UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Química

Este exemplar corresponde a redação final da Tese defendida por Ieda Spacino Scarminio e Aprovada pela Comissão Julgadora.

Campinas, 02 de março de 1989

Orientador: Prof.Dr. Roy E. Bruns

Edward

DESENVOLVIMENTO DE UM SISTEMA QUIMIOMÉTRICO PARA MICROCOMPUTADORES E ALGUMAS APLICAÇÕES

(mtb)

Ieda Spacino Scarminio (IC) (cat. Anna)

in José

Tese de Doutorado

CAMPINAS - 1989

AGRADECIMENTOS

- ao Dr. Roy E. Bruns, pela orientação, pelas discussões e pela amizade durante a execução deste trabalho.
- a Drª Sônia M.B. de Oliveira pela cessão dos dados referentes as amostras de bauxita.
- aos Drs. José Roberto Ferreira, L.A. Martinelli e Jefferson Mortatti pela cessão dos dados referentes as amostras de água.
- ao Dr. José Fernando G. Faigle e Ronei J. Poppi pela cessão dos dados cromatográficos.
- ao Dr. Benício B. Neto pela cessão dos dados referentes aos tensores polares atômicos, por suas importantes sugestões e pela ajuda nas correções da tese.
- à Diretoria do Instituto de Química da Universidade Estadual de Campinas, pela facilidades concedidas durante a realização deste trabalho.
- A Universidade Estadual de Londrina, em especial o Departamento de Química, pela oportunidade oferecida para realização deste trabalho.
- ao Jair, pela compreensão e incentivo durante todo este trabalho.
- A Marilza, pelo trabalho datilográfico.
- A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelas bolsas de estudo.
- A todos aqueles que de forma direta ou indireta contribuíram para este trabalho.
- E, em especial, aos amigos Moita, Pedro, Omar, Francisco, Paulo Sérgio e Glauco.

Ao Jair

e

Luigi.

ÍNDICE

Página

FIGURAS

Página

Figura

Página

Figura

TABELAS

Tabela

Página

Tabela

Página

Tabela

Página

Tabela

RESUMO

TÍTULO: Desenvolvimento de um sistema quimiométrico para microcomputadores e algumas aplicações.

AUTOR: Ieda Spacino Scarminio

ORIENTADOR: Roy Edward Bruns

INSTITUTIÇÃO: Universidade Estadual de Campinas

Instituto de Química

Caixa Postal 6154 - CEP - 13081

Campinas - SP

Para estimular o desenvolvimento de pesquisas em quimiometria no Brasil, o pacote computacional ARTHUR/75 operável em computadores de grande porte (PDP-10), e já utilizado no grupo de quimiometria da UNICAMP, foi modificado e adaptado por nós para microcomputadores com sistemas operacionais MS-DOS. Foram desenvolvidos ainda programas para regressão de componentes principais (PCR) e rotação ortogonal de componentes principais (VARIMAX). Todos estes programas foram usados para analisar diferentes conjuntos de dados provenientes de: 1) amostras de bauxita; 2) amostras de água da região amazônica; 3) tensores polares atômicos e 4) dados cromatográficos. Para os dados de bauxita, a rotação ortogonal VARIMAX mostrou-se útil na interpretação dos fatores geológicos subjacentes envolvidos. As amostras de água de diferentes regiões do Estado de Rondônia foram classificadas com mais de 80% de exatidão usando-se os métodos KNN e SIMCA. Para os dados de tensores polares dos halometanos, os escores da primeira componente principal têm um coeficiente de correlação de 0.996 com uma função simples da eletronegatividade dos substituintes. A regressão

em componentes principais mostrou-se uma ferramenta adequada para a análise cromatográfica quantitativa de tolueno, isoctano e etanol.

ABSTRACT

Title: Development of a chemometric system for microcomputers and some applications.

Author: Ieda Spacino Scarminio

Thesis advisor: Roy Edward Bruns

Institution: Universidade Estadual de Campinas
Instituto de Química
Caixa Postal 6154 - CEP- 13081
Campinas - SP

To stimulate research development in chemometrics in Brasil, the computacional program ARTHUR/75 applied to main frame computers (PDP-10), already in use by UNICAMP chemometric group, was changed and adapted for microcomputers, with the MS-DOS operational system. In addition programs for principal component regression (PCR) and to orthogonal rotation of principal component (VARIMAX) were developed. All these programs were used to analyze several data sets: 1) bauxite samples, 2) water samples from rivers in the Amazon region, 3) Atomic polar tensors from vibrational intensities and 4) chromatographic data organic mixtures. For the bauxite data, VARIMAX rotation proved to be useful in the interpretation of the underlying geological factors involved. Water samples from different regions in the state of Rondônia were classified with more than 80% accuracy using the K-nearest neighbor and SIMCA methods. For halomethanes, the first principal component scores of the polar tensor data have a correlation coefficient of 0.996 with a simple function of the substituent electronegativity. Principal component regression proved to be an adequate calculational tool for the quantitative chromatographic analysis of toluene, iso-octane and ethanol.

SÍMBOLOS MAIS IMPORTANTES

n - número de objetos

i - índice para as variáveis do bloco $\underset{\sim}{X}$

ℓ - índice para as variáveis do bloco $\underset{\sim}{Y}$

k - índice para objetos

$\underset{\sim}{x}$ - matriz de dados

$\underset{\sim}{\bar{x}}'$ - vetor linha das médias das variáveis

$\underset{\sim}{\varepsilon}$ - matriz dos resíduos de $\underset{\sim}{x}$

$\underset{\sim}{t}$ - vetor coluna dos escores

$\underset{\sim}{P}'$ - vetor linha dos pesos

$\underset{\sim}{T}$ - matriz dos escores de $\underset{\sim}{x}$

$\underset{\sim}{P}$ - matriz dos pesos de $\underset{\sim}{x}$

a - número de componentes principais

$\underset{\sim}{\lambda}_a$ - vetor dos 'a' maiores autovalores

$\underset{\sim}{U}$ - matriz dos autovetores

$\underset{\sim}{\Delta}_a$ - matriz dos autovalores

$\underset{\sim}{H}$ - matriz de associação

h^2 - comunalidade

$\underset{\sim}{Y}$ - matriz das variáveis dependentes

m - número de variáveis dependentes ($\underset{\sim}{Y}$)

$\underset{\sim}{b}$ - vetor dos coeficientes de regressão

$\underset{\sim}{B}$ - matriz dos coeficientes de regressão

$\underset{\sim}{Q}$ - matriz dos pesos de $\underset{\sim}{Y}$

$\underset{\sim}{F}$ - matriz dos resíduos de $\underset{\sim}{Y}$

CAPÍTULO I

INTRODUÇÃO

A química nos últimos anos tem sido fortemente influenciada pela automação na aquisição de dados e no controle de instrumentos. Com a automação de instrumentos, a disponibilidade de microcomputadores e a necessidade de trabalhar com problemas complexos, os químicos vêm redescobrindo a importância dos métodos estatísticos multivariados e estão começando a utilizá-los para auxiliar a resolver problemas químicos. Assim, vem-se desenvolvendo uma disciplina da química, batizada quimiometria (1).

A quimiometria pode ser definida como a disciplina química que utiliza métodos matemáticos e de estatística multivariada para, (a) definir ou selecionar as condições ótimas de medição e experiência, e (b) extrair de dados químicos o máximo de informações.

Embora seja um ramo interdisciplinar, a organização internacional dedicada ao uso e desenvolvimento de métodos quimiométricos (Chemometrics Society) é composta principalmente de dois grupos de químicos, um interessado em resolver problemas em química analítica e o outro voltado para os problemas de química orgânica.

Um dos objetivos da quimiometria é a conversão de dados em informação e finalmente em conhecimento que possa ser utilizado na redução de problemas. Para isso são utilizados sobretudo os métodos de reconhecimento de padrões, que permitem identificação rápida e eficiente de relações básicas porventura existentes em uma grande massa de dados.

Com os recentes avanços nas técnicas de coletas de dados automatizadas e digitalizadas, tornou-se possível medir rapidamente dezenas de variáveis em centenas de amostras. Neste caso as análises de dados tradicionais são lentas e limitadas porque tratam uma ou duas variáveis de cada vez. Já as técnicas de análise multivariada tratam todas variáveis simultaneamente, podendo rapidamente extrair as variáveis importantes e identificar as relações fundamentais para a identificação das amostras.

O reconhecimento de padrões consiste em duas fases: a análise exploratória dos dados e em seguida o uso de técnicas de reconhecimento de padrões propriamente ditas. A fase exploratória revela a existência (ou não) de: (1) amostras ou medidas anômalas, (2) relações entre as variáveis medidas, e (3) relações ou agrupamentos entre as amostras. A fase de aplicação de métodos de reconhecimento de padrões testa estas relações desenvolvendo modelos de classificação/previsão e determina a precisão destes modelos.

A quimiometria não se restringe aos métodos de reconhecimento de padrões. Os artigos indicados nas referências (2-6) listam a maioria dos aspectos ligados as atividades da quimiometria, entre eles: teoria de controle, teoria de amostragem, estimativa de parâmetros e estados, análise de agrupamentos, teoria dos grafos, otimização, calibração, resolução, planejamento de experimentos, análise de séries temporais, pesquisa operacional, teoria da informação, técnicas de transformação, inteligência artificial e controle de processos.

Várias áreas tem-se utilizado das técnicas de análise de dados multivariados: geologia (7), geoquímica (8), ecologia (9), etc...

O grupo de quimiometria da UNICAMP concentra suas atividades nos seguintes assuntos:

Reconhecimento de padrões: Estes métodos tem sido utilizados principalmente em problemas de classificação (10,11).

Calibração multivariada - Aplicação do método generalizado de adições padrão (GSAM), para detectar e quantificar efeitos interferentes e simultaneamente determinar as concentrações dos constituintes químicos de uma amostra (12). Aplicação do método de regressão dos mínimos quadrados parciais (PLS) para determinar simultaneamente as concentrações dos constituintes químicos de uma ou mais amostras (13).

Otimização - Aplicações dos métodos(14): a) Simplex em cinética química para ajuste de dados experimentais à modelos matemáticos; b) Planejamento fatorial em cromatografia e análise por injeção em fluxo para selecionar variáveis com efeitos nas respostas destes sistemas e c) Superfície de resposta para refinamento dos dados obtidos pelo Simplex.

O pacote computacional de reconhecimento de padrões utilizado pelo grupo de Quimiometria do Instituto de Química da UNICAMP é o sistema integrado de programas computacionais ARTHUR para computadores de grande porte (15), desenvolvido pelo grupo de pesquisa do Prof. Bruce Kowalski do Departamento de Química da Universidade de Washington, Seattle, W.A. Ao adquirir este sistema, o grupo de quimiometria o fez com liberdade de uso e disseminação, ou seja, sem as restrições de Copyright.

Devido a certas dificuldades no uso deste sistema, como a demora na execução devido a sobrecarga do computador central, e a exigência de memória disponível a partir de certos horários, foi necessário segmentar e adaptar os principais subprogramas pa

ra microcomputadores de 8 bits. Estes novos programas permitem um uso mais interativo e eficiente dos vários métodos de reconhecimento de padrões.

Para tornar a operação destes programas mais flexível e permitir que diferentes tipos de análise de dados sejam feitas de uma maneira mais simples, estabelecemos os seguintes objetivos para este trabalho:

1) Adaptar os principais subprogramas para microcomputadores tipo PC com sistema operacional MS-DOS.

2) Modificar os programas já adaptados para viabilizar seu uso em aplicações industriais.

3) Criar sistema de programas computacionais abertos (linguagem FORTRAN) para estimular o desenvolvimento da pesquisa em quimiometria no Brasil.

4) Aumentar os sistemas computacionais implementando novos programas, como Varimax e Regressão de Componentes Principais.

5) Desenvolver a metodologia da utilização integrada dos métodos quimiométricos através de aplicações em problemas reais, mesmo em áreas alheias à química analítica.

O capítulo II contém a descrição de alguns métodos de análise multivariada comumente usados para análise de dados químicos, entre eles métodos de reconhecimento de padrões e métodos de calibração multivariada. O capítulo III apresenta a descrição dos principais subprogramas do sistema ARTHUR adaptados e modificados por nós para microcomputadores com sistemas operacionais CP/M e MS-DOS. O capítulo IV contém as aplicações dos programas modificados a conjunto de dados obtidos por grupos de pesquisa estabelecidos em, ou perto de, Campinas. A primeira aplicação refere-se a um conjunto de dados geoquímicos proveniente do Institu

to de Geociências da Universidade de São Paulo, São Paulo. Setenta e três amostras de rocha foram coletadas na região de Miraí (MG), onde se localiza um importante depósito de bauxita. Os elementos mais abundantes nestas amostras são silício, alumínio, ferro, manganês, magnésio, cálcio, sódio, potássio, titânio e fósforo. Como elementos traço, bário, cério, crômio, gálio, lantânio, neodímio, vanádio, zinco, zircônio, cobre, níquel, nióbio, estrôncio e ítrio. Além destes elementos também foi feita a dosagem da água estrutural. Os métodos de reconhecimento de padrões foram utilizados para classificação destas amostras. Para distinguir os conjuntos de elementos de comportamento semelhantes foi utilizado o método varimax.

A segunda aplicação refere-se a um conjunto de dados químicos relativos a problemas ambientais, proveniente do Centro de Energia Nuclear na Agricultura (CENA), Piracicaba. Trinta e cinco amostras de água foram coletadas entre os anos 1983 e 1985 nas bacias dos rios Ji-Paraná, Jamari e Madeira e foram analisadas quanto aos teores de cálcio, magnésio, sódio, potássio, silício, alumínio, sulfato, cloreto, amônio, nitrato e ferro. Os métodos de reconhecimento de padrões foram utilizados para ver se havia alguma discriminação evidente entre estas bacias, devido a possíveis alterações decorrentes do desmatamento das florestas no Estado de Rondônia.

O conjunto de dados da terceira aplicação consiste de cinco quantidades invariantes de 158 tensores polares atômicos obtidos de intensidades de bandas na região do IV para aproximadamente 50 moléculas. Sabe-se que estas quantidades são altamente correlacionadas e análises foram feitas na tentativa de compreender a natureza destas relações. Além disso foram usados mé

todos de classificação para determinar agrupamentos naturais de tensores polares atômicos, os quais são importantes na previsão de intensidades vibracionais usando moléculas de referência.

Os resultados obtidos para estes conjuntos mostraram que a análise dos dados não poderiam ser feita por observação direta, sendo portanto necessário o tratamento com métodos estatísticos multivariados e de reconhecimento de padrões.

Na quarta aplicação baseamo-nos em conjuntos de amostras processadas por cromatografia gás-líquido de forma a obter picos com graus de superposição dependentes da temperatura da coluna. As amostras consistem de misturas de tolueno, isoctano e etanol em diferentes proporções. Como as variáveis são altamente correlacionadas, a regressão de componentes principais foi usada para estabelecer modelos de calibração. Também foram realizados estudos de superposição de bandas simuladas para verificar se o método de regressão de componentes principais é adequado para este tipo de estudo.

Finalmente no capítulo V, são apresentadas as conclusões gerais deste trabalho.

CAPÍTULO II

MÉTODOS DE RECONHECIMENTO DE PADRÕES

Entre os métodos de reconhecimento de padrões convêm distinguir o método de análise de componentes principais dos métodos de classificação. A finalidade do primeiro é gerar novas variáveis a partir das variáveis originais. Estas novas variáveis são geralmente usadas para analisar os dados usando algum tipo de projeção geométrica ou representação. A finalidade dos métodos de classificação, por outro lado, é desenvolver regras de decisão ou funções de classificação.

De uma maneira geral o conjunto de dados químicos (16, 17) consiste de n objetos, descritos por p variáveis. Os objetos químicos típicos são amostras analíticas ou compostos químicos. As variáveis são muitas vezes derivadas das quantidades de constituintes químicos nos objetos, por exemplo altura do pico em perfis cromatográficos, concentrações de elementos mais importantes e traço. Podem ser também derivadas de espectros (RMN, IV, UV, Raio-X, etc...), com os espectros convertidos em variáveis por digitalização direta em algumas frequências. Um aspecto importante é que as variáveis medidas têm que ser as mesmas para todos os objetos.

O conjunto de dados pode ser representado na forma de uma matriz, $\underline{X}$, com n linhas e p colunas, figura 1, onde cada elemento da matriz, X_{ki}, representa o valor da i-ésima variável para o k-ésimo objeto.

Os objetos da figura 1 podem ser representados graficamente como pontos num espaço p-dimensional onde cada coordenada cor-

responde a uma variável, como mostra a figura 2.

	1	2	i.....p
1	X_{11}	X_{12}	X_{1i}....X_{1p}
2	X_{21}	X_{22}	X_{2i}....X_{2p}
k	X_{k1}	X_{k2}	X_{ki}....X_{kp}
n	X_{n1}	X_{n2}	X_{ni}....X_{np}

Figura 1 - Matriz de dados, $\underset{\sim}{X}$. O elemento X_{ki} corresponde ao valor da variável i para a amostra k.

A medida de similaridade constitui uma grandeza fundamental nos métodos de reconhecimento de padrões (18). Para obter uma excelente estimativa da semelhança entre dois pontos k e j no espaço p basta calcular a distância euclidiana simples entre eles definida por:

$$d_{kj} = \left[\sum_{i=1}^{p} (X_{ki} - X_{ji})^2\right]^{\frac{1}{2}} \qquad |1|$$

onde o somatório é feito sobre as p medidas. Considera-se que quanto menor a distância entre os pontos, maior a similaridade entre as amostras representadas por eles. A similaridade é convenientemente definida por:

$$S_{kj} = 1 - d_{kj}/(d_{kj})_{MÁX} \qquad |2|$$

onde $(d_{kj})_{MÁX}$ é a maior distância entre dois pontos quaisquer no espaço p-dimensional. Para objetos idênticos $S_{kj} = 1$, enquanto

que $S_{kj} = 0$ corresponde a dois pontos separados pela maior distância constatada no conjunto estudado.

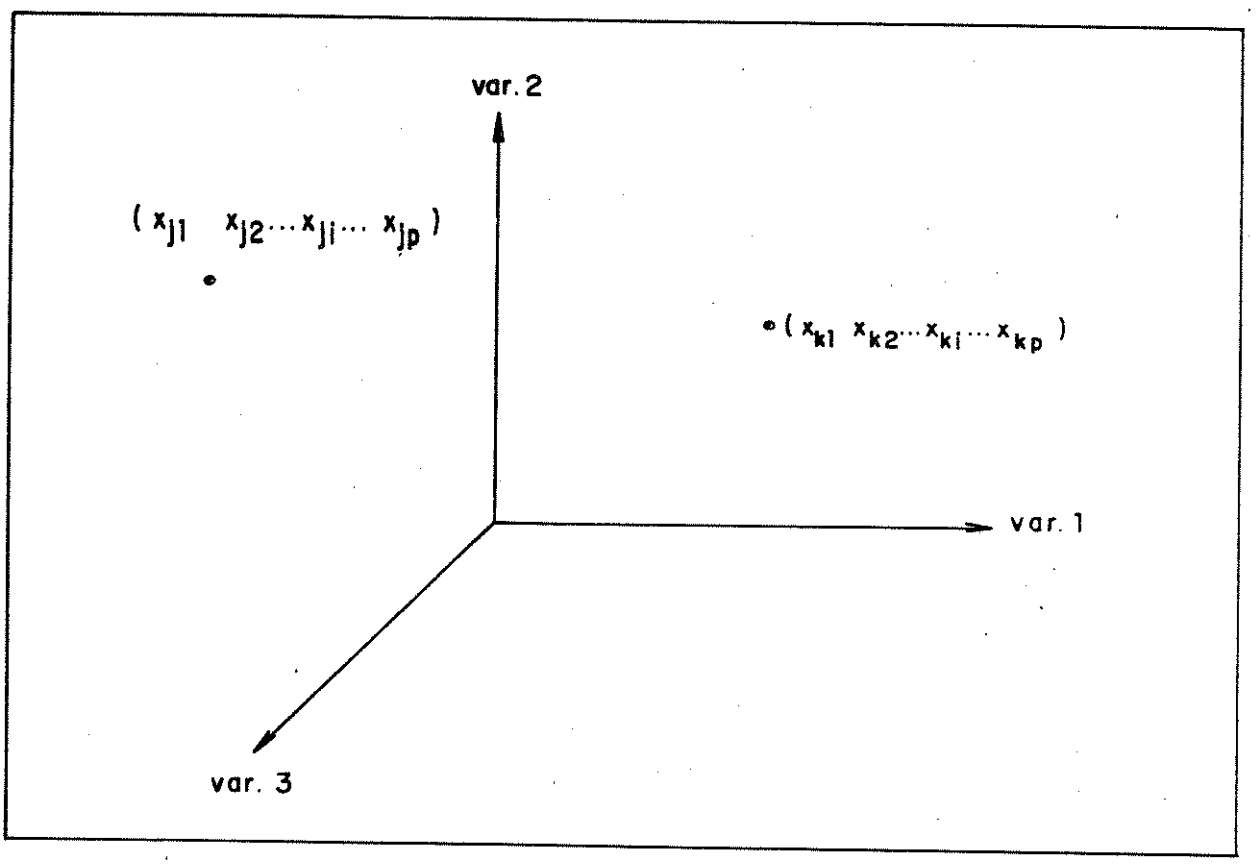


Figura 2 - Representação gráfica dos pontos relativos à k-ésima e à j-ésima amostras, com variáveis (1,2....i....p).

Análise de Componentes Principais (PCA)

Este método tem sido usado para estudar uma variedade de problemas de interesse na química analítica (19), entre eles problemas ambientais, cromatográficos, espectroscópicos, etc....

A análise de componentes principais tem muitos nomes alternativos (16,20) sendo os mais comuns, análise fatorial, projeção de autovetores, decomposição em valores singulares e expansão de Karhunen-Loeve. Neste trabalho adotamos o nome de análise de componentes principais, porque é o mais utilizado em química. A análise de componentes principais calcula a partir das variáveis originais novas variáveis, chamadas de componentes principais,que são combinações lineares das variáveis originais (21, 22).

Os dados originais são algumas vezes redundantes, porque contém várias correlações entre as variáveis e entre objetos. As novas variáveis são calculadas levando em conta as correlações presentes nos dados, mas elas mesmas não são correlacionadas entre si. Desta forma a estrutura dos dados torna-se aparente no espaço destas novas variáveis, podendo ser mais facilmente interpretada.

A finalidade principal do método é a redução da dimensão da matriz de dados, mas ele pode também ser usado para construir modelos de classificação para novos dados medidos para o mesmo sistema.

Dois métodos de análise de componentes principais serão discutidos. O primeiro, chamado de "tipo-R", refere-se ao estudo da relação entre as variáveis, e o segundo "tipo Q", ao estudo da relação entre objetos.

1. Análise do tipo R

1.1. Interpretação Geométrica

A matriz dos dados $\underline{X}$, figura 1, com n objetos e p variáveis, pode ser representada como um conjunto de n pontos em um espaço p-dimensional, figura 3. No espaço p-dimensional é difí-

cil a visualização quando p > 3 (16,22). Contudo, matematicamente tal espaço tem propriedades análogas às de um espaço com somente duas ou três dimensões. As entidades geométricas tais como pontos, linhas, planos, distâncias e ângulos, possuem as mesmas propriedades no espaço p e no espaço tridimensional.

Os passos para obter as componentes principais para um conjunto de dados são:

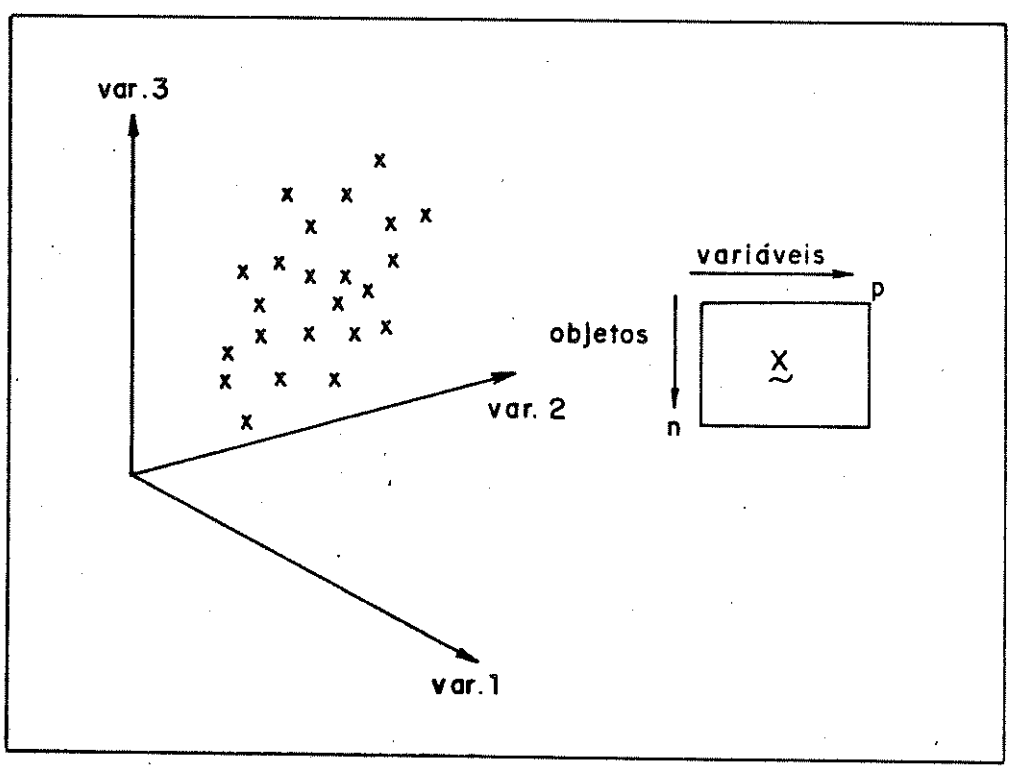


Figura 3 - Os n objetos no conjunto de dados constituem um aglomerado de pontos no espaço p.

1) Representar o conjunto de pontos pelo seu ponto central,

figura 4.

$\bar{X}_i$ = média da variável i para o conjunto que está sendo estudado

$\underset{\rightarrow}{\bar{X}}' = (\bar{X}_1\ \bar{X}_2 \ldots\ \bar{X}_i \ldots\ \bar{X}_p)$, $\underset{\rightarrow}{\bar{X}}'$ é um vetor linha.

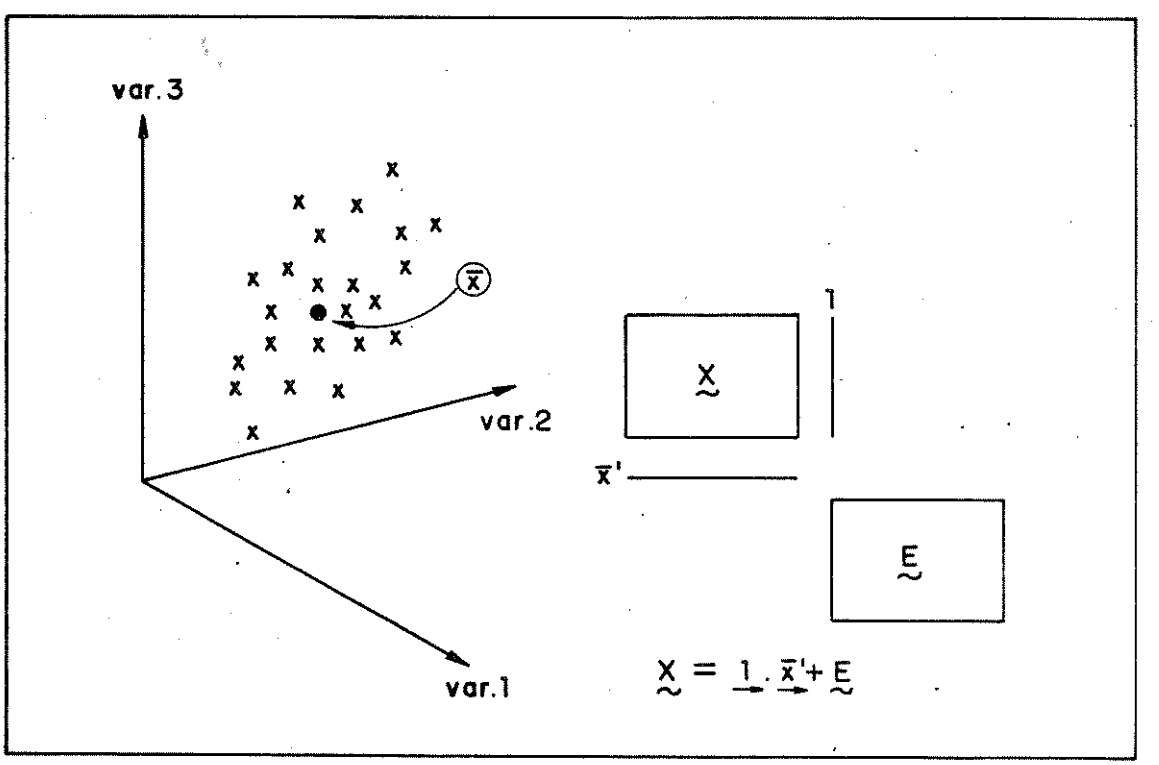


Figura 4 - O modelo mais simples é representar o conjunto de dados pelo seu ponto médio, $\bar{X}$, onde as coordenadas são definidas pelas médias das variáveis. A matriz $\underset{\sim}{X}$ é decomposta em $\underset{\sim}{X} = \underset{\rightarrow}{1} . \underset{\rightarrow}{\bar{X}}' + \underset{\sim}{\varepsilon}^{(0)}$. A matriz $\underset{\sim}{\varepsilon}^{(0)}$ contém os resíduos e_{ki}.

A matriz $\underset{\sim}{X}$ é decomposta em:

$$\underset{\sim}{X} = \underset{\rightarrow}{1} \cdot \underset{\rightarrow}{\overline{X}}' + \underset{\sim}{\varepsilon}^{(0)} \qquad |3|$$

onde $\underset{\sim}{\varepsilon}$ é a matriz n x p contendo os resíduos e_{ki}, ou seja, os desvios da média. Então

$$\underset{\sim}{\varepsilon}^{(0)} = \underset{\sim}{X} - \underset{\rightarrow}{1} \cdot \underset{\rightarrow}{\overline{X}}' \qquad |4|$$

Esta equação representa o modelo de zero componentes principais.

2) Continuar a análise usando a matriz $\underset{\sim}{\varepsilon}^{(0)}$ no lugar de $\underset{\sim}{X}$. Ajustar uma reta aos n pontos no espaço p de modo que os desvios sejam os menores possíveis no sentido dos mínimos quadrados, figura 5. Os cossenos diretores desta reta são os pesos de cada variável na primeira componente principal, e são denotados por P_{1i} formando o vetor linha P_1' (o primeiro vetor de pesos).

Cada ponto é projetado nesta reta, e assim obtemos os escores, t_{k1}, isto é, as cooordenadas dos pontos k ao longo do eixo P_1. Quando subtraímos $\underset{\rightarrow}{t}_k\underset{\rightarrow}{P}_i'$ de $\underset{\rightarrow}{e}_{ki}^{(0)}$ obtemos os novos resíduos $\underset{\rightarrow}{e}_{ki}^{(1)}$, os quais formam a nova matriz $\underset{\sim}{\varepsilon}^{(1)}$. Desta forma a matriz $\underset{\sim}{X}$ torna-se agora

$$\underset{\sim}{X} = \underset{\rightarrow}{1} \cdot \underset{\rightarrow}{\overline{X}}' + \underset{\rightarrow}{t}_1\underset{\rightarrow}{P}_1' + \underset{\sim}{\varepsilon}^{(1)}, \qquad |5|$$

o que representa o modelo de uma componente principal. Em geral, a matriz dos resíduos do modelo de uma componente principal é menor do que a matriz dos resíduos do modelo de zero componentes.

O modelo de uma componente principal corresponde portanto a uma reta no espaço p-dimensional. Esta reta contém o máximo de variância unidimensional dos dados, isto é, ela define a direção

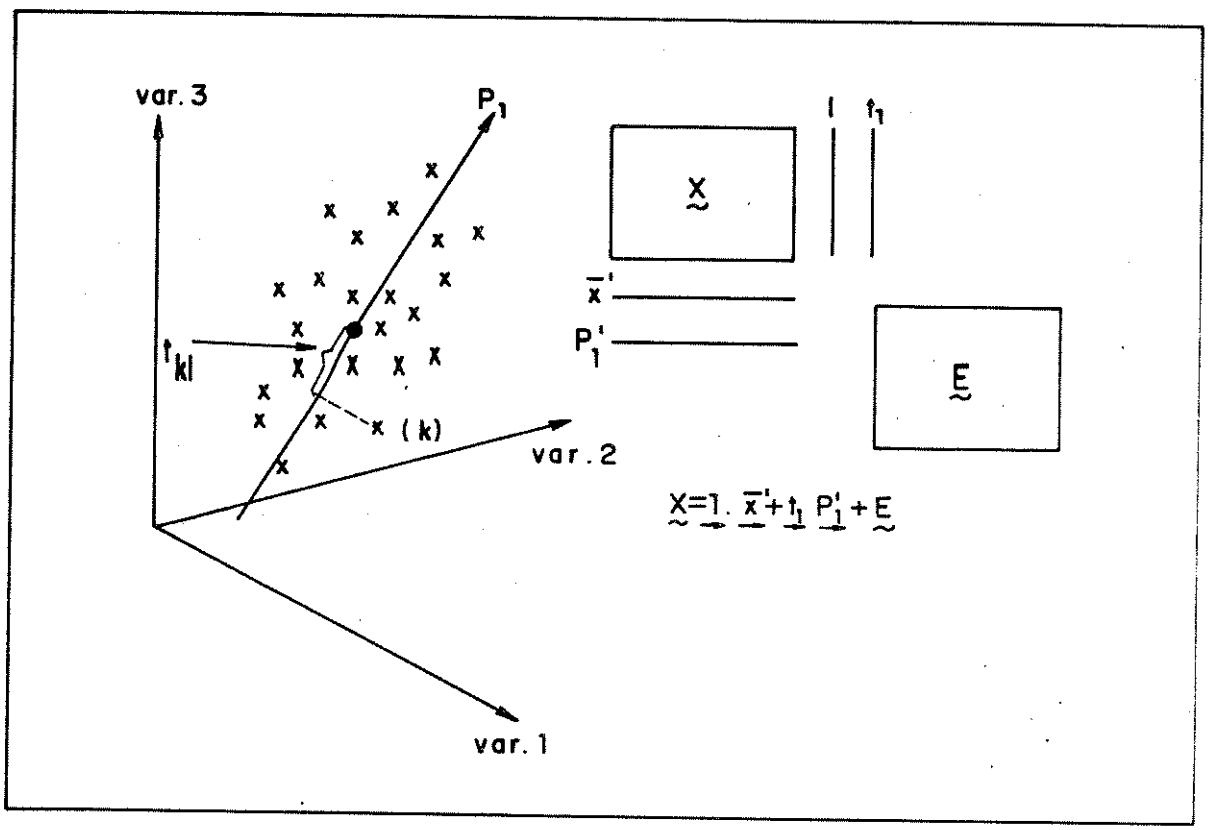


Figura 5 - Os dados são modelados (mínimos quadrados) por uma reta passando por $\bar{X}$. Esta reta é a primeira componente principal (CP_1) e sua equação é definida pelo vetor peso $\underset{\sim}{P}_1'$. Projetando o ponto k nesta reta encontra-se a distância t_{k1} de $\bar{X}$.

que explica o máximo da informação estatística contida nos dados.

3) Subtrair de $\underset{\sim}{X}$ o modelo de uma componente principal. Isto implica na remoção da direção P_1 dos dados. Assim,

$$\underset{\sim}{X} - \underset{\sim}{1} \cdot \underset{\sim}{\bar{X}}' - \underset{\sim}{t}_1 \underset{\sim}{P}_1' = \underset{\sim}{\varepsilon}^{(1)} \quad . \qquad |6|$$

Em seguida ajustar uma outra reta aos pontos de modo que os novos resíduos sejam os menores possíveis. Esta reta é perpendicular à primeira e representa a segunda componente principal, figura 6.

$$\underset{\sim}{X} = \underset{\rightarrow}{1} \cdot \underset{\rightarrow}{\overline{X}}' + \underset{\rightarrow}{t}_1 \underset{\rightarrow}{P}_1' + \underset{\rightarrow}{t}_2 \underset{\rightarrow}{P}_2' + \underset{\sim}{\varepsilon}^{(2)} \qquad |7|$$

A matriz $\underset{\sim}{X}$ torna-se:

$$\underset{\sim}{X} = \underset{\rightarrow}{1} \cdot \underset{\rightarrow}{\overline{X}}' + \underset{\sim}{T}\,\underset{\sim}{P}' + \underset{\sim}{\varepsilon} \qquad |8|$$

onde $\underset{\sim}{T} = (\underset{\rightarrow}{t}_1\ \underset{\rightarrow}{t}_2)$ e $\underset{\sim}{P}' = (\underset{\rightarrow}{P}_1'\ \underset{\rightarrow}{P}_2')$

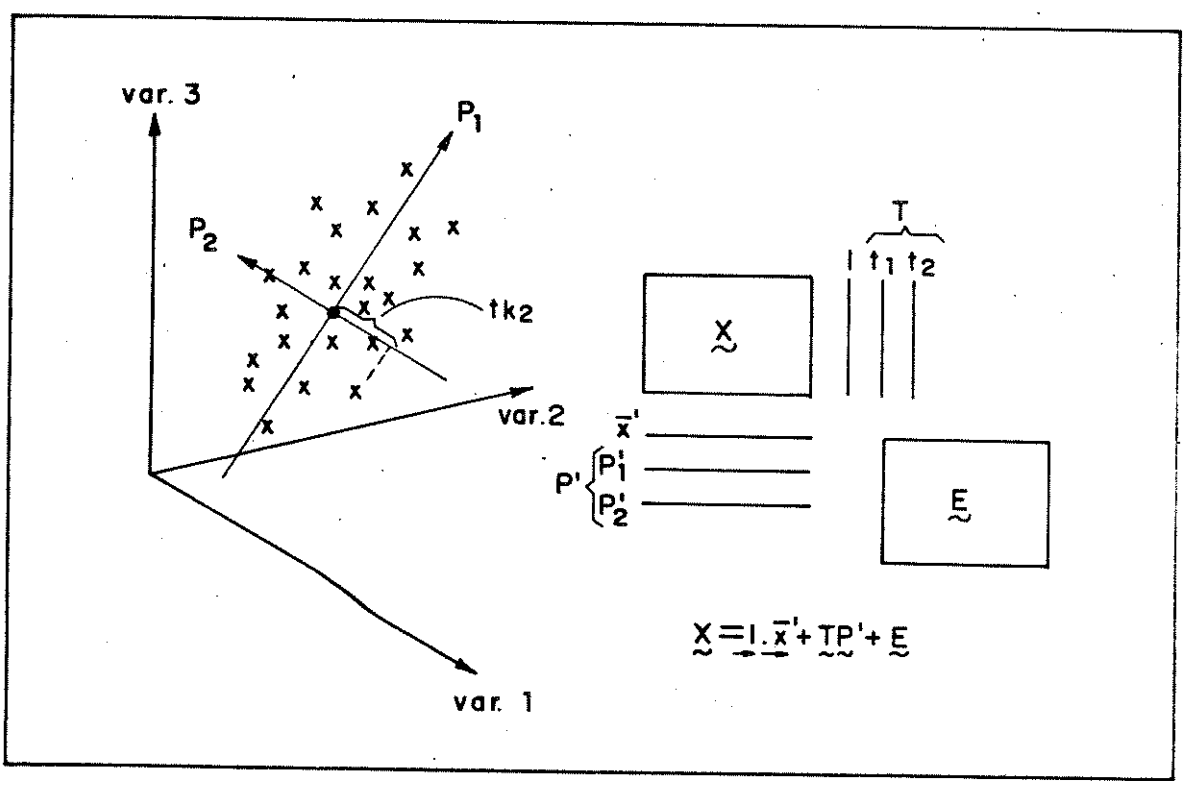


Figura 6 - A segunda componente principal, P_2, é uma reta que passa por $\overline{X}$, perpendicular a P_1.

O modelo com duas componentes principais é representada por um plano no espaço p-dimensional, definido por duas retas ortogonais. A reta da segunda componente está na direção da maior variância restante, depois de extraída a reta da primeira componente principal.

Este procedimento pode ser repetido até p vezes, quando então os resíduos se tornam iguais a zero.

A vantagem desta análise é que se pode usar somente as primeiras 'a' componentes para representar $\underset{\sim}{X}$. Estas 'a' componentes são escolhidas de tal forma que consigam explicar a parte mais importante da informação estatística, ou seja, a maior parte da variância total.

Com estas poucas componentes podemos projetar as colunas ta dos escores de $\underset{\sim}{T}$ umas contra as outras, obtendo um gráfico dos objetos. Este gráfico representa aproximadamente a configuração dos pontos no espaço p-dimensional, figura 7.

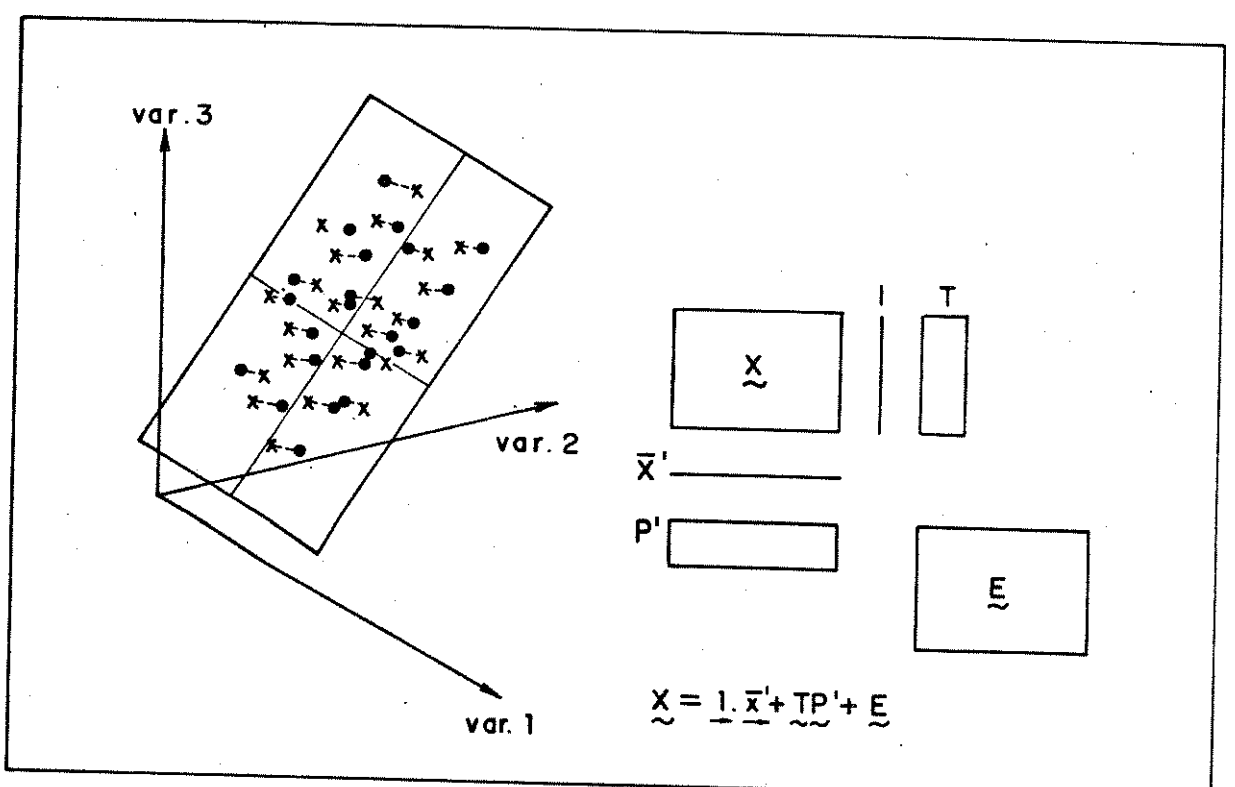


Figura 7 - A primeira e a segunda componente definem um plano. Assim tem-se uma janela que permite olhar o espaço p como uma imagem bidimensional da configuração dos pontos.

1.2. Definição Matemática

Os métodos do tipo R são baseados em modelos matemáticos que representam os dados de forma simplificada (7,23). Neste caso as variáveis formam os eixos das coordenadas para os objetos. A matriz de dados $\underset{\sim}{X}$ consiste de n linhas e p colunas, como ilustrada na figura 1.

Antes da análise estatística os dados são autoescalonados, produzindo variáveis com média zero e variância igual a um. Isto é necessário quando as variáveis são medidas em diferentes unidades (por exemplo, ppm, %, etc...). Com o autoescalonamento todas as variáveis passam a ter a mesma importância estatística.

O ponto de partida da análise do tipo R (7,19,20,23) para calcular as componentes principais é a matriz de covariância $\underset{\sim}{S}$ (igual à matriz de correlação para dados autoescalonados). Esta pode ser obtida pré-multiplicando-se a matriz de dados autoescalonados $\underset{\sim}{X}$ pela sua transposta $\underset{\sim}{X}'$.

$$\underset{\sim}{S} = \underset{\sim}{X}' \underset{\sim}{X} \qquad |9|$$

A matriz $\underset{\sim}{S}$ é quadrada, simétrica e de ordem p. Tem r autovalores positivos e (p-r) autovalores iguais a zero, sendo r o posto da matriz. Os autovalores positivos são $\lambda_1, \lambda_2, \ldots, \lambda_r$, com autovetores correspodentes $u_1, u_2, \ldots, u_r$, obtidos pela diagonalização da matriz de covariância:

$$\underset{\sim}{S}\,\underset{\sim}{U} = \underset{\sim}{U}'\,\underset{\sim}{\lambda} \qquad |10|$$

onde $\underset{\sim}{U}_a' = [u_1, u_2, \ldots, u_a]$

ou

$$\underset{\sim}{S} \simeq \lambda_1 u_1 u_1' + \lambda_2 u_2 u_2' + \ldots\ldots + \lambda_a u_a u_a'$$

Os termos 'a' ($a < p$) da decomposição de $\underset{\sim}{X}$, correspondem aos 'a' maiores autovalores $\lambda_1, \lambda_2, \ldots, \lambda_a$.

Na equação acima, os λ_i e $\underset{\rightarrow}{u}_i$ formam o melhor ajuste, usando o critério dos mínimos quadrados, para a matriz de covariância.

Existem dois métodos para obter as componentes principais que só diferem na forma de normalizar as componentes (23).

Primeiro método

1) Calcular a matriz de covariância, $\underset{\sim}{S}$.

2)Calcular os 'a' maiores autovalores $\underset{\rightarrow}{\lambda}_a$ e os autovetores correspondentes $\underset{\sim}{U}_a'$ da matriz de covariância.

3) Calcular $\underset{\sim}{P}' = \underset{\sim}{U}_a' \, \underset{\sim}{\Delta}_a^{1/2}$, onde Δ_a é uma matriz diagonal de ordem a x a, cujos elementos não nulos são $\lambda_1, \lambda_2, \ldots, \lambda_a$. Neste passo cada autovetor é normalizado de forma que o quadrado de seu módulo seja igual ao autovalor.

4) Calcular $\underset{\sim}{T} = \underset{\sim}{X} \, P' \, \underset{\sim}{\Delta}_a^{-1}$. Aqui as colunas são normalizadas pelos recíprocos dos autovalores.

Segundo método

1) Calcular a matriz de covariância S.

2) Calcular os 'a' maiores autovalores $\underset{\rightarrow}{\Delta}_a$ e os autoveto-

res associados $\underset{\sim}{U}'_a$.

3) $\underset{\sim}{P}' = \underset{\sim}{U}'_a$.

4) Calcular $\underset{\sim}{T} = \underset{\sim}{X}\,\underset{\sim}{P}'$.

