APLICAÇÃO DE ESTATÍSTICA MULTIVARIADA NO PROCESSAMENTO DIGITAL DE IMAGENS

Rubens Dias Humphreys
Instituto de Pesquisas Tecnológicas do
Estado de São Paulo S.A. - IPT
Caixa Postal 7141
01051 São Paulo, SP
BRASIL

RESUMO

Quatro técnicas de estatística multivariada: análise de componentes principais, análise discriminante, análise de agrupamento e análise de variável canônica aplicadas no processamento digital de imagens são discutidas. A primeira é usada para se reduzir a dimensionalidade dos dados digitais e eliminar a correlação entre bandas; a segunda é usada na classificação supervisionada; a terceira na classificação não supervisionada e a última é usada como uma forma de se discriminar as várias classes que compõe a imagem. O uso dessas técnicas é de grande valia para o processamento digital de imagens. O objetivo é dar uma introdução aos princípios teóricos dessas técnicas.

ABSTRACT

Four multivariable statistical techniques: principal components analysis, discriminant analysis, cluster analysis and canonical variate analysis applied to the digital processing of images are discussed. The first technique is used for dimensionality reduction and to eliminate the correlation between bands; the second is used for the supervised classification of an image, the third for unsupervised classification and the last for the discrimination. The use of these techniques is of great help in the digital processing of an image.

Estatística multivariada trata da análise de dados quando mais de uma variável é medida em um mesmo objeto ou amostra. Os dados são apresentados em várias dimensões, e o fato das observações serem originadas de um mesmo objeto ou amostra, gera dependência ou correlação entre as váriaveis medidas (Morrison, 1976; Mardia et al., 1979)

Essa condição é muito comum de ser encontrada em várias situações de pesquisa em diferentes disciplinas. Por exemplo, em estudos de vegetação, em botânica, um pesquisador dificilmente poderá explicar algum fenômeno através da medição de apenas uma variável

Como outro exemplo, podemos citar a interpretação visual de imagens de satélite. Um interpretador, a não ser que tenha larga experiência e um bom conhecimento da área de estudo, dificilmente poderá fazer uma interpretação adequada, discriminando todos os alvos, com o uso de apenas uma banda.

A tecnologia desenvolvida na construção dos satélites de observação terrestre, permite que se obtenha, em função do tipo de sensor, várias imagens de um mesmo alvo. Assim o sensor MSS produz quatro imagens que cobrem diferentes faixas do espectro da luz solar e o sensor TM produz sete imagens. Portanto, de cada elemento de resolução da imagem ("pixel") são gerados quatro e sete valores de brilhância para, respectivamente, o sensor MSS e o TM. Isto caracteriza a

representação digital de uma imagem de satélite como dados multivariados ou seja, de uma mesma amostra (elemento de imagem), é gerado um vetor de observações em quatro ou sete dimensões. Portanto, técnicas de estatística multivariada são aplicadas para o processamento automático da representação digital de uma imagem. Neste trabalho, serão brevemente discutidos aspectos teóricos das seguintes técnicas: análise de componentes principais, análise discriminante, análise de grupos e análise de variável canônica. Quando for disponível, será apresentado exemplo prático de uso das técnicas no processamento digital de imagens.

TÉCNICAS

1. Análise de Componentes Principais (ACP) - Análise de componentes principais é um procedimento usado para se reduzir a dimensionalidade de um conjunto de variáveis correlacionadas. Isto é obtido através de uma transformação dos dados originais, a partir da qual são gerados componentes principais, não correlacionados. Cada componente principal gerado é uma combinação linear dos dados originais. O número de componentes principais gerados é igual ao número de variáveis contidas nos dados originais. Assim, uma representação digital de uma imagem de sensor MSS gerará quatro componentes principais, pois possue quatro canais. A transformação dos dados é feita através de uma decomposição singular de valor da matriz de correlação ou de variância-covariância (Jolliffe, 1986). A variação total dos dados originais é mantida e cada componente gerado retem uma porcentagem dessa variação. O primeiro componente retem a maior porcentagem da variação total; o segundo retem a maior porcentagem do restante dessa variação, e assim sucessivamente (Marriot, 1976; Jolliffe, 1986). No caso de uma representação digital de uma imagem de satélite do sensor MSS, os dados originais seriam os valores de brilhância de cada elemento de resolução em cada uma das quatro bandas. Matematicamente, a transformação dos dados originais é obtida através da expressão (Jolliffe, 1986):

$$Y = AX$$ onde:

Y = matriz onde cada coluna representa um componente principal.

