UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA

Este exemplar corresponde à redação final da tese defendida por Ronei Jesus Poppi e aprovado pela Comissão Julgadora

28/04/89

Quantificação de picos Cromatográficos superpostos por Métodos de Calibração Multivariada

TESE DE MESTRADO

Ronei Jesus Poppi

Orientador : Prof. Dr. José Fernando Gregori Faigle

CAMPINAS

1989

AGRADECIMENTOS

- Ao Prof. Dr. José Fernando Gregori Faigle, pela orientação e amizade.

- Ao Prof. Dr. Roy Edwards Bruns, pela possibilidade da utilização dos seus computadores e programas.

- Ao Prof. Dr. Antonio Pires Valente, por toda a orientação na parte experimental em Cromatografia.

- À Profa. Dra. Ieda S. Scarminio, pelos seus ensinamentos em Quimiometria e pela possibilidade da utilização do PCR.

- Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa concedida.

- À UNICAMP pela bolsa de monitoria concedida.

- Aos amigos do andar térreo do "Bloco E" do Instituto de Química, em especial ao Paulo, Romeu e Bete.

- À todos aqueles que direta ou indiretamente contribuíram para a realização deste trabalho.

Aos meus pais

e à minha irmã

Dedico

ÍNDICE

página

RESUMO

TÍTULO : " Quantificação de Picos Cromatográficos Superpostos por Métodos de Calibração Multivariada ".

AUTOR : Ronei Jesus Poppi

ORIENTADOR : Prof. Dr. José Fernando Gregori Faigle

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

Este trabalho analisa a aplicação de Métodos de Calibração Multivariada a dados obtidos por técnicas instrumentais, onde as espécies que se deseja quantificar apresentam sinais superpostos.

Dois métodos matemáticos são apresentados, como uma alternativa ao trabalho experimental de separar instrumentalmente os sinais.

O Método dos Mínimos Quadrados Parciais (PLS) e o da Regressão de Componentes Principais (PCR) aqui estudados, fornecem excelente exatidão no cálculo de contribuições individuais das espécies para um sinal composto, quando se utiliza dados simulados.

Para a aplicação a um caso real, foi escolhida a técnica de Cromatografia Gasosa. Nesse caso, foram utilizadas misturas de Tolueno, Isooctano e Etanol, para três conjuntos de dados com diferentes superposições entre os picos.

Para os dois métodos utilizados foram efetuados estudos com diferentes processos de pré-tratamento dos dados.

Os resultados obtidos mostraram que esses métodos apresentam erros de previsão praticamente iguais aos obtidos por métodos manuais quando não havia nenhuma superposição entre os picos . A utilização de um método de Separação Linear, mostrou ser totalmente improdutivo nos casos estudados.

ABSTRACT

Title : "Quantification of overlapping chromatografic peaks by multivariate calibration methods".

Author : Ronei Jesus Poppi

Thesis advisor : Prof. Dr. José Fernando Gregori Faigle

Instituition : Universidade Estadual de Campinas

Caixa Postal 6154 - CEP 13081 - Campinas - S.P.

This work reports the application of multivariate calibration methods to instrumental analysis data, in conditions where the analytes of interest give overlapping signals.

Two mathematical methods were investigated, as possible alternatives to the experimental work of peak separation.

The partial least squares method (PLS) and the principal component regression method (PCR) showed excellent results in calculating the individual contribuitions of species to a composed signal, in simulated data.

For real applications the technique of gas chromatography was chosen. Mixtures of toluene, isoctane and ethanol for three different data sets, with increasing peak overlap were employed.

For both multivariate calibration methods, different data pre-processing methods were investigated.

The results contained errors of the same magnitude as those observed in treating isolated signals in the usual way. The traditional method of linear separation proved to be inadequate for the analytical determinations.

CAPÍTULO I

INTRODUÇÃO

As análises quantitativas que eram realizadas na maioria das vezes por "via úmida" como titulação, precipitação, reações específicas, etc. que são demoradas e muitas vezes pouco precisas, estão cada vez mais sendo substituídas por técnicas instrumentais como : RMN, IV , UV-vis., Espectrometria de massa, Polarografia, Cromatografia, Análise por Injeção em Fluxo, etc., que aliam a velocidade de análise com uma boa qualidade de resultados.

Nessas técnicas instrumentais não é obtida uma informação direta do resultado, mas sim uma grande quantidade de sinais (curvas, picos) que devem ser tratados matematicamente para então ser possível a quantificação das espécies presentes.

Muitas vezes, o sinal registrado pelo instrumento analítico está composto, ou seja, o sinal obtido não é devido a um único constituinte presente na amostra. O detector indica um sinal total cuja origem é indistinguível para ele, e que, frequentemente, deve ser resolvido em termos dos sinais dos constituintes.

A Resolução de sinais analíticos pode ser definida como *um processo no qual um sinal composto é reduzido para formas simples* [1].