Neste caso cada autovetor é normalizado para 1, tornando o procedimento computacional mais simples que o anterior.

As componentes definidas no ítem 3 ($\underset{\sim}{P}'$) para as duas soluções são chamados de componentes principais, e têm as seguintes propriedades:

- A primeira componente principal explica o máximo (possível em uma dimensão) da variância contida nos dados.

- A segunda componente principal não é correlacionada com a primeira e explica o máximo da variância restante, entre todas as componentes não correlacionados com a primeira componente principal.

- A terceira componente principal não é correlacionada com as duas primeiras e explica o máximo de variância restante, entre todas as componentes não correlacionadas com as duas primeiras.

A matriz $\underset{\sim}{X}$ é então descrita por

$$\underset{\sim}{X} = \underset{\sim}{T}\,\underset{\sim}{U}' + \varepsilon \qquad (11)$$

onde os autovetores contidos em $\underset{\sim}{U}'$ são os pesos, os elementos de $\underset{\sim}{T}$ são os escores e $\underset{\sim}{\varepsilon}$ é a matriz dos resíduos. Assim

$$\underset{\sim}{\varepsilon} = \underset{\sim}{X} - \underset{\sim}{T}\,\underset{\sim}{U}' \qquad (12)$$

onde $\underset{\sim}{T}$ é uma matriz de dimensão n x a, $\underset{\sim}{U}'$ é uma matriz a x p e $\underset{\sim}{\varepsilon}$ uma matriz n x p.

Em situações normais, o número de componentes principais extraídos, 'a', não é conhecido, mas é determinado de tal maneira que explique satisfatoriamente a variância total dos dados. Para determinar o número apropriado de componentes principais pode-se utilizar técnicas de "Cross-Validation" (24). Uma outra forma é calcular a percentagem de variância cumulativa obtida para os valores sucessivos de a = 1, 2, ..., A, e parar quando esta for suficientemente grande, por exemplo maior que 90%. A soma de todos os λ_p deve corresponder a 100% de variância total. Como o autovalor é a variância ao longo do autovetor correspondente, a percentagem da variância contida nos 'a' primeiros autovetores é dada por:

$$\%V = \sum_{i=1}^{a} \lambda_i .100 / \sum_{i=1}^{p} \lambda_i \qquad |13|$$

Assim a dimensão do espaço p é reduzida usando no lugar das variáveis originais os autovetores correspondentes aos autovalores com variância significativa. Neste caso o espaço p pode ser projetado usando esta transformação para um espaço de dimensão mais baixa, de fácil visualização para o pesquisador.

2. Análise do tipo Q de Imbrie

A análise do tipo Q é designada para descrever as interrelações entre objetos, enquanto que a análise do tipo R analisa as interrelações entre as variáveis. De uma certa forma, os escores derivados da análise do tipo R proporcionam um meio de descrever relações entre objetos; contudo, estas associações não

são usualmente baseadas em uma medida adequada da similaridade entre objetos. Isto é, a covariância ou correlação pode não ser o melhor critério para julgar o grau de similaridade entre objetos.

O ponto de partida da análise do tipo Q é construir uma matriz de associação n x n (7,23) contendo o grau de similaridade entre todos os possíveis pares de objetos. Uma vez que a matriz tenha sido construída, o procedimento segue passos similares aos da análise de componentes principais. As matrizes resultantes, no entanto, são interpretadas de forma completamente diferente. A medida de associação mais comumente utilizada é o coeficiente cosseno teta. Para dois objetos, k e j, esse coeficiente é determinado por:

$$\cos\theta_{kj} = \frac{\sum_{i=1}^{p} X_{ki} X_{ji}}{\sum_{i=1}^{p} X_{ki}^2 \sum_{i=1}^{p} X_{ji}^2} \qquad |14|$$

onde X_{ki} representa a i-ésima variável do k-ésimo objeto e p é o número de variáveis.

Esta medida dá o cosseno do ângulo entre dois vetores no espaço p-dimensional. Se as p variáveis forem escalonadas para média zero e variância igual a um, o coeficiente de correlação e o coeficiente cosseno teta serão semelhantes.

Depois de calcular o cosseno-teta para todos os possíveis pares de objetos, estes coeficientes são arranjados em uma matriz de associação $\underline{H}$. Esta operação pode ser realizada em dois passos. Primeiro definimos:

$$W_{ki} = \frac{X_{ki}}{\sum_{i=1}^{p} X_{ki}^2} \qquad (i = 1, \ldots, p;\ k = 1, \ldots, n). \qquad |15|$$

Isto corresponde a dividir cada elemento de uma linha, pela raiz quadrada da soma dos quadrados da linha, normalizando a matriz de dados tal que:

$$\sum_{i=1}^{p} W_{ki}^2 = 1, \text{ para } k = 1, \ldots n,$$

então $$\cos\theta_{kj} = \sum_{i=1}^{p} W_{ki}\, W_{ji} \quad . \qquad |16|$$

Usando notação matricial temos:

$$\underset{\sim}{W}_{(nxp)} = \underset{\sim}{D}^{1/2}\ \underset{\sim}{X} \qquad |17|$$

onde $\underset{\sim}{D}$ é a matriz nxn da soma dos quadrados das linhas de $\underset{\sim}{X}$. A matriz de associação torna-se então:

$$\underset{\sim}{H}_{(nxn)} = \underset{\sim}{W}\ \underset{\sim}{W}' = \underset{\sim}{D}^{-1/2}\ \underset{\sim}{X}\ \underset{\sim}{X}'\ \underset{\sim}{D}^{-1/2} \qquad |18|$$

A normalização da linha não afeta a relação de proporcionalidade entre as variáveis mas remove os efeitos dos diferentes tamanhos dos objetos. A matriz de associação $\underset{\sim}{H}$ contém a separação angular de todos os objetos (vetores linha) como eles estão situados no espaço p-dimensional.

Tendo calculado a matriz de associação, pode-se obter as componentes principais. Isto é feito através da diagonalização da matriz $\underset{\sim}{H}$, obtendo como solução os autovalores e autovetores, como no modelo tipo R.

Neste caso a componente é definida por:

$$\underset{\sim}{P} = \underset{\sim}{U}_a \; \Delta_a^{1/2} \qquad |19|$$

onde $\underset{\sim}{P}$ é uma matriz n x a, $\underset{\sim}{U}$ é uma matriz n x a e $\underset{\sim}{\Delta}$ é uma matriz a x a, onde 'a' é determinado pelo número de autovalores significativos.

Analogamente ao tipo R, a matriz dos escores pode ser definida como:

$$\underset{\sim}{T} = \underset{\sim}{W}' \; \underset{\sim}{P} \; \underset{\sim}{\Delta}^{-1} \qquad |20|$$

onde $\underset{\sim}{T}$ é uma matriz p x a, $\underset{\sim}{W}'$ é uma matriz p x n, $\underset{\sim}{P}$ é uma matriz n x a e Δ é uma matriz a x a.

Até agora enfatizamos a idéia conceitual da análise do tipo Q. Na prática, quando o número de objetos, n, for muito maior que o número de variáveis, p, é computacionalmente mais eficiente usar o seguinte procedimento:

1) Calcular a matriz de associação $\underset{\sim}{H}$ como $\underset{\sim}{H} = \underset{\sim}{W}'\underset{\sim}{W}$ ao invés de $\underset{\sim}{W}\,\underset{\sim}{W}'$.

2) Calcular os autovalores $\underset{\sim}{\Delta}$ e os autovetores correspondentes $\underset{\sim}{U}$.

3) A matriz dos escores é idêntica a $\underset{\sim}{U}_a$, isto é:

$$\underset{\sim}{T} = \underset{\sim}{U}_a$$

4) A matriz dos pesos pode ser calculada por:

$$\underset{\sim}{P} = \underset{\sim}{W} \; \underset{\sim}{T}$$

Neste caso nenhuma normalização é necessária nas colunas de $\underset{\sim}{T}$ e $\underset{\sim}{P}$.

Rotação Ortogonal de Componentes Principais (VARIMAX)

Como já vimos, a finalidade da análise de componentes principais é exibir a configuração das variáveis ou objetos da maneira mais simples possível. Embora a dimensionalidade se reduza, para um número de componentes 'a' com variância significativa, estas componentes nem sempre podem ser facilmente interpretadas.

A técnica de rotação varimax, descrita por Kaiser (7,23, 25,26), faz a rotação das componentes para encontrar uma posição mais facilmente interpretável. A matriz das componentes mais simples para interpretar obtida pela rotação, figura 8, contém poucos pesos grandes e muitos pesos próximos de zero.

O critério varimax envolve a maximização da variância dos pesos nas componentes. A quantidade que será maximizada é:

$$V = \sum_{i=1}^{a} \frac{\left[p \sum_{j=1}^{p} (u_{ji}^2/h_j^2)^2 - (\sum_{j=1}^{p} u_{ji}^2/h_j^2)^2 \right]}{p^2} \qquad |21|$$

onde $h_j^2 = \sum_{i=1}^{a} u_{ji}^2$ $(j = 1, 2, \ldots, p)$ e

u_{ji} é o peso da variável j na componente a, (veja figura 8), p é o número de variáveis e h_j^2 é a chamada comunalidade da variável j. Nesta rotação as componentes permanecem ortogonais.

A comunalidade h_j^2 é a fração da variância da variável j comum às componentes retidas no modelo. Esta fração é definida como a soma dos quadrados dos elementos de uma linha, correspondendo à variável daquela linha. A comunalidade total deve permanecer inalterada após a rotação.

A soma dos quadrados dos elementos de uma coluna é a variância ao longo da componente representada nesta coluna. Esta soma dividida pelo número de variáveis p dá a fração de variância explicada pela componente.

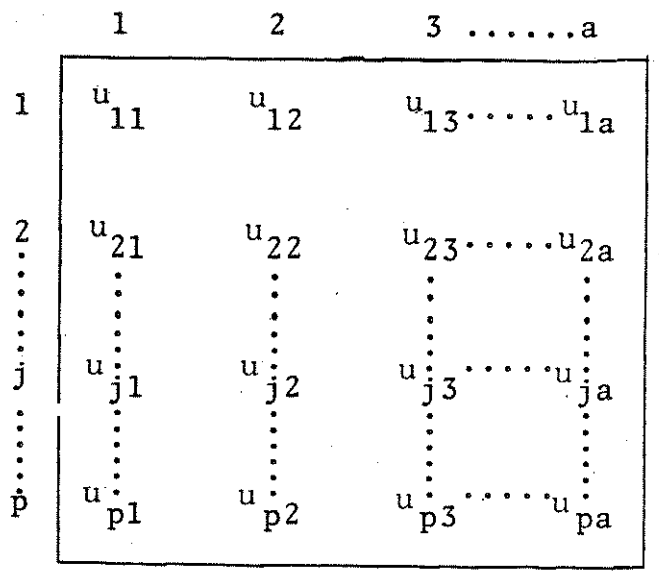


Figura 8 - Matriz dos pesos simplificada pela rotação varimax.

Métodos de Classificação

Uma finalidade importante da análise dos resultados para um conjunto de objetos é desenvolver regras de classificação de objetos de origem desconhecida, com base em um grupo de objetos de classificação conhecida, caracterizados por um número de medidas (variáveis). Por exemplo classificar produtos alimentícios de acordo com a origem ou qualidade, classificar vinho de acordo com a região e ano, etc. Se as regras de decisão enfatizam a diferença entre as classes, ou a similaridade dentro de cada classe, pode-se obter uma modelagem que permita a discriminação entre as

classes.

O conjunto de dados para classificação é formado por diferentes classes (ou categorias) com classificação conhecida. A este conjunto daremos o nome de conjunto de treinamento (10, 27, 28,29), Os dados são arranjados em uma matriz $\underset{\sim}{X}$, como na figura 1, onde o elemento X_{ki} representa o valor da i-ésima variável para o k-ésimo objeto. Com base neste conjunto de treinamento são desenvolvidas as regras de classificação, as quais tornam possível a classificação de novos objetos, que pertencem ao conjunto de teste, onde as variáveis são as mesmas obtidas para o conjunto de treinamento, figura 9.

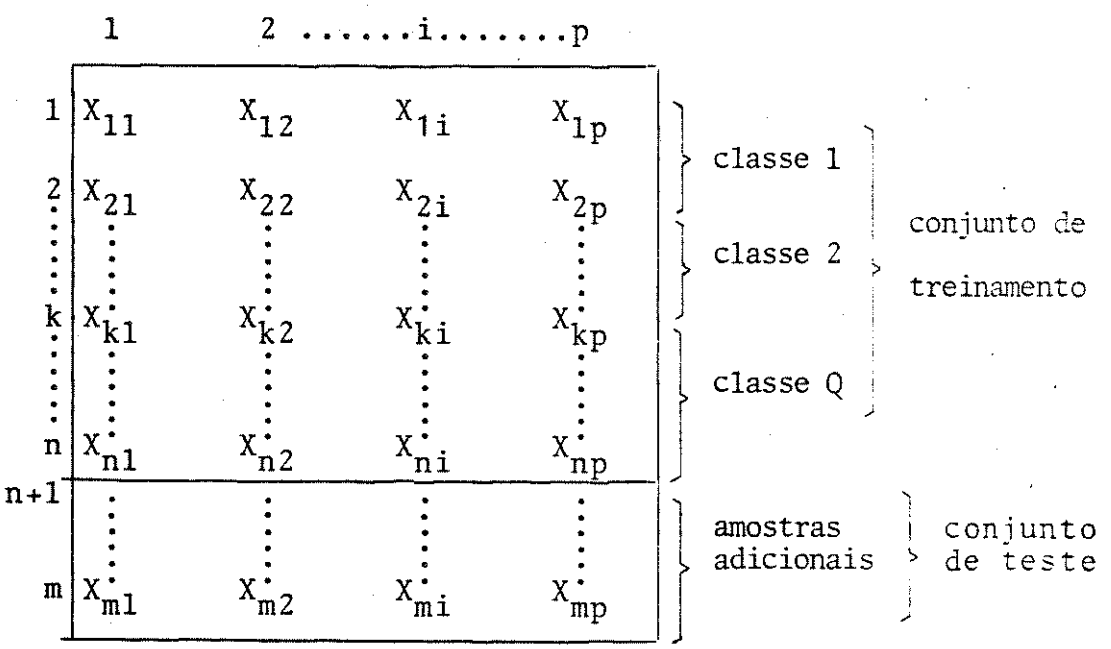


Figura 9 - Organização da matriz $\underset{\sim}{X}$, para fins de classificação.

Os objetos do conjunto de teste são comparados com os padrões característico de cada classe usando procedimentos matemáticos como meio de comparação. Cada objeto é atribuído à classe que mostra a maior similaridade com os dados do objeto, figura 10.

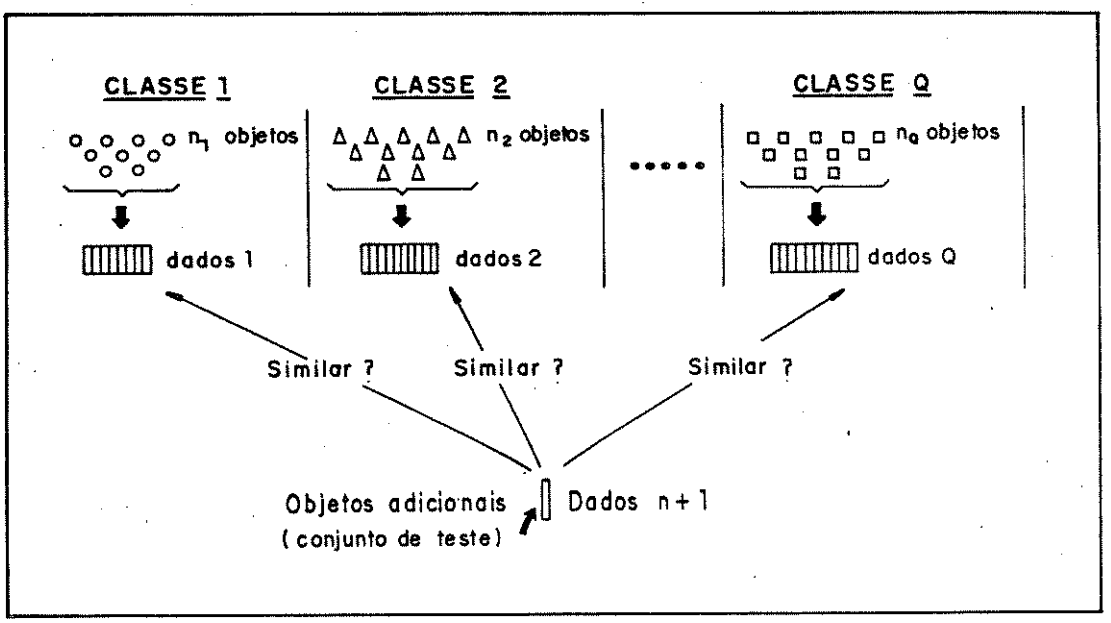


Figura 10 - Princípio metodológico do reconhecimento de padrões.

Neste trabalho discutiremos dois métodos de classificação particularmente eficazes em estudos químicos. Outros métodos podem ser consultados nas referências (29-32).

Modelos Independentes de Similaridade Utilizando Componentes Principais (SIMCA)

SIMCA é a sigla de "Soft Independent Modelling by Class Analogy" (33). Entretanto, um título mais descritivo seria 'Modelos Independentes de Similaridade Utilizando Componentes Principais (34). Este método usa a análise de componentes princi-

pais para construir separadamente um modelo para cada classe do conjunto de treinamento. Cada classe de objetos similares é descrito por um modelo com um pequeno número de componentes principais. O número ótimo de componentes para cada classe pode ser determinado pela técnica de 'cross-validation' (24).

Antes da análise estatística os dados devem ser autoescalonados. A análise então procede para encontrar a dimensionalidade correta (35) para os modelos de todas as classes.

Inicialmente, um modelo com zero componentes principais é ajustado para os dados da classe, quer dizer, cada variável é descrita pelo seu valor médio (34,36), como mostrado na equação 3. O modelo pode ser representado no espaço p por uma hiperesfera de raio $2S_o$, ilustrada na figura 11. S_o é o desvio padrão dos resíduos,

$$S_o = \sum_{i=1}^{p} \sum_{k=1}^{n} e_{ki}^2 / [(p-a)(n-a-1)]. \qquad |22|$$

O denominador da equação representa o número de graus de liberdade, sendo 'a' o número de componentes significativos.

O modelo de uma componente principal, descrito pela equação 5, é então representado geometricamente por um hipercilindro no espaço p, veja figura 11. O eixo do cilindro corresponde nesse modelo à primeira componente principal. Se a estrutura dos dados for relativamente simples, um modelo com uma componente é adequado para representar os dados. Caso a estrutura seja complexa mais componentes serão necessários.

A equação 7 descreve o modelo com duas componentes principais. Essa equação é representada geometricamente na figura 11 por uma hipercaixa no espaço p, cujo comprimento e largura são determinados pelos desvios padrão ao longo das duas componen

O número de componentes principais necessários para descrever os dados pode variar de uma categoria para outra, dependendo do grau de complexidade da estrutura dos dados em cada caso, como está ilustrado na figura 11.

Uma vez que os modelos de classe tenham sido estabelecidos, podem então ser usados para classificar objetos pertencentes a categorias desconhecidas, figura 12. Estes objetos (representados por $\underset{\rightarrow}{X}_{mi}$) são ajustados aos parâmetros das classes por regressão múltipla.

$$X_{mi} - \overline{X}'_i = \sum_{a=1}^{A} t_{ka} P'_{ai} + e_{im} \qquad |24|$$

O SIMCA (16,28,34) é capaz de indicar quando um ponto referente a um objeto não pertence a nenhuma das categorias definidas no conjunto de treinamento, representando-o como um ponto ou membro potencial de uma categoria não definida, figura 12.

Outra vantagem deste método é que ele pode ser aplicado a casos onde o número de amostras n seja bem menor que o número de variáveis p.

Finalmente o SIMCA permite que somente os pontos de uma categoria sejam modelados com componentes principais, enquanto que os outros pontos não são considerados, figura 13. Este problema de classificação binária é comumente encontrado em ciências dos materiais, controle de qualidade e em produtos que apresentam atividade biológica específica. Por exemplo, os objetos representados nesta figura por (●) podem ser relativos a um produto manufaturado de especificação bem definidas, resultantes de um bom controle de qualidade e seus pontos se dispõem no espaço definido pela caixa. Por outro lado, os produtos de má qualidade estarão

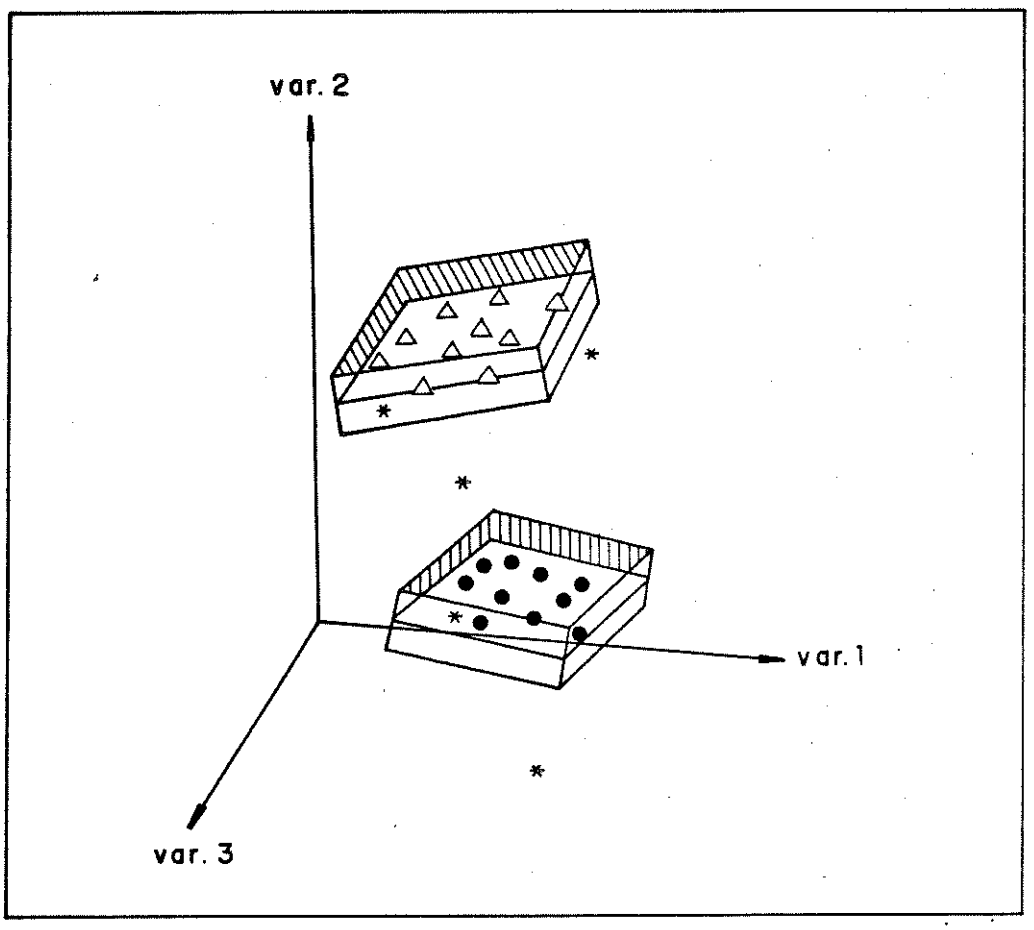


Figura 12 - As amostras assinaladas por (Δ) e (●) nas hipercaixas pertencem a classes definidas. As amostras do conjunto de teste (*), que se situam fora das hipercaixas são consideradas como pertencentes a classes ainda não definidas.

dispersos no espaço p. Este modelo é útil para decidir se uma amostra tem as especificações desejadas para o produto final.

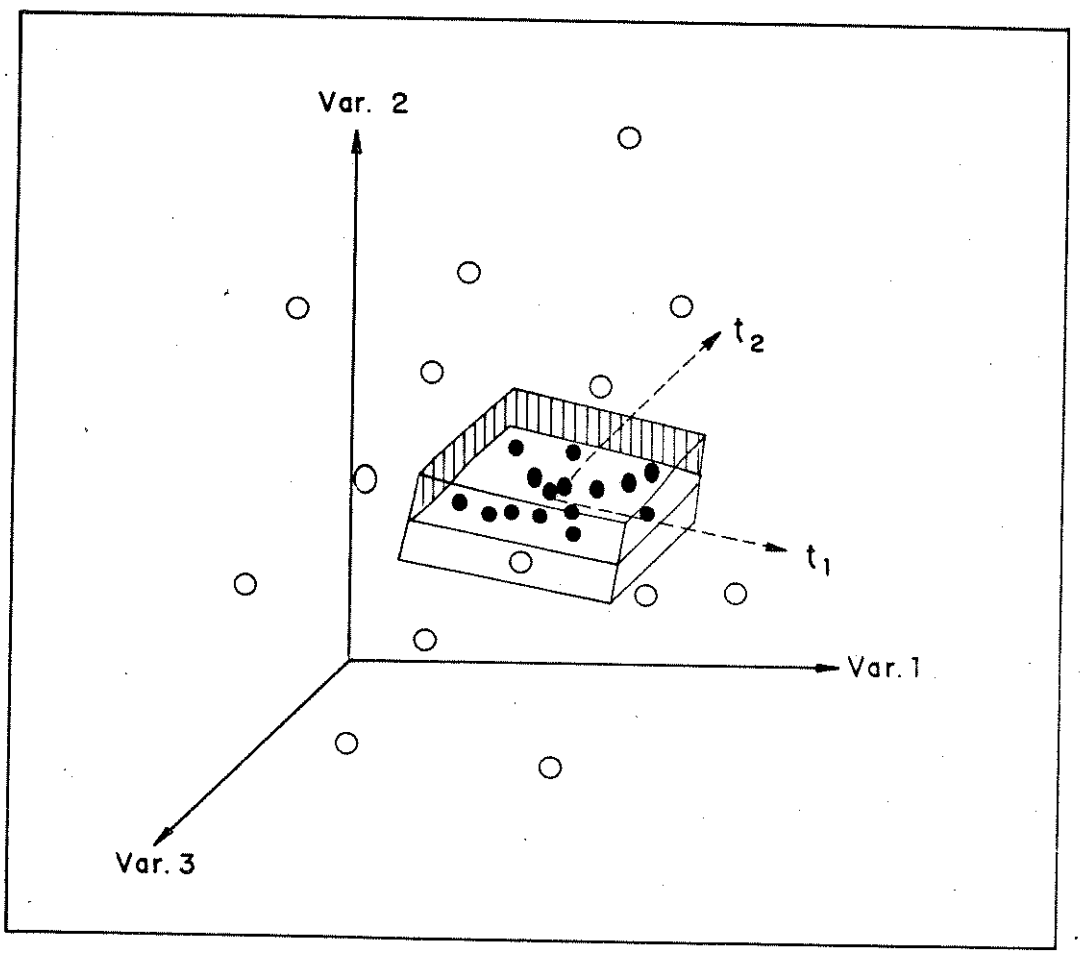


Figura 13 - Uma categoria bem definida (representando, por exemplo, um produto industrial de alta qualidade) circundada por uma outra, espalhada no espaço p (produtos de qualidade não controlada).

Regra dos K Vizinhos mais Próximos (KNN)

KNN é a sigla de 'K nearest-neighbors', regra dos K vizinhos mais próximos. Este método é computacionalmente muito simples e baseia-se nas similaridades dos objetos. A matriz das dis

tâncias entre os pontos dos conjuntos de treinamento e de teste é calculada para todos os pontos (18,37). A distância normalmente usada é a distância euclidiana definida na equação 1.

A finalidade do método (28,33,37,38) é classificar um objeto teste na classe (ou categorias) do conjunto de treinamento, a qual pertença a maioria dos seus vizinhos mais próximos, figura 14. Normalmente os k vizinhos mais próximos são tomados como K = 1,3,4... 10, e são determinados pelo conjunto de treinamento, onde a proximidade é definida com base nas distâncias entre os objetos.

Uma desvantagem do método é que ele não é capaz de indicar se um objeto pertence a uma categoria não definida, como no caso do SIMCA. Além disto, para um número muito grande de amostras o tempo de computação exigido é muito maior do que o do SIMCA.

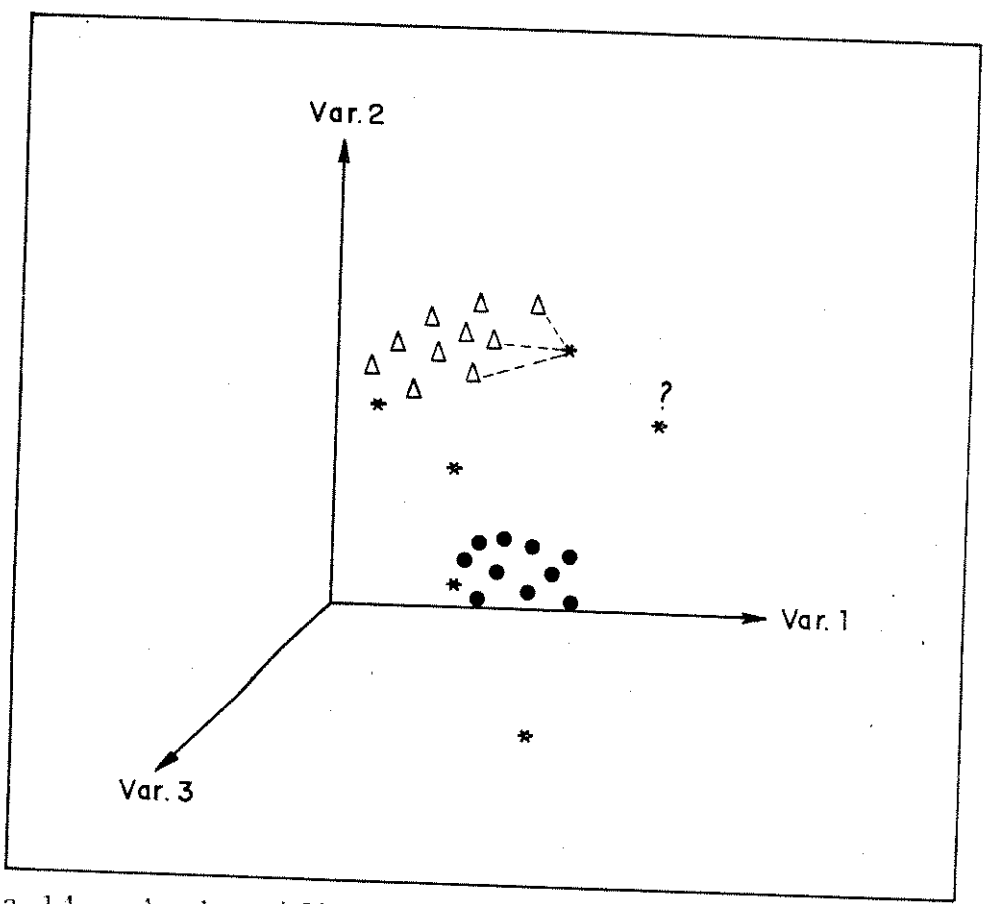


Figura 14 - A classificação de uma amostra com categoria desconhecida é feita com base nas categorias dos seus vizinhos mais próximos.

Calibração Multivariada

A pesquisa em calibração multivariada é uma das grandes atividades atuais da quimiometria (5). A calibração na química analítica relaciona as respostas do instrumento à concentrações químicas, e pode ser dividida em dois estágios (39). Primeiro, investigar as características de um método ou do instrumento, e então encontrar um modelo para sua atuação, expresso na relação $\underline{Y} = f(\underline{x})$, entre dois grupos de variáveis. Esta é a etapa de calibração ou treinamento. O conjunto de dados para esta etapa é chamado de conjunto de calibração ou conjunto de treinamento. O segundo estágio é o de previsão, em que as variáveis independentes X, normalmente respostas analíticas, são obtidas para uma ou mais amostras e são usadas para prever valores para as variáveis dependentes Y, as concentrações. O conjunto de dados usado neste estágio é chamado de conjunto de previsão ou conjunto de teste.

A tabela de dados é dividida em duas partes, os blocos $\underline{X}$ e $\underline{Y}$, figura 15. Para casos gerais, os conjuntos de treinamento e de teste são divididos em classes. Entretanto, em muitos casos práticos o número de classes é um.

Várias técnicas de regressão podem ser usadas na calibração multivariada. O método de calibração mais simples é a regressão linear múltipla (39-41). Para cada variável Y, são medidas p variáveis $X_i (i = 1,2... p)$, a fim de estabelecer uma relação linear entre elas. Para um sistema multivariado (39) isto pode ser representado matematicamente como

$$Y = b_o + b_1X_1 + b_2X_2 + ... b_pX_p + e \qquad |25|$$

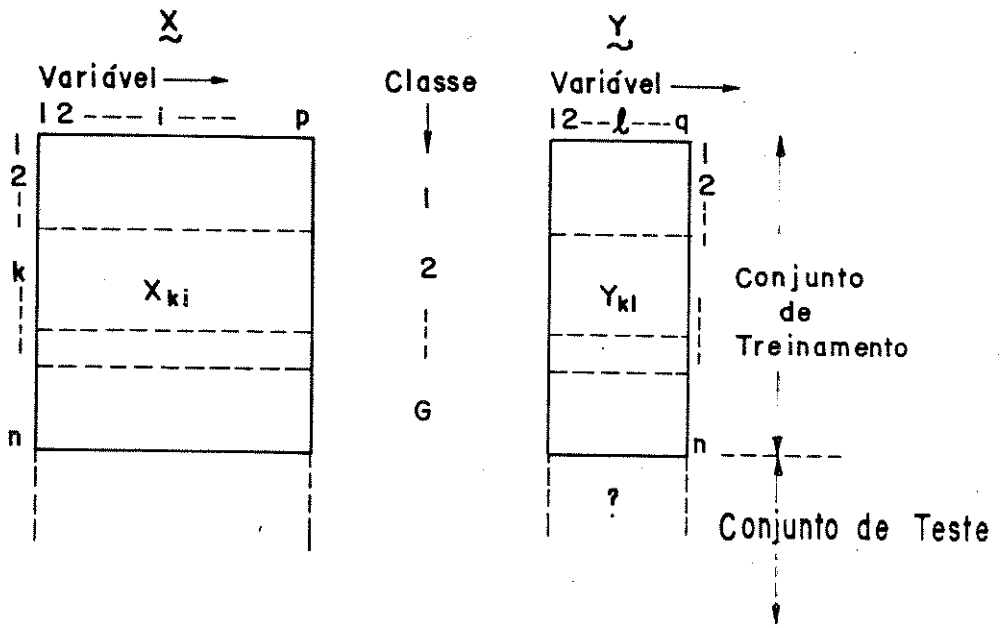


Figura 15 - Organização dos dados para um problema de calibração multivariada.

Com o autoescalonamento dos dados é possível eliminar a constante b_o na equação linear, isto é, $y = 0$ quando $X_i = 0$, para $i = 1,2,\ldots p$ e

$$Y = \sum_{i=1}^{p} b_i X_i + e \qquad |26|$$

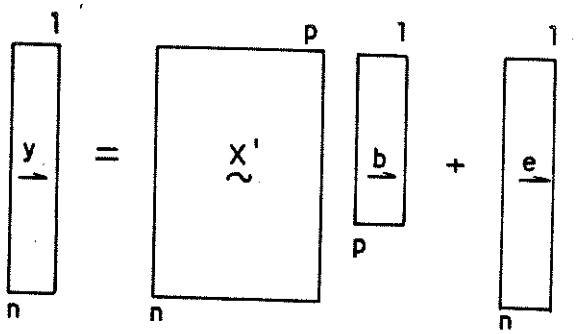
onde 'e' é o erro do modelo de regressão linear na estimativa do valor de Y. O tratamento de regressão linear fica mais simples usando notação vetorial

$$Y = \underset{\rightarrow}{X}' \underset{\rightarrow}{b} + e \qquad |27|$$

Esta equação descreve as dependências multilineares para somente uma amostra. Neste caso $\underset{\rightarrow}{b}$ é um vetor coluna e $\underset{\rightarrow}{X}'$ é um vetor linha. Para n amostras a equação 27 torna-se:

$$\underset{\rightarrow}{Y} = \underset{\sim}{X}' \underset{\rightarrow}{b} + \underset{\rightarrow}{e} \qquad |28|$$

Sua representação matricial é:



onde n é o número de amostras e p o número de variáveis independentes.

É possível distinguir três casos:

1. p > n. Existe um número infinito de soluções para b. Este caso não é útil em química analítica.

2. p = n. Esta situação não é normalmente encontrada na prática e não permite uma avaliação da qualidade estatística do modelo. Neste caso

$$\underset{\rightarrow}{e} = \underset{\rightarrow}{Y} - \underset{\sim}{X}' \underset{\rightarrow}{b} = 0.$$

3. p < n. Este caso não permite uma solução exata para $\underset{\rightarrow}{b}$ mas pode

mos chegar a uma solução para minimizar o comprimento do vetor $\underset{\rightarrow}{e}$. O método mais popular é o método dos mínimos quadrados, onde $\underset{\rightarrow}{e}'\underset{\rightarrow}{e}$ é minimizado com

$$\underset{\rightarrow}{e} = \underset{\rightarrow}{Y} - \underset{\sim}{X}'\underset{\rightarrow}{b} \qquad |29|$$

cuja solução é

$$\underset{\rightarrow}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\rightarrow}{Y} \qquad |30|$$

onde $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ representa a inversa do produto das matrizes $\underset{\sim}{X}'\underset{\sim}{X}$.

Um dos problemas mais frequentes na regressão linear múltipla é que, se existir correlações entre as variáveis independentes, a inversa de $\underset{\sim}{X}'\underset{\sim}{X}$ na equação 30 pode não existir. Colinearidade, determinante zero e singularidade são os nomes atribuídos a este problema. Se existir, os coeficientes de regressão tornam-se numericamente, bem como estatisticamente, incertos. Este é um problema em química onde as respostas analíticas são fortemente correlacionadas.

A regressão linear múltipla pode ser facilmente estendida para mais de uma variável dependente. Considerando duas variáveis dependentes y_1 e y_2, podemos escrever duas regressões e encontrar dois vetores $\underset{\rightarrow}{b}_1$ e $\underset{\rightarrow}{b}_2$.

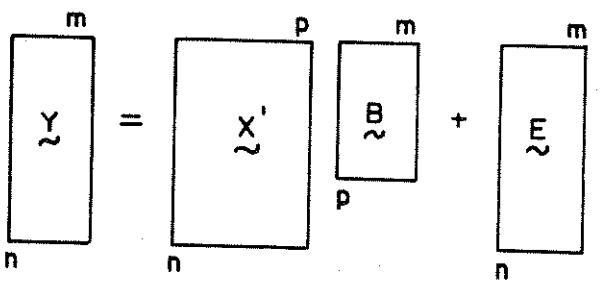
$$\underset{\rightarrow}{y}_1 = \underset{\sim}{X}'\underset{\rightarrow}{b}_1 + \underset{\rightarrow}{e}_1 \quad \text{e} \quad \underset{\rightarrow}{y}_2 = \underset{\sim}{X}'\underset{\rightarrow}{b}_2 + \underset{\rightarrow}{e}_2 .$$

Estes vetores podem então ser colocados em uma matriz, assim

$$\underset{\sim}{Y} = \underset{\sim}{X}'\ \underset{\sim}{B} + \underset{\sim}{\varepsilon} \qquad |31|$$

onde $\underset{\sim}{Y} = (\underset{\rightarrow}{y}_1\ \underset{\rightarrow}{y}_2)$, $\underset{\sim}{B} = (\underset{\rightarrow}{b}_1\ \underset{\rightarrow}{b}_2)$ e $\underset{\sim}{\varepsilon} = (\underset{\rightarrow}{e}_1\ \underset{\rightarrow}{e}_2)$. Este procedimento po-

de ser estendido a m variáveis independentes e a representação matricial é



Resumindo:

1. Para p > n, não existe solução única, a não ser que o número de variáveis seja reduzido.
2. Para p = n, existe uma solução única.
3. Para p < n, a solução dos mínimos quadrados é possível. Nos casos 2 e 3 a inversão da matriz pode causar problemas.
4. A regressão linear múltipla é possível para mais que uma variável dependente.

Para objetos onde conhecemos somente as variáveis independentes, $\underset{\sim}{X}_{teste}$, a previsão dos valores de $\underset{\sim}{Y}$ é feita multiplicando a matriz $\underset{\sim}{X}'_{teste}$ pela matriz dos coeficientes de regressão $\underset{\sim}{B}$ obtida no conjunto de treinamento. Assim,

$$\underset{\sim}{Y}_{prev} = \underset{\sim}{X}'_{teste}\ \underset{\sim}{B} \qquad |32|$$

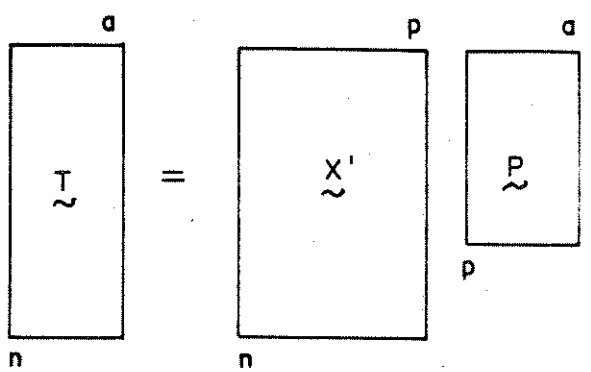
O método de regressão linear múltipla é um procedimento

adequado em situações ideais. Entretanto, o método é baseado em um passo problemático, que é a inversão da matriz $(X'X)$, podendo incorporar quantidades significativas de variância irrelevante no modelo (21).

Regressão em Componentes Principais - PCR

Alguns problemas associados com a regressão linear múltipla mencionados anteriormente podem ser resolvidos com a regressão em componentes principais. Além de apresentar variáveis independentes que são ortogonais, cada componente descreve uma fração de variância total. Como esta fração pode ser calculada para cada componente principal, fica mais fácil determinar quantas variáveis independentes devem ser incluídas na regressão.

O procedimento para construir o modelo de regressão em componentes principais consiste de dois passos (21,39). O primeiro é determinar os autovetores ou componentes com os 'a' autovalores mais significativos para a matriz X. Esta informação é usada para converter X na matriz de escores T. A transformação é $T = X P$ onde P é a matriz dos autovetores. Em representação matricial

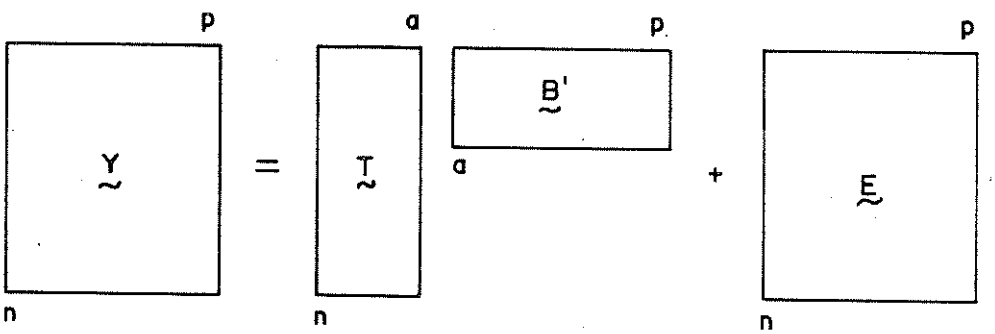


A matriz escore $\underset{\sim}{T}$ é composta pelos pontos dos dados originais no novo sistema de coordenadas descritos pelos autovetores.

O segundo é usar a regressão linear múltipla para fazer a regressão da matriz $\underset{\sim}{Y}$ na matriz dos escores:

$$\underset{\sim}{Y} = \underset{\sim}{T}\ \underset{\sim}{B}' + \underset{\sim}{\varepsilon} \qquad |33|$$

ou



onde $\underset{\sim}{B}'$ é a matriz transposta dos coeficientes de regressão e $\underset{\sim}{\varepsilon}$ é o erro de modelagem na estimativa dos valores de $\underset{\sim}{Y}$.

A previsão dos valores de $\underset{\sim}{Y}$ para o conjunto de teste é feita da seguinte forma: calcula-se os escores dos objetos da matriz $\underset{\sim}{X}_{teste}$ e multiplica-se pela matriz dos coeficientes de regressão, isto é:

$$\underset{\sim}{Y}_{prev} = \underset{\sim}{T}_{teste}\ \underset{\sim}{B}' \qquad |34|$$

A diferença entre a regressão linear múltipla e a regressão em componentes principais encontra-se na substituição da ma-

triz $\underset{\sim}{X}$ pela matriz escore $\underset{\sim}{T}$.

Usando a matriz escore $\underset{\sim}{T}$, as variâncias em $\underset{\sim}{B}$ serão pequenas

$$\underset{\sim}{B} = (\underset{\sim}{T}'\underset{\sim}{T})^{-1}\underset{\sim}{T}'\underset{\sim}{Y} \text{ e } \text{VAR-COV}(\underset{\sim}{B}) = (\underset{\sim}{T}'\underset{\sim}{T})^{-1}\sigma^2$$

Para uma matriz $\underset{\sim}{X}$ (nxp), se todos os p autovetores forem usados para formar $\underset{\sim}{P}$, os resultados das duas regressões serão idênticos (21). As vantagens da regressão em componentes principais estão baseadas na concentração de informação em $\underset{\sim}{X}$ com poucos componentes, usando assim um número mínimo de dimensões. Retirando-se alguns fatores que contêm mais ruído, não se reduz significantemente a quantidade de informação presente nos dados. Esta redução de dimensionalidade é muitas vezes importante para a estabilidade estatística dos valores de $\underset{\sim}{Y}_{prev}$. Outra característica importante da matriz $\underset{\sim}{T}$ é que as colunas serão mutuamente ortogonais, e assim pode-se sempre calcular a inversa $(\underset{\sim}{T}'\underset{\sim}{T})^{-1}$.

Resumindo:

1. A matriz de dados $\underset{\sim}{X}$ é representada pela matriz de escores $\underset{\sim}{T}$.
2. O número de variáveis é reduzido porque somente 'a' autovetores são usados na regressão.
3. O problema da singularidade da matriz de dados é eliminado porque as componentes principais são ortogonais.