A = matriz cujas colunas representam auto vetores da matriz de variância covariância ou de correlação.

X = matriz de dados originais.

Geometricamente, os autovetores calculados a partir da matriz de variância ou da matriz de correlação, são iguais ao coseno do ângulo de rotação dos eixos e a variância das variáveis nos eixos, é representada pelos autovalores (Isebrands et al., 1975; Morrison, 1976).

As tabelas 1 e 2 mostram um exemplo de aplicação de componentes principais em uma fração de uma imagem de satélite do sensor MSS, compreendendo 386 linhas e 265 colunas (Humphreys, 1989).

TABELA 1

Autovalores e Porcentagens correspondentes a cada autovetor

Autovetores	Autovalores	%
1	631.2	94.2
2	34.4	5.1
3	3.4	0.5
4	0.9	0.1

TABELA 2

Autovetores

1	2	3	4
-0.016	0.467	-0.069	0.881
-0.076	0.851	-0.221	-0.470
0.608	0.851	0.760	-0.049
0.790	-0.082	-0.607	0.010

A interpretação de cada componente é feita baseada nos escores de cada autovetor. Para auxiliar na interpretação, pode-se colocar em um gráfico os valores dos escores, como mostrado na Figura 1. O primeiro componente principal (CP-1) explica

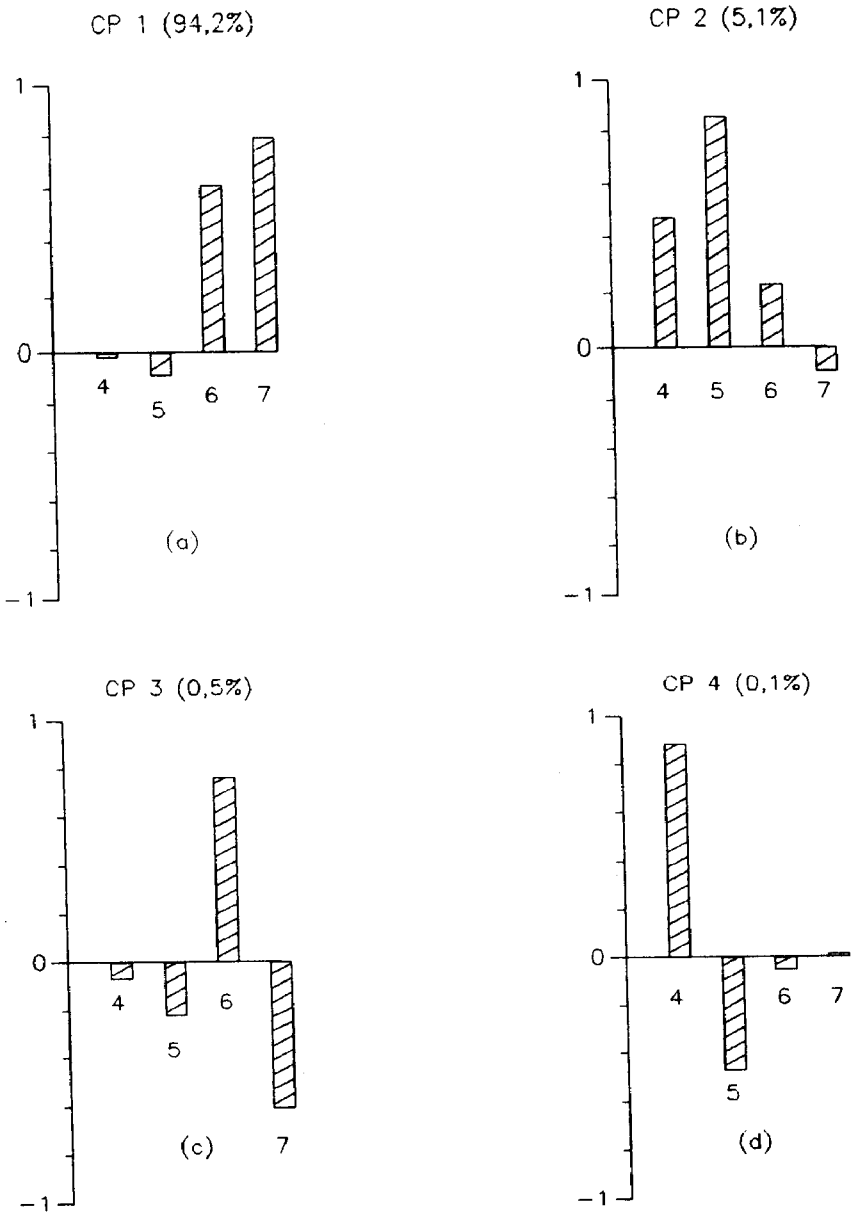

Fig. 1 – Gráficos dos Componentes Principais:
a) Primeiro Componente; b) Segundo Componente;
c) Terceiro Componente; d) Quarto Componente

94,2% da variância total dos dados originais e apresenta escores bastante elevados para as bandas 6 e 7, podendo-se interpretá-lo como sendo um "componente infravermelho". O segundo componente principal explica 5,1% da variação total dos dados originais e com escores elevados nas bandas 4 e 5 podendo-se interpretá-lo com um "componente do visual". Os outros dois componentes explicam uma fração insignificante da variação total e não serão considerados mais. Observa-se portanto que neste caso particular, uma imagem composta pelos componentes 1 e 2, que explicam 99,3% da variação total dos dados originais, pode ser considerada como uma representação efetiva em duas "bandas" de uma imagem multi espectral. Houve portanto uma redução na dimensionalidade dos dados com a vantagem adicional da ausência de correlação entre os componentes. Outros processamnetos digitais feitos com essa imagem seria mais rápido pois apenas duas "bandas" seriam utilizadas, o que significa menos tempo de uso de computador e portanto menos custo do trabalho, com muito pouca perda de informação.