O problema da resolução é frequente, e sua solução normalmente requer muito tempo de trabalho experimental, chegando em certos casos a ser impossível dentro das condições de trabalho existentes.

Na década passada um certo número de químicos "redescobriram" o poder dos métodos de estatística multivariada, estudando suas aplicações na solução de problemas químicos. Estatística Multivariada pode ser entendida como um conjunto de técnicas matemáticas que podem ser aplicadas à analise de dados, quando várias medidas são realizadas para uma mesma amostra.

Esses estudos levaram ao desenvolvimento de uma nova área na química : a *Quimiometria* [2]. Esta disciplina pode ser visualizada como a parte da química que utiliza métodos matemáticos e estatísticos para :

- definir ou selecionar as condições ótimas das medidas e experimentos,

- permitir a obtenção do máximo de informação a partir da análise dos dados químicos.

O número de métodos matemáticos utilizados na área da Quimiometria tem apresentado nos últimos anos considerável expansão, com um crescente número de pesquisadores interessados nessa área.

Ao lado dos procedimentos clássicos que vizavam essencialmente a classificação de amostras segundo suas propriedades, alguns métodos modernos permitem a quantificação de variáveis e mesmo a previsão de resultados em sistemas multivariados.

1. - OBJETIVOS :

O objetivo deste trabalho é investigar a aplicação de métodos de Estatística Multivariada, mais especificamente de Calibração Multivariada para tentar resolver o problema da superposição de sinais, como uma alternativa ao trabalho experimental.

Para que se pudesse demonstrar a aplicação desses métodos na prática, escolheu-se o caso de picos superpostos em Cromatografia Gasosa.

Este trabalho não pretende ser um estudo completo de todas as possibilidades da utilização da Calibração Multivariada; da mesma maneira também não se dispõe a ser um estudo específico em cromatografia. Ele pode ser encarado como uma investigação da possibilidade da utilização de alguns métodos computacionais para quantificar sinais analíticos superpostos, que por conveniência são cromatográficos.

2. - A RESOLUÇÃO DE PICOS CROMATOGRÁFICOS SUPERPOSTOS :

A Cromatografia Gasosa é atualmente uma das técnicas analíticas mais largamente empregada, sendo sua principal aplicação nas análises quantitativas.

Uma vez que a quantificação dos constituintes envolve a relação da área de cada pico com a quantidade de amostra presente, é desejável que os picos estejam totalmente resolvidos, para que seja possível estimar corretamente a área de cada pico de interesse.

A ocorrência de picos superpostos em cromatografia gasosa é um problema prático bastante comum. Normalmente procura-se contorná-lo otimizando alguns parâmetros cromatográficos, como a temperatura da coluna e a vazão do gás de arraste, o tamanho da coluna ou ainda a mudança do tipo de coluna. Esses procedimentos em certos casos são improdutivos.

Quando o grau de superposição dos picos é moderado, métodos baseados em diferentes técnicas de separação linear [3], têm sido utilizados.

Surgiram vários procedimentos para a separação linear, e dentre eles pode-se destacar o "Método da Perpendicular" [4] (figura 1).

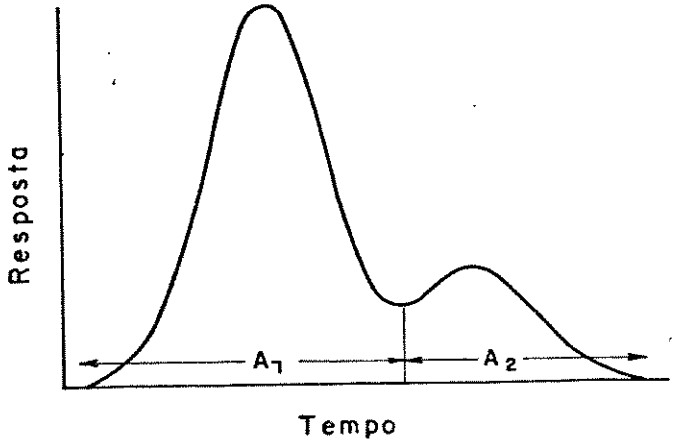


Figura 1 - O método da perpendicular. A área do pico da esquerda (A_1) é toda aquela à esquerda da linha perpendicular traçada no ponto mínimo. A área do pico da direita (A_2) é aquela à direita.

Neste método, os picos superpostos são resolvidos assumindo que a área de um pico é toda aquela tomada a partir de uma linha perpendicular traçada no ponto mínimo entre os picos. O método obviamente requer a existência de um vale entre os picos.

Outro método para a separação linear é o da "triangulação" [5]. A região da superposição é separada em proporção às áreas de triângulos retângulos, como ilustrado na figura 2.