Mínimos Quadrados Parciais (PLS)

O método dos mínimos quadrados parciais estima simultaneamente as componentes principais em ambas as matrizes $\underset{\sim}{X}$ e $\underset{\sim}{Y}$

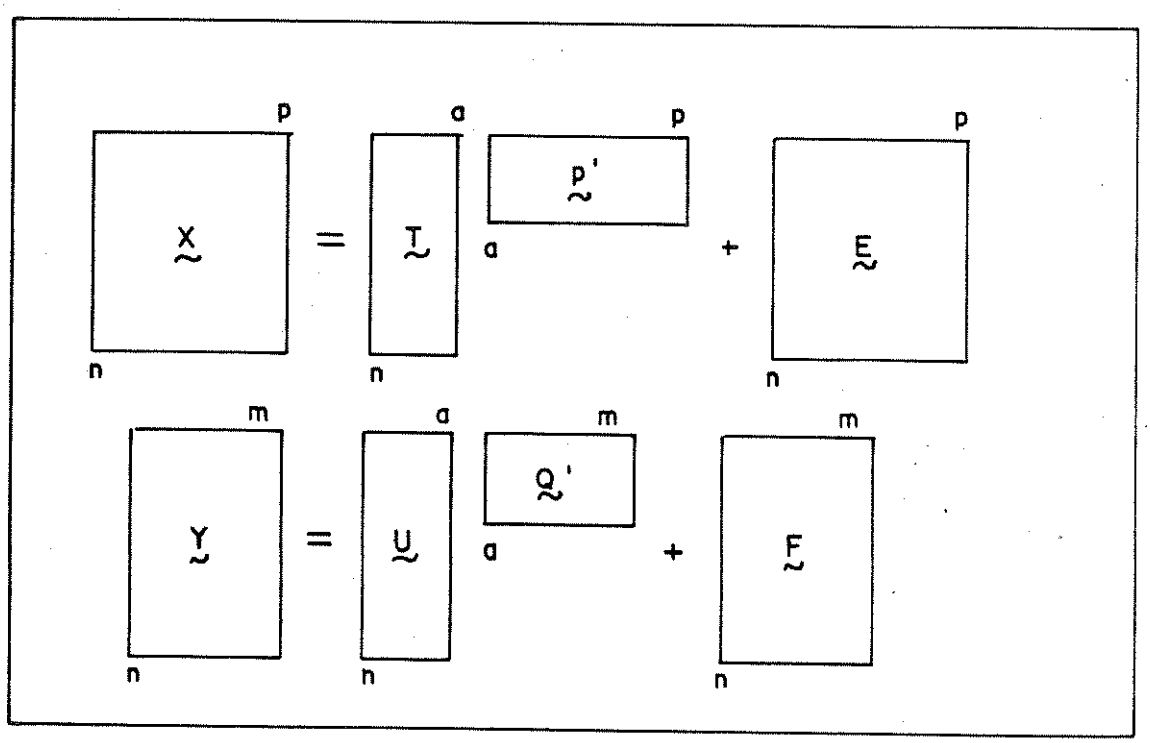
(39,41,42,43). O modelo simplificado consiste em uma regressão entre os escores das matrizes $\underset{\sim}{X}$ e $\underset{\sim}{Y}$. Os modelos resultantes são:

$$\underset{\sim}{X} = \underset{\sim}{T}\ \underset{\sim}{P}' + \underset{\sim}{\varepsilon} = \Sigma\ \underset{\rightarrow a}{t}\ \underset{\rightarrow}{P}'_a + \underset{\sim}{\varepsilon} \qquad |35|$$

$$\underset{\sim}{Y} = \underset{\sim}{U}\ \underset{\sim}{Q}' + \underset{\sim}{F} = \Sigma\ \underset{\rightarrow a}{u}\ \underset{\rightarrow}{q}'_a + \underset{\sim}{F} \qquad |36|$$

onde os elementos de $\underset{\sim}{T}$ e $\underset{\sim}{U}$ são os escores de $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ respectivamente, e os elementos $\underset{\sim}{P}'$ e $\underset{\sim}{Q}'$ são os pesos correspodentes. As matrizes $\underset{\sim}{\varepsilon}$ e $\underset{\sim}{F}$ são os erros associados à modelagem de $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ com o modelo dos mínimos quadrados parciais.

As equações 35 e 36 podem ser representadas matricialmente como:



A relação entre os blocos pode ser feita olhando-se um gráfico dos escores de $\underset{\rightarrow}{u}$, do bloco $\underset{\sim}{Y}$, contra os escores, $\underset{\rightarrow}{t}$, do bloco $\underset{\sim}{X}$. O modelo mais simples para esta relação é linear, figura 16.

$$\underset{\rightarrow}{u}_a = b_a \underset{\rightarrow}{t}_a \qquad [37]$$

onde b_a é o coeficiente de regressão.

Este modelo não é o melhor porque as componentes principais são calculadas para ambos os bloco separadamente; sendo assim,

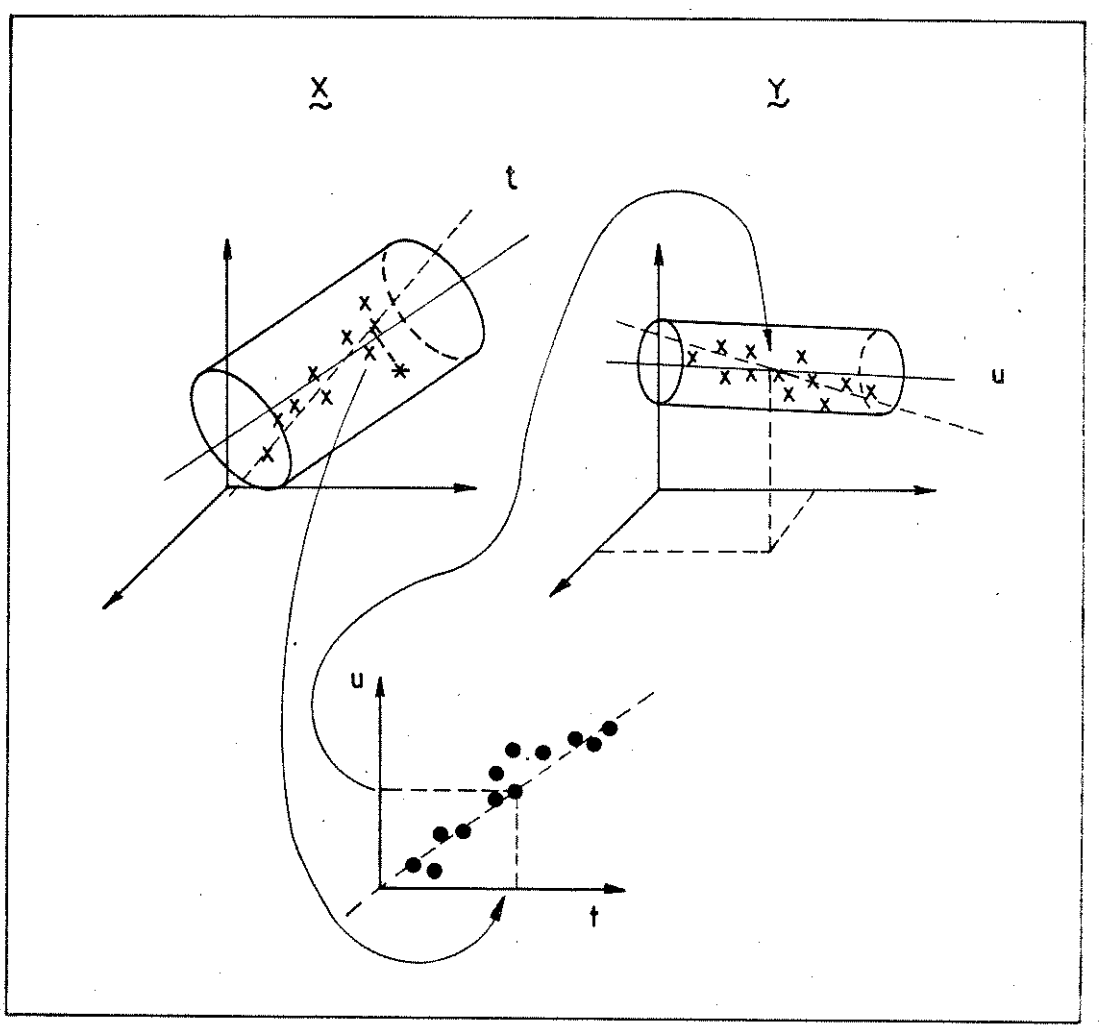


Figura 16-Ilustração geométrica do PLS com uma componente principal.

eles têm uma fraca relação entre si. Para melhorar a troca de informação entre os dois blocos, o PLS usa as componentes ligeiramente inclinadas, minimizando os resíduos para o bloco $\underset{\sim}{Y}$, ou seja os elementos da matriz $\underset{\sim}{F}$. Isto é feito ponderando-se os valores do bloco $\underset{\sim}{X}$ com os valores do bloco $\underset{\sim}{Y}$ e vice-versa. O algorítmo pode ser visto na referência 42.

Para um objeto k para o qual conheçamos somente os valores das variáveis independentes, a previsão dos valores de $\underset{\rightarrow}{Y}$ é feita do seguinte modo: o ponto representado pelo asterisco, figura 16, no espaço $\underset{\sim}{X}$ é projetado no eixo t. Este valor é transferido para o gráfico uxt, obtendo o valor de u correspondente ($u_{ka} = b_a t_{ka} + e_k$). Finalmente, o valor de u é localizado no eixo das componentes no espaço $\underset{\sim}{Y}$. Os valores das variáveis $\underset{\rightarrow}{Y}$ são determinadas então por projeções sobre os eixos originais.

Os modelos de mínimos quadrados parciais podem ser construídos em mais de uma relação entre blocos (44), onde cada par de componentes está relacionado pela equação 37.

CAPÍTULO III

PROGRAMAS

O sistema ARTHUR modificado oferece ao usuário alguns métodos matemáticos e de estatística multivariada coumumente usados na análise de dados químicos. Esta análise combina o processamento de dados com manipulações matemáticas ou estatíticas, condensando a informação contida nos dados originais de forma a ser mais facilmente compreendida e interpretada.

O pacote computacional está documentado no apêndice. Como é um sistema aberto (linguagem FORTRAN), o usuário com um pouco de conhecimento de programação pode facilmente expandir e modificar estes programas de acordo com sua necessidade, o que é impossível em programas disponíveis comercialmente.

O sistema é dirigido por uma sequência de programas, ver figura 17. Esta sequência pode ser alterada de acordo com o que se deseja da análise de dados. Os dados são organizados em forma matricial, como na figura 1 em unidades chamadas arquivos. O arquivo é preparado somente uma vez e a informação pode ser permanentemente salva ao longo dos cálculos. A matriz deve ser preparada em um formato especial para que seja reconhecida pelo sistema. A preparação e as descrições dos programas são discutidas com detalhes no apêndice.

Os dados originais devem ser gravados no arquivo FORT10.DAT. Passamos agora a descrever os diversos módulos:

SCAL - É usado para calcular média, desvio padrão, valor máximo e mínimo, curtose e assimetria de cada variável. Além

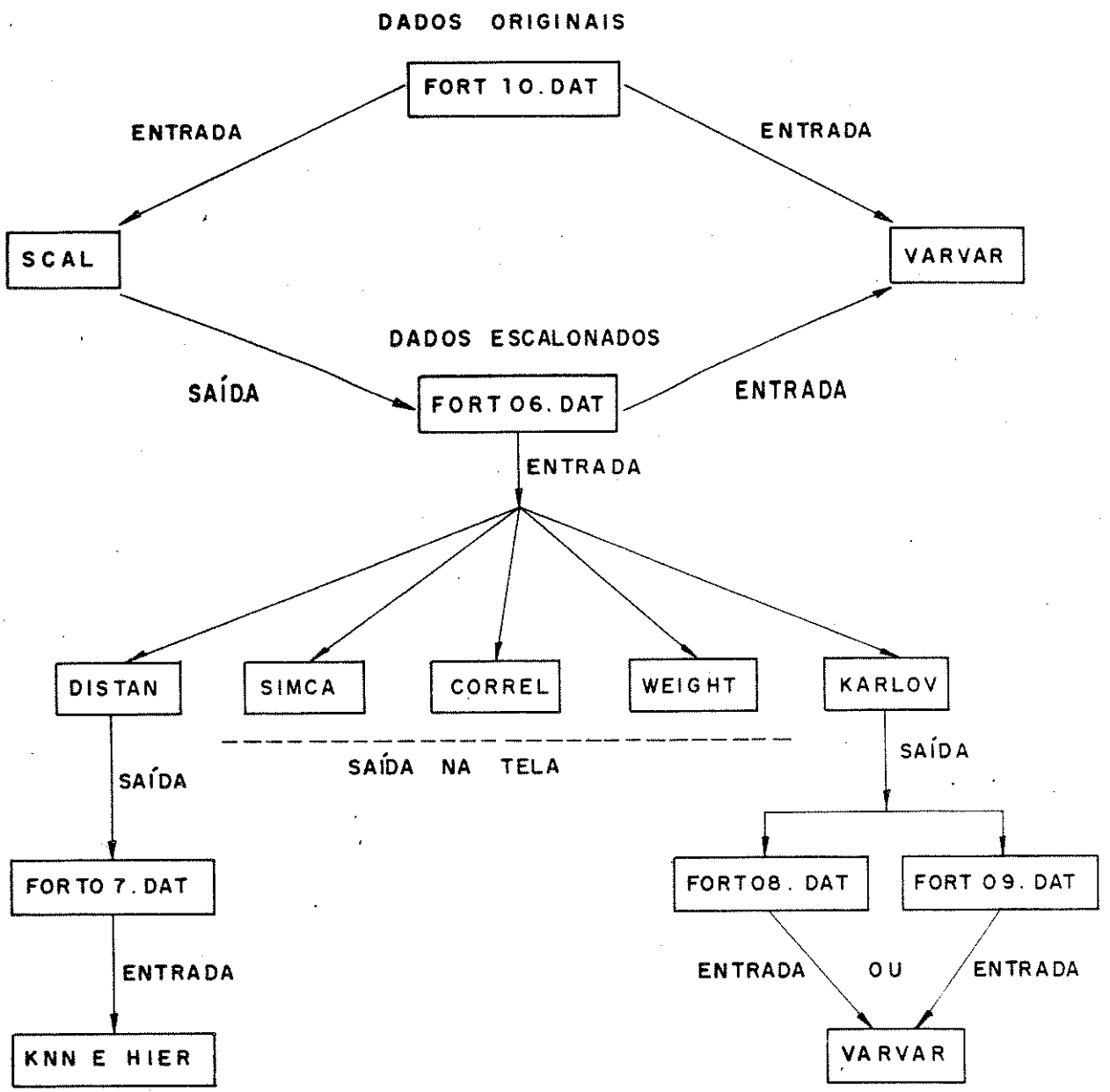


Figura 17 - Fluxograma: Programas e arquivos.

destas informações o programa faz o escalonamento dos dados. O autoescalonamento produz variáveis com média zero e variância igual a um. Quando todas as variáveis são medidas na mesma unidade, à vezes o escalonamento não é necessário.

DISTAN - Calcula a matriz das distâncias entre todos os objetos. Vários tipos de distâncias podem ser calculadas: Mahalanobis generalizada, Manhattan (por quarteirão) e distância de razão de Anders, além da euclidiana, que é mais empregada.

SIMCA - Constrói um modelo de componentes principais para cada classe (ou categoria), para fins de classificação. A classificação de um objeto de classe desconhecida é baseada em sua similaridade com os modelos que representam as classes.

CORREL- É usado para estudar a relação entre todas as variáveis, calculando as correlações entre todos os pares de variáveis.

WEIGHT- Avalia a importância individual (peso) de cada variável na discriminação entre cada par de categorias. É usado como método de redução de dimensionalidade. As duas técnicas mais importantes são o peso de Fisher e o peso de variância.

KARLOV- Este método faz análise de componentes principais. O método pode ser usado para reduzir dimensionalidade, obter informação sobre a estrutura dos dados no espaço p-dimensional num espaço bi-dimensional, identificar pontos

deslocados e delinear classes.

VARVAR- É usado para fazer gráficos em duas dimensões usando a impressora.

KNN - Faz classificação com base nas distâncias entre os objetos no conjunto de treinamento. Um objeto de categoria desconhecida é então classificado na categoria à qual pertence a maioria dos seus vizinhos mais próximos. O número de vizinhos mais próximos utilizado pelo programa (denotado pela letra K) é igual a 1, 3, até 10.

HIER - Produz um dendrograma que descreve o agrupamento hierárquico. No agrupamento do tipo R, obtém-se uma medida da similaridade das variáveis para todas as amostras. No agrupamento do tipo Q, o dendrograma liga os grupos de amostras com os mesmos níveis de similaridade.

VARIMAX- Faz a rotação ortogonal da matriz dos pesos. Esta rotação concentra-se na simplificação das colunas das componentes principais, tornando mínima a complexidade destas componentes.

PCR - Faz calibração multivariada utilizando a análise de componentes principais.

Os programas VARIMAX e PCR não faziam parte do sistema ARTHUR original e foram desenvolvidos nesta tese.

Os programas modificados estão sendo utilizados nas seguintes Universidades: UNICAMP, nos Institutos de Química e Geologia, UNESP de Araraquara, Universidade Federal de Pernambuco, Universidade Estadual de Londrina, Universidade Federal do Piauí,

Universidade Federal do Pará e Universidade Federal do Rio de Janeiro. Várias indústrias também estão utilizando estes programas: a Rio Doce Geologia e Mineração S/A. do Rio de Janeiro, Belém, Salvador e Belo Horizonte, a Cia. Souza Cruz do Rio de Janeiro e o Centro de Pesquisa da Copersucar em Piracicaba.

CAPÍTULO IV

APLICAÇÕES

Primeira Aplicação

Colaboradora:

Profa.Dra. Sonia M.B. Oliveira - Instituto de Geociências - Universidade de São Paulo, São Paulo, SP.

O conjunto de dados refere-se a materiais provenientes da alteração intempérica de granulitos ácidos e básicos da região de Miraí (MG), e constituem-se em bauxitas e argilas ricas em alumínio, intercaladas algumas vezes por níveis muito ricos em ferro.

DESCRIÇÃO EXPERIMENTAL

Foram selecionadas para este estudo 73 amostras, provenientes de 8 poços distribuídos numa área de cerca de 0,5 Km^2 e coletadas a profundidades variáveis entre 1 e 10 m.

A fluorescência de raios-X foi o método analítico utilizado na dosagem dos 10 elementos mais abundantes (Si, Al, Fe, Mg, Mn, Ca, Na, K, Ti e P), com resultados expressos em percentagem em peso dos óxidos (Fe total calculado Fe_2O_3), e de 14 elementos traço (Ba, Ce, Cr, Ga, La, Nd, V, Zn, Zr, Cu, Ni, Nb, Sr e Y), com resultados expressos em ppm dos elementos. A água estrutural foi dosada por métodos convencionais.

Finalidade

Estudar as vantagens do método de rotação ortogonal varimax das componentes principais em relação a análise de componentes principais sem rotação, para investigar as possíveis tendências diferenciais de concentração de certos elementos traço em materiais de alteração ricos em alumínio, materiais argilosos e materiais ferruginosos.

Uma classificação preliminar dos materiais foi feita pela Profa.Dra. Sônia M.B. de Oliveira, utilizando o diagrama triangular Fe_2O_3-Al_2O_3-SiO_2, figura 18, onde foram projetadas todas as amostras em função de seu conteúdo nos três óxidos, recalculados a 100%. Como critério utilizado para subdividí-las em três grupos, levou-se em conta os teores originais em Fe_2O_3 e SiO_2. Assim, amostras com teor de Fe_2O_3 maior que 25% são designadas lateritas, teor de Fe_2O_3 menor que 25% mas com SiO_2 maior que 16,1% argilas e teor de Fe_2O_3 menor que 25% mas SiO_2 menor que 16,1% as bauxitas.

As amostras do grupo das bauxitas estão próximas ao vértice direito, contendo alto teor de Al_2O_3. As amostras argilosas situam-se entre os polos Al_2O_3 e SiO_2 do diagrama. As lateritas correspondem a pontos cujo aumento no teor de Fe_2O_3 se dá em função da diminuição do teor de Al_2O_3, em relação as bauxitas.

Análise exploratória dos dados

O conjunto de treinamento consiste em 73 amostras e das

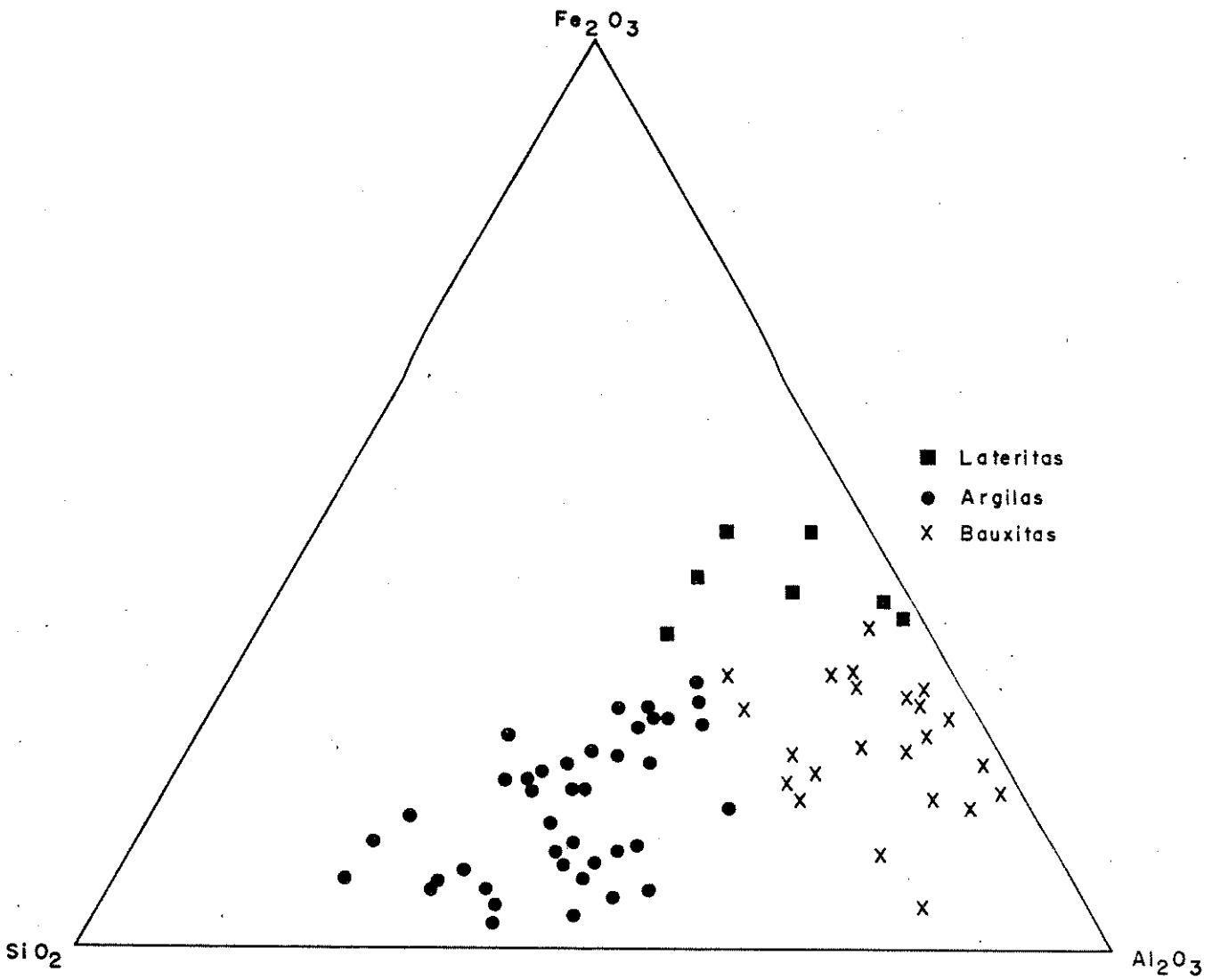


Figura 18 - Diagrama SiO_2-Al_2O_3-Fe_2O_3 para as amostras estudadas.

25 variáveis listadas na tabela 1. Todos os dados foram autoescalonados para obter variáveis com média zero e variância igual a um. Este procedimento tem a finalidade de atribuir a todas variáveis a mesma informação estatística, evitando distorções que possam ocorrer na análise dos dados devido a diferença de unidades de medidas, aqui percentagem e partes por milhão.

Este conjunto foi submetido a análise do tipo R, verificando-se qúe as 5 primeiras componentes explicam 76,7% da variância contida nos dados, ver tabela 1. Para tornar estas componentes mais fáceis para interpretar, foi então feita a rotação ortogonal varimax, na qual a quantidade de variância retida nestas componentes permanece inalterada, ver tabela 2.

Comparando as duas tabelas podemos ver, na tabela 1 que os elementos principais estão misturados, na terceira componente principal SiO_2 e Al_2O_3 e na quarta componente Al_2O_3 e Fe_2O_3, ou seja, há duas componentes principais com participação importante do Al_2O_3. A tabela 2 mostra a simplificação destas componentes após a rotação, tornando-as mais fáceis para interpretar. A primeira componente contém altos teores de Al_2O_3 e água estrutural, caracterizando a bauxita. A segunda componente representa altos teores em Fe_2O_3 e nos traços que o acompanham. Na terceira componente estão agrupados os elementos de transição, Mn, Zn, Ni e Cu, e a quarta componente está relacionada à associação dos alcalinos-terrosos Ba e Sr com os elementos de terras raras La, Nd e Ce. A quinta componente parece relacionada aos elementos de comportamento solúvel durante a alteração intempérica.

A figura 19 mostra o gráfico dos pesos das duas primeiras componentes principais rodadas, representando as variáveis no espaço bi-dimensional. Nesta figura aparecem claramente os

Tabela 1 - Pesos das variáveis nas cinco primeiras componentes principais (sem rotação).

	CP_1	CP_2	CP_3	CP_4	CP_5
Si	0,3216	-	-0,2367	-	-
Al	-	-	0,3288	-0,2548	-
Fe	-0,2619	-0,2473	-	0,2230	-
Mg	-	-0,2853	-	-	0,2106
Mn	-	-	0,3923	0,2531	-
Ca	0,2218	-	-0,3679	-	-
Na	-	-	-	0,3033	-0,3757
K	-	-	-0,2384	-	0,3844
Ti	-0,2428	-0,2875	-	-	-
P	-	-0,2359	-	-	-
H_2O	-0,2908	-	0,2334	-	-
Ba	-0,2314	-0,2616	-	-	-
Ce	-	-0,2214	0,3310	-	-0,2359
Cr	-	-	-	0,2742	-
Ga	-	0,2277	-	-	-0,3657
La	-	-0,3119	-	-0,3052	-0,2205
Nd	-	-0,3403	-	-0,3172	-
V	0,2579	-	-	0,2552	-
Zn	0,2893	-	0,2762	-	-
Zr	-	-	-	-	-0,3243
Cu	-	-	0,2189	0,2817	-
Ni	0,2893	-	0,2764	-	-
Nb	-	-0,3165	-	-	-
Sr	-	-0,2677	-	-0,3530	-
Y	-	-	-	-	0,3034
*	28,0%	21,8%	11,6%	9,2%	6,1%

*percentagem de variância explicada em cada componente.

Tabela 2 - Pesos das variáveis nas cinco primeiras componentes principais (com rotação).

	CP_1	CP_2	CP_3	CP_4	CP_5
Si	-	-	-	-	-
Al	-0,9026	-	-	-	-
Fe	-	-0,9530	-	-	-
Mg	-	-	-	-	-
Mn	-	-	0,9086	-	-
Ca	-	-	-	-	0,5375
Na	-	-	-	-	-
K	-	-	-	-	0,7723
Ti	-	-0,9248	-	-	-
P	-	-0,7035	-	-	-
H_2O	-0,8238	-	-	-	-
Ba	-	-	-	-0,5194	0,5689
Ce	-	-	0,6971	-0,5634	-
Cr	-	-	-	-	-
Ga	-	-	-	-	-
La	-	-	-	-0,8891	-
Nd	-	-	-	-0,9066	-
V	-	-0,8441	-	-	-
Zn	-	-	0,8360	-	-
Zr	-	-	-	-	-
Cu	-	-	0,7216	-	-
Ni	-	-	0,8757	-	-
Nb	-	-0,8242	-	-	-
Sr	-	-	-	-0,7916	-
Y	-	-	-	-	0,7503
*	13,8	20,9	16,0	14,7	11,4

* percentagem de variância explicada em cada componente.

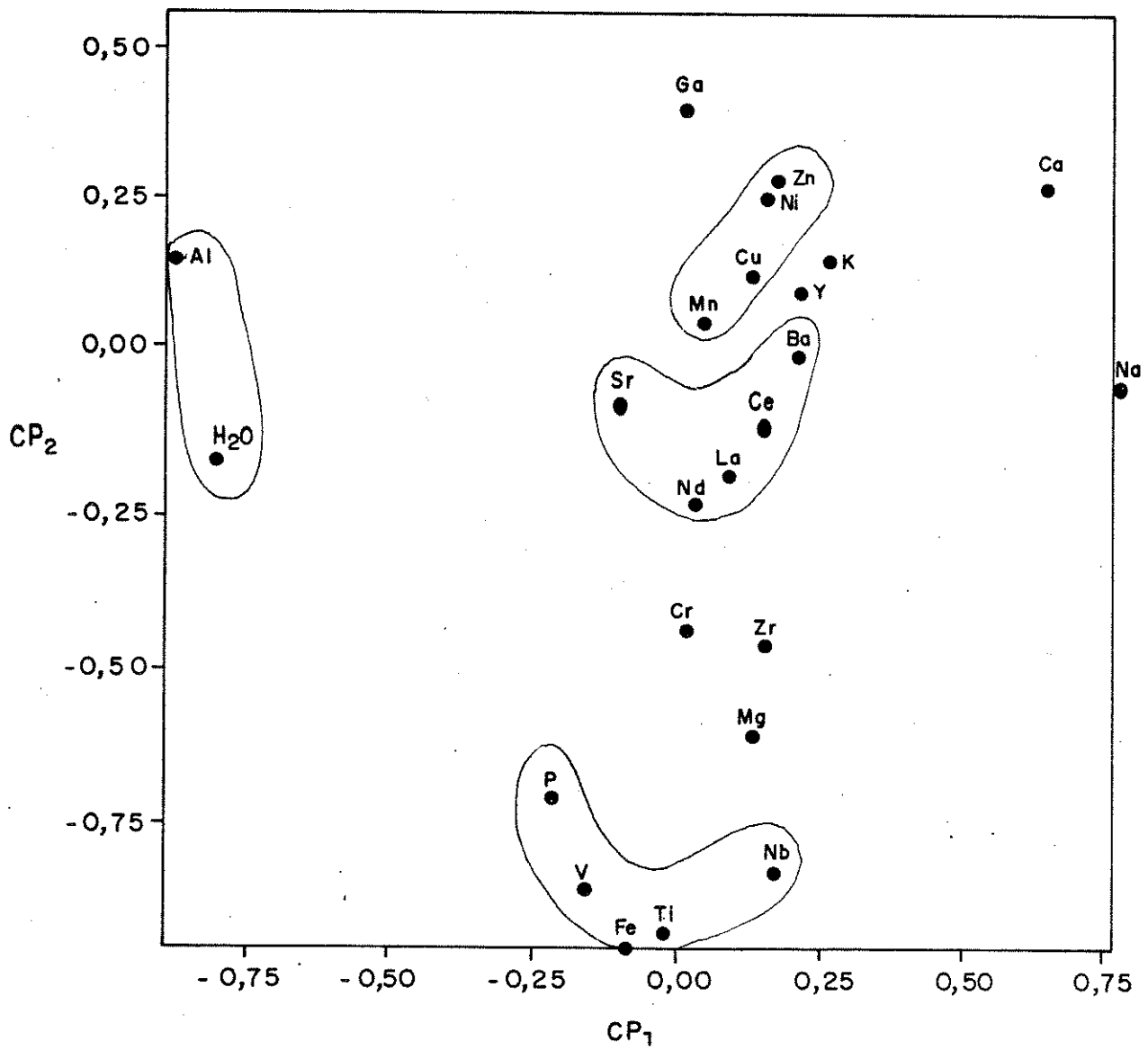


Figura 19 - Gráfico dos pesos das variáveis nas duas primeiras componentes principais (com rotação).

grupos de elementos que caracterizam as quatro primeiras componentes. O Mg, Na, K, Cr, Ga e Zr, Y, Ca são elementos que parecem estar fora das quatro principais associações.

Com a finalidade de relacionar os conjuntos destes elementos analisados pela rotação das componentes principais e os materiais definidos no diagrama triangular, foi feito então o gráfico dos escores das duas primeiras componentes principais, figura 20, ($CP_1 x CP_2$). A primeira componente principal discrimina as bauxitas e as argilas. Como os pesos do Al_2O_3 e da água estrutural são negativos, as amostras de bauxita situam-se na parte mais negativa da primeira componente principal e as amostras de argila na parte mais positiva. A segunda componente principal separa as lateritas da bauxita e argila. Mais uma vez os pesos nesta componente são negativos e por isto as lateritas estão situadas na parte mais negativa enquanto que as bauxitas e argilas estão situadas na parte mais positiva.

A figura 21 ($CP_2 x CP_3$) mostra que as lateritas são mais ricas em ferro (CP_2 mais negativo) que as bauxitas e argilas. Mostra também que a maior parte das amostras, tanto de laterita quanto de argila ou bauxita, contém baixos teores dos elementos de transição. As únicas amostras que se comportam diferentemente, indicando mais altos valores de CP_3, pertencem à classe das argilas.

Na figura 22 ($CP_1 x CP_4$), como já visto, a primeira componente principal separa bem as argilas das bauxitas, enquanto que, a quarta componente principal mostra que a dispersão dos teores de terras raras é maior nas bauxitas que nas argilas.

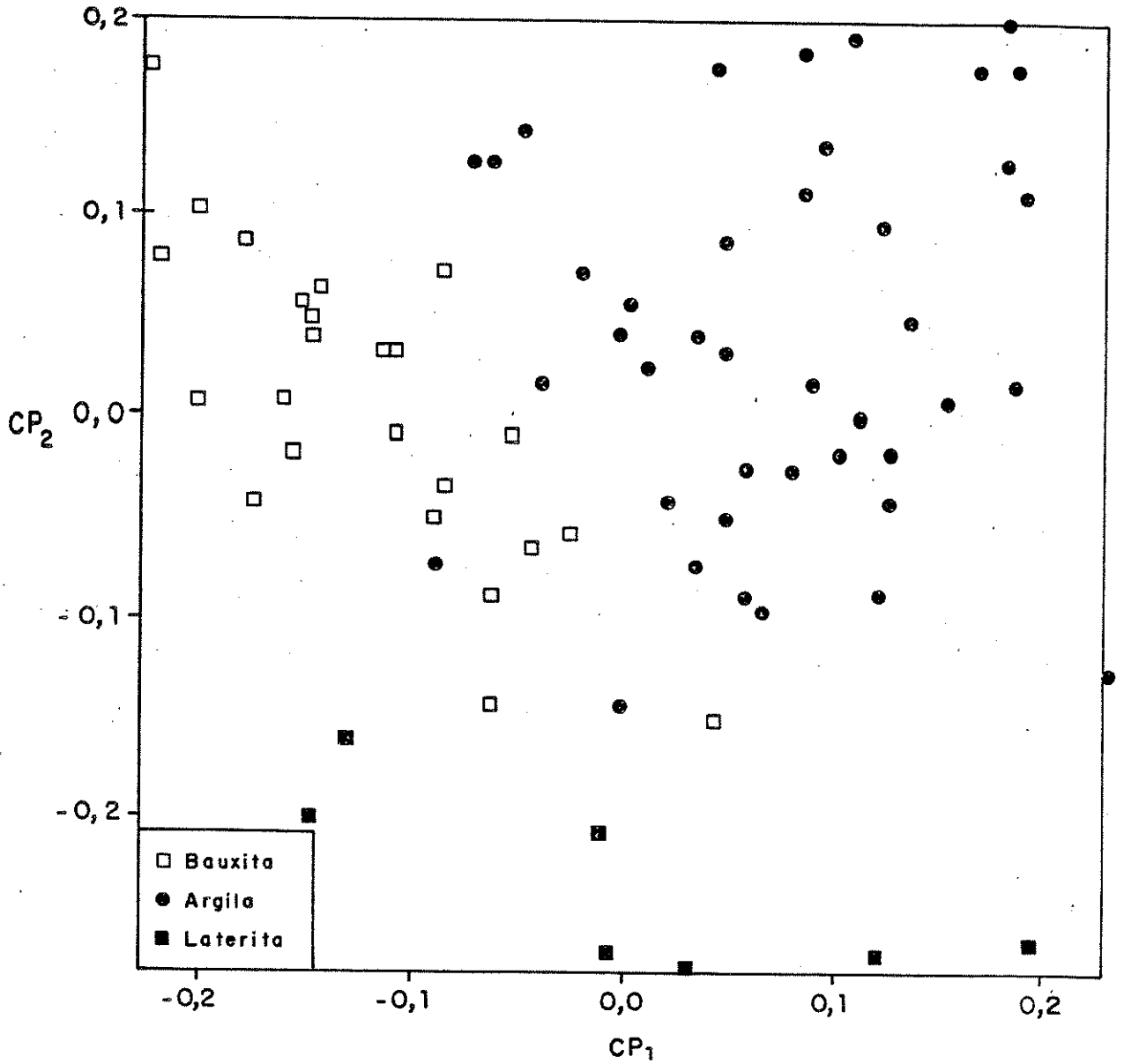


Figura 20 - Gráfico dos escores rodados para as componentes CP_1 e CP_2.

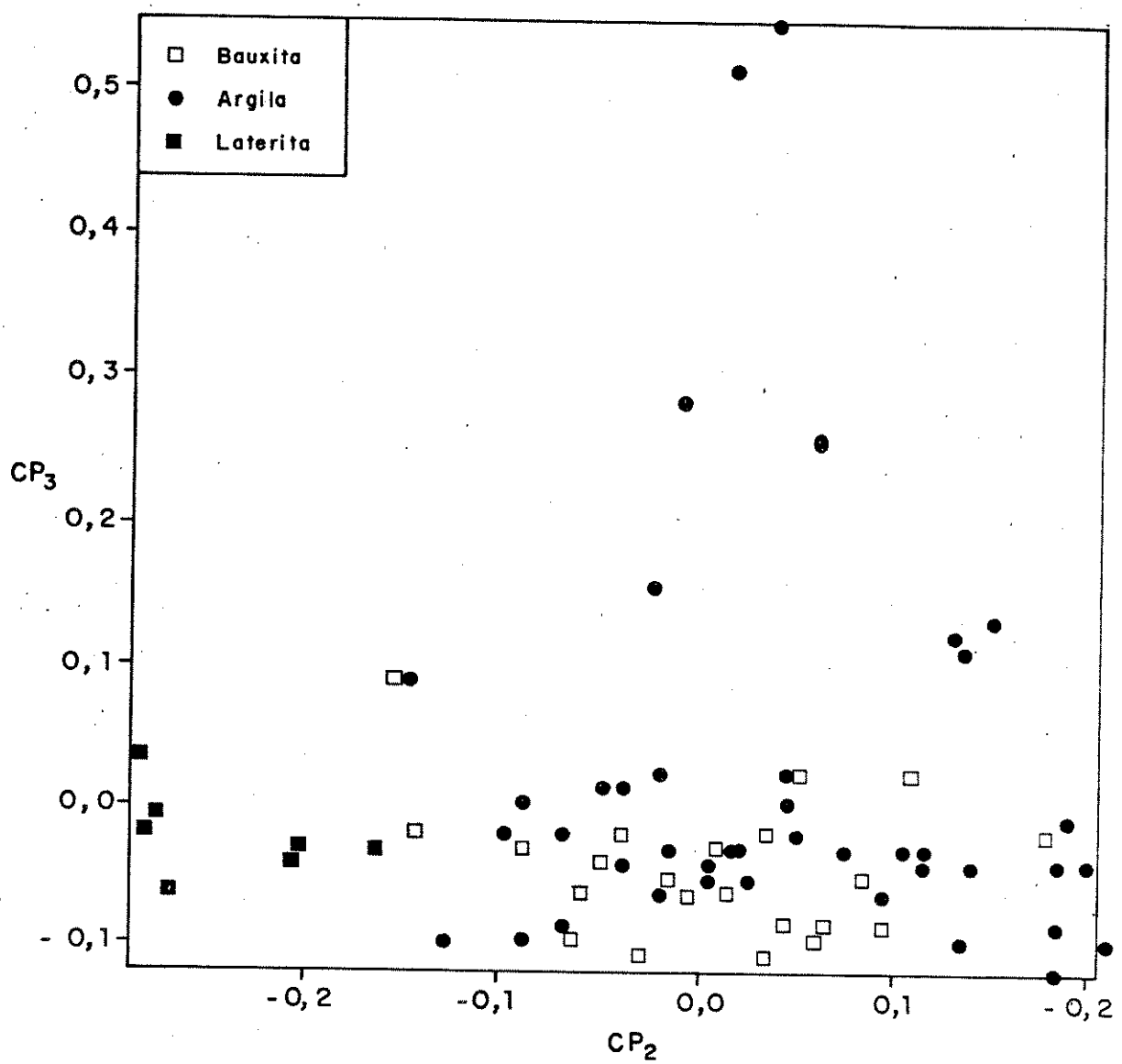


Figura 21 - Gráfico dos escores rodados para as componentes CP_2 e CP_3.

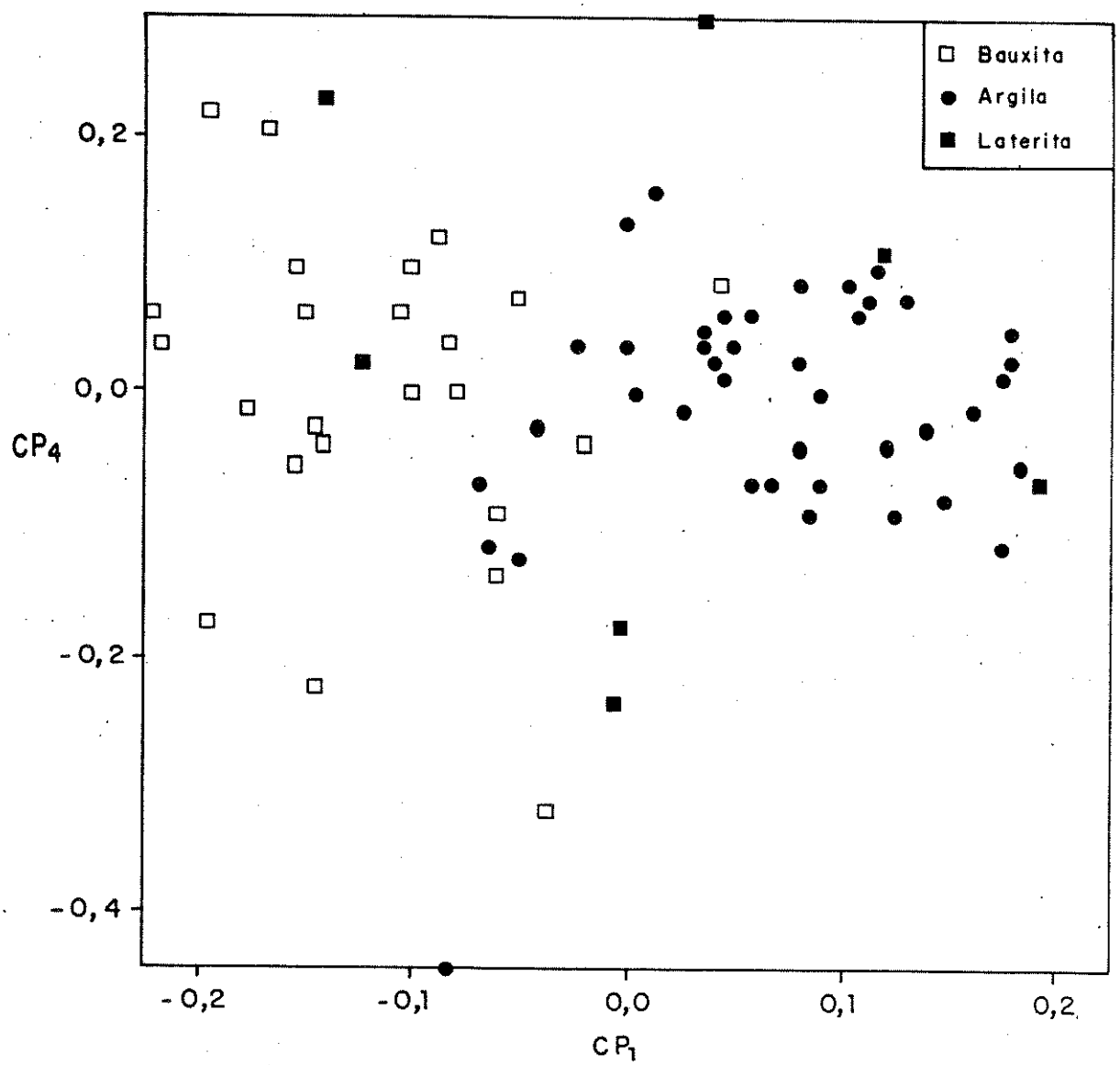


Figura 22 - Gráfico dos escores rodados para as componentes CP_1 e CP_4.

CONSIDERAÇÕES FINAIS

- A análise de componentes principais é capaz de reduzir o número de variáveis sem grande perda de informação. A rotação varimax tornou mais simples a interpretação das cinco componentes.

- A análise de componentes principais, juntamente com a rotação varimax, permitiu classificar corretamente as amostras em função de sua natureza bauxítica, argilosa ou ferruginosa. Também permitiu distinguir conjuntos de elementos de comportamento semelhantes.

- No caso estudado foi possível perceber que as bauxitas são ligeiramente mais ricas em Fe, V, Ti, P e Nb. Por outro lado, os elementos de transição Mn, Cu, Ni e Zn encontram-se mais concentrados nas argilas que nas bauxitas. Estas conclusões não eram evidentes numa simples inspeção das tabelas de análise química.

Segunda Aplicação

Colaboradores:

José Roberto Ferreira - Secretaria da Agricultura (Instituto de Pesca), São Paulo, S.P.

L.A. Martinelli e Jefferson Mortatti, Universidade de São Paulo (CENA), Piracicaba, S.P.

O conjunto de dados refere-se à concentração de espécies químicas dissolvidas presentes nas bacias dos rios Ji-Paraná, Jamari e Madeira no Estado de Rondônia.

DESCRIÇÃO EXPERIMENTAL

Os pontos de amostragem foram selecionados em função de sua distribuição, com vistas a uma melhor representatividade das bacias existentes, considerando-se concomitantemente as facilidades de acesso.

Desta forma, foram definidas as bacias dos rios Ji-Paraná, composta pelos rios Barão de Melgaço ou Comemoração, Pimenta Bueno, Machado ou Ji-Paraná e rio Jarú; bacia do rio Jamari, composta pelos rios Jamari e Candeias; bacia do rio Madeira, composta pelos rios Guaporé, Mamoré e Madeira, (figura 23). Os rios Ji-Paraná e Jamari são importantes afluentes da margem direita do rio Madeira.

As amostragens são referentes aos anos de 1983 (Abril-Maio; Julho-Agosto), 1984 (Janeiro-Fevereiro; Outubro-Novembro) e

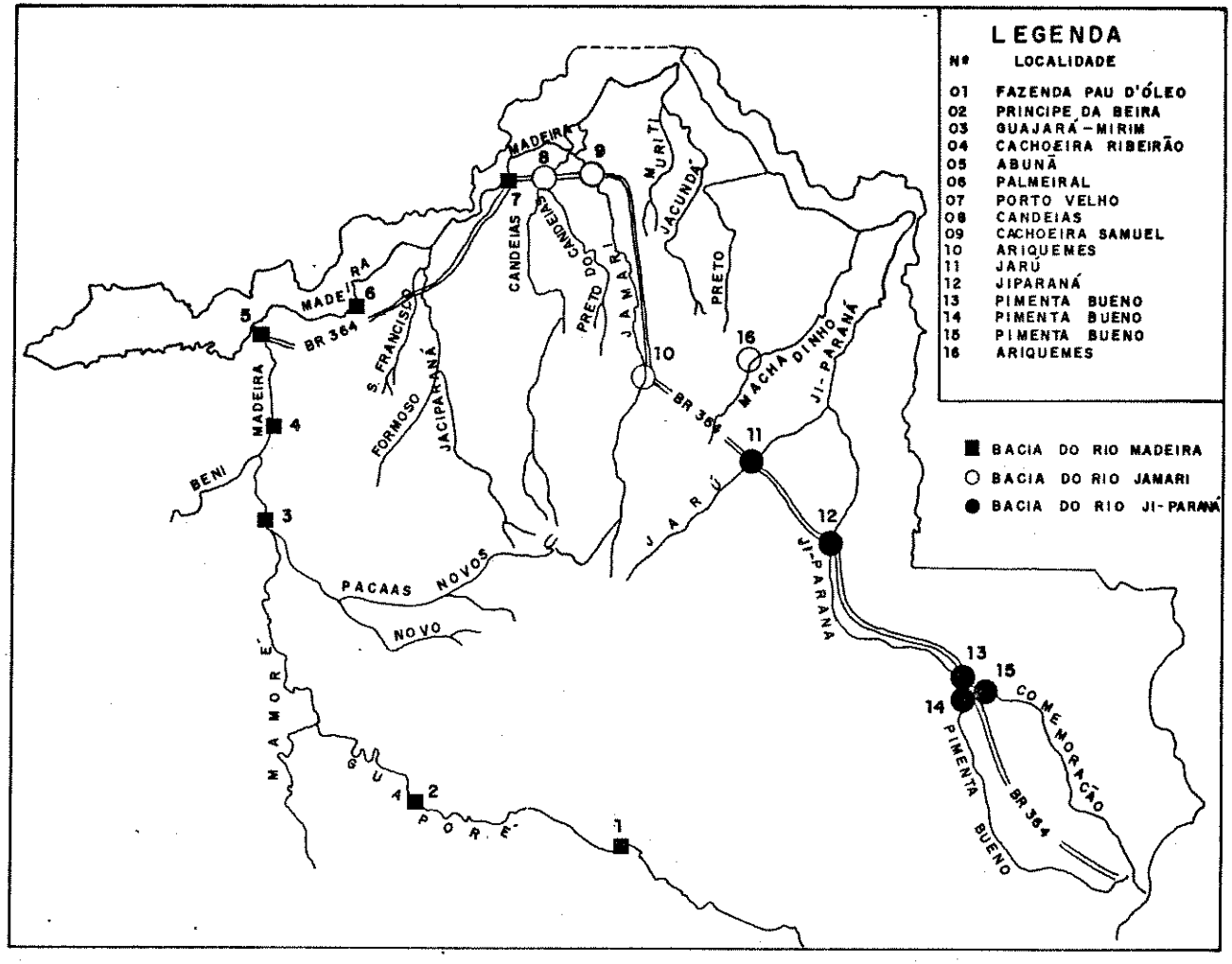


Figura 23 - Pontos de coleta de amostras de água distribuídos pelas bacias dos rios Ji-Paraná, Jamari e Madeira.