2. <u>Análise Discriminante</u>: Um problema frequentemente encontrado na prática é o de se classificar uma observação em uma população quando se tem uma série de populações. Em zoologia, por exemplo, um cientista mede várias características externas de três insetos pertencentes a diferentes espécies. Com essas medidas, o cientista desenvolve uma função discriminante e a utiliza para classificar um inseto desconhecido, mediante a medição das mesmas características, em uma das três espécies.

Esse mesmo procedimento é também utilizado em processamento digital de imagens. Neste caso, os dados iniciais para a classificação completa de uma área seriam obtidos a partir de áreas de treinamento, em um processo de classificação chamado de supervisionado. Este processo apresenta duas etapas: a primeira é a seleção de áreas de treinamento, que é a etapa mais crítica pois o resultado final da classificação depende diretamente dessa seleção; a segunda etapa é a classificação propriamente dita. Um método comumente utilizado para a classificação é o da máxima verossimilhança. Tendo em vista que este é um método paramétrico, há necessidade de se especificar a distribuição dos dados. É comum assumir-se a distribuição de Gaus ou a normal. (Hudson, 1987; Lillesand e Kiefer, 1979; Hajic e Simonett, 1976). A probabilidade de um elemento de imagem pertencer a uma certa classe, é determinada utilizando-se a função de densidade de probabilidade que para o classificador de máxima verossimilhança é dada por (Hudson, 1987):

$$P(E_{ij}|C)=(1/(2\pi)^{n/2}\,|Sc|^{1/2})*[\exp -1/2(E_{ij}-Xc)'Sc^{-1}(P_{ij}-Xc)]$$

onde:

$P(E_{ij}|C)$ = probabilidade do elemento de imagem E_{ij} pertencer à classe C

Sc = matriz de variância-covariância da classe C

Xc = vetor de médias espectrais para a classe C

n = número de bandas (n=4 para sensor MSS).

Pela característica desse classificador, nota-se que a quantidade de cálculos envolvidos é grande e isto constitue uma desvantagem desse processo (Lillesand e Kiefer, 1979).

Uma outra forma de classificação supervisionada é a que utiliza o classificador Bayesiano, que é um classificador ótimo que atribue dois pesos a uma probabilidade: uma probabilidade a priori de ocorrência de uma classe e uma função de perda ou de classificação errada (Hudson, 1987; Lillesand e Kiefer, 1979). O classificador Bayesiano aproxima-se do classificador de máxima verossimilhança se a perda devido à classificação incorreta for inversamente proporcional à probabilidade a priori (Hudson, 1987).