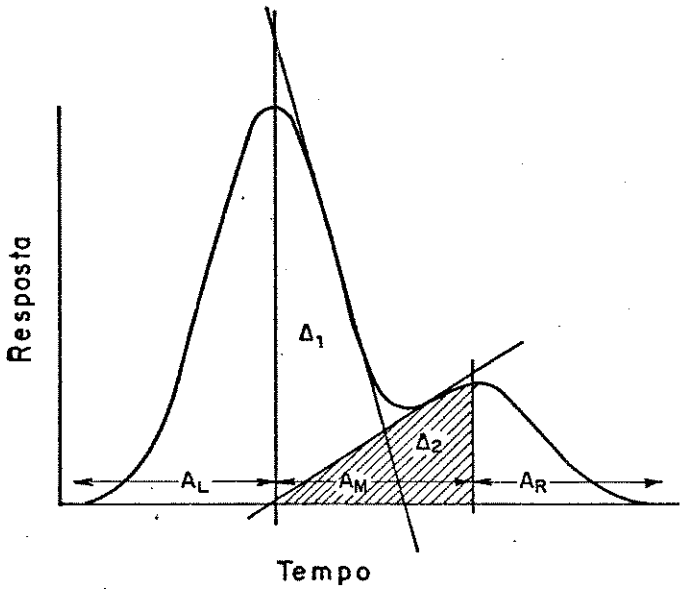


Figura 2 - O método da triangulação. A área "A_L" é para a parte do pico da esquerda sem superposição, e a área "A_R" para o pico da direita. A área "A_M" é a área superposta a ser resolvida.

São formados dois triângulos retângulos cujas áreas são Δ_1 e Δ_2, que correspondem à região superposta. Os picos são resolvidos de forma que :

$$A_1 = A_L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times A_M \qquad \text{p/ o pico da esquerda}$$

$$A_2 = A_R + \frac{\Delta_2}{\Delta_1 + \Delta_2} \times A_M \qquad \text{p/ o pico da direita}$$

Quando o grau de superposição aumenta muito ou quando há uma grande diferença nas áreas dos picos, os métodos da "perpendicular" e da "triangulação" dão resultados negativos e falham completamente.

Também existem outros métodos gráficos para a determinação da área de picos parcialmente superpostos, que podem ser encontrados na referência [6].

A grande limitação desses métodos motivou o desenvolvimento de outros mais sofisticados, e dentre eles, um que tem sido muito estudado para a resolução de picos superpostos é o *Método do Ajuste de Curvas* [1].

O procedimento consiste em inicialmente escolher uma função matemática para descrever cada pico cromatográfico. A função Gaussiana [7] foi a primeira a ser escolhida, mas como na prática muitas vezes não são obtidos picos deste tipo, funções gaussianas modificadas [8], funções de Poisson [9,10], funções bigaussianas [9,10], bem como outras geradas pela multiplicação ou combinações lineares de funções Gaussianas e de Cauchy [11] foram utilizadas para descrever os picos cromatográficos.

O pico cromatográfico é definido como uma função da posição "x", e de parâmetros estruturais "p" como : posição, altura, largura, etc. Assim para um pico com "k" parâmetros estruturais :

$$\text{PICO} = f(x, p_1, p_2, p_3, \ldots, p_k) \qquad (1)$$

A resposta cromatográfica "R", é definida como a soma das respostas de cada constituinte individualmente.

Para "L" constituintes :

$$R(x)=\sum_{i=1}^{L} f_i(x,p_1,p_2,p_3,\ldots,p_k) \qquad (2)$$

O computador faz repetidas combinações da soma dessas funções até encontrar os parâmetros que fazem com que o modelo melhor se ajuste aos dados experimentais. Esses parâmetros podem ser estimados por muitas maneiras, sendo o método dos mínimos quadrados o mais comum. Este método pode ser expresso como :

$$Y_j=R(x_j)+\epsilon_j \qquad (3)$$

onde : Y_j é o j-ésimo valor experimental, $R(x_j)$ é a respectiva resposta cromatográfica e ϵ_j é o erro do ajuste do j-ésimo ponto para o modelo. Os parâmetros são escolhidos de forma que para "n" valores experimentais :

$$\sum_{j=1}^{n} \epsilon_j^2 = \text{mínimo} \qquad (4)$$

Bons resultados foram obtidos com o ajuste de curvas [12], mas sua utilização fica limitada pela definição de uma função adequada para cada pico, além de ser sempre necessário estimar os parâmetros iniciais. Uma dificuldade adicional surge da necessidade de assumir que durante toda a análise o perfil do pico não sofrerá nenhuma modificação.

Devido a essas dificuldades, métodos de *Calibração Multivariada* podem ter sua utilização investigada. Estes métodos

não visam a resolução dos picos, uma vez que podem estimar as concentrações dos constituintes presentes sem a necessidade de se obter os picos devidamente separados.

A Calibração Multivariada corresponde a fundir em um único conjunto de dados, e dando um único tratamento, o que seria o resultado de várias curvas de calibração independentes se os picos estivessem separados.