1985 (Agosto-Setembro), abrangendo período de chuva e de seca.

A coleta foi realizada na parte central dos rios, em frente às réguas limnimétricas. Para tanto, fazia-se emborcar garrafas de Nisken a ± 30 cm de profundidade e distribuía-se seu conteúdo em frascos de polietileno de 500 ml, para se proceder em terra à filtragem a vácuo em filtro milipore 0,45 µm. A preservação adequada ao tipo de análise pretendida foi feita adicionando-se à amostra 1 ml HNO_3conc./l ou 1 ml H_2SO_4conc./l.

No Centro de Energia Nuclear na Agricultura, CENA-USP, Piracicaba, foram conduzidas as determinações de espécies químicas inorgânicas, as quais seguiram procedimentos estabelecidos (45). Os metais e a sílica foram determinadas em espectrômetro de emissão com plasma induzido em argônio - Jarrel Ash, modelo 975. As demais espécies químicas que constam deste trabalho foram determinadas através de métodos espectrofotométricos de análise pelo acoplamento de sistemas de injeção em fluxo (46) à espectrofotômetro Varian, modelo 634-S equipado com cubeta Hellma 178-OS de 80 µl. A propulsão dos fluídos foi obtida pelo emprego de uma bomba peristáltica Ismatec mp 13GJ4 e o sinal transiente medido, proporcional à concentração, foi registrado num aparelho REC 61 da Radiometer.

Finalidades

1. Comparar as concentrações de espécies químicas inorgânicas dissolvidas encontradas em três bacias hidrográficas pertencentes ao Estado de Rondônia.

2. Comparar através dos fluxos as quantidades de nutrien

tes (Ca, Mg, K e Na) perdidas nas três bacias.

Análise Exploratória dos dados

O conjunto de treinamento é composto de 35 amostras distribuídas em três classes (bacias). As variáveis consistem nos teores dissolvidos de Ca, Mg, Na, K, Si, Al, SO_4, Cl, NH_4, NO_3 e Fe, em mg/l. A partir das vazões estimaram-se os fluxos diários de nutrientes, e posteriormente a descarga específica ($Kg.d/km^2$), representada por (Ca)p, (Na)p, (K)p, (Mg)p, respectivamente, para cada elemento. Além destas estudou-se o somatório dos teores dos íons Ca, Mg, Na e K que, em solos, representa as bases trocáveis (T) e a descarga específica de T, (T)p.

Como no caso anterior, as variáveis foram autoescalonadas para evitar distorções na análise dos dados devidas à diferença de unidades existente entre as variáveis, aqui mg/l e $Kg.d/Km^2$.

A tabela 3 mostra as mais altas correlações entre alguns pares de variáveis, compostos principalmente pelas bases trocáveis do solo, Ca, Mg, K e Na, indicando haver uma contribuição em entrada de nutrientes deste substrato ao meio hídrico. Estiveram também altamente correlacionados os pares Fe-Al e Cl-SO_4. A alta correlação Fe-Al reforça a hipótese da existência de interação água $\rightleftarrows$ solo, mostrando que a área estudada é formada por solos constituídos de minerais altamente intemperizados, compostos de hidróxidos de Fe e Al (47). O Cl e o SO_4 são elementos que compõe minerais como os evaporitos, que estão presentes na região de estudo.

Tabela 3 - Principais correlações entre as variáveis.

VARIÁVEIS	CORRELAÇÃO	VARIÁVEIS	CORRELAÇÃO	VARIÁVEIS	CORRELAÇÃO
(Ca)p-(Mg)p	0,972	Ca-Na	0,860	SO_4-T	0,779
Ca-T	0,962	T-Mg	0,859	(Mg)p-(Na)p	0,778
(Mg)p-(T)p	0,951	K-(Na)p	0,857	Al-(Ca)p	0,773
(Ca)p-(T)p	0,951	SO_4-Ca	0,815	Cl-SO_4	0,771
Al-Fe	0,944	Ca-Mg	0,808	(Ca)p-(Na)p	0,756
(Na)p-(T)p	0,908	(K)p-(T)p	0,803	SO_4-Na	0,754
				Al-(Mg)p	0,739

A análise de correlação do conjunto total de dados revela similaridades no comportamento dos pares de variáveis mas não identifica convenientemente grupos de elementos com comportamento similares. A análise de componentes principais do tipo R foi então utilizada para estudar as interrelações entre variáveis e também o comportamento geral dos dados. A tabela 4 mostra os valores dos pesos das variáveis nas 5 primeiras componentes principais, que explicam 88,8% da informação contida nos dados. Os pesos com valores absolutos menores que 0,20 foram omitidos dessa tabela.

Pode-se observar na tabela 4 que as duas primeiras componentes principais, que explicam sozinhas 63,2% da variância total, têm contribuições significativas das variáveis que apresentaram as mais altas correlações (indicadas na tabela 3).

Para obter a configuração em duas dimensões dos pontos (referentes às amostras) do espaço de dimensão 17, foi feito o

Tabela 4 - Pesos das variáveis nas cinco primeiras componentes principais (sem rotação).

VARIÁVEIS	CP_1	CP_2	CP_3	CP_4	CP_5
Si	-	-	0,36	-0,60	-
Al	-0,28	-	0,21	-0,34	-0,22
Fe	-0,27	-	0,33	-0,25	-0,28
NH_4	-	-	-0,38	-	-0,45
Cl	-	0,39	-0,32	-	-
SO_4	-0,23	0,32	-0,22	-	-
NO_3	-	-	0,48	-	0,23
Ca	-0,32	0,25	-	-	-
(Ca)p	-0,33	-0,22	-	-	-
Mg	-0,25	0,27	-	-	0,24
(Mg)p	-0,33	-0,23	-	-	-
K	-	-	0,23	0,42	-0,65
(K)p	-	-0,32	-0,21	0,30	-
Na	-0,24	0,33	-	0,21	-
(Na)p	-0,25	-0,26	-0,23	0,25	0,21
T	-0,30	0,30	-	-	-
(T)p	-0,31	-0,27	-	-	-
*	40,1	23,1	11,4	8,9	5,3

* percentagem de variância explicada em cada componente.

gráfico dos escores para as duas primeiras componentes principais, figura 24. Este gráfico mostra que as composições químicas das amostras da bacia do rio Madeira podem ser discriminadas das composições químicas das demais bacias. Esta discriminação pode ser explicada pelo fato de que os rios que compõe esta bacia, com execeção do rio Guaporé, são mais ricos em nutrientes dissolvidos porque a sua nascente origina-se nos Andes (48). O Guaporé, situado a montante da confluência do rio Beni, é oriundo da Bolívia e sua águas possuem composição química semelhante à das outras duas bacias, ver figura 23.

A figura 24 indica também que as amostras do rio Jarú, que pertence a bacia do rio Ji-Paraná, podem ser discriminadas daquelas dos demais rios que fazem parte desta bacia. Isto deve-se muito ao fato de ser nesta sub-bacia que se encontram os grupos de solos mais férteis do Estado de Rondônia. Consequentemente, a procura por estes solos faz com que sejam verificadas as mais altas taxas de transformação de ecossistemas naturais em ecossistemas agro-pastoris.

A figura mostra ainda que há um aglomerado de pontos para valores positivos em CP_1 e para CP_2 aproximadamente zero.

A separação das amostras da figura 24 nas principais classes, pode ser feita com retas diagonais, o que significa que as duas componentes principais são importantes para que ocorra esta discriminação.

Para tornar as componentes principais mais simples e portanto mais fáceis de ser interpretadas foi feita a rotação ortogonal varimax das cinco primeiras componentes, que explicam 88,8% da variância total dos dados. A tabela 5 mostra os pesos das variáveis nestas componentes. Comparando as tabelas 4 e 5 pode-se

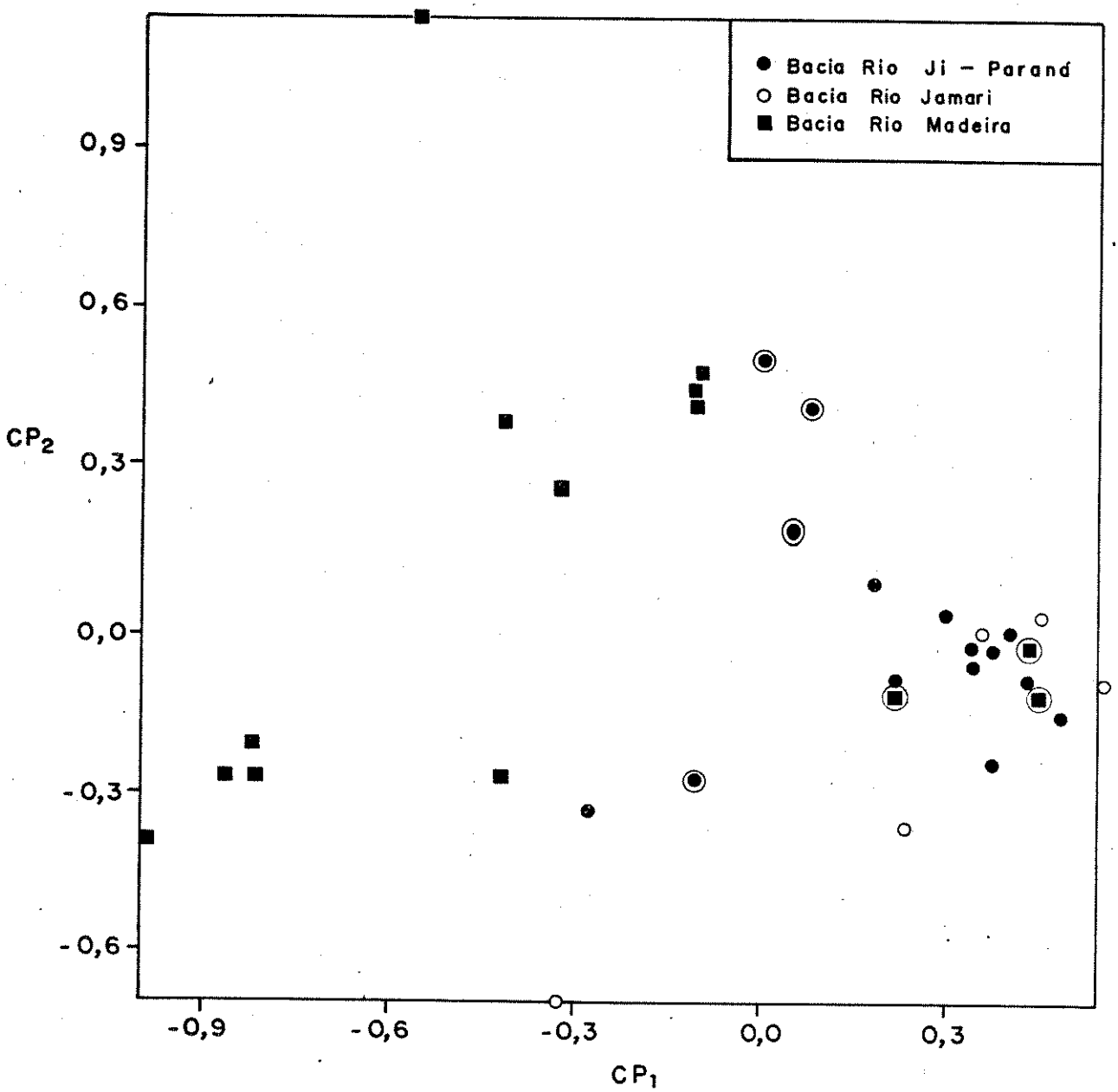


Figura 24 - Gráfico dos escores dos componentes CP_1 e CP_2 (não rodados). As amostras com os símbolos ◉ e ◉ pertencem aos rios Jarú e Guaporé, respectivamente.

Tabela 5 - Pesos das variáveis nas cinco primeiras componentes principais (com rotação).

VARIÁVEIS	CP_1	CP_2	CP_3	CP_4	CP_5
Si	-	-	-	0,58	-0,66
Al	-	-	0,89	-	-
Fe	-	-	0,92	-	-
NH_4	-	-	-	-0,67	-
Cl	-	0,77	-	-	-
SO_4	-	0,89	-	-	-
NO_3	-	-	0,66	-	-
Ca	-	0,89	-	-	-
(Ca)p	-0,76	-	0,54	-	-
Mg	-	0,83	-	-	-
(Mg)p	-0,78	-	0,53	-	-
K	-	-	-	-	-0,89
(K)p	-0,90	-	-	-	-
Na	-	0,93	-	-	-
(Na)p	-0,96	-	-	-	-
T	-	0,92	-	-	-
(T)p	-0,91	-	-	-	-
*	25,0	28,7	18,6	8,2	8,3

* percentagem de variância explicada em cada componente.

ver que na tabela 4 os pesos em CP_1 são todos negativos, enquanto que em CP_2 Cl, SO_4, Ca, Mg, Na, K e T são positivos e as descargas específicas são negativas. Na tabela 5 pode-se ver claramente a simplificação destas componentes, evidenciando-se a separação das variáveis. A primeira componente principal rodada, com aproximadamente 25% de informação estatística, ficou composta pelas descargas específicas de (Ca)p, (Mg)p, (K)p, (Na)p e (T)p, enquanto que a segunda componente rodada, que explica em torno de 28%, ficou composta por Cl, SO_4, Ca, Mg, Na e T.

A figura 25 mostra o gráfico dos escores destas componentes, as quais explicam sozinhas 53% da informação estatística contida nos dados. A primeira componente indica uma variação na concentração de nutrientes ao longo dos anos em que foram procedidas as amostragens nas três bacias consideradas. Estas alternâncias são relativas às hidrógrafas dos rios, uma vez que no intervalo compreendido entre -0,57 a -0,15 estão dispostas amostras coletadas em abril-maio (cheia-descendente) de 1983. Não estão disponíveis amostragens realizadas no mesmo período para os anos 84 e 85. Para o mesmo ano de 83, em períodos de seca (Julho-Agosto), observa-se que nas três bacias foram verificadas menores concentrações das variáveis que compõem esta componente (CP_1). No período de subida das águas (agosto-novembro) foram observadas concentrações similares, mesmo em anos diferentes (84 e 85). Assim, pelos dados disponíveis percebe-se que a variabilidade observada é devida mais a um fator de sazonalidade climática (períodos chuvosos) do que propriamente a alterações ocorridas nos ecossistemas que circundam essas bacias. Contudo, os dados indicam que as maiores amplitudes nas variações foram inversamente proporcionais ao tamanho (área) das bacias, as quais possuem diferen

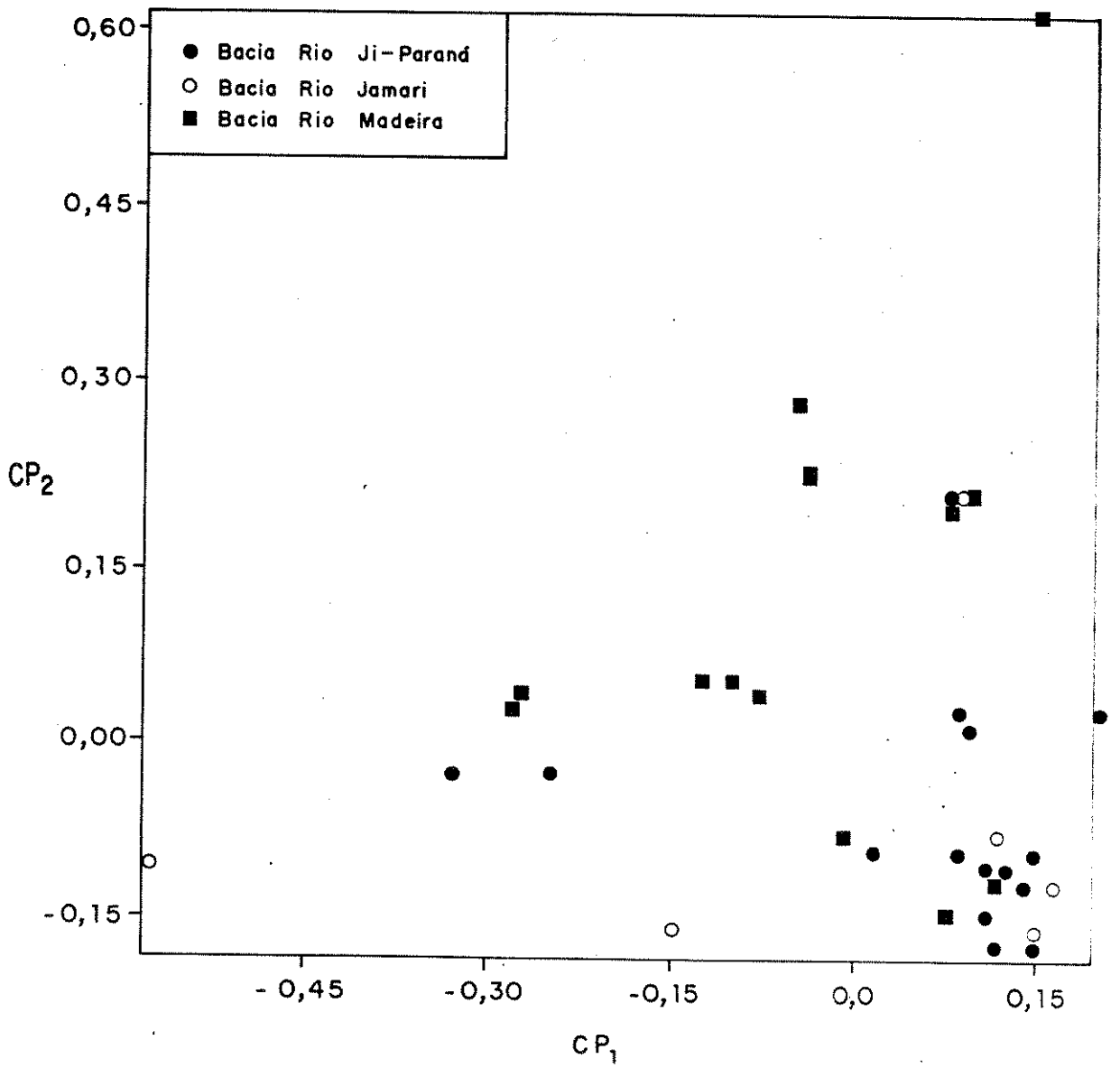


Figura 25 - Gráfico dos escores dos componentes CP_1 e CP_2 após a rotação.

tes "capacidade tampão". Estas variações temporais são mais facilmente visualizadas usando-se as componentes rodadas pelo método do varimax do que as componentes principais não rodadas.

A figura 25 mostra também que as variáveis que compõe a segunda componente são responsáveis pelas discriminações citadas anteriormente.

O método SIMCA, que gera um modelo para cada categoria, é capaz de classificar corretamente 80% das amostras de acordo com a bacia de origem, ver tabela 6. Para determinar o número ótimo de componentes principais para a modelagem de cada classe a técnica de Cross-Validation foi usada. É interessante notar que três das amostras classificadas incorretamente na bacia do rio Ji-Paraná pertencem ao rio Jarú, mostrando mais uma vez a dissimilaridade entre a composição desse rio e as demais pertencentes a essa bacia. O mesmo fato se verifica para a bacia do rio Madeira, onde as amostras classficadas incorretamente pertencem ao rio Guaporé.

Tabela 6 - Resultado do método SIMCA para classificação das amostras.

classes (bacias)	número de amostras	número de amostras classificadas incorretamente
Ji-Paraná	16	4
Jamari	5	0
Madeira	14	3

Os índices de Fisher e de variância indicaram que SO_4, Ca, Mg, Na e T são as variáveis que mais contribuíram para que es

ta diferenciação fosse verificada, confirmando o resultado apresentado pela rotação das componentes principais.

CONSIDERAÇÕES FINAIS

- Das três Bacias estudadas, as maiores diferenças nas concentrações das espécies químicas dissolvidas, especialmente Ca, Mg, Fe, SO_4 e Al, foram observadas na bacia do rio Madeira, enquanto que nas bacias dos rios Ji-Paraná (à exceção do rio Jarú) e Jamari as concentrações são semelhantes.

- As variações nas concentrações das bases trocáveis foram maiores do que as estimativas das descargas específicas destas mesmas variáveis.

- Na bacia do rio Ji-Paraná as maiores concentrações de íons dissolvidos foram verificadas no rio Jarú, onde são encontrados os solos mais férteis do Estado.

- O método de rotação ortogonal varimax permitiu a observação da dependência sazonal das concentrações das amostras.

Terceira Aplicação

Colaborador:

Prof.Dr. Benício B. Neto - Departamento de Química Fundamental - Universidade Federal de Pernambuco, Recife, PE.

O conjunto de dados consiste de 157 tensores polares atômicos para 50 moléculas (49), figura 26. Para permitir a comparação de valores oriundos de diferentes moléculas, os dados foram pré-processados, transformando-se os elementos dos tensores em parâmetros invariantes em relação à escolha de sistemas de coordenadas cartesianas.

FUNDAMENTO TEÓRICO

A intensidade integrada absoluta de uma transição no infravermelho é determinada experimentalmente usando-se a lei de Beer (50).

$$A_i = \frac{1}{c\ell} \int_{banda} \ln(Io/I)d\nu$$

onde c é a concentração molecular, ℓ o comprimento do caminho óptico, Io a intensidade da radiação incidente na amostra na frequência ν, I a intensidade transmitida. O intervalo de integração é a região espectral de interesse, normalmente a banda que está sendo estudada.

As intensidades vibracionais absolutas de bandas fundamen

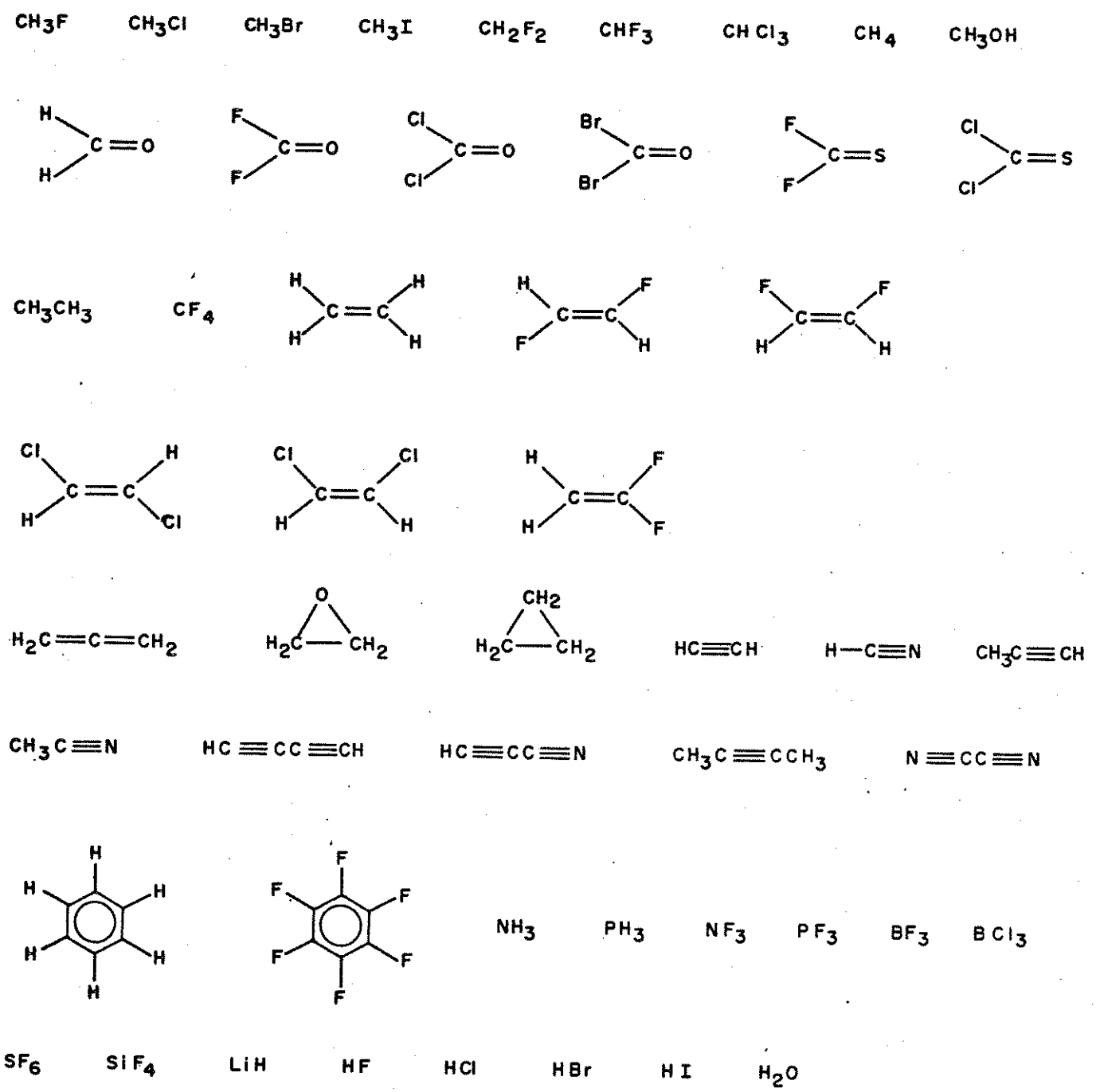


Figura 26 - Moléculas para os tensores polares considerados neste trabalho.

tais moleculares podem ser reduzidas a parâmetros moleculares através de

$$A_i = \frac{N\pi}{3c^2} \left(\frac{\partial \vec{p}}{\partial Q_i}\right)^2 \quad i = 1,2,\ldots,3N-6$$

onde Ai representa a intensidade da i-ésima banda fundamental e $\frac{\partial \vec{p}}{\partial Q_i}$ é a derivada do vetor momento dipolar em relação à i-ésima coordenada normal. O tensor polar molecular (51) em termos de coordenadas normais pode ser expresso pela matriz 3(3N-6)

$$\underset{\sim}{P}_Q = (\underset{\sim}{P}_{Q_1} : P_{Q_2} : \ldots\ldots : \underset{\sim}{P}_{Q_{3N-6}}),$$

onde cada linha corresponde à coordenada cartesiana X, Y ou Z da molécula e cada coluna a uma coordenada normal. Os tensores polares atômicos são calculados a partir de $\underset{\sim}{P}_Q$ pela equação

$$\underset{\sim}{P}_x = \underset{\sim}{P}_Q \underset{\sim}{L}^{-1} \underset{\sim}{U} \underset{\sim}{B} + \underset{\sim}{P}_\rho \underset{\sim}{\beta}$$

onde as matrizes $\underset{\sim}{L}$, $\underset{\sim}{U}$ e $\underset{\sim}{B}$ são comumente empregadas em análises de coordenadas normais e $\underset{\sim}{P}_\rho$ e $\underset{\sim}{\beta}$ são usadas para calcular a contribuição rotacional ao tensor polar. Esta equação transforma as variações do momento dipolar com deslocamentos ao longo de coordenadas normais em variações relativas a deslocamentos ao longo de coordenadas cartesianas atômicas. O tensor polar molecular, $\underset{\sim}{P}_x$, é dado por:

$$\underset{\sim}{P}_x = (\underset{\sim}{P}_x^{(1)} : \underset{\sim}{P}_x^{(2)} : \ldots\ldots \underset{\sim}{P}_x^{(\alpha)} \ldots\ldots \underset{\sim}{P}_x^{(N)})$$

para uma molécula com N átomos, sendo $\underset{\sim}{P}_x^{(\alpha)}$ o tensor polar do

α-ésimo átomo, isto é,

$$\underset{\sim}{P}_x^{(\alpha)} = \begin{bmatrix} \partial P_x/\partial_{x\alpha} & \partial P_x/\partial_{y\alpha} & \partial P_x/\partial_{z\alpha} \\ \partial P_y/\partial_{x\alpha} & \partial P_y/\partial_{y\alpha} & \partial P_y/\partial z_\alpha \\ \partial P_z/\partial_{x\alpha} & \partial P_z/\partial_{y\alpha} & \partial P_z/\partial_{z\alpha} \end{bmatrix}$$

Aqui $\partial P_\sigma/\partial_{\nu\alpha}$ são as derivadas das componentes cartesianas do momento dipolar em relação às coordenadas cartesianas atômicas de deslocamento ($P_{\sigma\nu} = \partial P_\sigma/\partial_{\nu\alpha}$ onde $\sigma,\nu = x, y, z$).

Para os tensores polares atômicos aqui estudados é sempre possível encontrar uma orientação do sistema de eixos na qual somente cinco dos nove elementos tenham valores diferentes de zero, e o tensor tenha a forma

$$\underset{\sim}{P}_x^{(\alpha)} = \begin{bmatrix} P_{xx} & o & P_{xz} \\ o & P_{yy} & o \\ P_{zx} & o & P_{zz} \end{bmatrix}$$

Como os elementos do tensor polar dependem da orientação da molécula relativa ao sistema de coordenadas cartesianas, prefere-se usar grandezas invariantes com a rotação da molécula. As invariantes usadas aqui são:

1. Derivada dipolar média, $\overline{\underset{\sim}{P}}_\alpha$,

$$\overline{\underset{\sim}{P}}_\alpha = \frac{1}{3}\,(P_{xx} + P_{yy} + P_{zz})\ .$$

2. Carga atômica efetiva, χ_α, definida por

$\chi_\alpha^2 = \frac{1}{3}$ Traço $[\underset{\sim}{P}_x^\alpha (\underset{\sim}{P}_x^\alpha)']$ onde $(\underset{\sim}{P}_x^\alpha)'$ é a transposta da tensor polar atômico para o átomo α.

3) Anisotropia, β_α^2,

$$\beta_\alpha^2 = \frac{1}{2}[(P_{xx} - P_{yy})^2 + (P_{yy} - P_{zz})^2 + (P_{zz} - P_{xx})^2] + 3(P_{xy}^2 + P_{yz}^2 + P_{xz}^2 + P_{zx}^2 + P_{yx}^2 + P_{zy}^2) .$$

4) A soma dos cofatores, C,

$$C = P_{yy}P_{zz} + P_{xx}P_{zz} + P_{xx}P_{yy} - P_{zx}P_{xz} .$$

5) O determinante do tensor polar, D,

$$D = \det \underset{\sim}{P}_x^{(\alpha)} .$$

ANÁLISE EXPLORATÓRIA

As variáveis para o conjunto dos 157 tensores polares atômicos das 50 moléculas da figura 26 são as cinco grandezas invariantes definidas acima. No lugar de empregar diretamente os valores de β^2, C e D usaremos β, $\sqrt{C}$ e $\sqrt[3]{D}$, para dar a todas as variáveis a mesma unidade, a saber, carga. Para β, χ e $\sqrt{C}$, que são definidas por equações quadráticas, serão usados os valores absolutos.

Neste estudo selecionamos três subconjuntos dos tensores polares atômicos, os tensores polares de hidrogênios, carbonos e heteroátomos.

HIDROGÊNIOS

Finalidade

Investigar se os valores dos tensores dos hidrogênios contêm informação suficiente para permitir a classificação dos carbonos ligados a eles como tendo hibridização sp^3, sp^2 ou sp.

Conjunto de treinamento - Trinta e dois tensores polares de átomos de hidrogênio foram escolhidos entre os tensores das moléculas apresentadas na figura 26. Somente foram incluídos os tensores de átomos de hidrogênio cujas ligações químicas com carbonos envolvem orbitais híbridos sp^3, sp^2 ou sp, ver figura 27.

Conjunto de teste - Oito tensores polares de átomos de hidrogênio, (ver figura 27) foram usados para testar a capacidade das regras de classificação desenvolvidas no conjunto de treinamento. Para alguns átomos de hidrogênio foram incluídos mais de um tensor, devido a diferentes intensidades experimentais ou a alguma incerteza na redução dos dados originais a tensores polares.

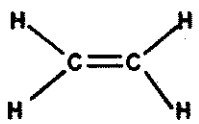
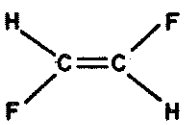
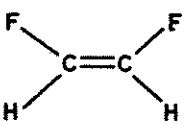
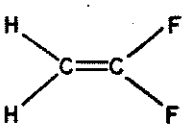
A análise de componentes principais mostra que as duas primeiras componentes principais explicam 93,2% da variância estatística. A primeira componente principal explica sozinha 82,1% da variância total e tem contribuições quase iguais de todas as variáveis. A segunda componente principal, que explica 11,1% de variância, tem maior contribuição de β e χ, com menores contribuições das outras variáveis, ver tabela 7.

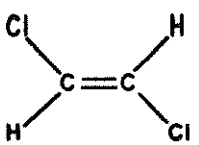
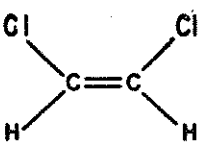
A figura 28 mostra o gráfico dos escores das duas primeiras componentes principais para os tensores polares atômicos deste subconjunto. Este gráfico mostra que a primeira componen

CONJUNTO DE TREINAMENTO

Sp^3 CH_3F CH_3Cl CH_3Br CH_3I CH_4 CH_3CH_3 CH_3CN

$CH_3C{\equiv}CCH_3$ $CHCl_3$ CHF_3 $CH_3C{\equiv}CH$

Sp^2  $H_2C{=}C{=}CH_2$   

Sp $CH_3C{\equiv}CH$ $HC{\equiv}CH$ $H{-}C{\equiv}N$ $HC{\equiv}CC{\equiv}CH$ $HC{\equiv}C{-}CN$

CONJUNTO DE TESTE

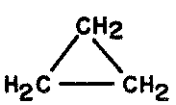
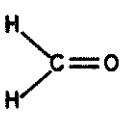
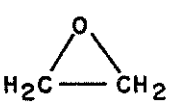
CH_3OH  CH_2F_2  

Figura 27 - Moléculas que formam os conjuntos de treinamento e de teste para os tensores polares dos átomos de hidrogênio.

Tabela 7 - Pesos das variáveis nas duas primeiras componentes principais.

VARIÁVEIS	CP_1	CP_2
$\overline{P}$	-0,47	-0,24
χ	-0,45	-0,55
β	0,40	-0,72
$\sqrt[3]{D}$	-0,43	0,32
$\sqrt{C}$	-0,47	-0,14
*	82,1%	11,1%

* percentagem de variância explicada em cada componente

te pode discriminar os hidrogênios ligados à carbonos sp dos outros hidrogênios orgânicos. As classes de hidrogênios ligados a carbonos sp^2 e sp^3 estão consideravelmente superpostas ao longo desta componente.

A segunda componente principal, que é dominada pela anisotropia, discrimina os ambientes com hibridização sp^2 dos totalmente saturados. A presença do grupo H-C(sp) perto da classe saturada e mais longe dos hidrogênios ligados a carbonos sp^2 é uma consequência da alta simetria destas moléculas. Embora os elétrons π nas ligações triplas CC sejam muito polarizáveis, a alta ordem do eixo de simetria impede a existência de elementos não nulos fora da diagonal e força dois dos elementos da diagonal a serem iguais nos tensores polares atômicos dos hidrogênios ácidos. Ambos os efeitos certamente tendem a abaixar a anisotropia do grupo H-C(sp).

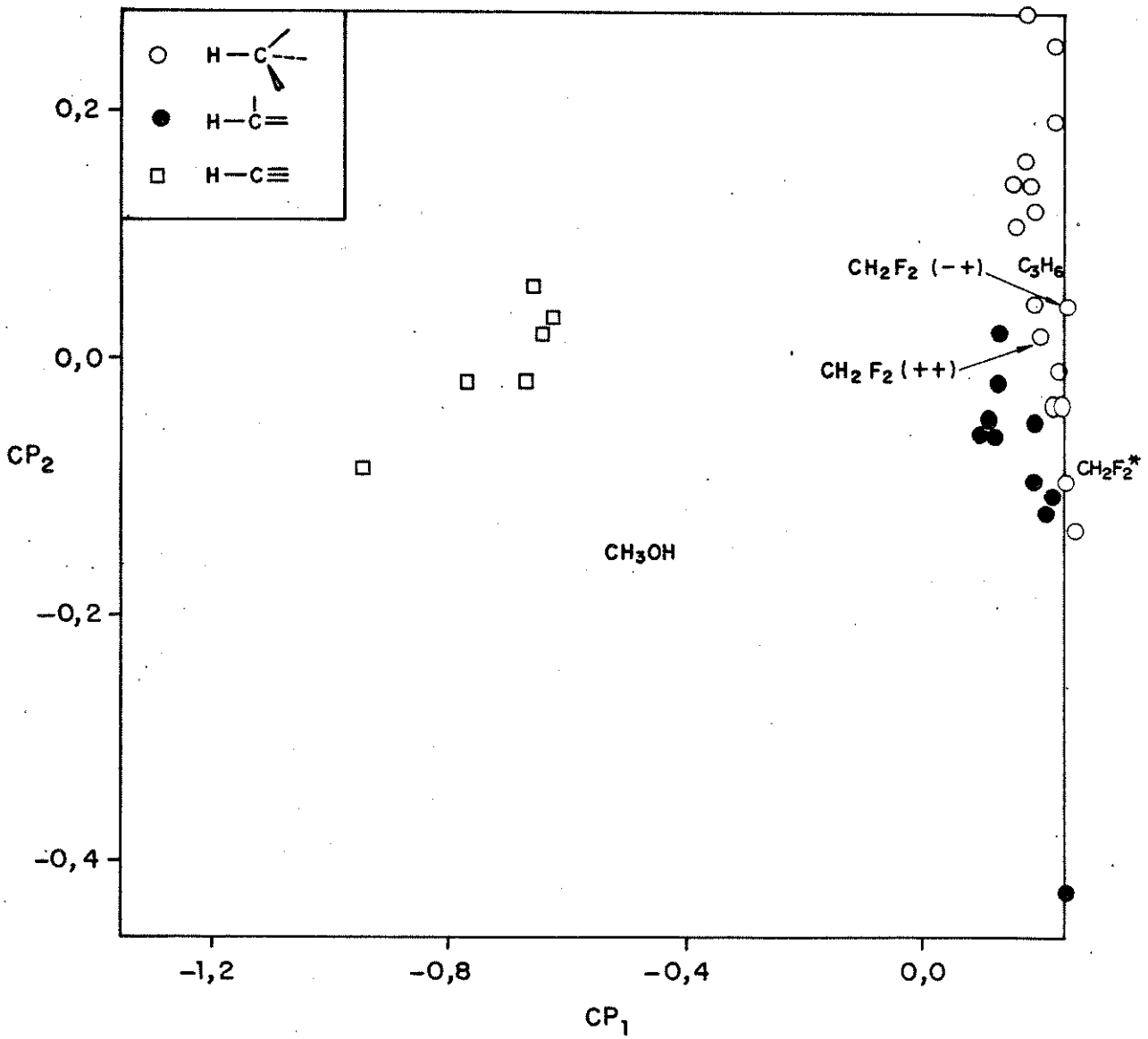


Figura 28 - Gráfico dos escores das duas primeiras componentes principais para os tensores polares dos átomos de hidrogênio. As amostras do conjunto de teste estão indicadas pela fórmula molecular (o ponto com * para o CH_2F_2 corresponde ao resultado mais antigo).

A tabela 8 mostra os resultados do SIMCA utilizando uma componente principal para a modelagem de cada classe. Este método mostrou-se capaz de classificar corretamente 93.7% dos hidrogênios ligados a carbonos sp^3, sp^2 e sp. A tabela 9 mostra os resultados para o conjunto de teste. Dois pontos do CH_2F_2 pertencem a diferentes escolhas de sinais para as derivadas do momento dipolar. O conjunto (++) mostra uma invariância isotópica para CH_2F_2 e CD_2F_2 ligeiramente superior aquela observada para o conjunto (- +), e é preferida na referência (52), em desacordo com resultados de química quântica.

A tabela 9 mostra que o ponto correspodente ao conjunto (++) e o outro ponto que corresponde ao resultado mais antigo para o CH_2F_2 foram classificados pelo método como sp^2, enquanto que o ponto (- +) foi classificado como sp^3. A figura 28 mostra que o ponto correspondendo aos sinais (++) e o do resultado mais antigo, estão mais afastados da região definida pelos outros tensores de hidrogênios ligados ao carbono com hibridização sp^3.

A tabela 9 mostra que para as duas alternativas o H_2CO foi classificado com sp^2 e para os anéis C_3H_6 e CH_2OCH_2 foram classificados como sp^3. Estas classificações poderiam ser esperadas.

A tabela 10 mostra a classificação feita através da regra do vizinho mais próximo para o conjunto de treinamento. Os melhores resultados foram obtidos com K = 1, 3, 4, 5, 6 e 7. A tabela 11 mostra que os resultados para o conjunto de teste é semelhante ao obtido pelo método SIMCA.

Tabela 8 - Resultado do método de classificação SIMCA para o conjunto de treinamento.

categoria	nº de amostras	pontos classificados incorretamente
H-C(sp^3)	15	2
H-C(sp^2)	11	0
H-C(sp)	6	0

Tabela 9 - Resultado do método de classificação SIMCA para o conjunto de teste.

	moléculas	classificação
	CH_3OH	sp
	C_3H_6	sp^3
(*)	CH_2F_2	sp^2
	H_2CO	sp^2
	H_2CO	sp^2
	CH_2OCH_2	sp^3
(++)	CH_2F_2	sp^2
(-+)	CH_2F_2	sp^3

* Resultado mais antigo

Tabela 10 - Resultado do método de classificação KNN para o conjunto de treinamento.

classes	nº de amostras	nº de pontos classificados incorretamente				
		1NN	3NN	5NN	7NN	9NN
H-C(sp^3)	15	1	1	1	1	4
H-C(sp^2)	11	0	1	0	4	9
H-C(sp)	6	0	0	0	0	0
% de informação correta		96,9	93,8	96,9	84,4	59,4

Tabela 11 - Resultado do método de classificação KNN para o conjunto de teste.

moléculas		Classificação obtida				
		1NN	3NN	5NN	7NN	9NN
	CH_3OH	sp^3	sp^3	sp^3	sp^3	sp^3
	C_3H_6	sp^3	sp^3	sp^3	sp^3	sp^3
	CH_2F_2	sp^3	sp^3	sp^3	sp^3	sp^3
	H_2CO	sp^3	sp^3	sp^3	sp^3	sp^3
	H_2CO	sp^2	sp^2	sp^2	sp^2	sp^2
	CH_2OCH_2	sp^3	sp^3	sp^3	sp^3	sp^3
(++)	CH_2F_2	sp^2	sp^2	sp^2	sp^2	sp^2
(-+)	CH_2F_2	sp^3	sp^3	sp^3	sp^3	sp^3

CARBONOS

Finalidade

Investigar se os valores dos tensores de carbono contêm informação suficiente para permitir a classificação dos carbonos de acordo com a hibridização: sp^3, sp^2 ou sp.

Conjunto de treinamento - Quarenta e sete tensores polares de átomos de carbono foram escolhidos entre os tensores das moléculas apresentadas na figura 26, ver figura 29.

Conjunto de teste - Sete tensores polares de átomos de carbono, (ver figura 29) foram usados para testar a capacidade das regras de classificação desenvolvidas no conjunto de treinamento.

Como no caso dos hidrogênios, para alguns átomos de carbono foi incluído mais de um tensor polar.

A análise de componentes principais mostra que as duas primeiras componentes explicam 96,4% da variância total. A primeira componente principal tem contribuição semelhante para todas as variáveis, enquanto que a segunda é dominada pela anisotropia, ver tabela 12.

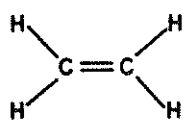
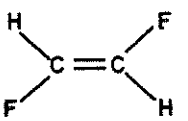
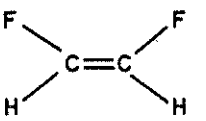
A figura 30 mostra o gráfico dos escores destas duas primeiras componentes principais. Comparando com a figura 28, a distribuição dos carbonos parece um pouco difusa e sem notáveis propriedades discriminatórias. Isto não é surpresa, em vista da diversidade de ambientes químicos aos quais os átomos de carbono nas moléculas aqui estudadas estão sujeitas, ver figura 30.

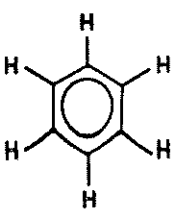
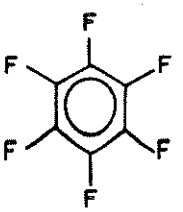
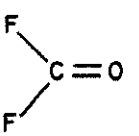
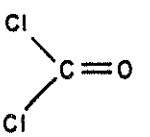
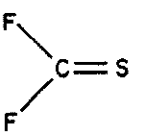
A figura mostra que para alguns grupos de moléculas geometricamente similares existe uma relação entre CP_1 e a eletro-

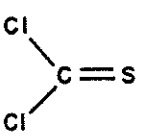
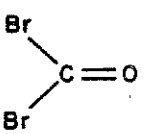
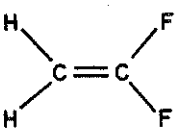
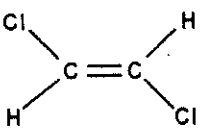
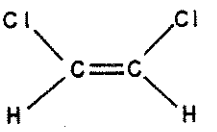
CONJUNTO DE TREINAMENTO

Sp^3 CH_3F CH_3Cl CH_3Br CH_3I CH_4 CH_3CH_3 CH_3CN $CH_3C{\equiv}CCH_3$

$CHCl_3$ CHF_3 $CH_3C{\equiv}CH$ CF_4

Sp^2  $H_2C{=}C{=}CH_2$  

Sp $CH_3C{\equiv}N$ $CH_3C{\equiv}CH$ $HC{\equiv}CH$ $H{-}C{\equiv}N$ $HC{\equiv}CC{\equiv}CH$

$N{\equiv}CC{\equiv}N$ $HC{\equiv}C{-}CN$

CONJUNTO TESTE

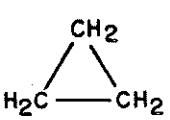
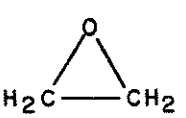
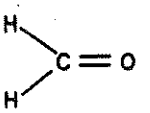
   CH_2F_2

Figura 29 - Moléculas que formam os conjuntos de treinamento e de teste para os tensores polares dos átomos de carbono.