A utilização do classificador de máxima verossimilhança proporciona uma melhor exatidão da classificação (Lillesand e Kiefer, 1979) e o uso da imagem gerada através da aplicação da transformação por componente principal, pode reduzir o tempo de processamento

por computador pois pode-se usar apenas os dois primeiros componentes

3. Análise de Agrupamento: O objetivo de uma análise de agrupamento, também chamada de classificação, é o de agrupar observações em variáveis ou indivíduos, em conjuntos ou categorias cujos componentes são próximos em termos de distância ou apresentem similaridades. Ou seja, os indivíduos dentro de um grupo são parecidos, têm características semelhantes ou são próximas e os indivíduos de grupos distintos, são dissemelhantes ou afastados (Johnston, 1978; Everitt, 1974). Há basicamente dois métodos para se proceder a uma análise de agrupamento (Greenacre , 1984): métodos hierárquicos e métodos não hierárquicos. Nos métodos hierárquicos obtem-se, como resultado final, um dendrograma. O analista tem que tomar a decisão no início do processo, sobre qual método de se determinar distância entre os grupos deve ser aplicado. Vários métodos para se determinar distâncias podem ser usados. Entre os mais comuns pode-se citar (Mardia et al., 1979) distâncias Euclidiana; Karl Pearson e Mahalanobis. Os métodos hierárquicos podem também ser usados com uma matriz de similaridade, em casos como taxonomia, por exemplo. A diferença entre similaridade e distância é que na primeira, os valores na matriz variam entre 0 e 1 (uns na diagonal) e na segunda, os valores podem assumir qualquer número positivo (zeros na diagonal) (Everitt, 1974). Métodos hierárquicos são úteis quando o analista não tem nenhuma idéia sobre o número de grupos que devem ser formados a partir dos dados originais (Grenacre, 1984).

Nos métodos não hierárquicos, o analista tem que tomar a decisão, antes do início do processo de agrupamento, sobre quantos grupos deverão ser formados. Os grupos são formados mediante a otimização de um critério de agrupamento, como por exemplo, minimizar a soma dos quadrados dentro de cada grupo. (Greenacre, 1984). Nos métodos não hierárquicos, as observações podem mudar de grupo para se atingir o critério de otimização. Nos métodos hierárquicos, as observações não mudam de grupo, ou seja, quando uma observação é alocada a um determinado grupo, ela permanece (Mardia , et. al, 1979).

Para a classificação de imagens de satélite pelo processo não supervisionado, geralmente usa-se um método não hierárquico. Chuvieco e Congalton (1988) e Hajic e Simonett (1976) apresentam uma discussão sobre o uso de análise de grupos e análise discriminante para o processamento digital de imagens de satélite.

4. Análise de Variável Canônica: Assim como em análise de componentes principais, a análise de variável canônica transforma os dados originais em uma combinação linear, a diferença é que em componentes principais utiliza-se a matriz de variância-covariância total ao passo que em análise de variável canônica esta matriz é desdobrada em duas componentes: a variação dentro da categoria e a variação entre categorias ou grupos (Merembeck et al, 1976).

O objetivo da análise de variável canônica é o de se obter uma representação gráfica mostrando as diferenças entre os grupos utilizando-se o menor número de dimensões possíveis (Gittins, 1984).

Para se conseguir esta separabilidade entre grupos, a transformação linear tem que conter a máxima quantidade de variação entre os grupos. Tal transformação linear é dada por (Gittins, 1984):

$$U = a'x \quad \text{onde:}$$

a é um autovetor correspondente ao máximo autovalor da equação:

$$Q_w - r^2 Q_t \; a = 0 \quad \text{onde:}$$

Q_w corresponde à soma de quadrados e produtos entre os grupos da matriz x e Q_t é a soma de quadrados e produtos total. A soma de quadrados e produtos dentro dos grupos é dada pela diferença entre Q_t e Q_w. A maximização da relação entre a soma de quadrados e produtos dentro dos grupos com a soma de quadrados e produtos total para a transformação linear U é o que se deseja. Isto é conseguido obtendo-se o primeiro autovalor e o correspondente autovetor da equação acima. A função linear U obtida com essa solução é

chamada de variável canônica ou função discriminante para os grupos.