Na Calibração Multivariada um conjunto de treinamento é obtido diretamente de cromatogramas de misturas de composição conhecida, a fim de gerar um modelo que descreve o comportamento das misturas como um todo.

Métodos baseados na Regressão Linear Múltipla a partir dos cromatogramas dos constituintes puros, ou em misturas desses, foram propostos [13]. A partir dos cromatogramas, são tomadas as alturas a vários tempos de eluição (digitalização) e seus valores são correlacionados com os dados de composição de amostras padrão.

Este método apresenta problemas quando há desvios da linearidade nas relações entre sinal e concentração. Ele também falha se componentes não identificados estão presentes, se há um grande ruído nos dados, ou quando há forte correlação nos dados obtidos a partir da digitalização dos cromatogramas.

A utilização da Regressão Linear Múltipla pode ser melhor realizada pela aplicação da *Análise de Componentes Principais* (PCA) [14] no conjunto de dados obtidos a partir dos cromatogramas digitalizados. Isto fornecerá um pequeno conjunto de novas variáveis, que são combinações lineares das variáveis originais, sendo também ortogonais entre si e, portanto, não correlacionadas,

que são utilizadas na regressão.

A Análise de Componentes Principais combinada com a Regressão Linear Múltipla foi denominada de *Regressão de Componentes Principais* (PCR) [15].

Um novo procedimento para Calibração Multivariada tem sido utilizado. É o *Método dos Mínimos Quadrados Parciais* - PLS - "Partial Least Squares" [16,17] desenvolvido por H. Wold [18], que tem alguns pontos em comum com o PCR, uma vez que os dados obtidos pela digitalização dos cromatogramas das misturas padrão são descritos pelos componentes principais. Entretanto, no PLS as informações obtidas a partir dos valores das concentrações nas amostras padrão também são descritas pelos componentes principais e são utilizadas para estimar as novas variáveis independentes, o que não é o caso do PCR.

Alguns artigos [19,20] mostram que o PLS é uma boa alternativa para os métodos mais clássicos como a Regressão Linear Múltipla e a Regressão de Componentes Principais, uma vez que ele é mais estável com respeito aos parâmetros da modelagem quando diferentes amostras são utilizadas para a construção da calibração.

O método dos Mínimos Quadrados Parciais (PLS) tem sido satisfatoriamente aplicado em análises quantitativas com Ultravioleta [21], Infravermelho [22,23], Difração de Raios-X [24], Eletroquímica [25], Análise por Injeção em Fluxo [26] e em Cromatografia Líquida de Alta Eficiência para determinações simultâneas de proteínas [27,28].

No presente trabalho será investigada a aplicação do PLS, bem como dos Métodos de Regressão Linear Múltipla e Regressão de

Componentes Principais (PCR).

Os programas para os cálculos utilizando a Regressão Linear Múltipla e o PCR [29], foram desenvolvidos em linguagem FORTRAN 77 para microcomputadores compatíveis com IBM-PC em ambiente MS-DOS pelo grupo de Quimiometria da UNICAMP, enquanto que para os cálculos com o PLS, foi utilizado o pacote computacional SIMCA-3B [30], desenvolvido em BASIC para microcomputadores de 8 bits em ambiente CP/M, pelo grupo de Quimiometria da Universidade de Umeå, Suécia.

A seguir, no capítulo 2, será apresentada uma descrição da Regressão Linear Múltipla, da Análise de Componentes Principais e da Regressão de Componentes Principais, que são necessários para um entendimento do PLS, que será apresentado no final deste mesmo capítulo.

No capítulo 3 será apresentado o estudo da aplicação desses métodos para dados simulados, onde não existem erros experimentais, permitindo verificar a adequação dos métodos à solução do problema formulado.

O capítulo 4 trata da aplicação dos métodos para os casos reais, onde foram quantificadas misturas de Tolueno, Isoctano e Etanol com diferentes tipos de superposição, sendo uma conclusão final dada no capítulo 5.

CAPÍTULO II

DESCRIÇÃO DOS MÉTODOS DE CALIBRAÇÃO MULTIVARIADA

1. - Introdução

Matrizes e sua notação

Os dados químicos multivariados podem ser arranjados na forma de uma tabela de dados, onde *objetos* são dispostos em linhas e *variáveis* em colunas.

Os objetos frequentemente são compostos químicos, e as variáveis são muitas vezes derivadas das propriedades dos constituintes químicos dos objetos, como por exemplo medidas de concentração, pH, condutividade, alturas (ou áreas) de picos cromatográficos ou dados espectrais (massa, RMN, Raman, Infravermelho, Raios-X, etc.) que são convertidos em variáveis por digitalização.

A tabela de dados pode então ser representada como uma matriz do tipo mostrado na figura 3, com "n" linhas e "p" colunas, onde o elemento da matriz, x_{ki}, indica o valor do k-ésimo objeto e da i-ésima variável. Em outras palavras, o índice das linhas indica o objeto e o das colunas, a variável.