Tabela 12 - Pesos das variáveis para as duas primeiras componentes principais.

variáveis	CP_1	CP_2
$\bar{p}$	-0,48	0,04
χ	-0,49	-0,06
β	-0,31	-0,87
$\sqrt[3]{D}$	-0,44	0,44
$\sqrt{C}$	-0,48	0,17
*	80,7	15,7

* percentagem de variância explicada em cada componente.

negatividade. Isto é mais perceptível para os metanos halogenados, cujos pontos, são individualmente identificados no gráfico. Estas moléculas estão distribuidas ao longo da CP_1, com os substituintes mais eletronegativos situados na CP_1 mais negativa e os menos eletronegativos na CP_1 mais positiva, sendo os casos extremos CF_4 e CH_4.

A figura 31 mostra a regressão dos escores CP_1 sobre a quantidade sigma definida abaixo para os metanos halogenados. O sigma é obtido subtraindo a eletronegatividade do carbono das eletronegatividades dos quatro átomos ligados a ele e então somando os resultados. Os valores conflitantes para as duas escolhas de sinais de CH_2F_2 não foram incluídos no cálculo de regressão. O ajuste é muito bom, com um coeficiente de correlação de 0.996. A equação da regressão é:

$$- PC_1(\text{escore}) = 0,221 + 0,996\ \Sigma$$

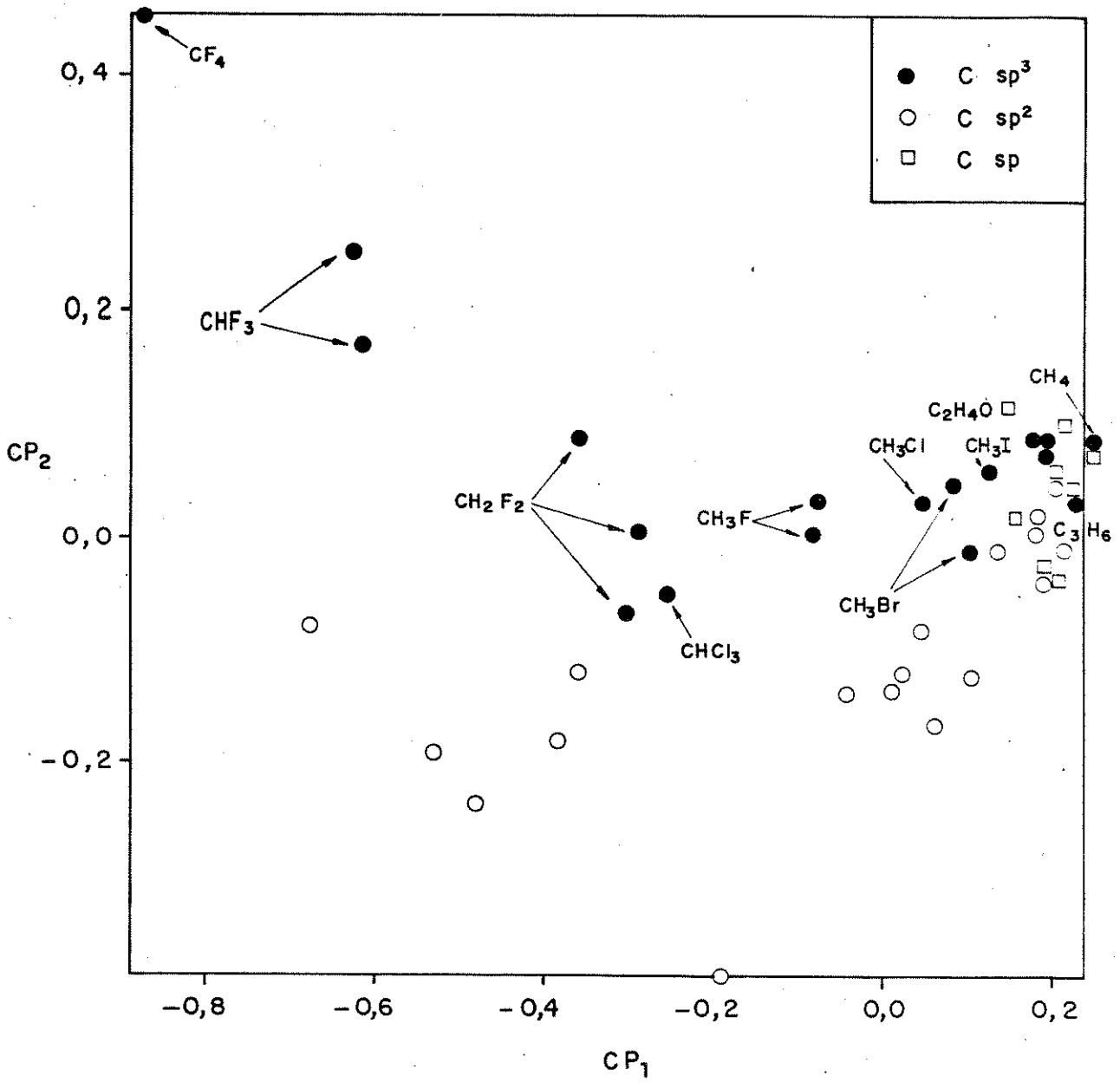


Figura 30 - Gráfico dos escores das duas primeiras componentes principais para os tensores polares dos átomos de carbono.

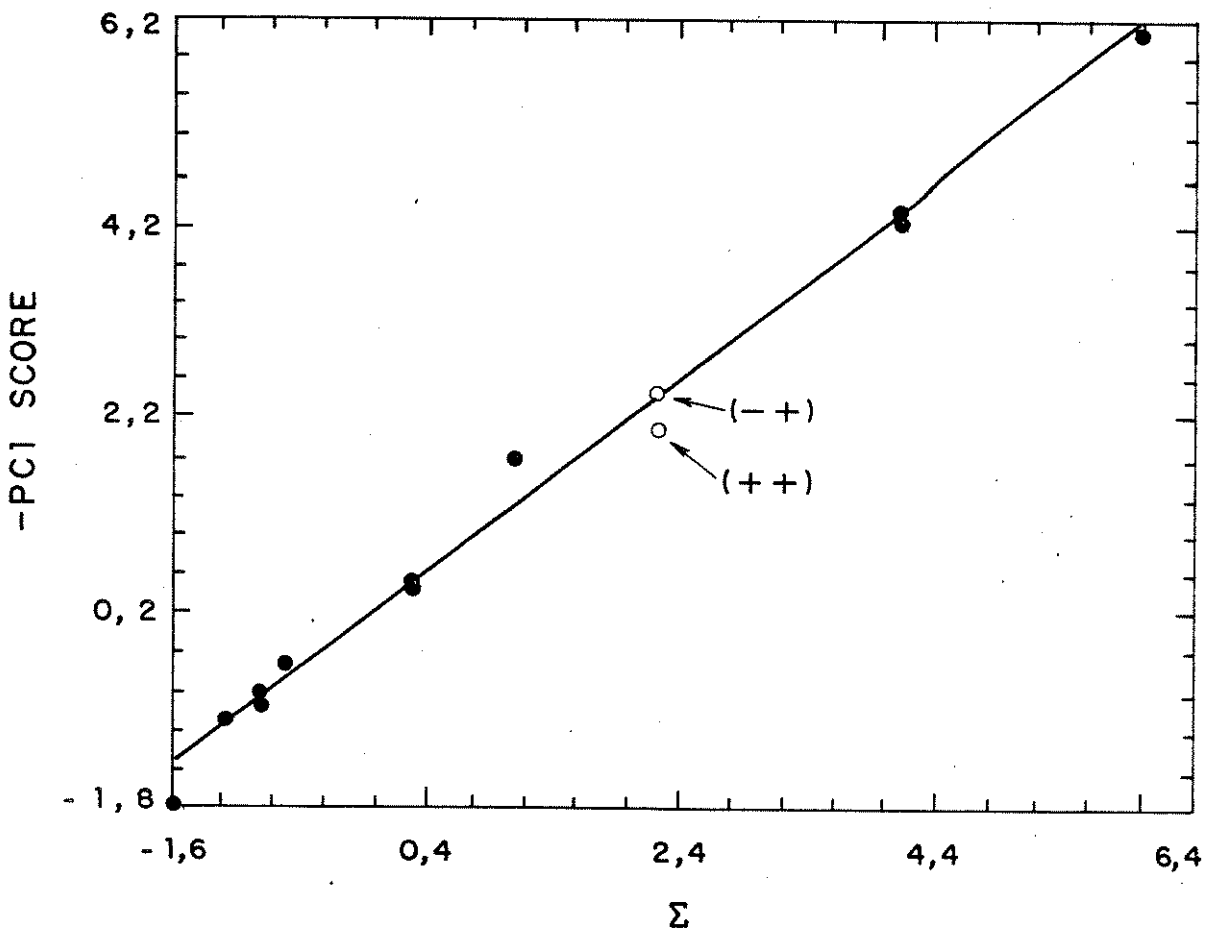


Figura 31 - Regressão de CP_1 na quantidade Σ para os átomos de carbono nos metanos halogenados.

As duas escolhas de sinais para CH_2F_2 foram também plotadas na figura 31. O ponto (++) está situado um pouco abaixo da linha de regressão, enquanto que a alternativa (-+), indicado por cálculos de química quântica e argumentos de similaridade, cai diretamente sobre ela.

Está claro que a CP_1 representa a eletronegatividade para os metanos halogenados, mas será que representa a eletronegatividade em geral? Investigações para responder esta pergunta estão sendo feitas.

A segunda componente na figura 30 mostra que os carbonos com hibridização sp^2 são mais anisotrópicas que aquelas com sp^3 e sp, e que para eles verifica-se a mesma ordem observada para seus hidrogênios associados ou seja,

$$\beta_{C,sp^2} > \beta_{C,sp^3} > \beta_{C,sp} .$$

As diferenças nos valores da anisotropia observadas para os tensores polares atômicos podem estar relacionadas aos ambientes eletrônicos dos átomos e ligações adjacentes aos átomos que estão sendo estudados. São encontrados maiores valores da anisotropia para os tensores polares atômicos de átomos adjacentes a uma ligação insaturada ou para um átomo com um par isolado polarizável, do que para tensores polares de átomos em um ambiente completamente saturado.

Devido à diversidade de ambientes químicos dos átomos de carbono nas moléculas, a classificação para este conjunto não foi muito boa. O método SIMCA foi capaz de classificar apenas 66% dos carbonos sp^3, sp^2 e sp para o conjunto de treinamento. O método KNN foi capaz de classificar para 1NN 74.2%; para 2NN e 3NN,

70.2% e para os demais vizinhos em torno de 66%.

Os resultados do conjunto de teste para o SIMCA e o KNN estão apresentados nas tabelas 13 e 14, respectivamente. Como as percentagens de classificação correta para o conjunto de treinamento foram baixas, não era de se esperar que os resultados apresentados nestas tabelas fossem confiáveis.

Tabela 13 - Resultado do método de classificação SIMCA para o conjunto de teste.

moléculas	classificação
C_3H_6	sp
CH_2OCH_2	sp
H_2CO	sp^2
H_2CO	sp^3
CH_2F_2	sp^3
CH_2F_2	sp^2
CH_2F_2	sp^3

Tabela 14 - Resultado do método de classificação KNN para o conjunto de teste.

moléculas	classificação obtida				
	1NN	3NN	5NN	7NN	9NN
CH_3H_6	sp^3	sp^3	sp^2	sp	sp
CH_2OCH_2	sp^3	sp^3	sp^3	sp^3	sp^3
H_2CO	sp^3	sp^3	sp^3	sp^3	sp^3
H_2CO	sp^3	sp^3	sp^3	sp^3	sp^3
CH_2F_2	sp^3	sp^3	sp^3	sp^3	sp^3
CH_2F_2	sp^3	sp^2	sp^2	sp^2	sp^2
CH_2F_2	sp^3	sp^2	sp^3	sp^3	sp^3

HETEROÁTOMOS

Finalidade

Investigar se os valores dos tensores polares de heteroátomos terminais contêm informação suficiente para discriminar entre heteroátomos de flúor, cloro, bromo, iodo e oxigênio.

Conjunto de treinamento - Este conjunto foi formado com quarenta e quatro tensores polares de heteroátomos. Os tensores polares de heteroátomos ligados a carbonos foram escolhidos entre os tensores para as moléculas apresentadas na figura 26.

A análise de componentes principais mostra que as duas primeiras componentes descrevem 98.5% da informação estatística dos dados. Como nos casos anteriores a primeira componente principal tem contribuição semelhante de todas as variáveis, enquantoque na segunda a maior contribuição é de β, ver tabela 15.

A figura 32 mostra os escores de todas as moléculas deste subgrupo. Os halogênios estão representados por dois tipos de símbolos geométricos. Os símbolos pretos representam os halogênios ligados a carbonos saturados, enquanto que os brancos correspondem aos halogênios ligados a átomos participantes de uma ligação insaturada ou contendo um par isolado polarizável. Os haletos de hidrogênio e os trihaletos de boro estão indicados pela fórmula molecular e os outros heteroátomos estão representados pelos seus símbolos atômicos.

A primeira componente principal (CP_1) separa mais ou menos os halogênios das moléculas poliatômicas em grupos; no lado mais negativo de CP_1 estão localizados os halogênios mais eletronegativos. O mesmo padrão, mas deslocado para a direita repete-se

Tabela 15 - Pesos das variáveis para as duas primeiras componentes principais.

variáveis	CP_1	CP_2
$\bar{p}$	0,51	-0,02
χ	-0,48	-0,29
β	-0,32	-0,75
$\sqrt[3]{D}$	0,42	0,50
$\sqrt{C}$	-0,48	-0,32
*	77,1	21,4

* percentagem de variância explicada em cada componente

para os haletos de hidrogênio. Para as invariantes dos tensores polares das moléculas restantes a situação não é tão clara. Os oxigênios, que se superpõe à classe dos fluoretos para as moléculas saturadas, ficam à esquerda dos nitrogênios. No meio ficam dois átomos de enxôfre, misturados com os cloretos.

A figura 32 mostra que estas duas componentes permitem uma boa discriminação do tipo de átomo ao qual um halogênio está ligado. Com somente uma exceção, todos os símbolos pretos, representando os halogênios ligados a átomos saturados, ficam na metade superior desta figura. Somente um ponto representando o CH_2F_2 e, com $CP_2 \tilde{-} -0,1$, representa um tensor polar de fluor ligado a um carbono saturado.

Sua presença entre os triângulos brancos representando os fluoretos em um ambiente saturado ou com par isolado sugere que o ponto correspondendo aos resultados experimentais mais recen-

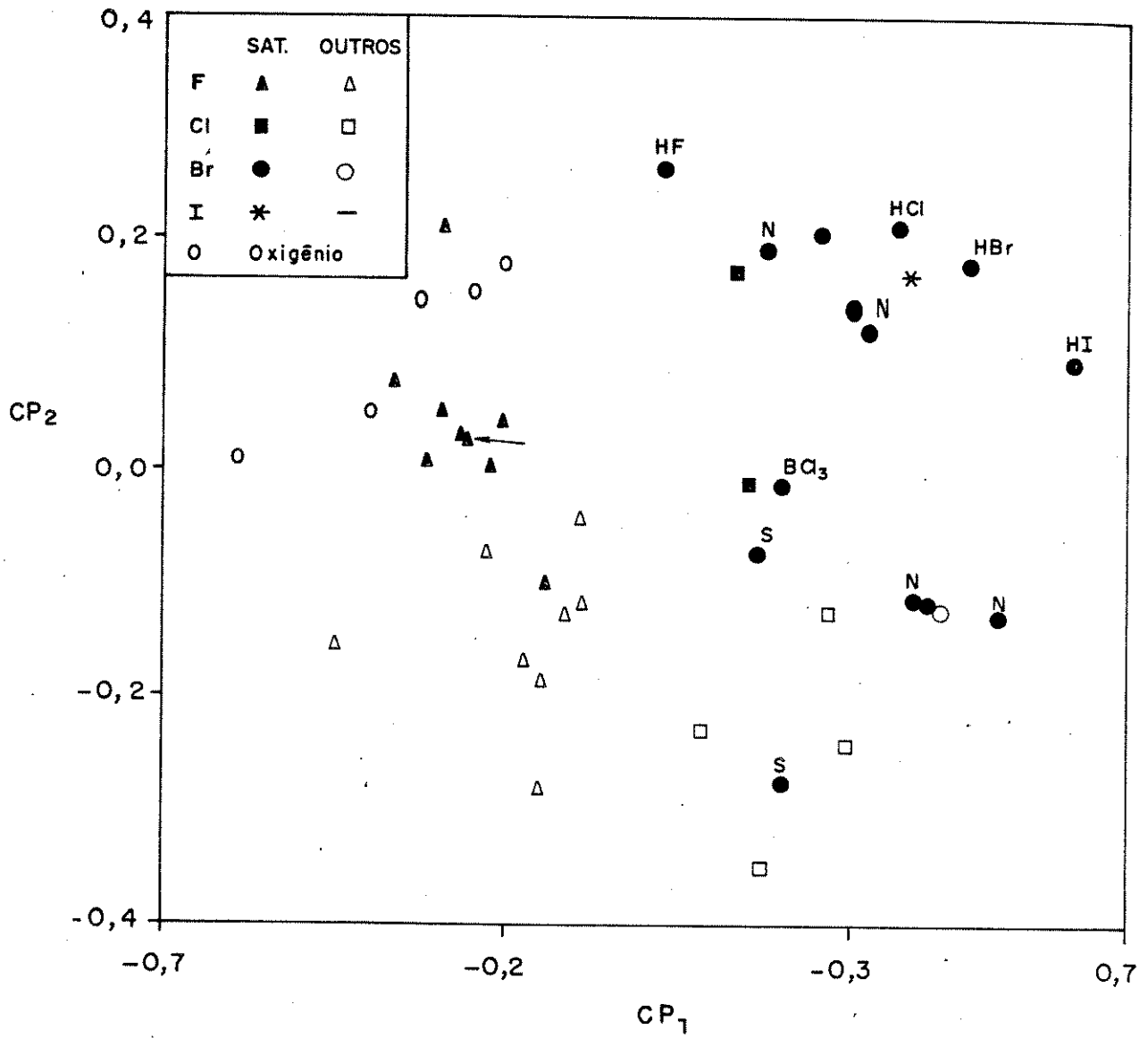


Figura 32 - Gráfico dos escores das duas primeiras componentes principais para os tensores polares dos heteroátomos terminais.

tes (52), indicado por uma flecha na figura 32, deve ser preferido caso a similaridade com os outros fluoretos da classe saturada seja levada em conta. Isto pode bem ser o caso, considerando-se a extensiva sobreposição verificada no espectro infravermelho do difluormetano (52). Vale a pena lembrar que esta conclusão é consistente com aquela obtida anteriormente para os tensores polares de hidrogênios.

A segunda componente mostra que para cada tipo de halogênio os valores da anisotropia são mais altos para ambientes não saturados ou de pares isolados do que para aqueles totalmente saturados. Isto pode ser explicado por uma consideração qualitativa da estrutura destas moléculas. Como os elétrons participantes de ligações duplas e pares isolados são mais polarizáveis, isto é, tem mais baixos potenciais de ionização do que aqueles de uma ligação saturada, pode esperar-se uma anisotropia mais baixa para um átomo em um ambiente completamente saturado.

CONSIDERAÇÕES FINAIS

- Possibilidade de usar projeções de componentes principais e os métodos de classificação SIMCA e KNN para determinar sinais de $\partial p/\partial_{Q_i}$ e para controle de qualidade de medidas experimentais.

- Para os metanos halogenados, a eletronegatividade dos átomos ligados ao carbono determinam a maior parte da variância total dos valores das invariantes dos tensores polares de carbono. O resto corresponde às propriedades dos átomos relacionadas a anistropia do tensor polar mais o erro experimental.

- Para as moléculas aqui estudadas os valores das cinco invariantes do tensor polar são altamente correlacionadas, tendo uma dimensionalidade intrínsica de aproximadamente dois. Isto deve ser devido a alta simetria das moléculas aqui estudadas.

- Os tensores polares de heteroátomos podem ser usados para classificação de acordo com o tipo de átomo F, O, Cl, N, Br e S. Além disto podem ser identificados ambientes de pares isolados ou ligações insaturadas na vizinhança do átomo que está sendo estudado.

Quarta Aplicação

Colaboradores:

Ronei J. Poppi - Instituto de Química - Universidade Estadual de Campinas, Campinas, SP.

Prof.Dr. José Fernando G. Faigle - Instituto de Química - Universidade Estadual de Campinas, Campinas, SP.

Finalidade

Verificar se o método de regressão em componentes principais, PCR, é viável para fazer análise quantitativa usando dados cromatográficos com diferentes graus de superposição de sinais. O estudo foi feito em duas etapas. Primeiro, dados simulados foram usados para estudar as propriedades matemáticas do método PCR relacionados com a regressão linear múltipla convencional. Depois, dados experimentais foram analisados para estudar o efeito do erro experimental na exatidão do método.

DADOS SIMULADOS

O pico cromatográfico foi definido como uma função do tipo Frazer-Suzuki (53),

$$f(t) = H_{exp}\{(-\ell n2/A^2)[\ln\{1+[A(t-tr)/d(2\ell n^2)^{1/2}]\}]^2\}$$

onde H é a amplitude do pico, σ é o desvio padrão, tr é o tempo

de retenção, e A é o fator de assimetria.

As curvas foram geradas somando-se duas destas funções (13). Os diferentes graus de superposição foram obtidos alterando-se o valor de tr. Neste caso foram estudados quatro graus de superposição de dois picos, como mostra a figura 33.

Para cada grau de superposição foi construído um conjunto de dados da seguinte forma: simulou-se 10 curvas de uma mistura de dois constituintes químicos. Estas curvas foram digitalizadas tomando-se 39 valores de alturas igualmente espaçadas. Desta maneira, foi formada uma matriz $\underset{\sim}{X}$ de ordem 10x39. Uma matriz $\underset{\sim}{Y}$ de ordem 10x2 foi formada com áreas conhecidas dos dois picos (área em unidades arbitrárias, u.a) que correspondem aos dois constituintes químicos para as 10 curvas, ver figura 15.

Três misturas adicionais foram simuladas e suas alturas foram tomadas da mesma forma descrita acima. Estas amostras não foram usadas no modelo de calibração, mas para testar a qualidade de previsão do modelo de regressão de componentes principais.

Antes de fazer a modelagem da matriz $\underset{\sim}{X}$, os dados foram autoescalonados. Na modelagem foram utilizadas duas componentes principais que conseguem explicar para o conjunto I, 100% da variância contida nos dados, e para os outros três em torno de 99%.

Os resultados previstos, os erros de previsão (EP) e o erro padrão de calibração (EPC) para os conjuntos de calibração e o erro padrão de previsão (EPP) para os conjuntos de testes são apresentados nas tabelas 16, 17, 18 e 19, para os conjuntos I, II, III e IV, respectivamente. Pode-se observar pelas tabelas que o EPP é maior para os casos onde há maior grau de superposição dos sinais.

Estes resultados foram comparados com a regressão linear

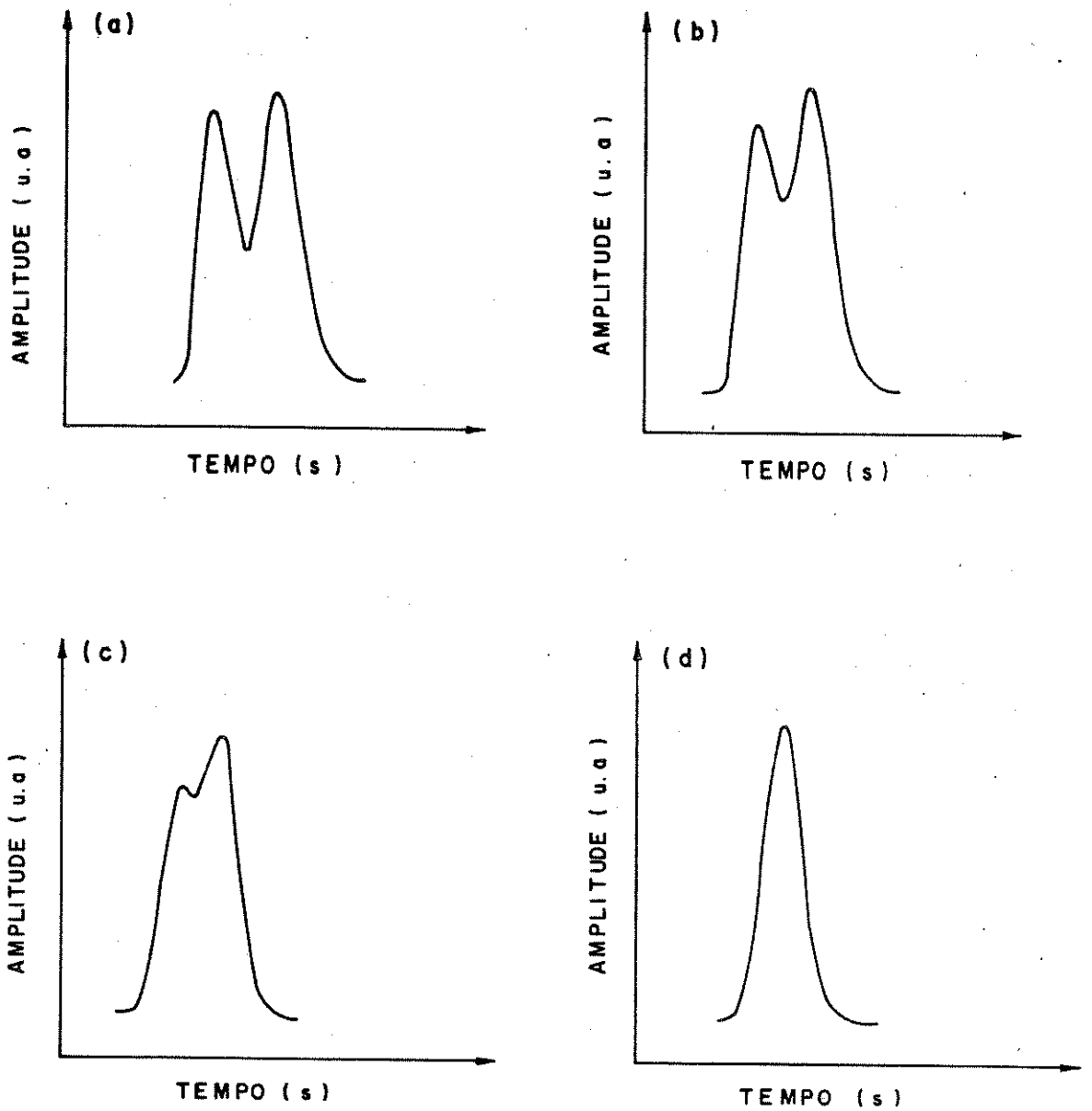


Figura 33 - Picos simulados construídos com o modelo de Frazer-Suzuki. A superposição foi obtida somando-se as funções para os dois constituintes puros, alterando-se o valor de tr. (a) conjunto I; (b) conjunto II; (c) conjunto III e (d) conjunto IV (superposição total).

Tabela 16 -Comparação entre o valor real e o previsto para os conjuntos de calibração e de teste, usando o modelo com duas componentes principais (área em unidades arbritárias), para o conjunto I.

Amostra nº	Pico 1			Pico 2		
	Real	Previsto	Erro de Previsão[a]	Real	Previsto	Erro de Previsão[a]
	Conjunto de treinamento					
1	36,086	36,086	0,000	34,558	34,558	0,000
2	45,108	45,108	0,000	34,558	34,558	0,000
3	45,108	45,108	0,000	25,918	25,918	0,000
4	40,597	40,597	0,000	34,558	34,558	0,000
5	36,086	36,086	0,000	25,918	25,918	0,000
6	40,597	40,597	0,000	30,238	30,238	0,000
7	40,597	40,597	0,000	43,197	43,197	0,000
8	27,065	27,065	0,000	38,878	38,878	0,000
9	31,576	31,576	0,000	34,558	34,558	0,000
10	27,065	27,065	0,000	34,558	34,558	0,000
EPC[b]			0,000			0,000
	Conjunto de teste					
11	36,086	36,086	0,000	30,238	30,238	0,000
12	31,576	31,576	0,000	38,878	38,878	0,000
13	31,576	31,576	0,000	43,197	43,197	0,000
EPP[c]			0,000			0,000

(a) $EP = (Y_{prev} - Y\ real)$

(b) $EPC = \left[\sum_{k=1}^{n} (Y_{real} - Y_{prev})^2/(n-a-1)\right]^{1/2}$, n é o número de amostras de calibração e 'a' o número de componentes principais

(c) $EPP = \left[\sum_{k=1}^{n} (Y_{real} - Y_{prev})^2/(n-1)\right]^{1/2}$, n é o número de amostras de teste.

Tabela 17 - Comparação entre o valor real e o previsto para os conjuntos de calibração e de teste, usando o modelo com duas componentes principais (área em unidades arbitrárias) para o conjunto II.

Amostra nº	Pico 1			Pico 2		
	Real	Previsto	Erro de Previsão[a]	Real	Previsto	Erro de Previsão[a]
	Conjunto de treinamento					
1	36,086	36,155	0,069	34,558	34,267	-0,291
2	45,108	45,106	-0,002	34,558	34,550	-0,008
3	45,108	45,101	-0,007	25,918	25,916	-0,002
4	40,597	40,582	-0,015	34,558	34,750	0,192
5	36,086	36,088	0,002	25,918	25,940	0,022
6	40,597	40,582	-0,015	30,238	30,245	0,007
7	40,597	40,592	-0,005	43,197	43,190	-0,007
8	27,065	27,055	-0,010	38,878	38,910	0,032
9	31,576	31,574	-0,002	34,558	34,582	0,024
10	27,065	27,051	-0,014	34,558	34,593	0,035
EPC[b]			0,028			0,134
	Conjunto de teste					
11	36,086	36,092	0,006	30,238	30,250	0,012
12	31,576	31,577	0,001	38,878	88,897	0,019
13	31,576	31,580	0,004	43,197	43,212	0,015
EPP[c]			0,005			0,018

(a), (b), (c)

Ver rodapé da tabela 16.

Tabela 18 - Comparação entre o valor real e o previsto para os conjuntos de calibração e de teste, usando o modelo com duas componentes principais (área em unidades arbitárias), para o conjunto III.

Amostra nº	Pico 1			Pico 2		
	Real	Previsto	Erro de Previsão[a]	Real	Previsto	Erro de Previsão[a]
Conjunto de treinamento						
1	36,086	36,376	0,290	34,558	34,474	-0,084
2	45,108	44,954	-0,154	34,558	34,603	0,045
3	45,108	44,911	-0,197	25,918	25,975	0,057
4	40,597	40,647	0,050	34,558	34,541	-0,017
5	36,086	36,333	0,247	25,918	25,847	-0,071
6	40,597	40,626	0,029	30,238	20,230	-0,008
7	40,597	40,691	0,044	43,197	43,171	-0,026
8	27,065	27,027	-0,038	38,878	38,888	0,010
9	31,576	31,312	-0,264	34,558	34,635	0,077
10	27,065	27,005	-0,060	34,558	34,575	0,017
EPC[b]			0,205			0,059
Conjunto de teste						
11	36,086	36,355	0,269	30,238	30,162	-0,076
12	31,576	31,333	-0,243	38,878	38,948	0,070
13	31,576	31,355	-0,221	43,197	43,263	0,066
EPP[c]			0,300			0,087

(a), (b), (c),

Ver rodapé da tabela 16.

Tabela 19 - Comparação entre o valor real e o previsto para os conjuntos de calibração e de teste, usando o modelo com duas componentes principais (área em unidades arbitrárias), para o conjunto IV.

Amostra nº	Pico 1			Pico 2		
	Real	Previsto	Erro de Previsão[a]	Real	Previsto	Erro de Previsão[a]
	Conjunto de treinamento					
1	36,086	36,013	-0,073	34,557	34,661	0,104
2	45,108	45,145	0,037	34,557	34,505	-0,052
3	45,108	45,198	0,090	25,918	25,790	-0,128
4	40,597	40,579	-0,018	34,557	34,582	0,025
5	36,086	36,065	-0,021	25,918	25,947	0,029
6	40,597	40,608	0,011	30,238	30,223	-0,015
7	40,597	40,525	-0,072	43,197	43,298	0,101
8	27,065	27,424	0,359	38,877	38,371	-0,506
9	31,576	31,447	-0,129	34,557	34,738	0,181
10	27,065	27,424	0,359	34,557	34,738	0,181
EPC[b]			0,205			0,228
	Conjunto de teste					
11	36,086	36,052	-0,034	30,238	30,284	0,046
12	31,576	31,420	-0,156	38,878	39,096	0,218
13	31,576	31,391	-0,185	43,197	43,457	0,260
EPP[c]			0,173			0,242

(a), (b), (c),

Ver rodapé da tabela 16.

múltipla convencional. Para o conjunto I obteve-se os seguintes resultados do erro padrão de previsão para o conjunto de teste: para o pico 1 o EPP foi de 0,008 u.a e para o pico 2 foi de 0,006 u.a. Para os outros três conjuntos não foi possível usar este método devido às altas correlações entre as variáveis, que impossibilitam a inversão da matriz de covariância. Como vimos no capítulo II, este é um dos problemas frequentes do método de regressão convencional.

DADOS EXPERIMENTAIS

As amostras foram preparadas por Ronei J. Poppi, e consistem de misturas de tolueno (Carlo Erba - p.a), isoctano (Carlo Erba - para cromatografia) e etanol (Merck - p.a), tratadas com peneira molecular para eliminar a água. Foi utilizado um cromatógrafo gás-líquido VARIAN (modelo 920) com detector de condutividade térmica, e uma coluna com 5,2% de SE-30 em cromossorb w (80-100 mesh) de 2,0 metros de comprimento e 1/8" de diâmetro.

Para obter os picos com diferentes graus de superposição alterou-se a temperatura da coluna. Para os três conjuntos estudados estabeleceu-se as seguintes temperaturas: 105°C, 120°C e 130°C.

As amostras foram preparadas por pesagem direta dos três constituintes químicos puros (13). Os cálculos das massas injetadas em mg, foram feitos considerando as densidades à temperatura ambiente de cada constituinte químico e o volune de injeção, o qual foi mantido constante, 2 µl.

Para cada temperatura foram obtidos 20 cromatogramas da

mistura dos três constituintes químicos, sendo 14 usados no conjunto de calibração e 6 no conjunto de previsão (ou teste).

Num estudo prévio digitalizou-se todos os cromatogramas inteiros medindo-se os valores das alturas para 41 tempos de eluição diferentes. Devido à semelhança entre os extremos destes cromatogramas para cada temperatura, o mesmo conjunto foi re-analizado e resolveu-se abandonar as alturas do início e do final de cada cromatograma, ficando-se apenas a parte central. A figura 34 dá uma idéia dos graus de superposição obtidos para cada temperatura, bem como o início e o final da digitalização. Mais detalhes sobre a parte experimental e a obtenção dos dados poderão ser vistos na referência (13).

No conjunto de calibração foram medidos os valores das alturas para 26 tempos de eluição diferentes, formando a matriz $\underset{\sim}{X}$, de ordem 14x26. A matriz $\underset{\sim}{Y}$, de ordem 14x3, foi formada com as massas (em mg) injetadas dos três constituintes químicos.

Os 6 cromatogramas do conjunto de teste da matriz $\underset{\sim}{X}$ foram digitalizados da mesma forma descrita acima.

A tabela 20 mostra a percentagem de variância contida nos dados para as 5 primeiras componentes principais, para os 3 conjuntos. Um fator importante que deve ser considerado é quantas componentes principais devem ser incluídas na modelagem da matriz $\underset{\sim}{X}$. Desejamos selecionar um número de componentes que permita modelar o sistema sem super ajustar os dados. Se o sistema for linear, o número de componentes principais necessários para representar os dados deverá ser três, uma vez que os conjuntos de dados das misturas contém três constituintes químicos. Contudo, a tabela 20 mostra que duas componentes principais contêm em torno de 90% de variância para os três casos. Isto ocor-

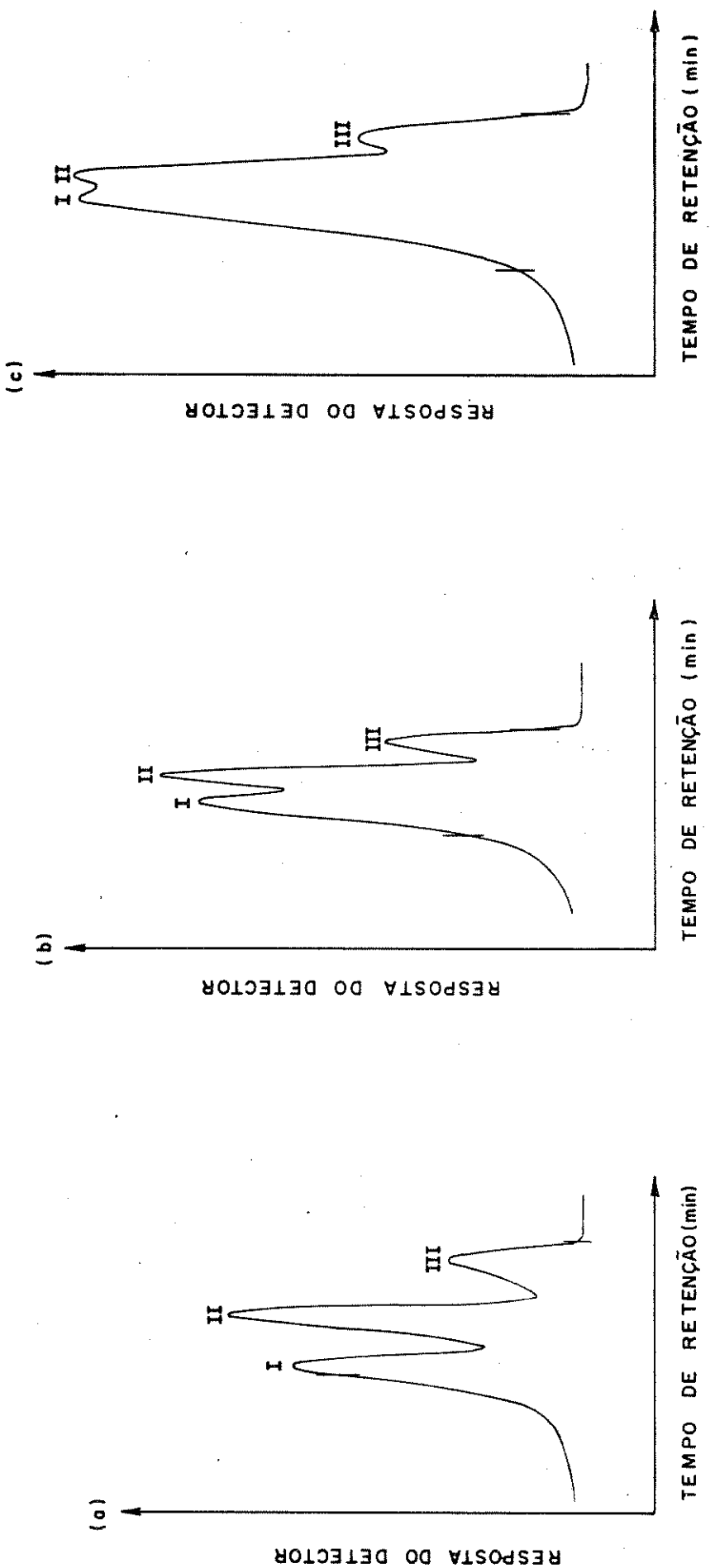


Figura 34 - Cromatogramas para os dados experimentais. (a) conjunto a 105°C; b) a 120°C e c) a 130°C. Pico I tolueno, II isooctano, III etanol.

re porque o volume injetado destas misturas foi mantido aproximadamente constante (2 µl), o que significa que existem somente dois constituintes químicos independentes. O terceiro é determinado pela soma dos outros dois.

Tabela 20 - Percentagem de variância explicada pelas cinco primeiras componentes, para os três conjuntos.

	105°C	120°C	130°C
CP_1	50,6	49,1	67,7
CP_2	43,9	38,7	24,1
CP_3	2,5	5,1	4,6
CP_4	1,8	4,2	2,3
CP_5	0,6	2,0	0,6
Total	99,4	99,1	99,3

Para estabelecer o melhor modelo de calibração foram feitas modelagens usando 3,4 e 5 componentes para cada temperatura, sendo os dados da matriz $\underline{X}$ primeiro autoescalonados.

As tabelas 21, 22 e 23 mostram as massas reais, as massas previstas e o erro de previsão, utilizando três componentes principais na modelagem dos conjuntos nas temperaturas 105°C, 120°C e 130°C, respectivamente. Estes resultados foram comparados em termos de erro padrão de calibração com os modelos utilizando quatro e cinco componentes principais, tabela 24. Para o conjunto 105°C não houve diferença significativa nos erros padrão para as três modelagens, sendo assim, três componentes parecem ser suficientes para descrever os dados. Para o conjunto 120°C

Tabela 21 - Comparação entre a massa real e prevista (em mg) para o conjunto de calibração, usando o modelo de três componentes principais para o conjunto a 105°C.

Amostra nº	Tolueno			Isoctano			Etanol		
	Real	Prevista	Erro de Previsão[a]	Real	Prevista	Erro de Previsão[a]	Real	Prevista	Erro de Previsão[a]
1	0,724	0,727	0,003	0,481	0,479	-0,002	0,353	0,354	0,001
2	0,817	0,801	-0,016	0,447	0,460	0,013	0,308	0,308	0,000
3	0,630	0,641	0,011	0,512	0,508	-0,004	0,403	0,400	-0,003
4	0,850	0,855	0,005	0,315	0,319	0,004	0,428	0,419	-0,009
5	1,014	1,019	0,005	0,201	0,194	-0,007	0,409	0,413	0,004
6	0,483	0,486	0,003	0,685	0,689	0,004	0,340	0,334	-0,006
7	0,413	0,417	0,004	0,769	0,761	-0,008	0,308	0,314	0,006
8	0,503	0,500	-0,003	0,615	0,613	-0,002	0,402	0,409	0,007
9	0,300	0,300	0,000	0,769	0,788	0,019	0,391	0,391	0,000
10	0,343	0,354	0,011	0,480	0,482	0,002	0,701	0,691	-0,010
11	0,321	0,323	0,002	0,401	0,398	-0,003	0,810	0,814	0,004
12	0,416	0,409	-0,007	0,506	0,519	0,013	0,606	0,599	-0,007
13	0,387	0,339	-0,048	0,215	0,216	0,001	0,963	0,962	-0,001
14	0,530	0,514	-0,016	0,403	0,398	-0,005	0,619	0,640	0,021

(a) Ver rodapé da tabela 16.

Tabela 22 - Comparação entre a massa real e prevista (em mg) para o conjunto de calibração, usando o modelo de três componentes principais para o conjunto a 120°C.

Amostra nº	Tolueno			Isoctano			Etanol		
	Real	Prevista	Erro de Previsão[a]	Real	Prevista	Erro de Previsão[a]	Real	Prevista	Erro de Previsão[a]
1	0,723	0,709	-0,014	0,481	0,498	0,017	0,353	0,346	-0,007
2	0,816	0,751	-0,065	0,447	0,479	0,032	0,308	0,330	0,022
3	0,629	0,575	-0,054	0,512	0,562	0,050	0,403	0,396	-0,007
4	0,849	0,894	0,045	0,315	0,307	-0,008	0,428	0,396	-0,032
5	1,012	1,025	0,013	0,201	0,176	-0,025	0,409	0,425	0,016
6	0,482	0,493	0,011	0,685	0,691	0,006	0,340	0,323	-0,017
7	0,412	0,394	-0,018	0,769	0,759	-0,010	0,308	0,336	0,028
8	0,502	0,501	-0,001	0,615	0,636	0,021	0,401	0,378	-0,023
9	0,300	0,338	0,038	0,769	0,750	-0,019	0,391	0,397	0,006
10	0,343	0,395	0,052	0,480	0,424	-0,056	0,701	0,717	0,016
11	0,321	0,315	-0,006	0,401	0,435	0,034	0,810	0,778	-0,032
12	0,415	0,484	0,069	0,506	0,447	-0,059	0,606	0,610	0,004
13	0,387	0,333	-0,054	0,215	0,259	0,044	0,963	0,964	0,001
14	0,529	0,512	-0,017	0,403	0,394	-0,009	0,619	0,644	0,025

[a] Ver rodapé da tabela 16.

Tabela 23 - Comparação entre a massa real e prevista (em mg) para o conjunto de calibração, usando o modelo de três componentes principais para o conjunto a 130°C.

Amostra nº	Tolueno			Isoctano			Etanol		
	Real	Prevista	Erro de Previsão[a]	Real	Prevista	Erro de Previsão[a]	Real	Prevista	Erro de Previsão[a]
1	0,722	0,736	0,014	0,480	0,468	-0,012	0,353	0,354	0,001
2	0,814	0,800	-0,014	0,446	0,446	0,000	0,307	0,321	0,014
3	0,628	0,624	-0,004	0,511	0,466	-0,045	0,402	0,457	0,055
4	0,848	0,848	0,000	0,314	0,275	-0,039	0,427	0,471	0,044
5	1,010	1,016	0,006	0,201	0,236	0,035	0,408	0,363	-0,045
6	0,482	0,452	-0,030	0,684	0,712	0,028	0,339	0,334	-0,005
7	0,412	0,398	-0,014	0,767	0,779	0,012	0,308	0,307	-0,001
8	0,501	0,510	0,009	0,614	0,619	0,005	0,401	0,386	-0,015
9	0,299	0,321	0,022	0,785	0,745	-0,040	0,391	0,416	0,025
10	0,342	0,352	0,010	0,479	0,498	0,019	0,700	0,671	-0,029
11	0,362	0,331	0,005	0,401	0,401	0,000	0,809	0,801	-0,008
12	0,414	0,422	0,008	0,505	0,503	-0,002	0,605	0,602	-0,003
13	0,386	0,367	-0,019	0,215	0,212	-0,003	0,961	0,983	-0,022
14	0,528	0,535	0,007	0,402	0,446	0,044	0,618	0,562	-0,056

(a) Ver rodapé da tabela 16.

Tabela 24 - Comparação dos erros padrão de calibração[a] em (mg) para os conjuntos de calibração, utilizando 3, 4 e 5 componentes para a modelagem.

número de componentes	Tolueno	Isoctano	Etanol
	105°C		
3	0,018	0,010	0,009
4	0,009	0,010	0,009
5	0,009	0,010	0,008
	120°C		
3	0,047	0,039	0,023
4	0,030	0,032	0,024
5	0,030	0,033	0,025
	130°C		
3	0,016	0,031	0,035
4	0,017	0,025	0,030
5	0,018	0,011	0,016

(a) Ver rodapé da tabela 16.

os maiores erros padrão para o tolueno, isoctano e etanol foram encontrados usando o modelo com 3 componentes principais, enquanto que para os modelos de quatro e cinco componentes a diferença nos erros não é significativa. Neste caso parece que são necessárias pelo menos quatro componentes para descrever os dados. Este aumento no número de componentes pode ocorrer devido a várias razões: 1) Presença de interações químicas (54) ou seja, efeitos de matriz; 2) variações na linha de base (54); 3) presença de interferentes e 4) não linearidade nos dados. No caso de não linearidade, a referência 42 mostra com dados simulados que o número de componentes principais não aumenta, mas causa erros de previsão mais altos.

No conjunto 130°C o erro padrão de calibração para o tolueno é praticamente constante, enquanto que para o isoctano e etanol ele diminui com o aumento do número de componentes principais incluídos na modelagem.

A tabela 25 mostra as massas reais, as massas previstas e o erro de previsão para as amostras dos conjuntos de teste para os três graus de superposição, utilizando-se o modelo de três componentes principais. Estes resultados foram comparados em termos de erro padrão de previsão com modelos construídos utilizando-se quatro e cinco componentes, como mostra a tabela 26.

Vemos que para os conjuntos 105°C e 130°C as diferenças nos erros padrão para os três modelos não são tão significativas. Para o conjunto 120°C, a tabela mostra que o erro padrão é maior para o modelo com três componentes, enquanto que nos modelos com quatro e cinco componentes, a diferença nos erros padrão não é significativa. Estes resultados mostram que para este conjunto são necessárias quatro componentes para descrever os dados, en-

Tabela 25 - Comparação entre a massa real e prevista (em mg) para os conjuntos de teste, usando 3 componentes principais.