A transformação dos dados para se achar os valores do vetor a é feita tendo-se como restrição que a variância da variável canônica (U) é unitária, condição necessária para se chegar a uma solução única para U (Gittins, 1984). Outra característica resultante da transformação é que a correlação entre duas variáveis canônicas é zero, ou seja, elas são ortogonais. Geometricamente falando, o que ocorre é uma rotação dos eixos de tal forma que a maior porcentagem da soma dos quadrados das distâncias entre as médias das categorias é maximizada no primeiro eixo. No segundo eixo, que é ortogonal ao primeiro, o restante da soma de quadrados é maximizada (Merembeck et al, 1976). Portanto, com duas ou três variáveis canônicas, correspondentes àquelas com os maiores valores de r, é possível obter-se uma boa discriminação entre os grupos, ocorrendo também uma redução da dimensionalidade (Gittins, 1984).

Na aplicação da técnica de componentes principais, assume-se que os dados apresentam uma distribuição multivariada normal (Jolliffe, 1986). No caso de análise de variável canônica, assume-se que a distância entre dois grupos é Mahalanobis. Porém, em geral, essa suposição não é restritiva. Multinormalidade não é requerida em aplicações descritivas mas é desejável que a distribuição conjunta das variáveis seja razoavelmente simétrica e não muito alongada (Gittins, 1984).

CONCLUSÕES: Devido à forma como os dados de uma representação digital de imagem é apresentada, com grande quantidade de informações por elemento de imagem, o uso de estatística multivariada para o processamento é imprescindível. Procurou-se neste trabalho dar alguns esclarecimentos sobre algumas das técnicas de estatística multivariada mais utilizadas no processamento digital de imagens. Deve-se ter em mente que o resultado obtido de uma análise de imagens via computador, não descarta o uso de fotografias aéreas ou da imagem propriamente dita. Isto se deve pelo fato de que a precisão que se obtem em uma classificação computadorizada de imagem dificilmente atinge a 100%. Portanto, sempre ocorrem elementos de imagem que são classificados erroneamente ou que simplesmente não são classificadas sendo incorporados na classe "outros". As técnicas de estatística multivariada são, na realidade, uma "arma" que o técnico faz uso para que seu trabalho de processamento digital de imagens seja facilitado.

Referências Bibliográficas

CHUVIECO, E. CONGALTON, R.G. Using Cluster Analysis to Improve the Selection of Training Statistics in Classifying Remotely Sensed Data, Photogrammetric Engineering and Remote Sensing, 54(9):1275-1281, setembro, 1988

EVERITT, B. Cluster Analysis, Heinemann Educational Books, 121p., 1974

GITTINS, R. Canonical Analysis A Review with Applications in Ecology, Springer - Verlag, 351p., 1984.

GREENACRE, M.J. Theory and Apllications of Correspondence Analysis, Academic Press, 364p., 1984.

HAJIC, E. J.; SIMONETT, D. S. Comparisons of Qualitative and Quantitative Image Analysis, Remote Sensing of Environment, Editors Joseph Lintz Jr. e David S Simonett, Adison-Wesley Publishing Co, 374:410, 1976

HUDSON, D. W. Digital Classification of Landsat Multispectral Scanner Data - An Introduction, Michigan State University Agricultural Experimental Station, Research Report 483, 12p., 1987

HUMPHREYS, R. D. Evaluation of Multilevel Sampling Techniques for Forest Inventory in Northern Michigan, PhD Dissertation, Michigan State University, Forestry Department, 174p., 1989

ISEBRAND, J. G.; THOMAS, R. C. Introduction to Uses and Interpretation of Principal Component Analysis, General Technical Report NC-17, US Forest Service, 19p., 1975.

JOHNSTON, R. J. Multivariate Statistical Analysis in Geography, Longman Group, 280p., 1978.

JOLLIFFE, I. T. Principal Component Analysis, Springer-Verlag, 271p., 1986

LILLESAND, T. M.; KIEFER, R.W. Remote Sensing and Image Interpretation, John Willey & Sons, 612p., 1979.

MEREMBECK, B.F.; BORDEN, F. Y.; PODWYSOCKI, M.H.; APPLEGATE, D. N. Application of Canonical Analysis to Multispectral Scanner Data, Application of Computer Methods in the Mineral Industry, Proceedings of the 14th APCOM Symposium, Edited by R. V. Ramani, 867:879, October, 1976.

MORRISON, D. F. Multivariate Statistical Methods, McGraw Hill Book Company, 415p., 1976.