A notação normalmente empregada em álgebra linear e utilizada pelos principais autores de textos sobre Quimiometria [1,17] será utilizada durante a explanação dos métodos matemáticos utilizados.

Serão utilizadas letras maiúsculas em negrito para

representar matrizes, e letras minúsculas também em negrito para representar vetores (linhas ou colunas das matrizes). Por exemplo, $\mathbf{X}_{ixj}$ refere-se a uma matriz de dimensão ixj, ou seja tem "i" linhas e "j" colunas. Os vetores coluna serão indicados como $\mathbf{x}_j$; $\mathbf{x}_2$ por exemplo representa a segunda coluna da matriz X.

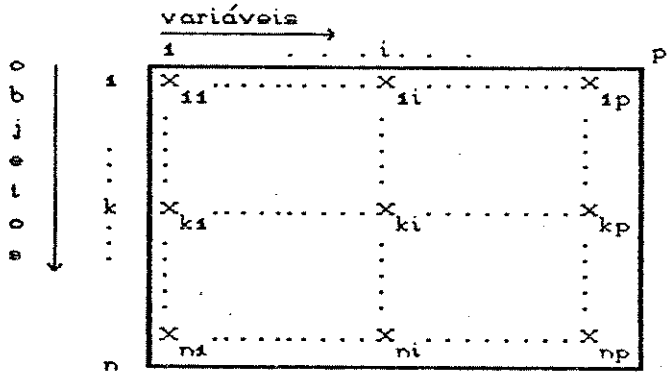


Figura 3 - A matriz de dados químicos

A transposta de uma matriz ou vetor será representada por um superescrito "T" , como por exemplo $\mathbf{X}^T$. A operação de transposição é dada pela troca das linhas pelas colunas. Por esta convenção $\mathbf{x}_2^T$ representa a segunda coluna da matriz X, escrita como um vetor linha. Todos os vetores serão vetores coluna. Os correspondentes vetores linha serão designados como vetores transpostos.

As letras comuns representarão os escalares, que podem ser elementos de matrizes (a_{ij}), de vetores (a_j), ou outras constantes como coeficientes de regressão (b).

A Relação entre duas Matrizes de Dados :

É possível estabelecer uma relação entre duas matrizes de dados X e Y, quando houver uma dependência entre as propriedades que descrevem cada uma delas. Sinais analíticos podem ser relacionados com quantidades de substâncias, intensidades de absorção no infravermelho com valores de momento de dipolo e assim por diante. A forma de estabelecer esta relação é a base da Calibração Multivariada.

A Calibração Multivariada consiste basicamente de duas fases : a calibração e a previsão.

Na fase de Calibração, tomando como exemplo o caso da cromatografia a ser estudado neste trabalho, é montada uma matriz de dados (X) das respostas instrumentais (pela digitalização dos cromatogramas), para um conjunto de amostras com composição conhecida. Também uma matriz com os valores de concentração (Y) em cada amostra é formada. Os dados utilizados nesta etapa constituem o *Conjunto de Treinamento*.

O próximo passo é escolher um método matemático apropriado que melhor possa reproduzir Y a partir dos dados da matriz X.

O modelo matemático desenvolvido na fase de calibração é então utilizado na fase de previsão para estimar as concentrações dos constituintes de novas amostras, a partir de seus cromatogramas digitalizados. Os dados utilizados nesta fase formam o *Conjunto Teste*.

Os dados para a Calibração Multivariada podem ser organizados conforme a figura 4. Os cromatogramas digitalizados são as

variáveis independentes, e as concentrações das substâncias nas amostras, as variáveis dependentes.