Amostra nº	Tolueno			Isoctano			Etanol		
	Real	Prevista	Erro de Previsão (a)	Real	Prevista	Erro de Previsão (a)	Real	Prevista	Erro de Previsão (a)
105°C									
15	0,631	0,646	0,015	0,403	0,403	0,000	0,529	0,515	-0,014
16*	0,202	0,254	0,052	0,895	0,842	-0,053	0,359	0,371	0,012
17	0,396	0,394	-0,002	0,615	0,593	-0,022	0,501	0,527	0,026
18	0,415	0,422	0,007	0,303	0,289	-0,014	0,839	0,850	0,011
19	0,613	0,600	0,013	0,608	0,609	0,001	0,311	0,321	0,010
20	0,521	0,508	-0,013	0,496	0,513	0,017	0,523	0,515	-0,008
120°C									
15	0,629	0,586	-0,043	0,402	0,461	0,059	0,528	0,501	-0,027
16*	0,201	0,338	0,137	0,893	0,802	-0,091	0,358	0,339	-0,019
17*	0,395	0,438	0,043	0,614	0,567	-0,047	0,500	0,515	0,015
18	0,415	0,338	-0,077	0,303	0,326	0,023	0,838	0,835	-0,003
19	0,612	0,699	0,087	0,607	0,540	-0,067	0,310	0,308	-0,002
20	0,520	0,539	0,019	0,495	0,472	-0,023	0,522	0,531	0,009
130°C									
15	0,629	0,611	-0,018	0,402	0,441	0,039	0,527	0,500	-0,027
16*	0,201	0,243	0,042	0,892	0,846	-0,046	0,358	0,373	0,015
17	0,395	0,382	-0,013	0,613	0,622	0,009	0,500	0,502	0,002
18	0,414	0,404	-0,010	0,302	0,286	-0,016	0,837	0,865	0,028
19*	0,611	0,573	-0,038	0,606	0,690	0,084	0,310	0,249	-0,061
20	0,520	0,519	-0,001	0,494	0,521	0,027	0,522	0,491	-0,031

(*) As amostras não fazem parte dos modelos

(a) Ver rodapé da tabela 16.

Tabela 26 - Comparação dos erros padrão de previsão[a] em (mg) para os conjuntos de teste, utilizando 3, 4 e 5 componentes para a modelagem[b].

número de componentes	Tolueno	Isoctano	Etanol
	105°C		
3	0,012	0,016	0,017
4	0,013	0,015	0,017
5	0,009	0,015	0,017
	120°C		
3	0,059	0,055	0,017
4	0,016	0,030	0,021
5	0,020	0,030	0,020
	130°C		
3	0,014	0,029	0,029
4	0,013	0,031	0,035
5	0,013	0,023	0,023

(a) Ver rodapé da tabela 16

(b) As amostras com (*) na tabela 25 não entraram no cálculo do erro padrão de previsão porque não fazem parte dos modelos.

quanto que para os outros dois podem ser usadas apenas três componentes.

Uma das vantagens do uso da regressão em componentes principais é que o método pode ser usado para identificar amostras que não se ajustam ao modelo de calibração, como é o caso das amostras com asterísco na tabela 25. Estas amostras não foram incluídas no cálculo dos erros padrão de previsão na tabela 26.

CONSIDERAÇÕES FINAIS

A capacidade do método de regressão de componentes principais para prever as composições das amostras de teste para cromatogramas com alto grau de superposição foi demostrada.

Para os conjuntos de dados simulados demonstrou-se que nos casos de alto grau de superposição não é possível usar a regressão linear múltipla convencional, devido às altas correlações entre as variáveis. Nesta situação a inversa de $\underset{\sim}{x}'\underset{\sim}{x}$ na equação $\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$ não existe. No caso da regressão de componentes principais isto não é problema, porque as componentes principais são ortogonais entre si, ou seja,não são correlacionadas e as redundâncias matemáticas são eliminadas do cálculo.

Outra vantagem é que, uma vez que o modelo tenha sido determinado, é possível classificar novas amostras como similares ou não ao conjunto de calibração. Isto significa que a informação é obtida mesmo que o conjunto de calibração não seja qualificado para determinar a composição de uma nova amostra.

Um aspecto crítico na calibração multivariada é determinar o número ótimo de componentes principais para a regressão.

Alguns métodos foram discutidos nas referências (54,55). Neste trabalho usamos um outro método discutido em (56), onde o número ótimo de componentes principais é escolhido a fim de minimizar os erros padrão de previsão.

CAPÍTULO V

CONCLUSÃO GERAL

A relação das Universidades e indústrias apresentada no capítulo III mostra que a aplicação dos métodos quimiométricos no Brasil está-se expandindo. Com a adaptação e modificação dos programas para microcomputadores tornou-se perfeitamente viável o uso destes métodos nas pequenas Universidades, uma vez que na maioria delas não havia um computador de grande porte a que o pesquisador pudesse ter acesso. O pacote computacional ARTHUR adaptado e modificado para microcomputadores agiliza a aplicação destes métodos porque tendo os programas em linguagem FORTRAN o pesquisador pode adequá-los ao tipo de microcomputador disponível em seu laboratório.

Isto é importante porque os métodos de reconhecimento de padrões cada vez mais estão ganhando aceitação entre os químicos analíticos, o que fica claro pelo grande número de publicações relatando aplicações destes métodos.

As três primeiras aplicações efetuadas neste trabalho mostram a importância do emprego destes métodos para fazer análise multivariada. As informações obtidas destes conjuntos não poderiam ser extraídas da simples inspeção dos dados resultantes da análise química. Embora a aplicação destes métodos pareça muito complicado à primeira vista, pode-se chegar a relações muito simples como no caso da relação da eletronegatividade para os metanos halogenados. Aliás, um dos objetivos de quimiometria é determinar procedimentos simples para resolver problemas difíceis.

A implementação do programa VARIMAX foi muito importante,

pois ajudou a extrair dos dados informações úteis que não podiam ser visualizadas apenas com a análise de componentes principais.

O programa PCR, também implementado neste trabalho, pode ser considerado um método alternativo de calibração multivariada, próprio para casos onde a regressão linear múltipla convencional não possa ser usada devido às restrições já discutidas neste trabalho. Com o uso crescente de aparelhos de aquisição automática de dados multivariados em laboratórios químicos, estes casos tendem a ser mais numerosos no futuro.

REFERÊNCIAS

1. B.R. Kowalski, Trends Anal. Chem., 1 (1981) 71.
2. B.R. Kowalski, Anal. Chem., 52 (1980) 112R.
3. I.E. Frank e B.R. Kowalski, Anal. Chem., 54 (1982) 232R.
4. M.F. Delaney, Anal. Chem., 56 (1984) 261R.
5. L.S. Ramos, K.R. Beebe, W.P. Carey, E. Sãnchez M., B.C. Erickson, B.E. Wilson, L.E. Wangen e B.R. Kowalski, Anal. Chem., 58 (1986) 294R.
6. S.D. Brown, T.Q. Barker, R.J. Lariver, S.L. Monfre e H.R. Wilk, Anal. Chem. 60 (1988) 252R.
7. J.C. Davis, "Statistics and Data Analysis in Geology", John Wiley & Sons, New York, 1973.
8. K. Esbensen, L. Lindqvist, I. Lundholm, D. Nisca e S. Wold, Chemometrics and Intelligent Laboratory Systems, 2 (1987) 161.
9. E.C. Pielou, "The Interpretation of Ecological Data", John Wiley & Sons, New York, 1984.
10. I.S. Scarminio, Tese de Mestrado, Universidade Estadual de Campinas, 1981.
11. P.S. de Souza, Tese de Mestrado, Universidade Estadual de Campinas, 1986.
12. M.C.U. de Araũjo, Tese de Doutorado, Universidade Estadual de Campinas, 1987.
13. R.J. Poppi, Tese de Mestrado, Universidade Estadual de Campinas, 1989.
14. R. Vergili Junior, Tese de Mestrado, Universidade Estadual de Campinas, 1988.

15. D.L. Duewer, J.R. Koskinen e B.R. Kowalski, ARTHUR, disponível por B.R. Kowalski, Department of Chemistry, BG-10, University of Washington, Seattle, WA 98195.

16. S. Wold, C. Albano, W.J. Dunn III, U. Edlund, K. Ebensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg e M. Sjöström, "Proceedings of NATO Advanced Study Institute on Chemometrics Mathematics and Statistics in Chemistry", (B.R. Kowalski ed.), Riedel, 1983, 17.

17. S. Wold, Pattern Recognition, 8 (1976) 127.

18. B.R. Kowalski e C.F. Bender, J. Am. Chem. Soc., 9 (1972) 5632.

19. D.L. Duewer, B.R. Kowalski e J.L. Fashing, Anal. Chem., 48 (1976) 2002.

20. M. Mellinger, Chemometrics and Intelligent Laboratory Systems, 2 (1987) 29.

21. K.R. Beebe e B.R. Kowalski, Anal. Chem., 59 (1987) 1007 A.

22. S. Wold, K. Esbensen e P. Geladi, Chemometrics and Intelligent Laboratory Systems, 2 (1987) 37.

23. K.J. Jöreskog, J.E. Klavan e R.A. Reyment, "Geological Factor Analysis", Elsevier Scientific Publishing Company, Amsterdam, 1976.

24. S. Wold, Technometrics, 20 (1978) 397.

25. W.W. Cooley e P.R. Lohnes, "Multivariate Data Analysis", John Wiley & Sons, New York, 1971.

26. M.A. Sharaf, D.L. Illman e B.R. Kowalski, "Chemometrics", John Wiley & Sons, New York, 1986.

27. W.O. Kwan e B.R. Kowalski, J.Food Sci., 43 (1978) 1320.

28. C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Nordēm, M. Sjöström, S. Wold, Anal. Chim. Acta, 103 (1978) 429.

29. M. Sjöström e B.R. Kowalski, Anal. Chim. Acta, 112 (1979) 11.

30. D. Coomans, D.L. Massart e L. Kaufman, Anal. Chim. Acta, 112 (1979) 97.

31. D. Coomans, D.L. Massart, e I. Broeckaert e A. Tassin, Anal. Chim. Acta, 133 (1981) 215.

32. D. Coomans e D.L. Massart, Anal. Chim. Acta, 133 (1981) 225.

33. H. Van Der Voet e D.A. Doornbos, Anal. Chim. Acta, 161 (1984) 115.

34. R.E. Bruns e J.F.G. Faigle, Quim. Nova, 2 (1985) 84.

35. M.P. Derde, D. Coomans e D.L. Massart, Anal. Chim. Acta, 141 (1982) 186.

36. C. Albano, G. Blomqvist, D. Coomans, W.J. Dunn III, U. Edlund, B. Eliasson, S. Hellberg, E. Johansson, B. Nordém, D. Johnels, M. Sjöström, B. Söderström, H. Wold e S. Wold, "Proceedings of Symposium on Applied Statistics", (A. Hoskuldsen, K. Esbensen et al eds.), Copenhagen, 1981, 183.

37. K. Varmuza, Anal. Chim. Acta, 122 (1987) 227.

38. D. Coomans e D.L. Massart, Anal. Chim. Acta, 136 (1982) 15.

39. P. Geladi e B.R. Kowalski, Anal. Chim. Acta, 185 (1986) 1.

40. T. Naes e H. Martens, Trends Anal. Chem., 3 (1984) 266.

41. M. Sjöström, S. Wold, W. Lindberg, J. Persson e H. Martens, Anal. Chim. Acta, 150 (1983) 61.

42. P. Geladi e B.R. Kowalski, Anal. Chim. Acta, 185 (1986) 19.

43. W. Lindberg, J. Persson e S. Wold, Anal. Chem., 55 (1983) 643.

44. S. Wold, P.Geladi, K. Esbensen e J. Öhman, Journal of Chemometrics, 1 (1987) 41.

45. "Standard Methods for the Examination of Water and Waste Water, 14ª ed., APHA/AWWA/WPCE, New York, 1975.

46. J. Ruzicka e E.H. Hansen, "Flow Injection Analysis", 2ª ed., John Wiley & Sons, New York, 1988.

47. L.C.R. Pessenda, J.R. Ferreira, A.C.F.N.S. Tancredi, L.A. Martinelli, R. Hirata e J. Mortatti, Acta Limnológica Brasileira, 1 (1986) 179.

48. J. Mortatti, Tese de Doutorado, Escola Superior de Agricultura Luiz de Queiroz - USP, 1986.

49. B.B. Neto, M.M.C. Ferreira, I.S. Scarminio e R.E. Bruns, J. Phys. Chem., aceito para publicação.

50. J. Overend, "Infrared Spectroscopy and Molecular Structure", (M. Davis ed.), Elsevier, Amsterdam, 1963.

51. W.B. Person e J.H. Newton, J. Chem. Phys., 61 (1974) 1040.

52. S. Kondo, T. Nakanaga e S. Saeki, J. Chem. Phys., 73 (1980) 5409.

53. R.D.B. Frazer e E. Suzuki, Anal. Chem., 41 (1969) 37.

54. D.M. Haaland, Anal. Chem., 60 (1988) 1208.

55. D.M. Haaland e E.V. Thomas, Anal. Chem., 60 (1988) 1193.

56. J.A. Cowe, S. Koester, C. Paul, J.W. McNicol e D.C. Cuthbertson, Chemometric and Intelligent Laboratory Systems, 3 (1988) 233.

APÊNDICE

MANUAL DO USUÁRIO

SISTEMA COMPUTACIONAL ARTHUR
PARA MICROCOMPUTADORES

PROF.DR. ROY E. BRUNS

PROFA. IEDA SPACINO SCARMINIO

GRUPO DE QUIMIOMETRIA
INSTITUTO DE QUÍMICA
UNIVERSIDADE ESTADUAL DE CAMPINAS
CAIXA POSTAL 6154
13083 - CAMPINAS - SP

ÍNDICE

PREFÁCIO

Os programas que fazem parte deste pacote originaram-se do sistema ARTHUR, desenvolvido pelo grupo de pesquisa do Prof. Bruce Kowalski do Departamento de Química da Universidade de Washington, Seattle, WA. Ao adquirir o sistema ARTHUR, o grupo de quimiometria do Instituto de Química da UNICAMP o fez com liberdade de uso e disseminação, ou seja, sem as restrições de "copyright".

O grupo de quimiometria da UNICAMP segmentou o sistema e modificou os principais subprogramas adaptando-os para microcomputadores com sistemas operacionais CPM e DOS.

Sendo este pacote computacional aberto (linguagem FORTRAN) consideramos que seu valor principal é didático, permitindo a divulgação de métodos quimiométricos no Brasil. O pacote é distribuido gratuitamente pelo grupo de quimiometria nas versões FORTRAN e em módulos executáveis. Programas comerciais contendo os métodos que estão incluídos em nosso pacote podem ser adquiridos de:

INFOMETRIX, INC.
P.O. Box 25888
Seattle, WA 98125

ou

Principal Data Components
2505 Shepard Blvd.
Columbia, MO 65201

Os autores gostariam de agradecer o estímulo e apoio de Rio Doce Geologia e Mineração S.A., durante o desenvolvimento deste trabalho.

DEFINIÇÕES

CATEGORIA - Conjunto de objetos tendo uma mesma propriedade. Para dados com propriedades contínuas, NCAT = 1.

CONJUNTO DE TREINAMENTO - É um subconjunto do conjunto de dados tendo propriedades e categorias conhecidas, usadas para desenvolver regras de classificação.

CONJUNTO DE TESTE - É o subconjunto dos dados tendo propriedades e categorias conhecidas, usadas para testar a eficácia das regras de classificação desenvolvidas a partir do conjunto de treinamento.

ARQUIVO DE DADOS

Os dados originais devem ser armazenados no arquivo FORT10.DAT, que deve conter para cada amostra, com os formatos indicados entre parenteses, o seguinte:

a. ID, um número para identificação (I4);

b. NAME, um nome com até 8 caracteres para identificação (2A4);

c. CN, o número da categoria (F2.0);

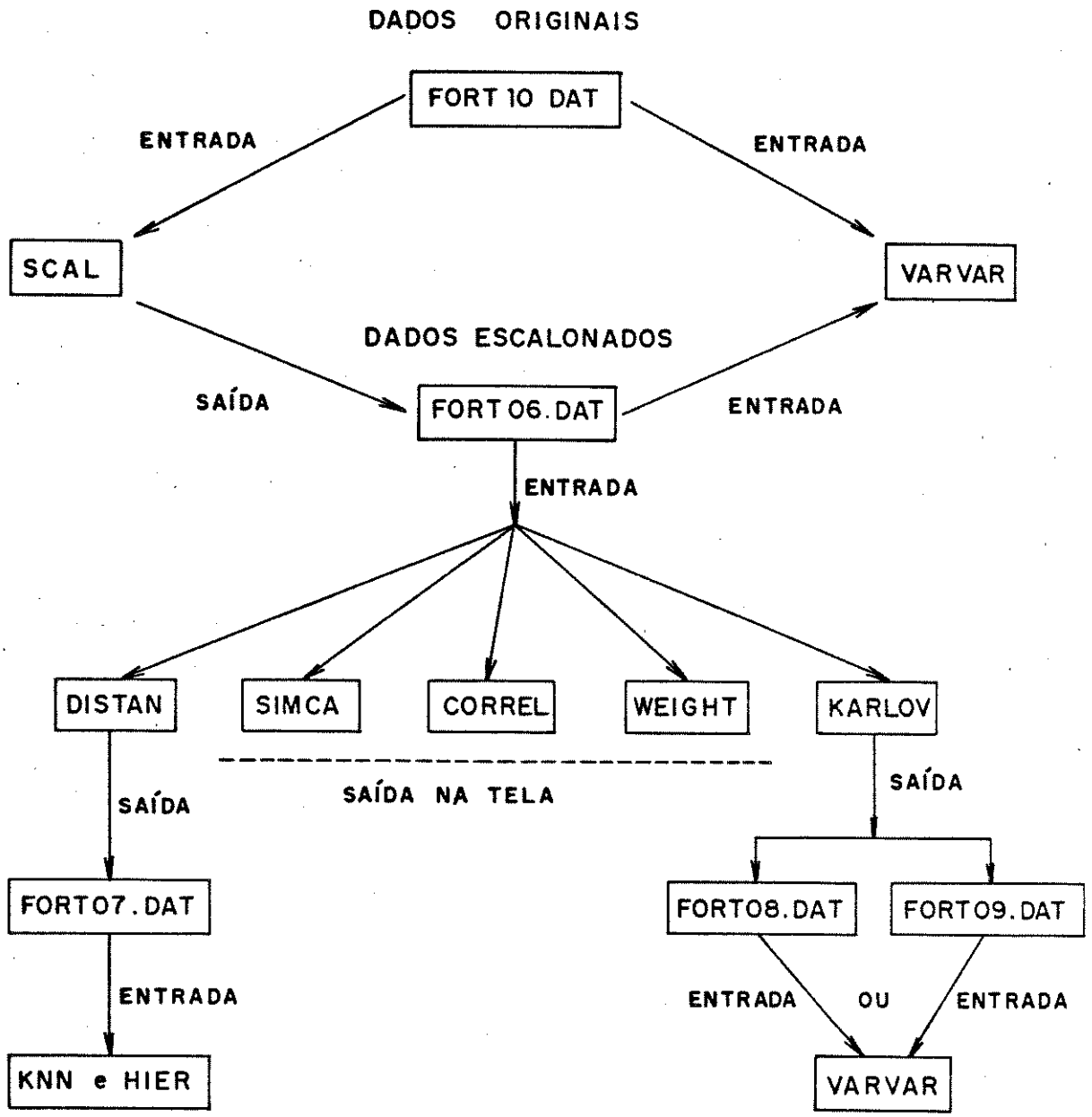
d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

Notar que no formato 6F10.5 e 8F10.5 existe uma barra, então no caso de completar um registro com 6, 14, 22, etc. variáveis, deve-se deixar um registro completo (80 colunas) em branco.

FLUXOGRAMA

PROGRAMAS E ARQUIVOS



CORREL

FINALIDADE DO PROGRAMA

Este programa gera todas as correlações, característica-característica (variável-variável) e característica-propriedades, com intervalos de confiança para as correlações e uma estimativa da probabilidade de que os dados sejam originários de uma população sem nenhuma correlação. As covariâncias inter-características também podem ser listadas. As correlações podem ser obtidas usando o conjunto original ou o escalonado. Pode-se optar também por usar todo o conjunto de dados, somente os dados de uma categoria ou os dados de um grupo de categorias.

BIBLIOGRAFIA

Qualquer manual de estatística contendo os conceitos básicos de covariância e correlação.

O.L. Davies e P.L. Goldsmith, "Statistical Methods for Research and Production", Hafner (1972) p. 234.

P.R. Bevington. "Data Reduction and Error Analysis for the Physical Sciences", McGraw-Hill (1969) p. 123.

DADOS DE ENTRADA

Digitar os seguintes parâmetros na tela:

IARQ = número do arquivo de entrada;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto teste;
NVAR = número de variáveis;
NCAT = número de categorias;
NPNT = 1 se quizer imprimir a matriz de covariância; 0 se não;
NCO = 1 correlação para todo o conjunto de dados; 0 correlação por categoria ou grupo de categorias;
TS = o valor de t de Student para o intervalo de confiança dos valores da correlações;

NCC = número de categoria desejada; para terminar digitar zero.

(NA(J), J = 1, NVAR) = nome das características ou variáveis.

O arquivo FORT10.DAT (dados originais) ou FORT06.DAT (dados escalonados), é lido automaticamente pelo programa e contém para cada amostra:

a. ID, um número para identificação (I4);

b. NAME, um nome com até 8 caracteres para identificação (2A4);

c. CN, o número da categoria (F2.0);

d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

DEFINIÇÕES

1 - Covariância

$$COV_{i,j} = \sum_{k=1}^{NPAT} (x_{i,k} - \bar{x}_i)(\bar{x}_{j,k} - \bar{x}_j)/(NPAT - 1)$$

$$\bar{x}_i = \sum_{k=1}^{NPAT} x_{i,k}/NPAT$$

2 - Correlação de Propriedade (Correlação a Propriedade)

$$COR(p)_i = \frac{\sum_{k=1}^{NPAT} (x_{i,k} - \bar{x}_i)(p_k - \bar{p})}{\left\{ \sum_{k=1}^{NPAT} (x_{i,k} - \bar{x}_i)^2 \sum_{k=1}^{NPAT} (p_k - \bar{p})^2 \right\}^{1/2}}$$

$$\bar{p} = \sum_{k=1}^{NPAT} p_k / NPAT$$

3 - Correlações Inter-Variáveis ou Inter-Características

$$COR_{i,j} = \frac{\sum_{k=1}^{NPAT} (x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)}{\left\{ \sum_{k=1}^{NPAT} (x_{i,k} - x_i)^2 \sum_{k=1}^{NPAT} (x_{j,k} - \bar{x}_j)^2 \right\}^{1/2}}$$

4 - LO, HI (Intervalo de Confiança em torno da Correlação

Este intervalo de confiança é obtido por meio da transformação z de R.A. Fisher (Veja referência de Davies).

LO = tanh (z - sig)
HI = tanh (z + sig)

$$z = \tanh^{-1}(COR_{i,j})$$

$$sig = TS\sigma_z$$

$$\sigma_z = (1.0/(NPAT - 3)^{1/2}$$

5 - Probabilidade de que os dados pertençam a uma população com correlação zero (Veja referência de P.R. Bevington).

$$OPROP = 2\int_{|r|}^{1} P_r(r,v)dr$$

$$P_r(r,v) = \frac{1}{\sqrt{\pi}} \frac{\sqrt{(v+1)/2}}{\sqrt{(v/2)}} (1 - r^2)^{(v-2)/2}$$

$v = NPAT - 2$

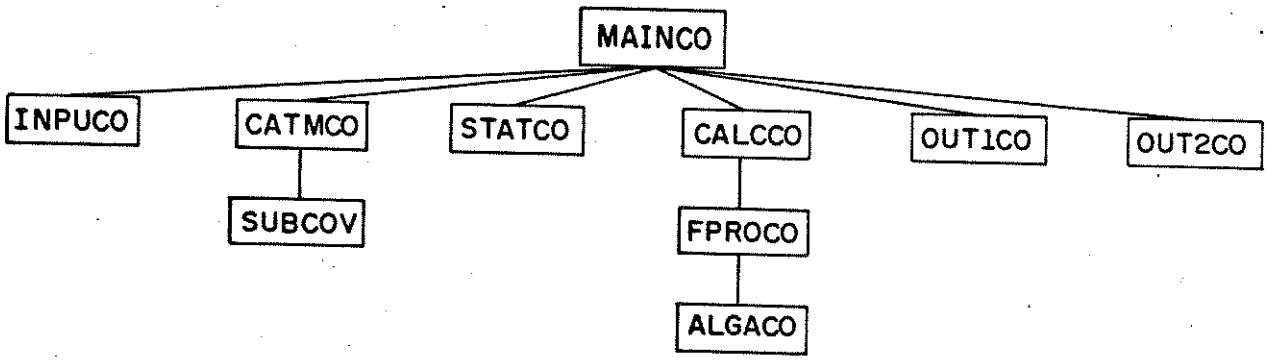
$r = COR_{i,j}$

ESTRUTURA DO PROGRAMA

1 - Subrotinas

MAINCO: faz chamadas para as outras subrotinas
INPUCO: lê os dados de entrada
CATMCO: seleciona as categorias
STATCO: cálculo da covariância
CALCCO: cálculo das correlações e dos intervalos de confiança para estas correlações
ALGACO: cálculo do logarítmo natural da função gama
FPROCO: cálculo da probabilidade, OPROP
OUT1CO: saída da matriz de covariância
OUT2CO: saída da matriz de correlação e de seus valores associados
SUBCOV: recompõe o conjunto de dados original

2 - Organização



3 - Detalhes

a. Os arquivos FORT02.DAT e FORT03.DAT são usados como arquivos de trabalho.

b. O número de variáveis que o programa pode tratar (NVAR) é determinado pelo parâmetro MAX:

$$MAX \geq 0.5 * NVAR * NVAR + 8.5 * NVAR + 6$$

A memória de CPU usada pelo programa pode ser expandida, modificando-se as dimensões das matrizes X e Na e o valor de MAX no programa MAINCO. No PC as dimensões X e Na devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX.

DISTAN

FINALIDADE DO PROGRAMA

Este programa calcula a matriz das distâncias, que tem ordem NPAT x (NPAT + NTEST). A matriz é armazenada em um arquivo no disco para ser usada por outros subprogramas. Vários tipos de distância podem ser calculados: Mahalanobis generalizada, Manhattan (por quarteirão) e distância de razão de Anders. As distâncias de Mahalanobis e de Manhattan serão mais siginificativas se forem calculadas a partir de dados autoescalonados.

BIBLIOGRAFIA

B.R. Kowalski e C.F. Bender, J. Am. Chem. Soc. 94, 5632 (1972).

O.V. Anders, Anal. Chem. 44, 1930 (1972).

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada;
NARQ = número do arquivo de saída;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;
NCAT = número de categorias.

- < 0 para calcular a distância de razão de Anders
- = 0 para calcular a distância Euclidiana (distância de Mahalanobis de ordem 2)
- = 1 para calcular a distância de Manhattan
- > 1 para calcular a distância de Mahalanobis de ordem N

XLF = Limite inferior para a distância de razão de Anders. Estimativa inicial = 0.667

XHF = Limite superior para a distância de razão de Anders. Estimativa inicial = 1.5

O arquivo FORT06.DAT, que é lido automaticamente pelo programa, contém para cada amostra:

a. ID, um número para identificação (I4);
b. NAME, um nome com até 8 caracteres para identificação (2A4);
c. CN, o número da categoria (F2.0)
d. X(J), J = 1, NAVR; os valores das variáveis; da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

Os resultados são armazenados no arquivo FORT.07 DAT, que serve de entrada para os subprogramas KNN (regra do vizinho mais próximo) e HIER (análise de agrupamentos), na seguinte forma:

ID_k, $NAME_k$, CN_k, $d_{k1}d_{k2}d_{kj}d_k$, NPAT

onde d_{kj} representa a distância entre a k-ésima e j-ésima amostras, e k = 1,2..., (NPAT + NTEST).

FORMAT (1X, I5, 2X, 2A4, 2X, F6.2, 10E10.3/10 (1X, 10E10.3/)).

DEFINIÇÕES

1 - Distância de Mahalanobis de Ordem N

$$D\ (N)_{k,j} = \left\{ \sum_{i=1}^{NVAR} (X_{i,k} - X_{i,j})^N \right\} 1/N$$

2 - Distância de Manhattan

$$D_{k,j} = \sum_{i=1}^{NVAR} |x_{i,k} - x_{i,j}|$$

3 - Distância de Razão de Anders

$$D_{k,j} = \left\{ \sum_{i=2}^{NVAR} \sum_{1=1}^{i-1} M_{k,j,i,1} \right\} / (NVAR(NVAR - 1)/2)$$

$M_{k,j,i,1}$ = 0 quando $XLF < R_{k,j,i,1} < XHF$

= 1 quando $(XLF > R_{k,j,i,1})$ ou $(XHF < R_{k,j,i})$

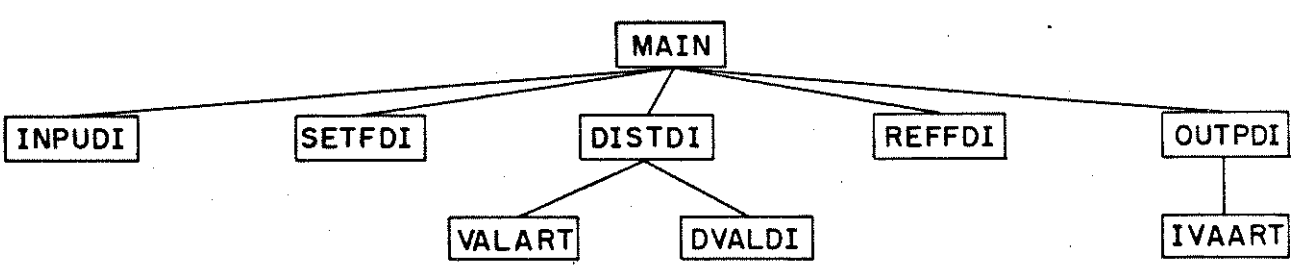
$$R_{k,j,i,1} = \frac{(x_{i,k}/x_{1,k})}{(x_{i,j}/x_{1,j})}$$

ESTRUTURA DO PROGRAMA

1 - Subrotinas e Funções

MAINDI = chama as subrotinas
INPUDI = lê os dados de entrada
SETFDI = manipulação dos arquivos em disco
DISTDI = cálculo da matriz das distâncias
DVALDI = cálculo dos elementos da matriz das distâncias
REFFDI = manipulação de arquivos
OUTPDI = saída da matriz das distâncias

2 - Organização do Programa



3 - Detalhes

a. Os arquivos FORT02.DAT, FORT03.DAT e FORT04.DAT são usados como arquivos de trabalho.

b. o número de variáveis (NVAR) e amostras (NPAT) que o programa pode tratar é determinado pelo parâmetro MAX:

$$MAX \geq NPAT + 3 * NVAR$$

A memória de CPU usada pelo programa pode ser expandida, modificando-se as dimensões das matrizes X e NA e o valor de MAX no programa MAIN. No PC as dimensões X e NA devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX.

HIER

FINALIDADE DO PROGRAMA

Este programa produz um dendrograma que descreve o agrupamento hierárquico (as vezes chamado agrupamento de modo Q) das amostras de um conjunto de treinamento. Este dendrograma liga grupos de amostras com mesmos níveis de similaridade. O dendrograma pode ser usado para agrupar as amostras em qualquer número de agrupamentos e também para determinar quantos agrupamentos existem para qualquer nível de similaridade. A similaridade entre objetos é definida usando-se as distâncias entre os pontos representando as amostras no espaço.

PRÉ-REQUISITOS

A matriz das distâncias inter-amostras deve ser gravada no arquivo FORT07.DAT (pelo programa DISTAN).

BIBLIOGRAFIA

B.R. Kowalski e C.F. Bender, J. Am. Chem. Soc. 94 5632 (1972).

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada

IWAIT =
- 0 se cada amostra tiver peso igual na determinação dos níveis de "linkage" independente do tamanho do grupo a que pertença.
- 1 se cada grupo tiver peso igual na determinação dos níveis de "linkage" independente do número de amostras que estejam contidas no grupo.

IPULL = { 0 se o número de seções em que o dendrograma será impresso for determinado pelo programa.
1 se o dendrograma vai ser impresso em IPULL seções (máximo de 3).

O aquivo de dados FORT07.DAT produzido pelo programa DISTAN é automaticamente lido pelo programa HIER.

DEFINIÇÕES

1 - Similaridade = 1.0 - $D_{i,j}$/DMAX

Onde DMAX = o valor máximo na matriz das distâncias

2 - Método de Agrupamento Usando Pesos Iguais para as Amostras

$$S_{new} = \frac{(NUM_{1,old})(S_{1,old}) + (NUM_{2,old})(S_{2,old})}{(NUM_{1,old} + NUM_{2,old})}$$

Onde:

$NUM_{i,old}$ = número de amostras dentro do agrupamento representado por $S_{i,old}$

$S_{i,old}$ = grupos escolhidos para serem aglomerados neste ciclo do cálculo.

3 - Método de Agrupamento Usando Pesos Iguais para os Grupos

$$S_{new} = \frac{S_{1,old} + S_{2,old}}{2}$$

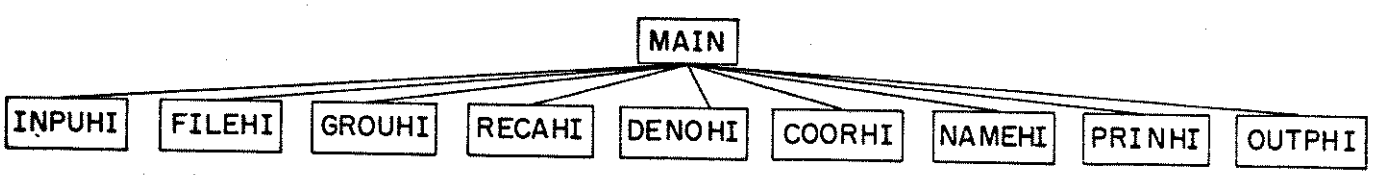
ESTRUTURA DO PROGRAMA

1 - Subrotinas

MAIN = dados de entrada e chamadas para as subrotinas

INPUHI = lê os dados de entrada inseridos através do teclado
FILEHI = iniciação dos arquivos
GROUHI = agrupamentos
RECAHI = novo cálculo das distâncias
DENOHI = formação do dendrograma
COORHI = coordenadas do dendrograma para impressora
NAMEHI = leitura, em matrizes, dos identificadores das amostras
PRINHI = saída do dendrograma
OUTPHI = saída

2 - Organização



3 - Detalhes

a. Somente as primeiras NPAT amostras serão aglomeradas. Certifique-se de que você definiu o conjunto de treinamento de modo a incluir todas as amostras que voce gostaria de aglomerar. O algoritmo implementado neste programa usa algumas simplificações computacionais para reduzir o tempo de cálculo. Os agrupamentos serão quase iguais aqueles formados por agrupamento hierárquico verdadeiro, mas é possível que os níveis de similaridade sejam diferentes.
b. O número de variáveis (NVAR) ou amostras (NPAT) que o programa pode tratar é determinado pelo parâmetro MX:

MAX ≥ 12 * NPAT

A memória de CPU, usada pelo programa pode ser expandida, modificando-se as dimensões das matrizes X e NA e o valor de MAX no programa MAIN. No PC as dimensões X e NA devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX

KARLOV

FINALIDADE DO PROGRAMA

Novas variáveis são geradas como combinações lineares das variáveis originais. As novas variáveis são linearmente independentes e são colocadas em ordem decrescente de variância. As novas características podem ser grafadas, dando projeções dos autovetores ou componentes principais dos dados. Os componentes principais também podem ser usados para reduzir o número de variáveis perdendo o mínimo de informação estatística.

REQUISITOS

Para projeções de dados não tendenciosos use dados autoescalonados. As vezes, projeções de dados ponderados com os pesos de Fisher ou de variância podem ser interessantes.

BIBLIOGRAFIA

E.R. Malenowski e D.G. Howery, "Factor Analysis in Chemistry", Wiley - Interscience, New York, 1980.

J.B. Kruskal, "Factor Analysis and Principal Components Bilinear Methods" in J.B. Kruskal e Tannurs, edit., "International Encyclopedia of Statistics", Vol. I, McMillan Press, New York, 1978.

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;
NVAR = número de variáveis
NARQ = número do arquivo para a saída dos escores;

NSARQ = número do arquivo para a saída dos "loadings".

Para aguardar os "loadings" dos primeiros (no máximo 10) N componentes principais no arquivo FORT09.DAT, digite N.

FORMAT (I4, 8X, I2, 6X, 6F10.5/12 (8F10.5/)).

O arquivo FORT06.DAT é lido automaticamente pelo programa, e contém para cada amostra:

a. ID, um número para identificação (I4);
b. NAME, um nome com até 8 caracteres para identificação (2A4);
c. CN, número da categoria (F2.0);
d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

Os dados transformados pelo método de Karhunen-Loeve são armazenados no arquivo FORT08.DAT.

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

DEFINIÇÕES

1 - Autovalor

λ_k = o k-ésimo autóvalor da matriz de covariância. Os autovalores são colocados em ordem decrescente

2 - Informação preservada num componente

$$INFO\ (C)_k = (\lambda_k)\ (100.0)/\sum_{i=1}^{NVAR} \lambda_i$$

3 - Informação preservada - total acumulado

$$INFO(T)_k = \left(\sum_{i=1}^{k} \lambda_i\right)\ (100.0)/\sum_{i=1}^{NVAR} \lambda_i$$

4 - Autovetores

β_k^t = o k-ésimo autovetor da matriz de covariância colocado em ordem decrescente do autovalor associado.

5 - Transformação Karhunen-Loeve

$\underset{\sim}{\theta} = \underset{\sim}{X}\underset{\sim}{\beta}^t$ onde

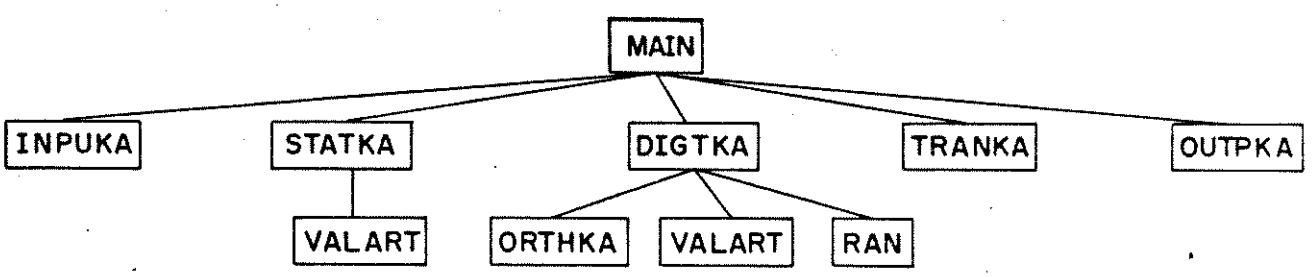
$\underset{\sim}{\beta}^t$ = matriz de transformação com os autovetores postos em colunas na ordem decrescente de autovalor.

ESTRUTURA DO PROGRAMA

1 - Subrotinas

MAIN = faz chamadas das subrotinas
STATKA = calcula a matriz de covariância
DIGTKA = cálculo de autovalores e autovetores
ORTHKA = faz parte de método de tridiagonalização para DIGTKA
TRANKA = faz a transformação Karhunen-Loeve
OUTPKA = saída dos resultados (autovalores e autovetores)
VALART = preenche o vetor A de comprimento N com valores reais.
RAN = gerador de números aleatórios

2 - Organização



3 - Detalhes

a. Este programa usa os arquivos FORT02.DAT, FORT03.DAT, FORT04.DAT e FORT05.DAT como arquivos de trabalho.

b. O número de variáveis (NVAR) que o programa pode tratar é determinado pelo parâmetro MAX:

$MAX \geq 12.5 * NVAR + 0.5 * NVAR * NVAR.$

A memória de CPU usada pelo programa pode ser expandida, modificando-se as dimensões das matrizes X e NA e o valor de MAX no subprograma MAIN. No PC as dimensões X e NA devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX.

DESCRIÇÃO DO MÉTODO DE COMPONENTES PRINCIPAIS

A matriz de dados é representada por X:

		variáveis			
		1	2 ..... i		P = NVAR
amostras	1	X_{11}	X_{12}	X_{1i}	X_{1P}
	2	X_{21}	X_{22}	X_{2i}	X_{2P}
	k	X_{k1}	X_{k2}	X_{ki}	X_{kp}
	r = NPAT	X_{r1}	X_{r2}	X_{ri}	X_{rp}

onde os elementos x_{ki} representam os valores da i-ésima variável e da k-ésima amostra. A matriz de covariância (correlação para os dados autoescalonados) é obtida multiplicando-se a matriz de dados pela sua transporta.

$$(\underset{\sim}{COV}) = \underset{\sim}{X}^t \underset{\sim}{X}$$

Essa matriz é posteriormente diagonalizada, for-

necendo como solução:

$$(\underset{\sim}{COV})\ \underset{\sim}{\beta}^t = \underset{\sim}{\beta}^t \underset{\sim}{\lambda}$$

ou

$$(\underset{\sim}{COV})\ \beta_j^t = \lambda_j \vec{\beta}_j^t$$

onde $\vec{\underset{\sim}{\beta}}$ é o autovetor correspondente ao j-ésimo autovalor, compondo a j-ésima coluna na matriz β^t. O j-ésimo autovalor é proporcional a fração da variância original expressa pelo j-ésimo autovetor. Assim a diagonalização da matriz de covariância permite a projeção de um espaço de p variáveis em um outro de ordem inferior, em geral de duas ou três dimensões, a fim de facilitar a visualização, o que é de grande importância em problemas de classificação de dados. Os "A" primeiros autovetores, quando ordenados de forma decrescente segundo seus autovalores, expressam uma percentagem de variância dada por:

$$\left(\sum_{i=1}^{A} \lambda_i / \sum_{i=1}^{P} \lambda_i\right) 100$$

O valor de A costuma ser pequeno nos problemas classificatórios, ficando em geral em torno de dois ou três, dependendo da complexidade do conjunto de dados. Em problemas numéricos, o valor de A é determinado pelo número de componentes principais necessários para descrever o conjunto de dados. Os outros (A-p) autovetores não são siginificativos, uma vez que expressam uma percentagem de variância comparável aos valores dos erros experimentais. A matriz $\underset{\sim}{X}$ é então descrita por:

$$\underset{\sim}{X} = \underset{\sim}{\theta}\ \underset{\sim}{\beta}$$

onde os autovetores contidos em $\underset{\sim}{\beta}$ são chamados "loadings" e os valores de $\underset{\sim}{\theta}$ são chamados "scores", ou seja, valores das variáveis de n amostras, após a transformação para o espaço descrito pelos autovetores. Os valores de $\underset{\sim}{\theta}$ podem ser calculados pela equação:

$$\underset{\sim}{X}\ \underset{\sim}{\beta}^t = \underset{\sim}{\theta}$$

como a matriz $\underset{\sim}{\beta}$ é ortogonal, $\underset{\sim}{\beta}^t = \underset{\sim}{\beta}^{-1}$.

As matrizes θ e β apresentam alguns fatores abstratos, cuja importância está em predizer o número de componentes principais que determina a estrutura dos dados da matriz X. Normalmente, procura-se obter uma matriz de transformação apropriada, mediante a rotação dos fatores abstratos no espaço p ou através do "target testing", o que converte tais fatores em fatores reais.

KNN

FINALIDADE DO PROGRAMA

Este programa faz a classificação de amostras baseada na regra do vizinho mais próximo. As amostras devem pertencer a categorias discretas. O número de vizinhos mais próximos utilizado pelo programa é igual a K = 1, 3, 4, 5, 6, 7, 8, 9 e 10. A proximidade é definida com base nas distâncias entre amostras.

PRÉ-REQUISITOS

A matriz das distâncias inter-amostras deve estar armazenada no arquivo FORT07.DAT.

BIBLIOGRAFIA

T.M. Cover e P.E. Hart, IEEE Trans. on Info. Theory, IT-13, 21 (1967).

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;
NCAT = número de categorias
NVAR = número de variáveis.

O arquivo de dados FORT07.DAT produzido pelo programa DISTAN é lido automaticamente pelo programa KNN.

DEFINIÇÕES

1 - 1-NN = o número da categoria da amostra mais próxima da amos-

tra sendo classificada (isto é, a que corresponde ao menor D_{ij}, onde D_{ij} = distância entre as i-ésima e j-ésima amostras.

2 - K-NN = o número da categoria que é representada com mais frequência nas K amostras mais próximas à amostra sendo classificada, sendo K = 3, 4, 5, 6, 7, 8, 9 e 10. Nos casos em que as frequências de duas ou mais categorias são iguais, a categoria com a menor soma das distâncias será escolhida.

3 - Total Missed = Para um dado K-NN, o número total de amostras que foram classificadas erroneamente. Este resultado é dado para os conjuntos de treinamento e de teste.

4 - "Percent Correct"

a. Para o conjunto de treinamento:

% = [NPAT - (número de erros)] (100.0)/NPAT

Onde:

NPAT = número de amostras no conjunto de treinamento.

b. Para o conjunto teste:

% = [NTEST - (número de erros)] (100.0)/NTEST

Onde:

NTEST = número de amostras no conjunto de teste.

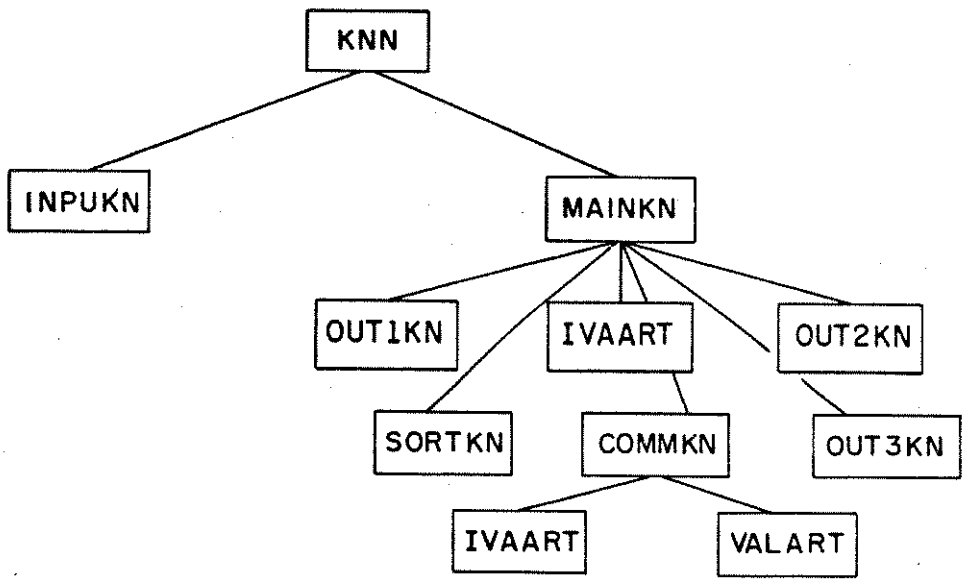
ESTRUTURA DO PROGRAMA

1 - Subrotinas

KNN = chama as subrotinas
INPUKN = lê os dados de entrada
MAINKN = chama as subrotinas
OUT1KN = dados de saída
SORTKN = sorteia os dez vizinhos mais próximos
COMMKN = contagem das categorias dos vizinhos mais próximos

OUT2KN = saída, resultado para cada amostra
OUT3KN = saída, resumo dos resultados

2 - Organização



3 - Detalhes

a. Além dos resultados de classificação também são impressos os índices e a distância dos dez vizinhos mais próximos.

b. O número de amostras (NPAT) e categorias (NCAT) que o programa pode tratar é determinado pelo parâmetro MAX:

MAX ≥5 * NPAT + 2 * NCAT + 48

A memória de CPU, usada pelo programa pode ser expandida, modificando as dimensões das matrizes X, NA e o valor MAX no subprograma MAIN. No PC as dimensões X e Na devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX.

