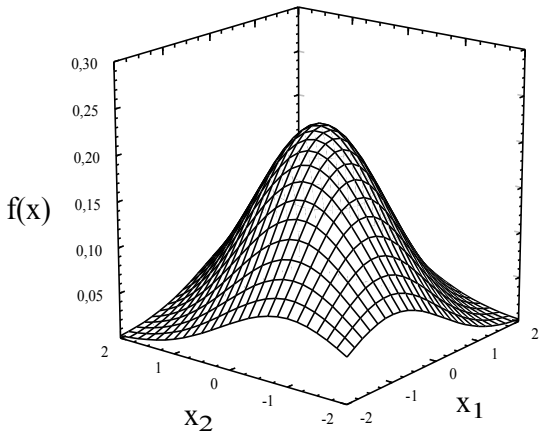

Fig. 11 Função densidade de probabilidade da distribuição normal bivariada

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 $\chi^2_{(v)}$, 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

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

sendo *v* o número de graus de liberdade e □ 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 $W_m(\cdot \mid \Sigma)$. Considerando uma matriz S, teríamos sua função densidade de probabilidade definida em função da matriz considerada, do número de variáveis e das observações e da matriz de covariância.

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

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).

Embrapa

Roraima

MINISTÉRIO DA AGRICULTURA,
PECUÁRIA E ABASTECIMENTO

BRASIL
UM PAÍS DE TODOS
GOVERNO FEDERAL