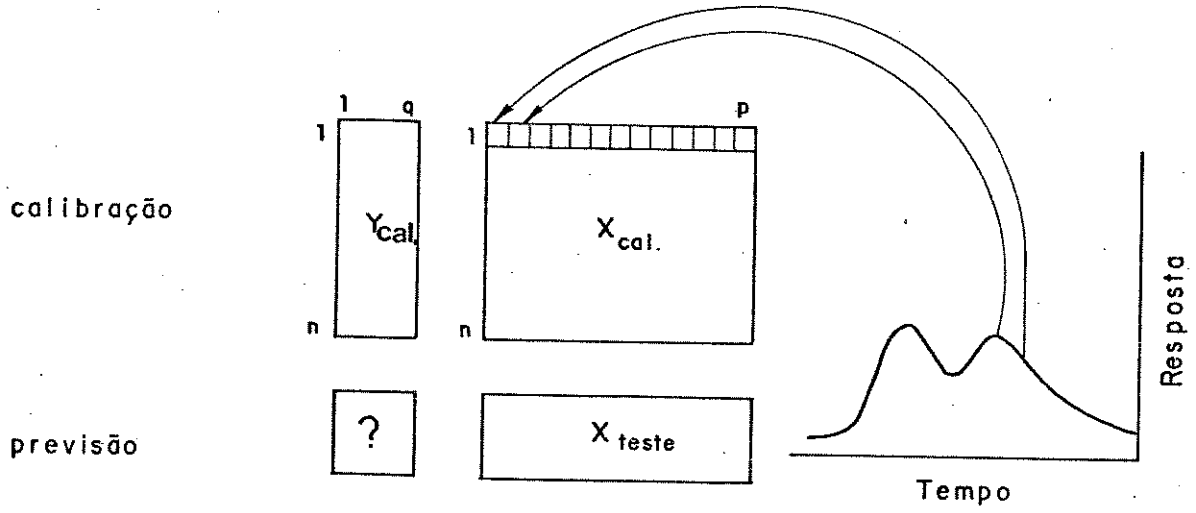


Figura 4 - Organização dos dados para a Calibração Multivariada. Neste exemplo, $X_{cal.}$, que é a matriz das variáveis independentes, tem "n" linhas de cromatogramas e "p" colunas correspondentes as alturas medidas em diferentes tempos de eluição (digitalização). A matriz das variáveis dependentes, $Y_{cal.}$, tem dimensão nxq, o que corresponde a "n" amostras com "q" concentrações dos contituintes químicos em cada uma delas.

Os termos *Bloco Dependente* e *Bloco Independente* são muitas vezes utilizados para descrever os blocos de variáveis dependentes e independentes respectivamente.

Pré-tratamento dos Dados :

Muitas vezes antes que o modelo seja desenvolvido é conveniente tratar o conjunto de dados originais de modo a tornar os cálculos computacionais melhor condicionados.

Normalmente, o primeiro passo envolve um *escalonamento* [31] dos dados a fim de que cada variável tenha a mesma influência no estágio inicial dos cálculos. O escalonamento é realizado pela divisão de todos os termos de uma certa variável pelo desvio padrão para esta variável, de forma que a variância (o quadrado do desvio padrão) torna-se unitária.

Isso assegura que as influências relativas das diferentes variáveis sobre os cálculos sejam independentes das suas unidades, já que todas as variáveis passam a ser expressas em unidades de desvio padrão.

Esse escalonamento é recomendado quando não há nenhuma informação prévia a respeito da importância das diferentes variáveis para a modelagem. Entretando, se algum tipo de informação é conhecida, pode-se decidir que certas variáveis são mais (ou menos) importantes que outras e atribuir a elas pesos proporcionais a esta importância.

Além de tornar a variância unitária, os dados são *centrados na média* [31], que corresponde a fazer com que para cada variável seus valores tenham média zero. O valor médio para cada variável é calculado a partir do conjunto de treinamento e então subtraído de cada um dos termos da variável correspondente. Isto significa mover o sistema de coordenadas para o centro dos dados.

Durante toda a descrição dos métodos matemáticos que serão discutidos a seguir, assume-se que todas as variáveis tem média zero.

Existem muitas outras formas de pré-tratamento dos dados, e uma delas consiste na normalização de cada espectro (cromatograma). Este pré-tratamento é aplicado às linhas da matriz de dados (objetos), ao contrário dos anteriores que eram aplicados às colunas da matriz (variáveis).

A normalização empregada consiste em dividir cada variável pela soma das variáveis correspondentes àquele objeto. Desta forma, para qualquer vetor linha da matriz de dados, a soma de todas as variáveis é sempre igual a um, ou seja, cada espectro (ou cromatograma) é normalizado à unidade.

A decisão da aplicação do pré-tratamento de dados costuma ser muito complicada, podendo ainda ser incluída nessa fase a *redução de variáveis* [32]. A redução de variáveis permite eliminar os termos que não sejam relevantes na modelagem, retirando-os da matriz de dados. Em nosso caso, como se verá adiante, a redução de variáveis obedeceu a critérios particulares, em função da natureza dos dados.

2. - A Regressão Linear Múltipla :

Na Regressão Linear Múltipla assume-se que a concentração (y) é uma função das respostas instrumentais ($x_1, x_2, \ldots, x_q$). Para respostas lineares, a concentração de um composto pode ser escrita como :

$$y = x_1 b_1 + x_2 b_2 + \ldots + x_q b_q + e \qquad (5)$$

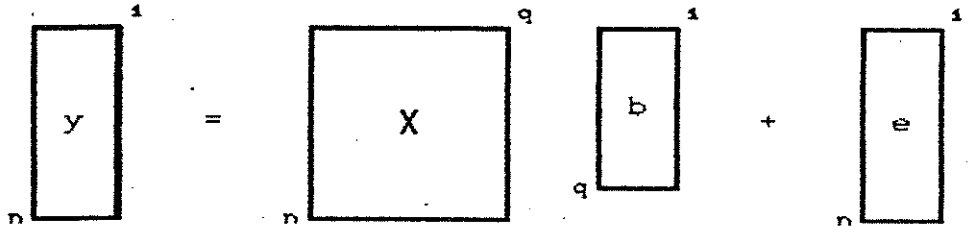
$$y = x^T b + e \qquad (6)$$

onde "b" é o vetor dos coeficientes de regressão e "e" é o resíduo ou erro.