DESCRIÇÃO DO MÉTODO "REGRA DO VIZINHO MAIS PRÓXIMO" - KNN - (K-NEAREST NEIGHBOR)

A idéia básica do KNN está ilustrada na Figura KNN-1. Utiliza-se um conjunto de treinamento para distribuir os pontos entre as classes. Neste estágio de treinamento, não se de

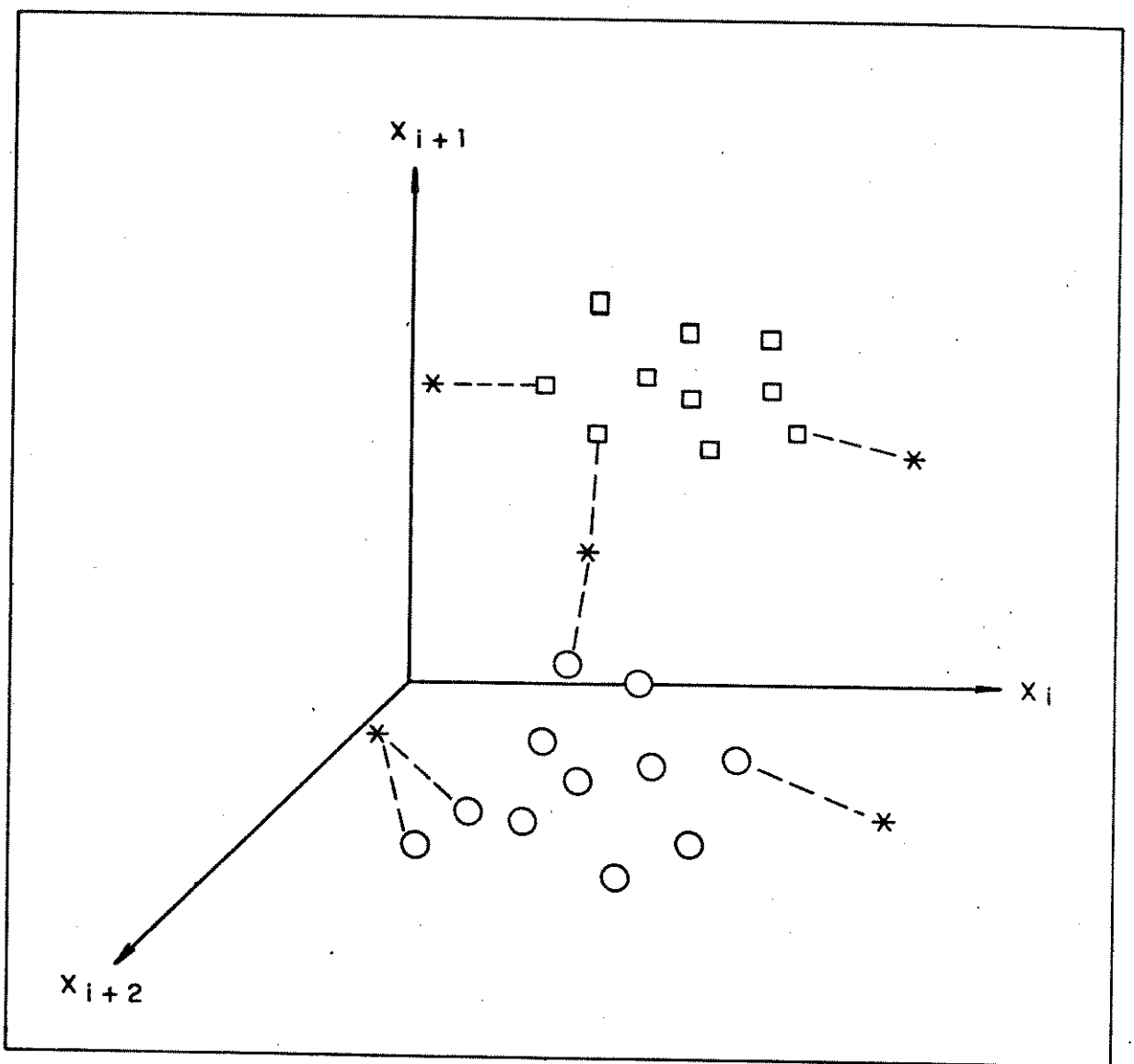


FIGURA KNN-1

termina nenhuma função matemática, ao contrário do que se faz nos métodos da análise discriminante linear (LDA) e da máquina de aprendizagem linear (LLM). A classificação de um ponto é feita na categoria de seu vizinho mais próximo. É interessante utilizar mais que um vizinho mais próximo, já que as rotinas de classificação fornecem normalmente K vizinhos mais próximos, onde K varia de 1 a 10.

Existem muitas vantagens no KNN em relação aos outros métodos de reconhecimento de padrões. Por exemplo, distribuições bimodais que não são linearmente separáveis podem ser classificadas com sucesso pelo KNN. Além disso não é necessário manter a relação entre o número de amostras e variáveis igual a ou maior que 5 como nos casos anteriores. Entretanto o KNN não é capaz de indicar se uma amostra pertence a uma categoria ainda não definida.

A filosofia do KNN é diferente do LDA e do LLM. Estes dois métodos fazem a separação entre categorias, enquanto o KNN é um exemplo de modelo de similaridade. As amostras do conjunto de treinamento mais similares a amostra em estudo servem como referência para sua classificação.

SCAL

FINALIDADE DO PROGRAMA

Este programa transforma os valores de cada variável de acordo com a escala especificada. O autoescalonamento produz variáveis com média zero e variância um. A escala baseada no intervalo (range) dos valores da variável produz uma transformação em que cada variável passa a ter um valor mínimo de zero e o valor máximo de um (range-scaled). Ambos os escalonamentos servem para tornar todas as variáveis iguais em grandeza sem destruir sua informação de discriminação (consideramos o autoescalonamento a melhor transformação). As características das distribuições estatísticas de cada variável (média, desvio padrão, terceiro e quarto momentos, curtose e assimetria) são calculadas. O escalonamento pode ser feito para todo o conjunto de dados, ou então selecionando uma categoria ou um grupo de categorias. No caso do escalonamento por grupo de categorias as amostras do conjunto de treinamento são automaticamente colocadas em ordem crescente de categorias pelo programa. Pode-se também usar esta opção para reordenação de categorias para todo o conjunto de dados.

BIBLIOGRAFIA

B.R. Kowalski e C.F. Bender, J. Am. Chem. Soc. 94, 5632 (1972).

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;
NVAR = número de variáveis;
NCAT = número de categorias;
NESC = = 1 escalonamento de todos os dados (global)
= 0 escalonamento para uma categoria ou um grupo de categorias;

NSCL = 0 se os dados serão escalonados;
= 1 se os dados serão escalonados de acordo com o intervalo de valores (range-scaled);
NARQ = número do arquivo de saída;
NCC = número da categoria desejada; para terminar digitar zero.

O arquivo FORT10.DAT é lido automaticamente pelo programa e deve conter para cada amostra:

a. ID, um número para identificação (I4);
b. NAME, um nome com até 8 caracteres para identificação (2A4);
c. CN, número da categoria (F2.0);
d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

Os dados globais são guardados no arquivo FORT06.DAT, e os dados para categoria ou grupos de categorias no arquivo CATSCA.DAT. Para uso posterior em outros programas deste sistema computacional, o arquivo CATSCA.DAT deve ter o nome mudado para FORT06.DAT. Os formatos dos dados nos arquivos FORT10.DAT e FORT06.DAT são iguais.

DEFINIÇÕES

1 - Média

$$\bar{X}_i = \sum_{k=1}^{NPAT} X_{i,k}/NPAT$$

2 - Desvio padrão

$$\sigma_i = \left\{ \sum_{k=1}^{NPAT} (X_{i,k} - \bar{X}_i)^2 \right\}^{1/2} / (NPAT - 1)^{1/2}$$

3 - Desvio padrão normalizado

$$s_i = \sigma_i / \bar{X}_i$$

4 - Mínimo

$xmin_i$ = o valor mínimo de x_i no conjunto de treinamento.

5 - Máximo

$xmax_i$ = o valor máximo de x_i no conjunto de treinamento.

6 - Intervalo

$$R_i = xmax_i - xmin_i$$

7 - Terceiro momento

$$m_{3,i} = \left\{ \sum_{k=1}^{NPAT} (x_{i,k} - \bar{X}_i)^3 \right\} / NPAT$$

8 - Quarto momento

$$m_{4,i} = \left\{ \sum_{k=1}^{NPAT} (x_{i,k} - \bar{x}_i)^4 \right\} / NPAT$$

9 - Assimetria (Skewness)

$$\alpha_{3,i} = m_{3,i} / (m_{2,i})^{3/2}$$

$$m_{2,i} = \left\{ \sum_{k=1}^{NPAT} (x_{i,k} - \bar{x}_i)^2 \right\} / NPAT$$

10- Curtose

$$\alpha_{4,i} = m_{4,i} / (m_{2,i})^2$$

11- Escala de intervalo

$$x'_{i,k} = (x_{i,k} - xmin_i)/R_i$$

12- Auto-escala

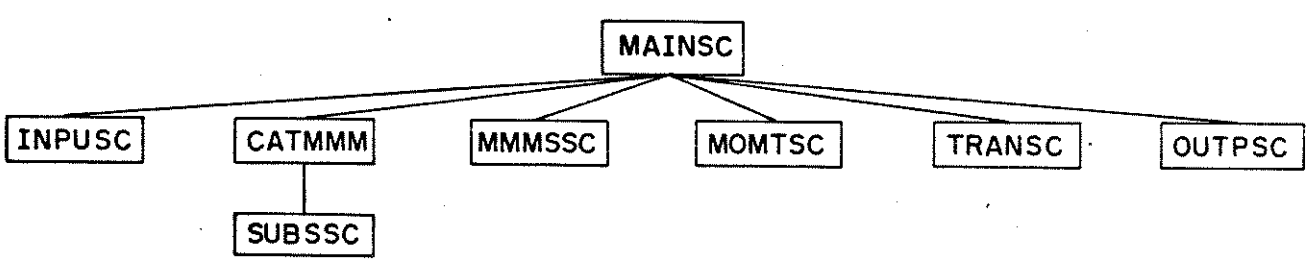
$$x'_{i,k} = (x_{i,k} - \bar{x}_i)/(NPAT.m_{2,1})^{1/2}$$

ESTRUTURA DO PROGRAMA

1 - Subrotinas

MAIN = dados de entrada e chamadas para as subrotinas
INDUSC = dados de entrada
CATMMM = seleciona as categorias
MMMSSC = somatório dos momentos
MOMTSC = cálculo dos momentos e coeficientes estatísticos
TRANSC = transformação dos dados
OUTPSC = saída dos resultados
SUBSSC = recompõe o conjunto de dados original

2 - Organização



3 - Detalhes

a. Os arquivos FORT04.DAT e FORT02.DAT são usados como arquivos de trabalho.

b. O número de variáveis (NVAR) que o programa pode tratar é

determinado pelo parâmetro MAX.

MAX ≥ 14 * NVAR - 1

A memória de CPU, usado pelo programa pode ser expandida modificando as dimensões das matrizes X e NA e o valor de MAX no subprograma MAIN. No PC as dimensões X e NA devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX.

SIMCA

FINALIDADE DO PROGRAMA

Este programa classifica as amostras a partir de suas similaridades com os modelos de componentes principais. Os componentes principais são calculados para cada categoria. O número de componentes pode ser definido pelo usuário ou ser determinado usando "cross validation".

PRÉ-REQUISITO

Para classificações não tendenciosas use dados autoescalonados. Nossa experiência mostra que melhores índices de classificação são obtidos usando dados autoescalonados. As amostras do conjunto de treinamento devem ser colocadas em ordem crescente de categoria; isto pode ser obtido através do programa SCAL.

BIBLIOGRAFIA

S. Wold, J. Pattern Recognition 8, 127 (1976).

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número de arquivo de entrada;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;
NVAR = número de variáveis;
NCAT = número de categorias;
NCOM = número de componentes principais,

a. Se todas as categorias tiverem o mesmo número de componentes principais, simplesmente inserir este número.

b. se todas as categorias tiverem números diferentes de componen-

tes principais inserir 0. Depois inserir ordenadamente o número de componentes principais para cada categoria, da categoria 1 até NCAT.

c. Se o "cross-validation" for usado para determinar o número de componentes principais inserir -1. Depois inserir o valor de NCVS, que é o número máximo de componentes principais a ser considerado no "cross-validation". Este número deve ser maior que três.

SLIM - Inserir o número de desvios padrões para determinar o tamanho dos hipervolumes usados na classificação das amostras. Este valor depende do critério usado para uma classificação aceitável - Normalmente, é sugerido o valor 2.0.

NPNT - Se for igual a 1, os componentes principais de cada categoria serão impressos, se for 0 não.

O arquivo FORT06.DAT é lido automaticamente pelo programa, e contém para cada amostra:

a. ID, um número para identificação (I4);
b. NAME, um nome com até 8 caracteres para identificação (2A4);
c. CN, número da categoria (F2.0);
d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

DEFINIÇÕES

1 - Símbolos entre Categoria e Distância

****, ***, **, *

Estes símbolos dão uma idéia rápida da qualidade da classificação de uma amostra como membro das diversas categorias. Quanto maior o número de *, maior a probabilidade de classificação correta.

2 - A distância de uma amostra ao modelo de componentes principais associado a uma categoria é dada por:

$$D_{i,k} = \left\{ \sum_{j=1}^{NVAR} EPS^2_{i,j} / NVAR \right\}^{1/2}$$

$$EPS_{i,j} = \sum_{j=1}^{NCOM_k} x_{i,j} - (C_m)\ (\beta_{j,m,k})$$

$$C_m = \sum_{j=1}^{NVAR} (x_{i,j})\ (\beta_{j,mk})$$

Onde:

NCOM = número de componentes usados para descrever a K-ésima categoria.

$\beta_{m,n,k}$ = o m-ésimo valor do n-ésimo componente principal para a k-ésima categoria.

3 - MISSES - o número de amostras no conjunto de treinamento classificados erroneamente para cada categoria.

4 - Percent correct

% = (100.0) (NPAT - número de erros)/NPAT

5 - Matriz de distâncias entre categorias

CD_{ij} = "distância" média das amostras da i-ésima categoria em relação ao modelo de componentes principais da j-ésima categoria.

$$CD_{ij} = \left(\sum_{i=1}^{N_k} \sum_{j=1}^{NVAR} EPS^2_{i,j} \right) / (N_k)\ (NVAR))^{1/2}$$

Onde:

N_k = número de amostras na i-ésima categoria.

6 - MISSES - o número de amostras no conjunto teste que foram classificadas erroneamente.

7 - Percent correct - Conjunto teste

% = (100.0) (NTEST - número de erros)/NTEST

8 - Componentes Principais

$\lambda k,m$ - o m-ésimo autovalor da matriz de covariância para as amostras da k-ésima categoria.

STOT = a variância total para todas as variáveis de todas as amostras de uma dada categoria.

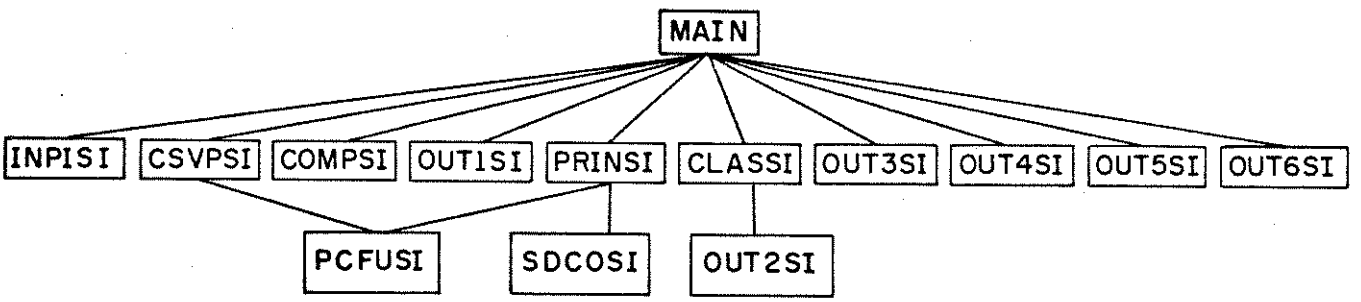
SFIN = a variância total dos componentes principais de uma dada categoria.

ESTRUTURA DO PROGRAMA

1 - Subrotinas

MAIN = faz chamada das rotinas
INPISI = dados de entrada
CSVPSI = faz o "cross-validation"
PCFUSI = determina os componentes principais
COMPSI = calcula os valores de F usados na "cross-validation"
PRINSI = chama as subrotinas para calcular os componentes principais
SDCOSI = desvios padrões das linhas e colunas
CLASSI = faz a classificação
OUT2SI = imprime resultados
OUT3SI = imprime resultados
OUT4SI = imprime resultados
OUT5SI = imprime resultados
OUT6SI = imprime resultados

2 - Organização



3 - Detalhes

a. Os arquivos FORT02.DAT, FORT03.DAT e FORT04.DAT são usados como arquivo de trabalho.

b. O número de amostras (NPAT), variáveis (NVAR) e categorias (NCAT) que o programa pode tratar é determinado pelo parâmetro MAX.

$$MAX \geq NCAT * NCAT + 3 * NCAT + 5 * NVAR + 4 * NPAT$$

A memória de CPU usada pelo programa pode ser expandida, modificando-se as dimensões das matrizes X e NA e o valor de MAX no subprograma MAIN. No PC as dimensões X e Na devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual á MAX e NA deve ser 2* MAX.

DESCRIÇÃO DO MÉTODO SIMCA

Entre todos os métodos de reconhecimento de padrões discutidos no Seminário de Estudos Avançados em Quimiometria, patrocinado pela OTAN em Setembro de 1983 o SIMCA foi o que despertou maior interesse.

SIMCA é a sigla para "Soft Independent Modelling by Class Analogy". Entretanto, um título mais descritivo seria "Modelos Independentes de Similaridade Utilizando Componentes Principais". Esse método apresenta notáveis vantagens para a classificação por categorias, quando comparado aos outros métodos descritos. Baseia-se na análise de componentes principais, já discutida

anteriormente.

Considere o arranjo de pontos ilustrado na Figura SIMCA-1. No sistema de coordenadas originais tanto a dimensão X1 como a X2 são importantes, uma vez que uma parte siginificativa da variância das amostras está representada em cada eixo. Por outro lado, os pontos se dispõem segundo uma relação linear, o que permite o uso de um modelo de componente principal, utilizando apenas um eixo capaz de descrever adequadamente a estrutura dos dados. Isso é uma forma de pré-processamento dos dados, podendo também ser estendida como um método de redução de variáveis. O novo sistema de coordenadas obtido pode ser relacionado com o original, por uma simples rotação.

A posição dos eixos do novo sistema de coordenadas é determinada através da diagonalização da matriz de covariância, como descrito anteriormente; os valores dos dados no novo sistema de coordenadas, θ, são fornecidos pelo autovetor $\vec{\beta}_j$, obtido de uma das colunas da matriz $\underset{\sim}{\beta}^t$, em:

$$\theta = \underset{\sim}{X}\underset{\sim}{\beta}^t$$

Para o caso bidimensional ilustrado na Figura SIMCA-1 o autovalor $\lambda 1$ é cerca de 10 vezes maior que 2, ou seja, cerca de 90% da variância total é expressa pelo eixo do componente principal, $\vec{\beta}_1$, sendo os restantes 10% atribuídos a $\vec{\beta}_2$. Como $\vec{\beta}_1$ e $\vec{\beta}_2$ são autovetores, devem ser mutuamente ortogonais e os eixos devem guardar entre si um ângulo de 90^o.

O método SIMCA é um esquema de classificação por modelo de similaridade. Se repetirmos medidas multivariadas de um mesmo objeto, os valores obtidos para uma variável devem diferir entre si apenas pelo erro experimental, e podem ser expressos por:

$$x_{ki} = \alpha_i + e_{ki}$$

$$k = 1,2,...r$$
$$i = 1,2,...p$$

onde α_i representa o valor médio para a variável i e e_{ki} é o erro experimental da variável i para o objeto k. Esse modelo, de pouca aplicação prática, é um modelo de zero componentes e pode ser representado no espaço p por uma hiperesfera de raio S_o conforme se vê na Figura SIMCA-2. S_o é o desvio padrão dos resíduos,

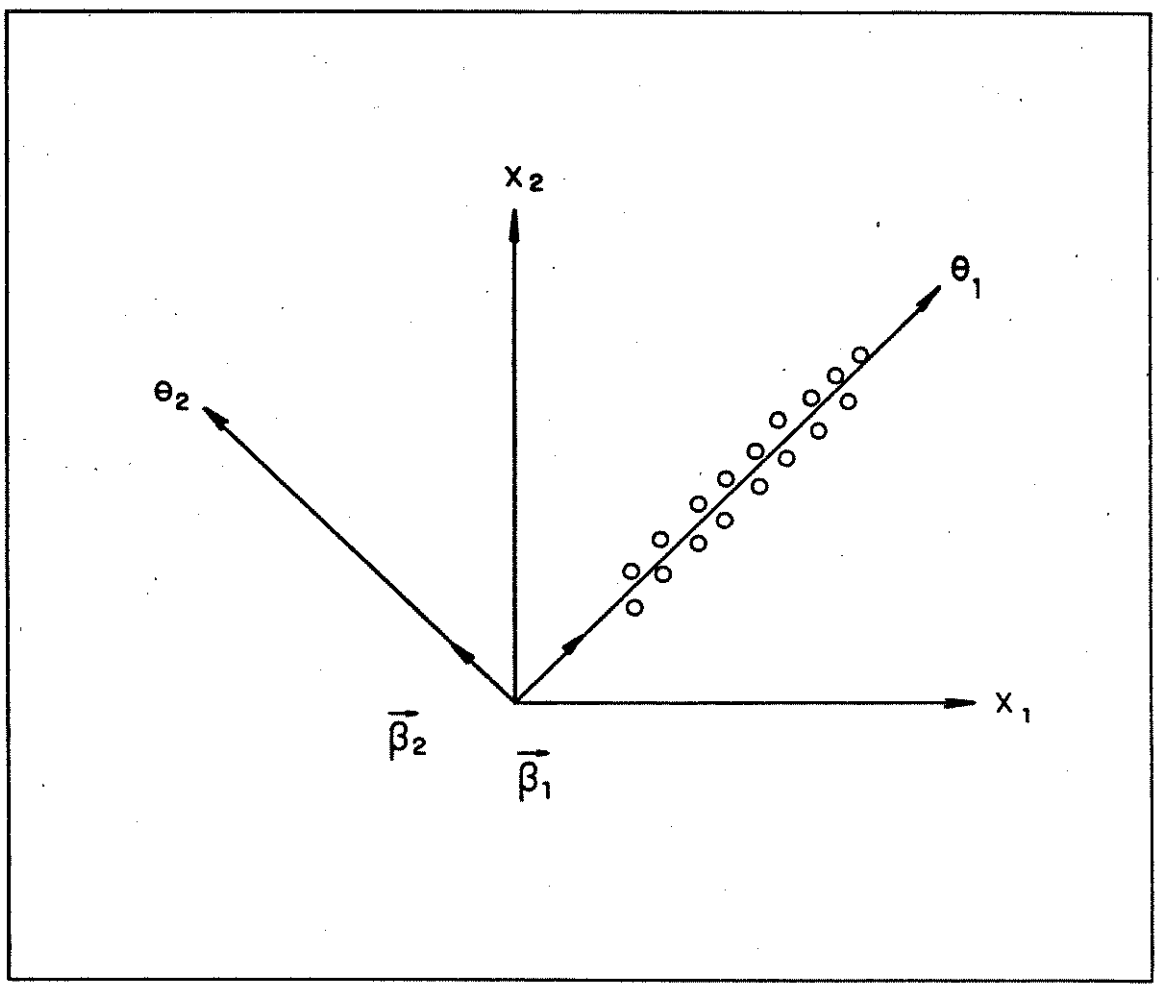


FIGURA SIMCA-1

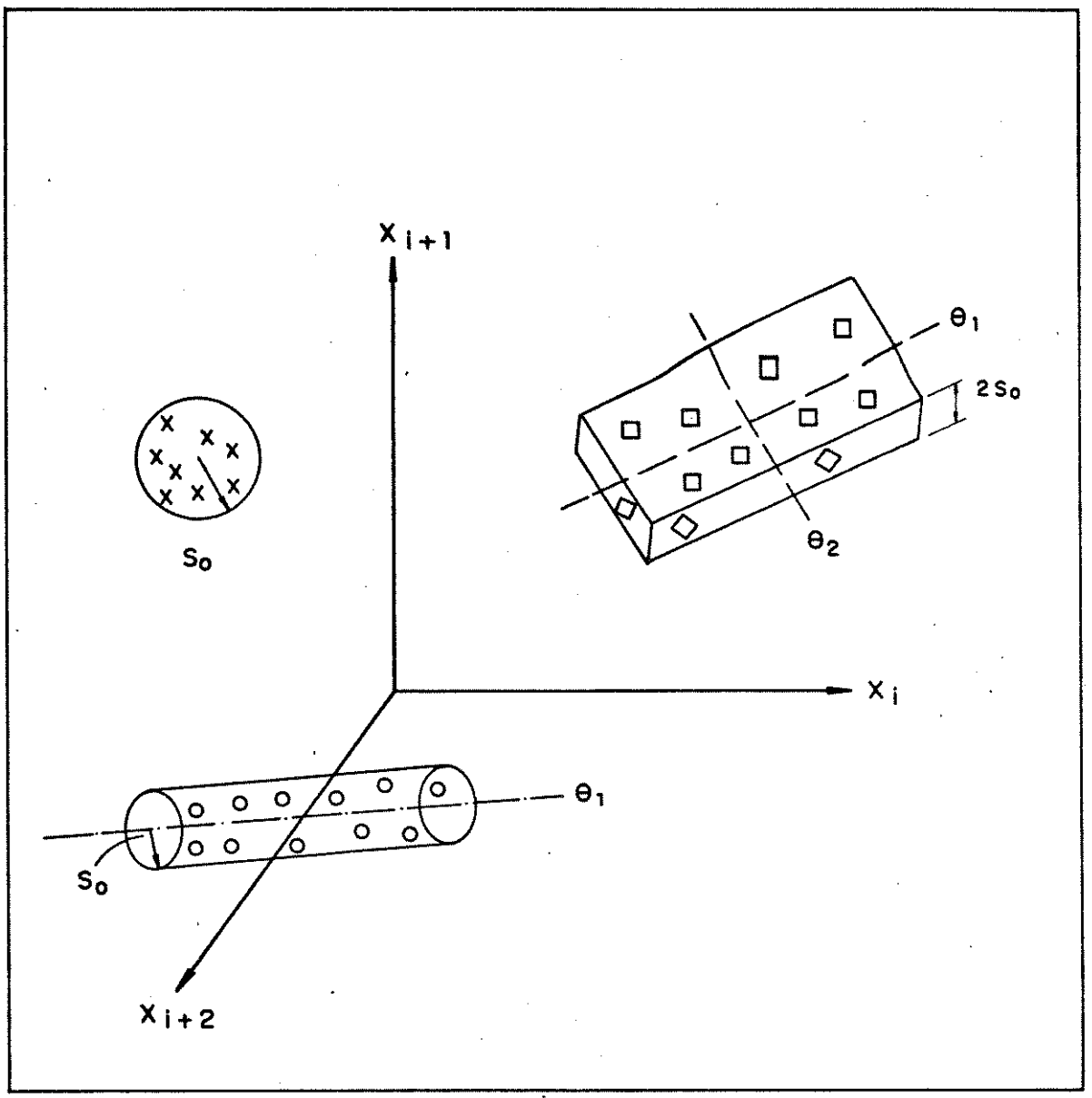


FIGURA SIMCA-2

$$S_o^2 = \sum_{i=1}^{p} \sum_{k=1}^{r} e_{ki}^2 \; / \; [(p - A)\ (r - A - 1)]$$

onde o denominador da equação representa o número de graus de liberdade. Se este modelo for aplicado para objetos similares mas não idênticos e_{ki} e S_o aumentarão porque os resíduos conterão erros de modelagem além de erros de medida. Nesse caso é necessário um modelo mais sofisticado de um componente, a fim de descrever a estrutura mais complexa dos dados:

$$x_{ki} = \alpha_i + \theta_k \beta_i + e_{ki}$$

Os valores de θ_k e β_i nessa equação compõem inicialmente as matrizes (nx1) $\underset{\sim}{\theta}$ e $\underset{\sim}{\beta}^t$. Geometricamente, o modelo pode ser representado por um hipercilindro no espaço p (FIGURA SIMCA-2). O eixo do cilindro corresponde nesse modelo ao componente principal, que é o eixo de maior variância, $\vec{\beta}_j$; o raio do cilindro é dado por S_o. A posição relativa de cada ponto ao longo do eixo do componente principal é dada pelo valor de θ_k.

Para dados cuja estrutura é ainda mais complexa, é necessário um modelo de dois componentes.

$$x_{ki} = \alpha_i + \theta_{k1}\beta_{1i} + \theta_{k2}\beta_{2i} + e_{ki}$$

$i = 1,2,...p$
$k = 1,2,...r$

Essa equação é representada geometricamente na Figura SIMCA-2 por uma caixa no espaço p cujo comprimento e largura são determinados pelos desvios padrões dos pontos ao longo dos dois componentes principais, e a profundidade é $2S_o$.

Em geral, para estruturas de dados muito complexas, utiliza-se um método com A componentes:

$$x_{ki} = \alpha_i + \sum_{a=1}^{A} \theta_{ka} \beta_{ai} + e_{ki}$$

Essa equação representa uma hipercaixa no espaço p. O valor do objeto k projetado no eixo do componente principal

a é dado por θ_{ka}, enquanto β_{ai} representa a taxa de variação da variável i devida a variações unitárias no a-ésimo componente principal.

Os dados do conjunto de treinamento são utilizados para determinar o modelo de cada categoria. Para a determinação do número apropriado de componentes principais pode-se utilizar técnicas de "Cross-Validation". O modelo aplicado a cada categoria é totalmente independente do das outras. Também o número de componentes necessários para descrever os dados pode variar de uma categoria para outra, dependendo do grau de complexidade da estrutura dos dados em cada caso, como se vê na Figura SIMCA-2.

Terminada a modelagem, resta classificar os pontos correspondentes à amostras desconhecidas. Esse processo é exemplificado na Figura SIMCA-3, por projeções bidimensionais de dois modelos de um componente. Os valores $S_p^{(1)}$ e $d_p^{(2)}$ mostrados na figura, representam a menor distância entre um ponto de teste e o eixo do componente principal de cada categoria. Esses valores, por conveniência computacional, são determinados por técnicas de regressão linear.

Podemos agora, uma vez descrito o método SIMCA, ilustrar suas vantagens com relação aos outros métodos. O SIMCA, é capaz, por exemplo, de indicar quando um ponto referente a uma amostra não pertence a nenhuma das categorias classificadas, indicando-o com um ponto deslocado (outlier) ou membro potencial de uma categoria ainda não definida. Esse é o caso mostrado na Figura SIMCA-4. Todos os pontos situados dentro das hipercaixas são classificados como pertencentes a categoria definida por seu hipervolume; o ponto indicado pelo triângulo é acusado como um ponto deslocado e não pertence a nenhuma categoria definida. Todos os métodos de reconhecimento de padrões descritos classificariam incorretamente esse ponto, dando-o como pertencente a uma das categorias definidas no conjunto de treinamento.

Outra vantagem do SIMCA é a sua aplicação a problemas do tipo ilustrado na Figura SIMCA-5. Os pontos (indicados por "0") aqui representados pode ser relativos a um produto manufaturado de especificações bem definidas, ou a um produto que apresente atividade biológica específica. É razoável esperar que os produtos de boa qualidade sejam resultantes de um controle de qualidade severo e seus pontos se disponham no espaço definido

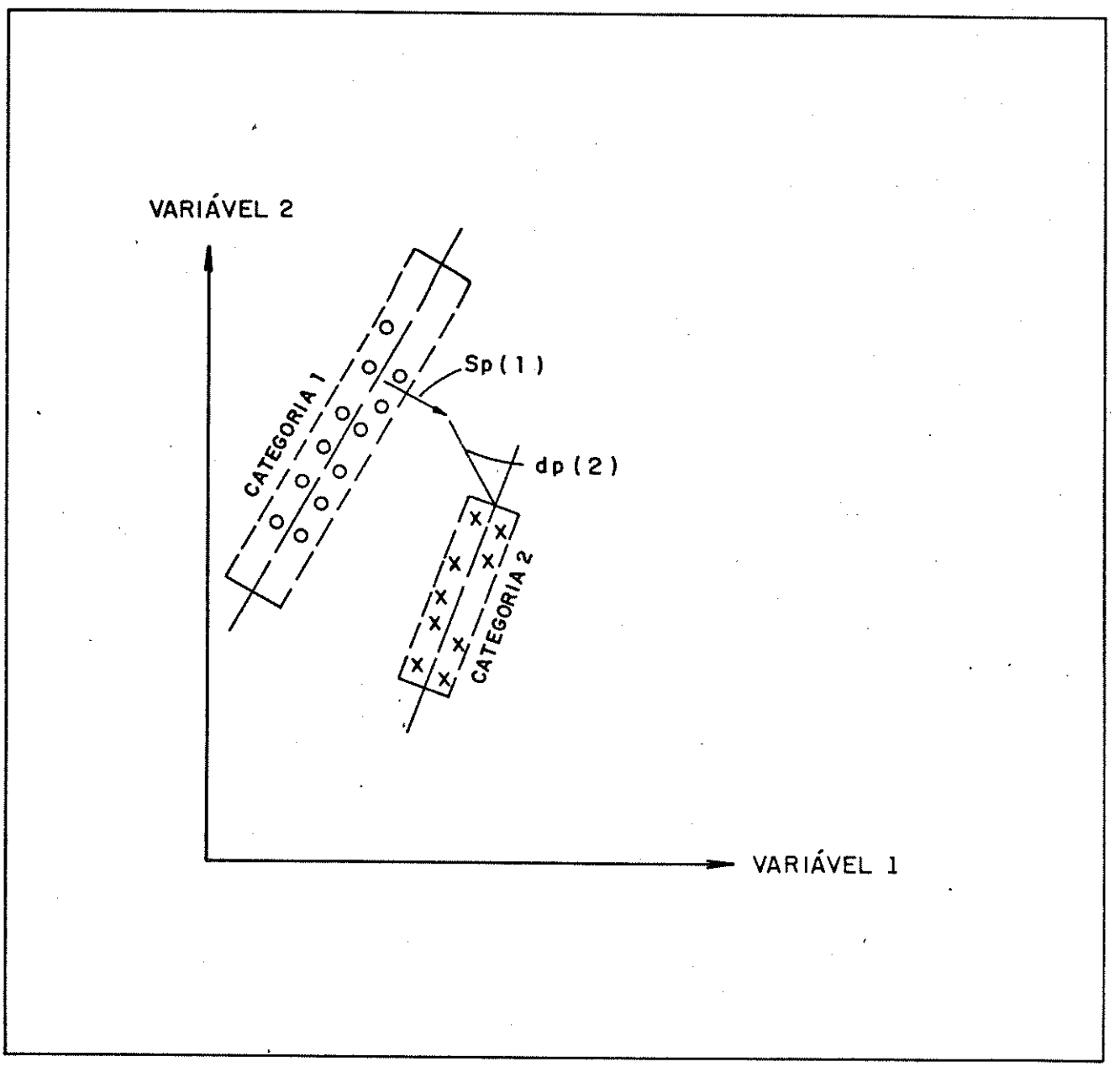


FIGURA SIMCA-3

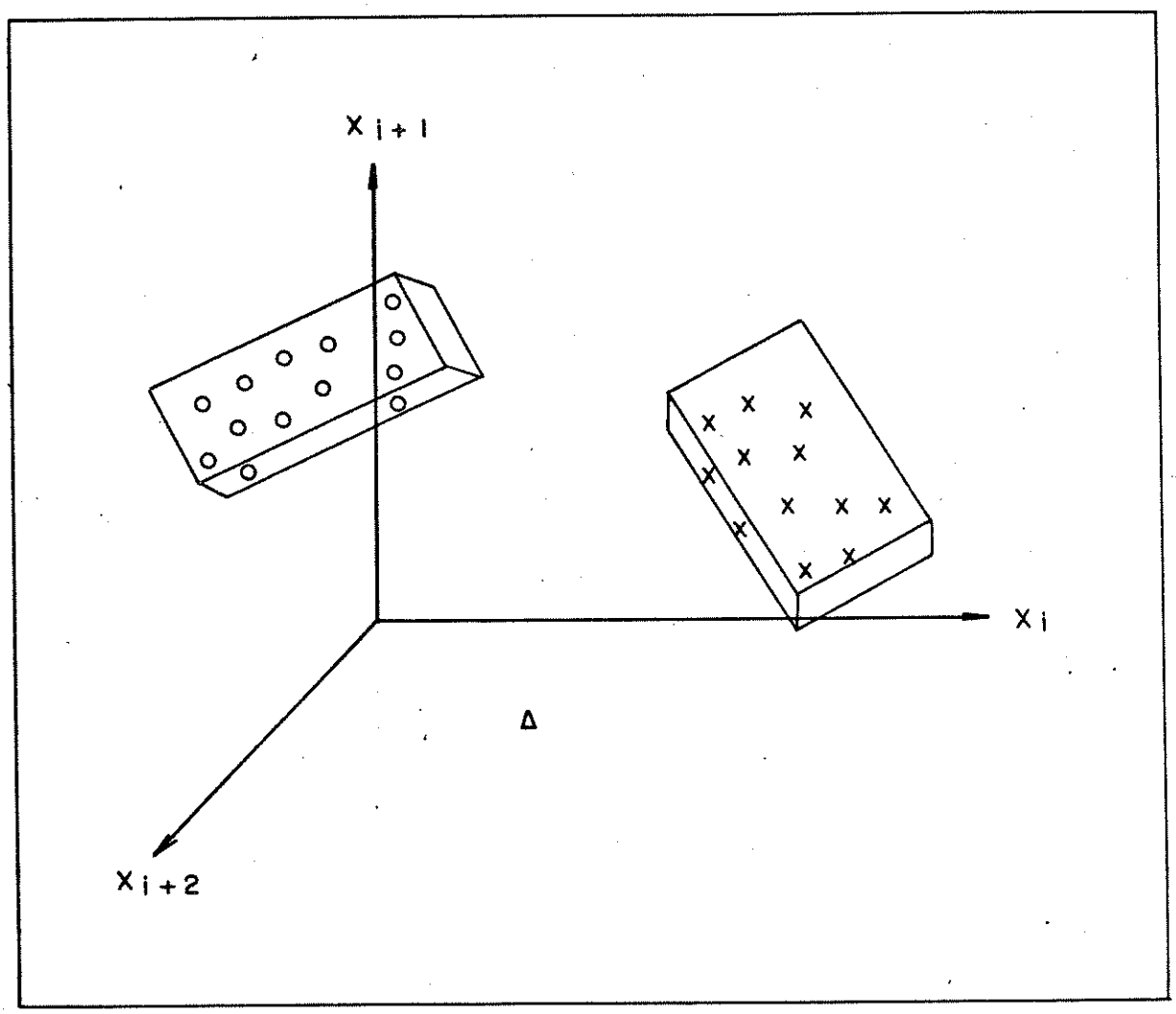


FIGURA SIMCA-4

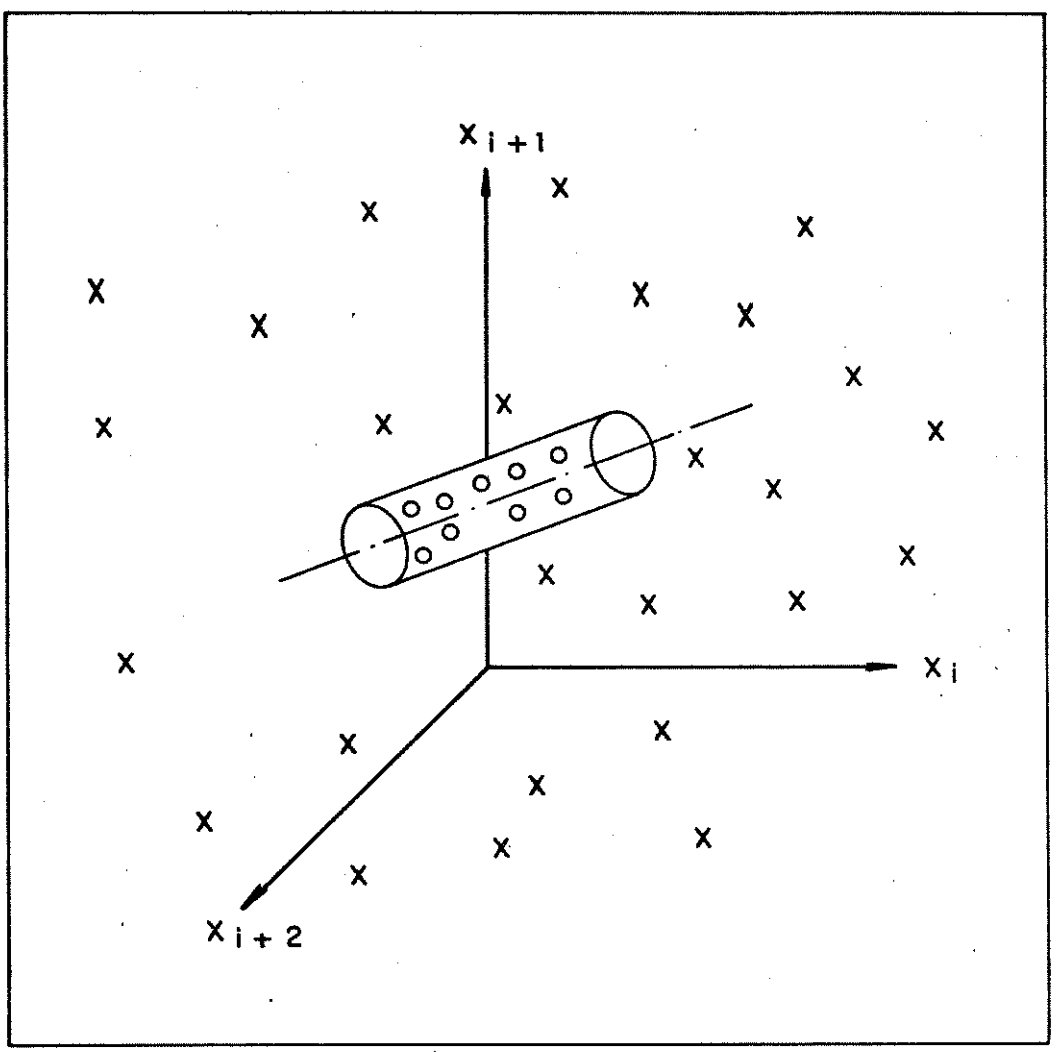


FIGURA SIMCA-5

por um hipercilindro ou uma hipercaixa. Por outro lado, os produtos de má qualidade ou de baixa atividade biológica devem estar dispersos numa faixa ampla, e seus pontos (indicados por "x") muito provavelmente deverão se situar fora dos limites especificados pelo modelo do SIMCA. Nos métodos de reconhecimento de padrões descritos anteriormente, só é possível definir uma classificação quando o conjunto de treinamento apresenta duas ou mais categorias bem definidas.

Finalmente, o SIMCA pode ser aplicado a casos onde o número de amostras, r, é bem menor que o de variáveis, p, uma vez que o número de graus de liberdade, dado por (p - A) (r-A-1), aumenta tanto com o número de amostras como com o de variáveis. O LDA e o LLM são especificamente limitados nesse sentido.

TRANS9

FINÁLIDADE DO PROGRAMA

Este programa gera a transposta da matriz de dados, com dimensão NVAR por NPAT. Esta matriz pode ser utilizada para calcular o dendograma para as variáveis.

BIBLIOGRAFIA

- Nenhuma

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada;
NARQ = número do arquivo de saída;
NPAT = número de amostras (ou objetos);
NVAR = número de variáveis.

O arquivo FORT06.DAT é lido automaticamente pelo programa, e contém para cada amostra:

a. ID, um número para cada identificação (I4);

b. NAME, um nome com até 8 caracteres para identificação (2A4);

c. CN, um número da categoria (F2.0);

d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR.

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

Os resultados são armazenados no arquivo TRANS9.DAT que, para uso posterior, deve ter sua identificação mudada para FORT06.DAT.

A memória de CPU usada pelo programa pode ser expandida modificando as dimensões das matrizes X e B no programa.

VARIMAX

FINALIDADE DO PROGRAMA

Este programa faz a rotação ortogonal da matriz dos loadings. Na solução VARIMAX ocorre a simplificação das colunas, enquanto na solução QUARTIMAX ocorre a simplificação das linhas. No método VARIMAX a variância dos quadrados dos loadings são maximizadas.

PRÉ-REQUISITO

A matriz dos dados autoescalonados deve estar gravada no arquivo FORT06.DAT.

BIBLIOGRAFIA

W.W. Cooley e P.R. Lohnes, "Multivariate data Analysis", John Wiley & Sons, Inc, New York, 1971.

H.H. Harman, "Modern Factor Analysis", The University of Chicago Press, Chicago, 3ª ed., 1976.

H.F. Kaiser, Psychometrika, 23, 187 (1958).

J.C. Davis, Statistics and Data Analysis in Geology", Wiley, New York, 2ª ed., 1986.

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada;
NARQ = número do arquivo de saída;
N = número de fatores (deve ser igual ao número de componentes principais armazenados no arquivo FORT09.DAT);
NVAR = número de variáveis;
NPAT = número de amostras
METHOD = 0 rotação VARIMAX
= 1 rotação QUARTIMAX

FLUXOGRAMA E ARQUIVOS

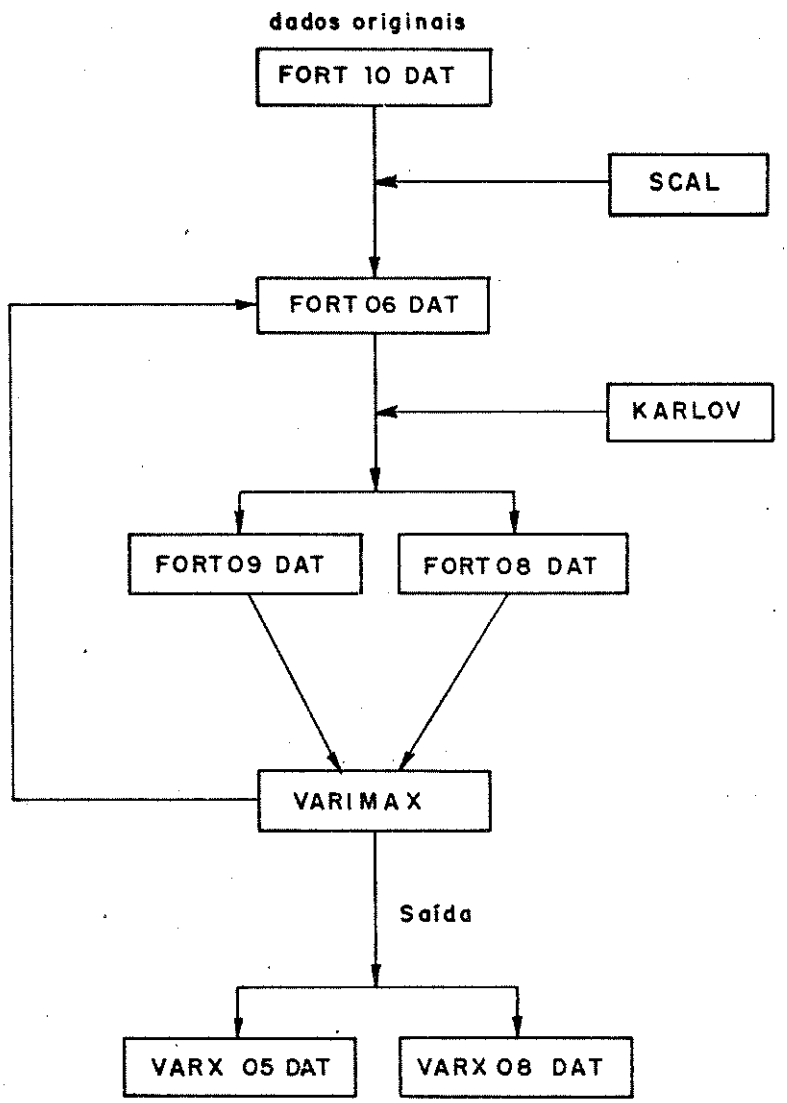


FIGURA VARIMAX 1

O arquivo FORT09.DAT, que é lido automaticamente pelo programa, contém para cada variável:

a. ID, um inteiro para identificação (I4);
b. NAME, o mesmo que ID, abreviado se necessário;
c. X(J), J= 1,N; os valores dos loadings dos fatores, do fator 1 até o fator N (F10.5).