A equação (5) descreve as dependências multilineares para apenas uma amostra. Se no conjunto de calibração existirem "n" amostras, os y_i (i=1,2,...,n) podem ser escritos como um vetor coluna y, permanecendo o vetor b, e os vetores x^T formando as linhas de uma matriz X :

$$y = X b + e \qquad (7)$$

Para um melhor entendimento dessas equações matriciais, pode-se fazer uma representação gráfica :



Os coeficientes de regressão que formam o vetor b, são estimados minimizando o valor do vetor de resíduos "e", na seguinte equação :

$$e = y - Xb \qquad (8)$$

Durante a fase de calibração, utilizando o critério dos mínimos quadrados, os coeficientes de regressão são encontrados por :

$$b = (X^T X)^{-1} X^T y \qquad (9)$$

Durante a fase de previsão, a concentração de uma amostra desconhecida pode ser obtida simplesmente por :

$$y = x_o^T b \qquad (10)$$

onde x_o^T representa o vetor das respostas instrumentais (variáveis independentes) da amostra cuja concentração será determinada.

Na equação (9) aparece o problema mais frequentemente encontrado na utilização deste método. Se existirem correlações altas entre as variáveis independentes, a inversa de $(X^T X)$ pode não existir. Colinearidade ou singularidade são nomes para este problema. Mesmo que $(X^T X)^{-1}$ exista, com correlações altas os valores previstos para "b" terão erros muito grandes.

Outra desvantagem deste método é que o número de variáveis independentes não pode ser maior que o número de misturas de calibração usado na análise. Se isso acontecer resultará em um número infinito de soluções, e somente será possível encontrar uma solução reduzindo-se o número de variáveis independentes.

Nesses casos, a *Regressão Linear Múltipla por Passos* [32] aparece como uma alternativa. Este método apresenta um problema muito sério, que é eliminar-se variáveis que contém informação relevante para a descrição do sistema.

A Regressão Linear Múltipla também pode ser extendida para o caso de mais de uma variável dependente. Por exemplo para o caso de duas variáveis dependentes, y_1 e y_2, pode-se simplesmente montar duas regressões e encontrar dois vetores de coeficientes b_1 e b_2 :

$$y_1 = Xb_1 + e_1 \quad ; \quad y_2 = Xb_2 + e_2 \qquad (11)$$

Arranjando y_1 e y_2 em uma única matriz e fazendo o mesmo para b_1 e b_2, e para e_1 e e_2, resulta :

$$Y = XB + E \qquad (12)$$

Uma explanação completa da Regressão Linear Múltipla pode ser encontrada na referência [32].

Este método de Calibração Multivariada baseado na Regressão Linear Múltipla, onde se assume que a concentração é função das respostas instrumentais, é conhecido como *Calibração Indireta* [33] ou *Inversa* [34]. Também existe a *Calibração Clássica* [34], com a resposta instrumental sendo colocada como função da concentração.

3. - Métodos baseados nos Componentes Principais :

3.1 - Análise de Componentes Principais (PCA) :

Na química, a *Análise de Componentes Principais* foi introduzida por Malinowski no final dos anos 60 com o nome de

"Análise de Fatores Principais", e a partir da década seguinte começaram a aparecer um grande número de aplicações. A cada dois anos publicam-se artigos de revisão sobre Quimiometria e as principais aplicações do PCA podem ser encontradas nas referências de [35] a [39]. Serão abordadas aqui apenas as partes fundamentais do PCA que devem ser compreendidas para que a Regressão de Componentes Principais (PCR) e o Método dos Mínimos Quadrados Parcipais (PLS) possam ser introduzidos. Maiores detalhes à respeito podem ser encontrados na referência [14].

Estruturas Geométricas no espaço-p :

As principais idéias do PCA são facilmente demonstradas em termos geométricos, devido à facilidade para se visualizar estruturas em espaços de duas e três dimensões. Uma vez entendida a idéia básica, é fácil generalizá-la para espaços de dimensões mais altas.

Cada objeto ou amostra descrito por um conjunto de "p" variáveis, pode ser representado como um ponto em um espaço p-dimensional obtido pela colocação de cada variável como um dos eixos de coordenadas. Assim, uma matriz de dados X com "n" objetos e "p" variáveis pode ser representada como um conjunto de "n" pontos em um espaço p-dimensional.

Uma vez que espaços com p>3 não são possíveis de visualizar, os de duas e três dimensões podem ser usados como ilustração. Conceitos geométricos como pontos, retas, planos, distâncias e ângulos têm o significado equivalente em espaços com muitas

dimensões ou em espaços menores.