FORMAT (I4, 8X, I2, 6X, 6F10.5/12(8F10.5/))

Os loadings dos fatores rodados são armazenados no arquivo VARX05.DAT e os escores correspondentes no arquivo VARX08.DAT. Para uso posterior (programa VARVAR deste sistema computacional), estes arquivos devem ter os nomes mudados para FORT09.DAT e FORT08.DAT, respectivamente.

FORMAT (I4, 8X, I2, 6X, 6F10.5/12(8F10.5/)) para o arquivo VARX05.DAT e

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12(8F10.5/)) para o arquivo VARX08.DAT.

DEFINIÇÕES

Variância Maximizada

$$V = \sum_{k=1}^{n} \left\{ \frac{\left[p \sum_{j=1}^{p} (S_{jk}^2/h_j^2)^2 - (\sum_{j=1}^{p} S_{jk}^2/h_j^2)^2\right]}{p^2} \right\} \max$$

onde:

S_{kj} é o novo loading para a variável j no fator k;

j= 1,2...., p e k = 1,2 ....,n.

h_j^2 é a comunalidade da variável j.

Comunalidade

$$h_j^2 = \sum_{k=1}^{n} S_{jk}^2$$

onde j = 1,2,...., p

Detalhes

a. Este programa usa o arquivo VARM04.DAT como arquivo de trabalho.
b. A memória de CPU usada pelo programa pode ser expandida, modificando-se as dimensões das matrizes no programa principal.

DESCRIÇÃO DO MÉTODO

Para conhecer melhor as propriedades de um sistema é desejável tentar interpretar a importância química dos autovetores (componentes principais) pela análise de seus loadings. A técnica de rotação VARIMAX ajuda neste processo, e tem como objetivo rodar o eixo de cada fator para posições tais que as projeções de cada variável sobre os eixos dos fatores estejam próximas às extremidades ou próximos à origem. O critério VARIMAX consiste na maximização da variância dos quadrados dos loadings nos fatores, e com isto decresce o número de variáveis com valores de loadings intermediários e aumenta o número de variáveis com grandes e pequenos loadings em cada fator. Assim a interpretação em termos de variáveis originais é feita mais facilmente. A variância V_k dos quadrados dos loadings sobre o k-ésimo fator é dada por:

$$V_k = \left\{ \frac{p \sum_{j=1}^{p} (S_{jk}^2)^2 - (\sum_{j=1}^{p} S_{jk}^2)^2}{p^2} \right\}$$

onde p é o número de variáveis, S_{jk} é o loading da variável j no fator k. A quantidade que será maximizada é

$$V = \sum_{k=1}^{p} \left\{ \frac{\left[p \sum_{j=1}^{p} (S_{jk}^2 / h_j^2)^2 - \left(\sum_{j=1}^{p} S_{jk}^2 / h_j^2 \right)^2 \right]}{p^2} \right\}$$

onde h_j^2 é a comunalidade da j-ésima variável.

Maximizar a variância implica maximizar a variância dos loadings, o que tende a produzir um dos extremos (positivo ou negativo) ou loadings próximo de zero, satisfazendo o objetivo da rotação do fator.

O critério QUARTIMAX maximiza esta variância dos quadrados dos loadings sem normalizar pelas comunalidades

$$\left| \sum_{j=1}^{p} \sum_{k=1}^{n} S_{jk}^4 \right|_{\max}$$

Já que em cada linha a soma dos quadrados deve permanecer constante, o que o método faz é diminuir para cada linha ou variável, o número de fatores, maximizando dentro de cada linha as variâncias dos quadrados dos loadings.

VARVAR

FINALIDADE DO PROGRAMA

Este programa produz na impressora gráficos de característica vs. característica e/ou característica vs. propriedade, usando o conjunto original, escalonado, escores dos componentes principais ou "loadings" dos componentes principais.

PRÉ-REQUISITOS

Nenhum. Os gráficos podem ser produzidos usando o número de identificação (índice), nome da amostra ou o número da categoria. A posição X, Y da amostra é então indicada pelo seu índice, pelos dois caracteres iniciais do nome e pelo número da categoria, respectivamente. Muitas vezes os gráficos usando o número da categoria são mais fáceis de interpretar.

BIBLIOGRAFIA

- Nenhuma

DADOS DE ENTRADA

IARQ = número do arquivo de entrada;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;
NVAR = número de variáveis;
NCAT = número de categorias;

NPRS =
- 1, se os gráficos de característica vs. característica forem especificados pelo usuário.
- 0, se todos os gráficos de característica vs. característica forem gerados.
- -1, se nenhum gráfico de característica vs. característica for produzido.

NPRO = { 1, se os gráficos de características vs. propriedades forem especificados pelo usuário.
-1, se todos os gráficos de característica vs. propriedades forem gerados.
0, se nenhum gráfico de característica vs. propriedade for produzido

IFIN = { =1, se gráficos usando o "número" de identificação da amostra devem ser produzidos.
=0, se não.

IFNAM = { =1, se os gráficos usando o nome de identificação da amostra forem produzidos.
=0, se não.

IFCAT = { =1, se os gráficos usando o número da categoria da amostra forem produzidos.
=0, se não

NX,NY = número da variável/característica a ser colocada na abscissa e ordenada respectivamente.

NX = número da variável/característica a ser colocada na bascissa contra a propriedade na ordenada.

(Para indicar o término das especificações o usuário deverá digitar 0 para NX, quando NPRS = 1 e/ou NPRO = 1).

O arquivo FORT10.DAT (dados originais), FORT06.DAT (dados escalonados) ou FORT08.DAT (escores dos componentes principais), é lido automaticamente pelo programa e contém para cada amostra:

a. ID, um índice para identificação (I4);

b. NAME, um nome para identificação com até 8 caracteres (2A4);

c. CN, número da categoria (F2.0);

d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12 (8F10.5/)).

Se a opção for o arquivo FORT09.DAT, este também

é lido automaticamente pelo programa, só que no lugar do nome para identificação, a amostra tem um número igual ao do índice, e a categoria, não é especificada.

FORMAT (I4, 8X, I2, 6X, 6F10.5/12 (8F10.5/)).

DEFINIÇÕES

1 - YMAX e YMIN

Os valores máximo e mínimo da característica/propriedade na ordenada.

2 - XMAX e XMIN

Os valores máximo e mínimo da característica/propriedade na abscissa.

3 - INDEX PLOT

Gráfico em que o número de identificação da amostra é colocado nas coordenadas (X,Y) da amostra.

4 - NAME PLOT

Gráfico em que o nome da amostra (os 2 caracteres iniciais) é colocado nas coordenadas (X, Y) da amostra.

5 - CATEGORY PLOT

Gráfico em que o número da categoria da amostra é colocado nas coordenadas (X, Y) da amostra.

6 - PLOT/NOT

Indicam respectivamente, para cada linha de saída, os número de amostras colocadas no gráfico e o número de amostras que se sobrepõem a estas e que por isso são suprimi-

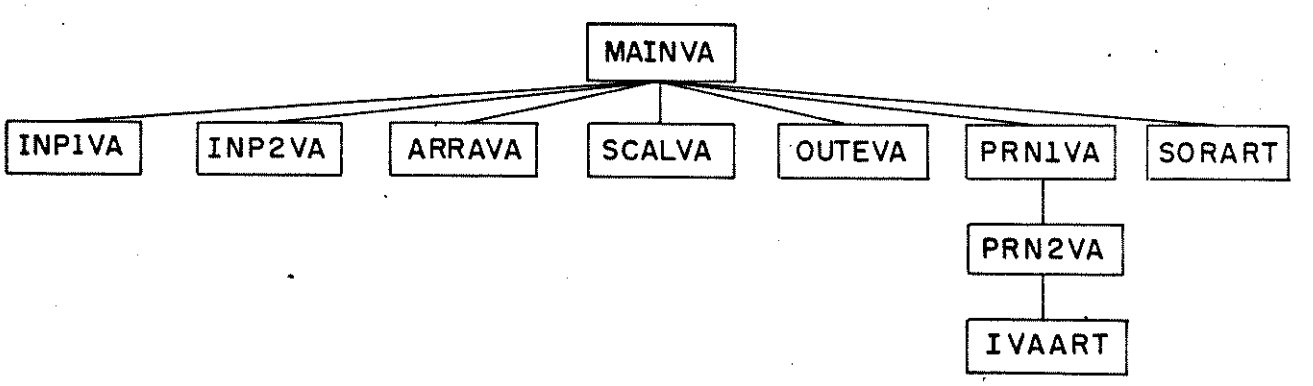
dos do gráfico.

ESTRUTURA DO PROGRAMA

1 - Subrotinas

MAINVA = chama as outras subrotinas
INP1VA = dados de entrada
INP2VA = dados de entrada e especificação dos gráficos
INITVA = coloca os símbolos das amostras em (matrizes)
ARRAVA = puxa os parâmetros de gráficos, número de variável/característica e/ou propriedade em (matrizes) para grafar
SCALVA = escalona as tabelas de dados para grafar
OUTEVA = dados de saída, caso haja erro no cálculo
PRN1VA = inicia os gráficos para a impressora
PRN2VA = faz os gráficos usando a impressora

2 - Organização



3 - Detalhes

a. Os arquivos FORT02.DAT e FORT03.DAT são usados como arquivos de trabalho.

b. O número de variáveis (NVAR) ou amostras (NPAT) que o programa pode tratar é determinado pelo parâmetro MAX:

MAX ≥ 9 * NTOT para NTOT > NVAR e 50 > NVAR

ou

$$MAX \geq 8 * NTOT + NVAR \text{ para } 50 > NVAR > NTOT$$

A memória de CPU usada pelo programa pode ser expandida, modificando-se as dimensões das matrizes X e NA e o valor de MAX no subprograma MAIN. No PC as dimensões X e Na devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX.

WEIGHT

FINALIDADE DO PROGRAMA

Este programa determina a importância individual (peso) de cada variável ou característica na discriminação entre cada par de categorias. Quando se tem três ou mais categorias, as médias destes pesos são também calculadas para cada variável. Duas regras de avaliação (funções de peso) são incluídas no programa: peso de variância e peso de Fisher.

PRÉ-REQUISITOS

Para os pesos de variância e Fisher, são necessárias pelo menos duas amostras por categoria e duas ou mais categorias.

BIBLIOGRAFIA

B.R. Kowalski e C.F. Bender; J. Am. Chem. Soc., 94, 5632 (1972).

R.A. Fisher, Ann. Eugen., 7, 179 (1936).

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

IARQ = número do arquivo de entrada;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;
NVAR = número de variáveis;
NCAT = número de categorias

- Os dados ponderados serão calculados?

Sim = 1 Não = 0

Para dados ponderados pela variância digite 0,

ponderados pelo peso de Fisher digite 1.

Digite o número do arquivo de saída para os dados ponderados: 5.

O arquivo FORT06.DAT é lido automaticamente pelo programa e contém para cada amostra:

a. ID, um número para identificação (I4);

b. NAME, um nome com até 8 caracteres para identificação (2A4);

c. CN, número da categoria (F2.0);

d. X(J), J = 1, NVAR; os valores das variáveis, da variável 1 até a variável NVAR (F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12(8F10.5/)).

DEFINIÇÕES

$W_{j,m,n}$ = medida do poder da variável j para separar a m-ésima categoria da n-ésima categoria.

1 - Peso de Variância

$$w(V)_{j,m,n} = \frac{SSQ_{j,m} + SSQ_{j,n} - (2)\ (SUM_{j,m})\ (SUM_{j,n})}{VARIN_{j,m} + VARIN_{j,n}}$$

$$SSQ_{j,i} = \sum_{k=1}^{N_i} (x_{k,i,j})^2/Ni$$

$$SUM_{j,i} = \sum_{k=1}^{N_i} (x_{k,i,j})/N_i$$

$$VARIN_{j,i} = \frac{\sum_{k=1}^{N_i} (x_{k,i,j})^2 \; (N_i) - \left(\sum_{k=1}^{N_i} x_{k,i,j}\right)^2}{(N_i)^2}$$

N_i = número de amostras na i-ésima categoria

$x_{k,j,i}$ = j-ésima variável da k-ésima amostra da i-ésima categoria

$W(V)_j$ = a média geométrica de $W(V)_{j,m,n}$ para todos os pares de categorias.

$$= \prod_{m=1}^{NCAT-1} \prod_{n=m+1}^{NCAT} w(V)_{j,m,n}/(NCAT)\;(NCAT-1)$$

2 - Peso de Fisher

$$w(F)_{j,m,n} = \frac{(SUM_{j,m} - SUM_{j,n})^2}{VARIM_{j,m} + VARIN_{j,n}}$$

$$VARIM = \frac{\left\{\sum_{k=1}^{N_i} (x_{k,i,j})^2\right\} (N_i) - \sum^{N_i} x_{k,i,j})^2}{(N_i - 1)}$$

$W(F)$ = a média aritmética de $W(F)_{j,m,n}$ para todos os pares de categorias

$$= \left\{\sum_{m=1}^{NCAT-1} \sum_{n=m+1}^{NCAT} W(F)_{j,m,n}\right\} / (NCAT)\;(NCAT-1)/2$$

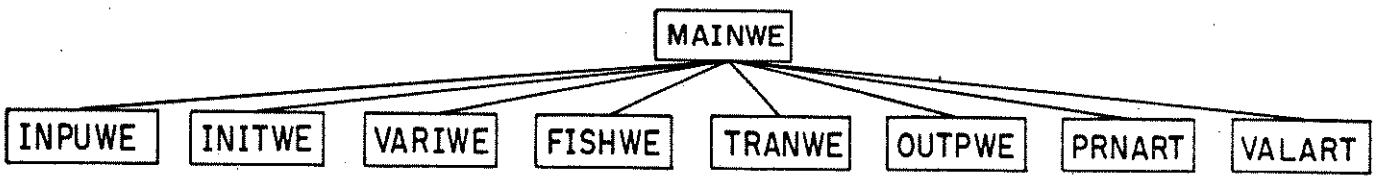
ESTRUTURA DO PROGRAMA

1 - Subrotinas

MAIN = lê os dados de entrada e chama as subrotinas
INITWE = inicialização para o cálculo dos pesos

VARIWE = cálculo dos pesos de variância
FISHWE = cálculo dos pesos de Fisher
TRANWE = nova ordenação das variáveis e/ou transformação das variáveis
OUTPWE = saída dos resultados

2 - Organização



3 - Detalhes

a. O arquivo FORT05.DAT é reservado para o conjunto de dados ponderados pelo peso de variância ou de Fisher.
Os arquivos WEIG08.DAT, WEIG09.DAT e WEIG10.DAT são usados como arquivos de trabalho.

b. O número de variáveis (NVAR) e categorias (NCAT) que o programa pode tratar é determinado pelo parâmetro MAX.

MAX ≥6 * NVAR + 2 * NCAT + 4 * NCAT * NCAT

A memória de CPU, usada pelo programa pode ser expandida, modificando as dimensões das matrizes X, NA e o valor MAX no subprograma MAIN. No PC as dimensões X e NA devem ser iguais a MAX. Para máquinas de 8 bits a dimensão de X deve ser igual a MAX e NA deve ser 2 * MAX.

DESCRIÇÃO DO MÉTODO WEIGHT

A redução de variáveis permite eliminar os valores que não são relevantes para a classificação desejada. Uma maneira de se fazer essa redução consiste em decidir se uma valor é mantido ou eliminado da matriz de dados. Pode-se também aplicar

um método mais suave, calculando o peso de variância ou o peso de Fisher para cada variável, num par de categorias.

O peso de Fisher para a variável i e para amostras das categorias p e q, pode ser calculado da seguinte maneira: divide-se o quadrado da diferença dos valores médios da variável i para as classes p e q, pela soma das variâncias dessa variável nas duas categorias, ou seja:

$$w_{pq}(i) = \frac{\left[\bar{x}_i(p) - \bar{x}_i(q)\right]^2}{s_i^2(p) + s_i^2(q)}$$

O conceito do peso de Fisher está ilustrado na Figura WEIGH-1. Seu valor aumenta com o aumento do numerador da expresão acima. Em outras palavras, quanto mais separados os valores de uma variável para as duas classes, mais importante será essa variável para discriminar os objetos das classes p e q. O valor do peso de Fisher aumenta também quando diminui a soma das variâncias dentro de cada classe. Variâncias pequenas correspondem a picos mais estreitos na Figura WEIGH-1. As variáveis que exibem picos mais estreitos apresentam maior resolução na separação das categorias ou classes. Assim, quanto maior o peso de Fisher, maior o potencial discriminatório da variável i para as classes p e q. Pode-se calcular o peso de Fisher para cada variável simplesmente tomando a média dos pesos obtidos na equação anterior, para todos os pares de categorias. Esses pesos médios podem então ser multiplicados pelos valores das variáveis autoescalonadas:

$$X'_{ki} = W_i X_{ki}$$

para se obter os valores ponderados das variáveis, X'_{ki}, o que pode facilitar a separação das amostras em suas categorias apropriadas.

O peso de variância tem a mesma função que o de Fisher. Para as classes p e q, o peso de variância é calculado dividindo a variância interclasse pela soma das variâncias intraclasse.

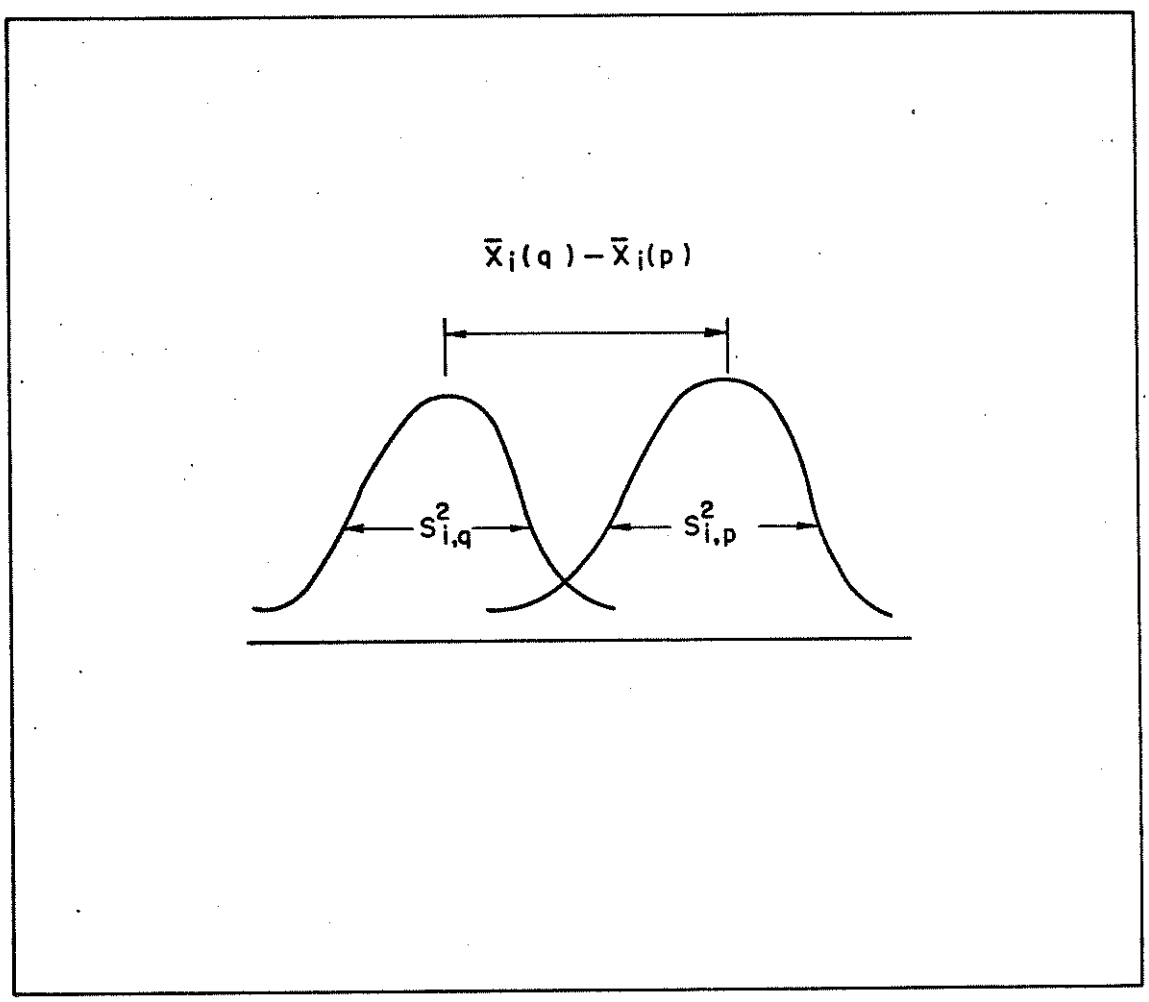


FIGURA WEIGH-1

PCR

FINALIDADE DO PROGRAMA

Este programa é usado para previsões quantitativas em problemas complexos, por exemplo, a determinação das concentrações de diversos componentes em uma ou mais amostras em que isto não possa ser feito diretamente. O programa permite decidir se uma amostra adicional faz parte ou não do conjunto de amostras usado na calibração, e fornece o erro padrão de calibração e o erro padrão de previsão.

ORGANIZAÇÃO DOS DADOS

Os dados estão organizados em duas matrizes, como mostra a figura PCR1. A matriz $\underset{\sim}{X}$ contém a descrição das variáveis independentes, normalmente respostas analíticas, e a matriz $\underset{\sim}{Y}$ contém as variáveis dependentes, normalmente concentrações. Os dados das matrizes $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ são divididos em dois conjuntos: 1) o conjunto de calibração e 2) o conjunto de previsão.

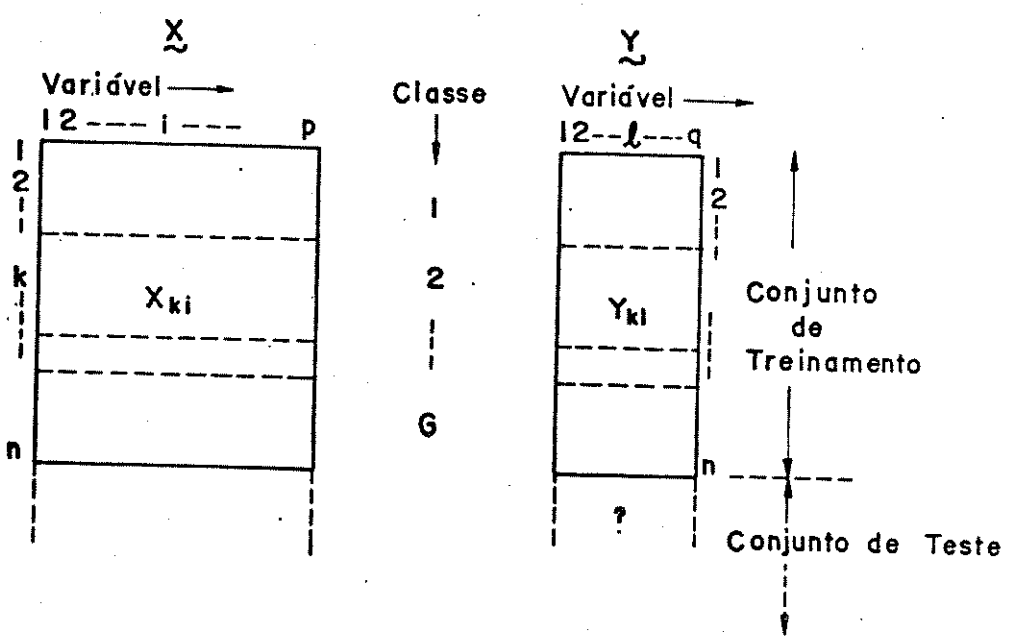


FIGURA PCR-1

REQUISITOS

Antes da análise de regressão em componentes principais deve-se fazer o autoescalonamento dos dados da matriz $\underline{X}$ e depois faz-se a análise de componentes principais, como mostra a figura PCR2.

REFERÊNCIAS

K.R. Beebe e B.R. Kowalski, Anal. Chem., 59 (1987) 1007A.

S.Wold, P. Geladi, K. Esbensen e J. Öhman, Journal of Chemometrics, 1 (1987) 41.

P. Geladi e B.R. Kowalski, Anal. Chim. Acta, 185 (1986) 1.

D.M. Haaland e E.V. Thomas, Anal. Chem., 60 (1988) 1193.

D.M. Haaland e E.V. Thomas, Anal. Chem., 60 (1988) 1202.

D.M. Haaland, Anal. Chem., 60 (1988) 1208.

DADOS DE ENTRADA

Digitar os seguintes parâmetros usando o teclado:

NARQ = número de arquivo de entrada para as variáveis dependentes;
NIN = número de arquivo de entrada para as variáveis independentes;
NOUT = número de arquivo de saída;
NND = número de variáveis dependentes;
NVAR = número de variáveis independentes;
NNC = número de componentes principais armazenados no arquivo FORT09.DAT;
NC = número de componentes principais para a regressão;
NPAT = número de amostras no conjunto de treinamento;
NTEST = número de amostras no conjunto de teste;

FLUXOGRAMA E ARQUIVOS

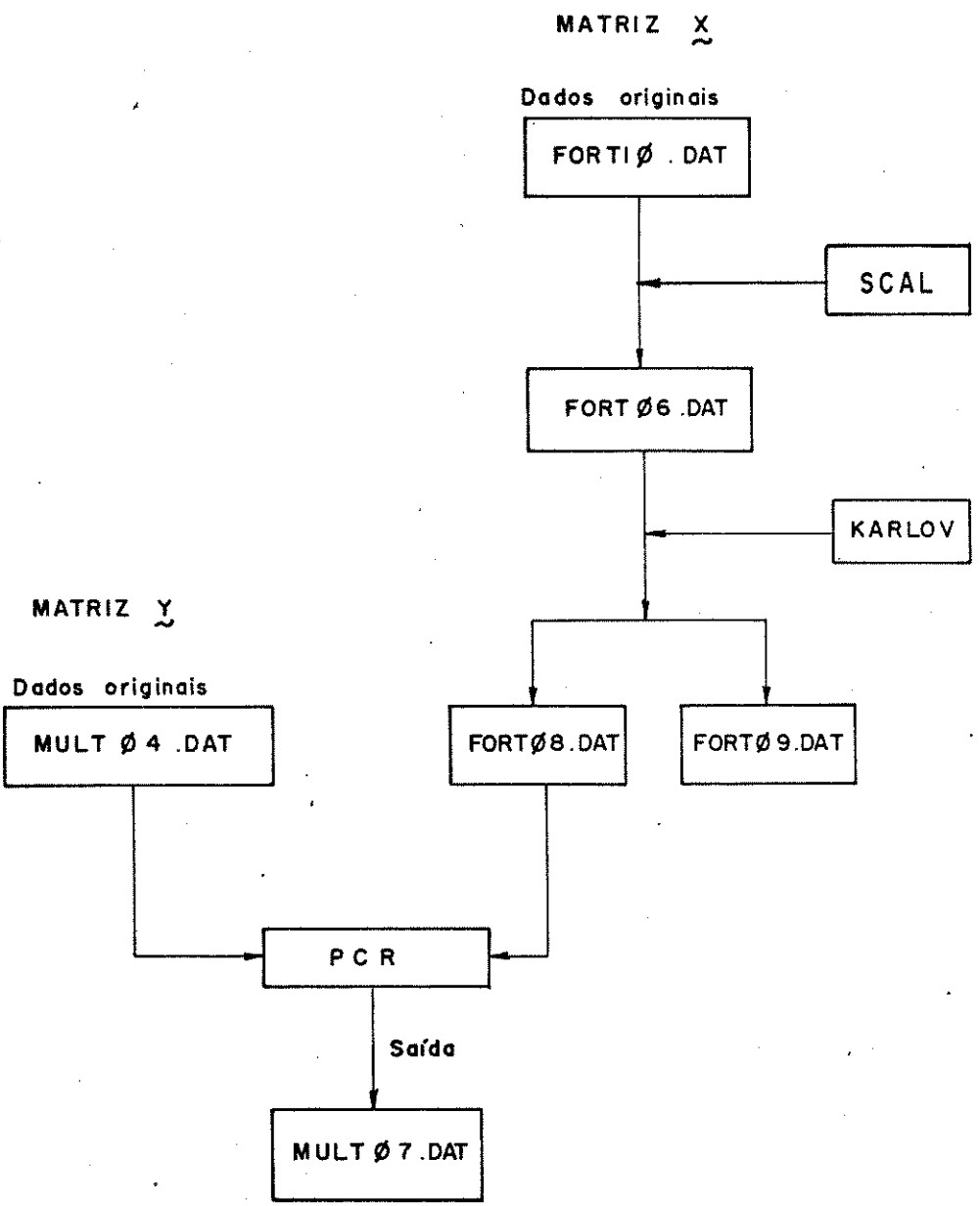


FIGURA PCR - 2

Os arquivos MULT04.DAT, FORT10.DAT são lidos automaticamente pelo programa e devem conter, para cada amostra:

a. ID, um número para identificação (I4);
b. NAME, um nome com até 8 caracteres para identificação (2A4);
c. CN, número da categoria (F2.0);
d. X(J), J=1, NVAR; os valores das variáveis independentes, da variável 1 até NVAR, para o arquivo FORT10.DAT (F10.5);
e. Y(J), J=1, NND; os valores das variáveis dependentes, da variável 1 até NND, para o arquivo MULT04.DAT(F10.5).

FORMAT (I4, 2X, 2A4, 2X, F2.0, 2X, 6F10.5/12(8F10.5/))

Os formatos dos dados nos arquivos FORT08.DAT, FORT06.DAT, FORT10.DAT e MULT04.DAT são iguais.

DEFINIÇÕES

Ver SCAL, KARLOV e SIMCA.

DETALHES

A memória de CPU usada pelo programa pode ser expandida, modificando-se as dimensões de todas as matrizes no programa principal e nas subrotinas.

DESCRIÇÃO DO MÉTODO DE REGRESSÃO DE COMPONENTES PRINCIPAIS

As possibilidades de usar métodos de calibração convencional univariado para quantificação de propriedades de amostras complexas são limitadas. A calibração univariada supõe que o método de medida seja específico ou seletivo, isto é, que nenhum outro constituinte da amostra influencie na medida do sinal do constituinte desejado.

Na abordagem multivariada um conjunto de calibração de n amostras com composições conhecidas (m) é analisado para dar a um certo número de sinais (p) para cada amostra. Desta forma é obtida uma matriz $\underline{Y}$ (nxm) que contém as composições conhecidas das

amostras e uma matriz $\underline{X}$ (nxp) que contêm as respostas instrumentais das amostras, ver figura PCR1. Esta relação linear entre as respostas e as concentrações pode ser usada para prever a composição de uma amostra desconhecida, onde se conhece apenas as variáveis da matriz $\underline{X}$, conjunto de teste.

O método multivariado mais comum é a regressão linear múltipla. Este método apresenta algumas desvantagens que limitam o seu uso: 1) Supõe que as variáveis independentes tenham pouco ou nenhum erro experimental; 2) é rigorosamente apropriada somente para sistemas onde as variáveis independentes sejam ortogonais entre si, e 3) o número de amostras tem que ser maior que o número de variáveis.

Uma maneira de reduzir a dimensionalidade dos dados é aplicar a análise de componentes principais na matriz $\underline{X}$, como mostra a figura PCR2. A análise de componentes principais produz um conjunto de novas variáveis não correlacionadas que pode ser usado juntamente com a regressão linear múltipla para dar a regressão de componentes principais.

O resultado da seção de componentes principais (ver KARLOV), podem ser usados para interpretar a análise de componentes principais de uma matriz de dados $\underline{X}$, que é uma representação de $\underline{X}$ por sua matriz escore θ. A tranformação é

$$\underline{\theta} = \underline{X}\ \underline{\beta}$$

A fórmula de regressão pode ser escrita como

$$\underline{Y} = \underline{\theta}\ \underline{B}' + \underline{\varepsilon} \quad \text{(solução: } \underline{B}' = (\underline{\theta}'\underline{\theta})^{-1}\ \underline{\theta}'\underline{Y}\ .$$

Neste caso a inversão de $\underline{\theta}'\underline{\theta}$ não causa problemas, uma vez que as componentes principais são são correlacionados.

Um aspecto crítico na calibração multivariada é determinar o número ótimo de componentes principais, as dimensões da PCR, sem ajustar excessivamente os dados da calibração. Se o sistema seguir a lei de Beer, o número de componentes principais na modelagem deve ser igual ao número de constituintes químicos na amostra. Em aplicações reais o número de componentes principais é muitas vezes maior que o número de constituintes químicos na amostra. Vários fatores podem causar este aumento do número de componentes principais: 1) determinação da linha de base; 2) interações químí-

cas, e 3) ruídos aleatórios. Para dados simulados mostrou-se que a não linearidade dos dados não aumenta o número de componentes principais, mas aumenta o erro de previsão das amostras de teste. Alguns métodos tem sido propostos para a escolha do número de componentes principais, sendo os mais comuns o "Cross-Validation" e a comparação entre os erros padrão de calibração e de previsão.

CROSS-VALIDATION (CSV)

Para cada nova dimensão da análise de componentes principais (SIMCA), é calculado o valor CSV dividindo-se o conjunto de calibração em B partes iguais, sendo B-1 partes usadas para calibração e a outra para previsão. Este passo é repetido até que todas as partes tenham sido usadas para previsão. O programa KARLOV indica o número ótimo de componentes usando este método.

COMPARAÇÃO ENTRE OS ERROS PADRÃO DE CALIBRAÇÃO E PREVISÃO

O erro padrão de calibração (SEC) é definido como

$$SEC = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2/(n-a-1)$$

onde Y_i é o valor real para a i-ésima amostra, $\hat{Y}_i$ o valor previsto para a i-ésima amostra, n o número de amostras incluídas no conjunto de calibração e 'a' o número de componentes principais incluídos no modelo.

O erro padrão de previsão (SEP) é definido como

$$SEP = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)/(n-1)$$

onde n é o número de amostras usadas no conjunto de previsão.

O número de componentes principais que deverá ser incluído na modelagem deve ser aquele para o qual os erros de previsão das amostras de composição desconhecida difiram muito pouco do erro padrão de calibração.

DETECÇÃO DE OBJETOS ANÔMALOS

A idéia básica do método é que os dados multivariados, X_{ki}, observados em um grupo de objetos similares, podem ser representadas aproximadamente pelo modelo de componentes principais com A componentes.

$$X_{ki} = \alpha_i + \sum_{a=1}^{A} \theta_{ak} \beta'_{ai} + e_{ki} \quad ,$$

onde os parâmetros α, β e θ descrevem a parte sistemática de $\underline{X}$.

Na equação acima os resíduos e_{ki} descrevem a parte aleatória de $\underline{X}$. Esta parte consiste de 1) erros de medidas e outras imprecisões e 2) erro do modelo, isto é, imperfeição da aproximação.

Os resíduos e_{ki} têm desvio padrão residual S_o

$$So = \left(\sum_{i}^{p} \sum_{k}^{r} e_{ki}^2/(P-A)(n-A-1)\right)^{1/2} .$$

O desvio padrão residual S_o mede a distância entre o modelo da classe e um objeto pertencente à classe. Os resíduos são também usados para calcular a importância de cada variável i e detectar objetos anômalos no conjunto de treinamento.

A segunda idéia do método é que os objetos no conjunto de teste podem ser classificados de acordo com o seu grau de ajuste ao modelo da classe. Este é calculado ajustando-se cada objeto (com dados X_{ij}) ao modelo de classe por uma regressão múltipla.

$$|X_{ji} - \alpha_i| = \sum_{a=1}^{A} t_{aj} \beta'_{ia} + e_{ji} .$$

Aqui os coeficientes de regressão t_{aj} são determinados de forma a minimizar os resíduos e_{ji}. O grau de ajuste entre o objeto e o modelo é medido pelo desvio padrão residual S_j com p-A graus de liberdade,

$$S_j = \left(\sum_{i=1}^{p} e_{ji}^2/p\text{-}A\right)^{1/2} .$$

Este desvio padrão residual é proporcional à distância entre o ponto representando o objeto j e o modelo de classe no espaço p. ver

figura PCR3.

Se um objeto faz parte da classe do conjunto de treinamento, o desvio padrão residual acima é definido como

$$S_j = [\sum_{i=1}^{p} e_{ji}^2 [n/(p-A)(n-A-1)]]^{1/2} .$$

O objeto j é classificado como provavelmente pertencente à classe modelada se S_j não for muito maior que o desvio padrão residual, S_o. Isto pode ser avaliado por meio de um teste-F aproximado, com (P-A) e (P-A)(n-A-1) graus de liberdade

$$F = S_j^2/S_0^2 ,$$

Isto corresponde à construção de um intervalo de confiança em redor do modelo de classe, como mostra a figura PCR4. O topo e a base do cilindro são determinados com base na distribuição dos objetos ao longo do eixo θ.

Depois de estabelecida a modelagem os objetos do conjunto de teste são classificados ajustando-se cada um ao modelo da classe. O desvio padrão resultante contém informação sobre a similaridade entre o objeto e a classe.

Se um valor t do objeto ajustado está fora da variação normal do θ correspondente, o desvio padrão do objeto é aumentado pelo termo $(t_a-\theta_{a,lim})$, isto é,

$$d_j = S_{j\ aum} = [S_j + \Sigma \phi_a^2 (ta-\theta_{a,lim})^2]^{1/2} \quad \text{onde}$$

$$\theta_{a,lim} = \begin{cases} \theta_{a,max} + 0,5\ S_{\theta,a} \\ \theta_{a,min} - 0,5\ S_{\theta,a} \end{cases} ,$$

$$S_{\theta,a} = [\sum_{k=1}^{n} \theta_{ak}^2/n]^{1/2} \quad \text{e}$$

$$\phi_a = S_j/S_{\theta,a} .$$

O termo ϕ_a foi introduzido para tornar os dois termos S_j e $ta-\theta_{a,lim}$, comparáveis.

Isto corresponde a medir a distância entre o objeto e o ponto mais próximo dentro do cilindro correspondente ao mode

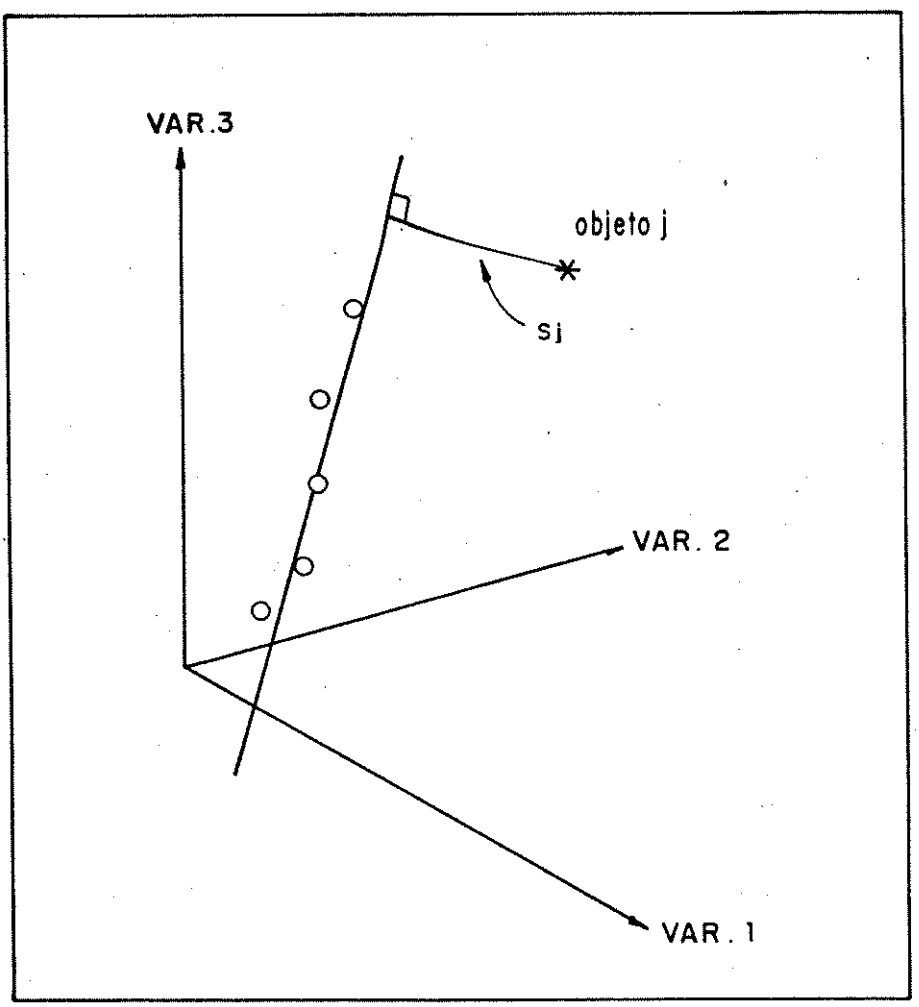


FIGURA PCR - 3

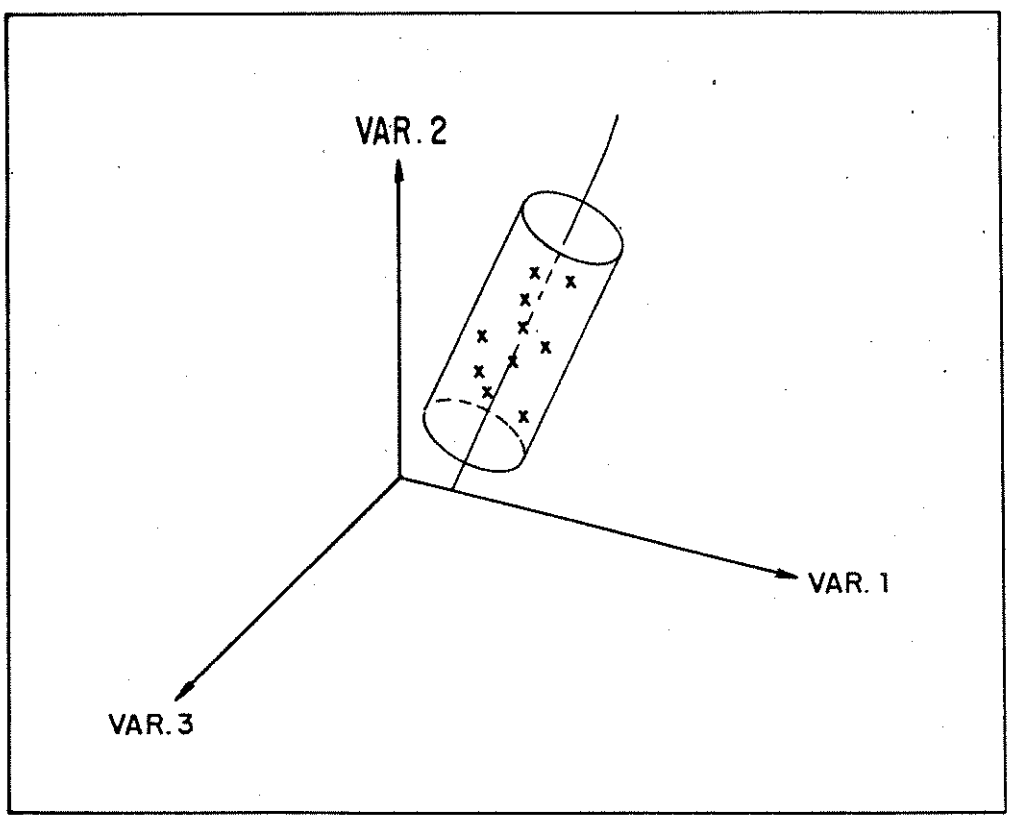


FIGURA PCR-4

lo da classe (d_j na figura PCR5).

O resultado do desvio padrão pode ser usado para calcular a probabilidade relativa de que o objeto pertença a esta classe.

Se o valor de d_j para um objeto do conjunto de teste for maior que $2xS_o$ o objeto está fora do modelo da classe e o resultado do PCR não é fidedigno.

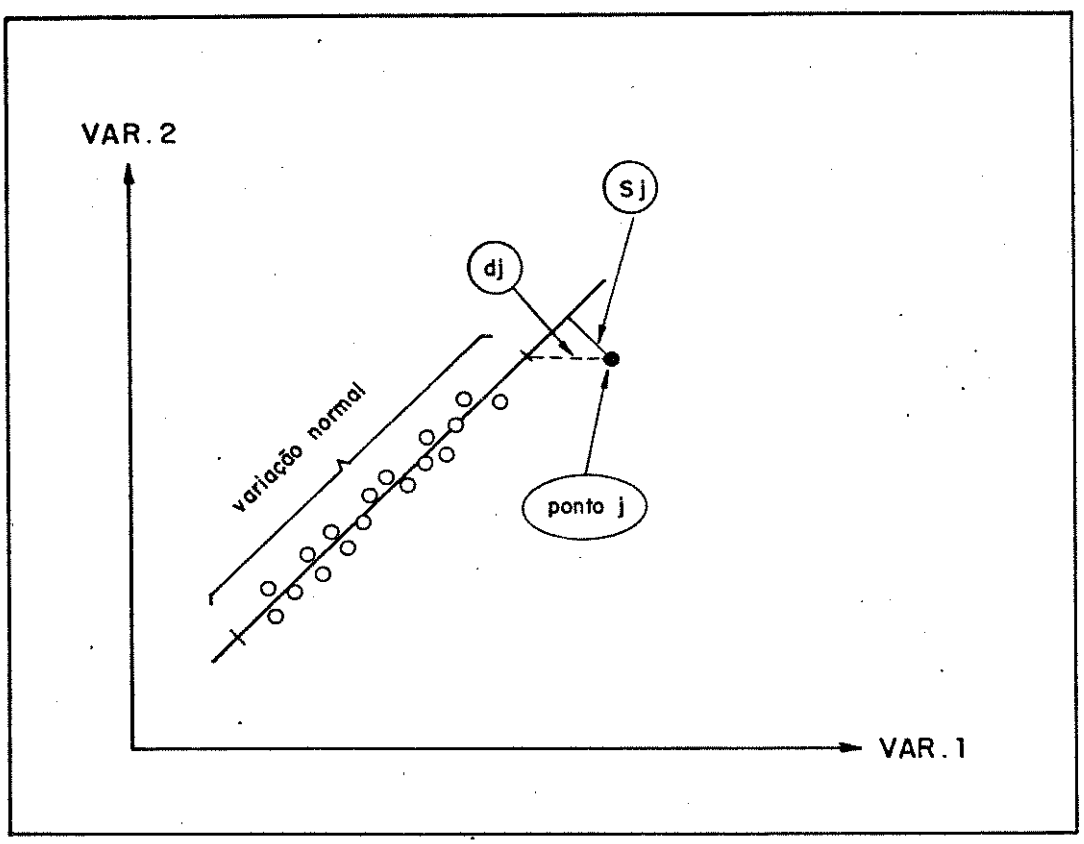


FIGURA PCR - 5