Considere-se a seguinte matriz 2x3, com 2 objetos e 3 variáveis :

$$X = \begin{bmatrix} 2 & 3 & 2 \\ 1 & 1 & 3 \end{bmatrix}$$

A matriz X pode ser representada como dois pontos (os 2 objetos, ou seja, as 2 linhas) em um espaço tridimensional (3 variáveis, ou seja, 3 colunas). A figura 5 ilustra a representação geométrica da matriz X.

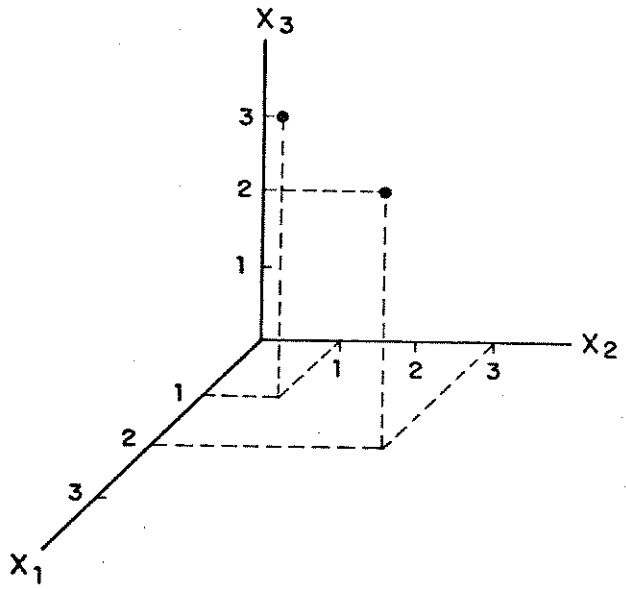


Figura 5 - Representação Geométrica da matriz X.

No caso a ser estudado neste trabalho os cromatogramas são digitalizados, onde se obtém "p" variáveis. Os "n" cromatogramas a serem analisados, cada um com "p" variáveis, fornecem um conjunto de dados muito extenso, o que impede sua análise direta. Se trabalharmos num espaço p-dimensional cada cromatograma torna-se um único ponto e o problema se reduz a analisar os "n" pontos nesse novo espaço.

Um reconhecimento visual dos pontos no espaço permite compreender mais facilmente o problema, mas isso só pode ser feito diretamente quando $p \leq 3$.

A Análise de Componentes Principais pode ser entendida geometricamente como um método para encontrar uma estrutura linear (reta, plano, hiperplano) que adequadamente modele o conjunto de pontos, fazendo com que um espaço de dimensão alta possa ser reduzido a um outro com menor dimensão, tornando possível a visualização dos pontos.

O homem, embora sua memória deixe muito a desejar, excede em muito o computador na capacidade de reconhecer tendências ou padrões de objetos no espaço. O computador é eficiente e rápido para cálculos. Desta forma, com o PCA, torna-se possível aliar a grande capacidade de memória do computador com a habilidade humana para reconhecer essas tendências ou padrões.

Conceitos Matemáticos da Análise de Componentes Principais

Matematicamente o PCA [40] corresponde à decomposição da matriz de dados **X** numa soma do produto de dois vetores, **t** e **p**,

mais uma matriz de resíduos E :

$$X = t_1 p_1^T + t_2 p_2^T + \ldots + t_a p_a^T + E \qquad (13)$$

Para ilustrar o que t e p significam, um exemplo para duas variáveis em um plano bidimensional é mostrado na figura 6. Extensões para mais dimensões são facilmente realizáveis, porém muito difíceis de serem visualizadas.

No exemplo da figura 6.A, é construída uma reta que melhor se ajusta aos pontos, de tal forma que os desvios perpendiculares são os menores possíveis no sentido dos mínimos quadrados. Esta reta de mínimos quadrados é aquela (entre todas as possíveis) que contém o máximo de variância dos dados, ou seja, a reta que explica o máximo de informação estatística. Esta reta corresponde ao *primeiro componente principal*.

Os coeficientes de direção desta reta (cossenos dos ângulos entre uma variável e o componente principal) são chamados de *loadings* (figura 6.B) , um para cada variável, e são representados pelo vetor p^T. Assim p^T é um vetor linha 1x2 com os elementos p_1 e p_2.

Projetando cada ponto na reta do componente principal, obtém-se os *scores*, um para cada objeto, que são representados pelo vetor t, um vetor coluna de dimensão 6x1 neste exemplo. Os *scores* são as coordenadas de cada ponto ao longo da reta do componente principal.

Também uma matriz de resíduos E, representando a parte não descrita pelo modelo, é formada pela subtração de cada elemento da

matriz original, x_{ki} de $t_k p_i$. Os elementos da matriz E são desta maneira formados por :

$$e_{ki} = x_{ki} - t_k p_i \qquad (14)$$

Desta forma, o modelo para primeiro Componente Principal é representado por uma reta no espaço p-dimensional, ou seja, uma equação linear com uma variável :

$$X = t_1 p_1^T \qquad (15)$$

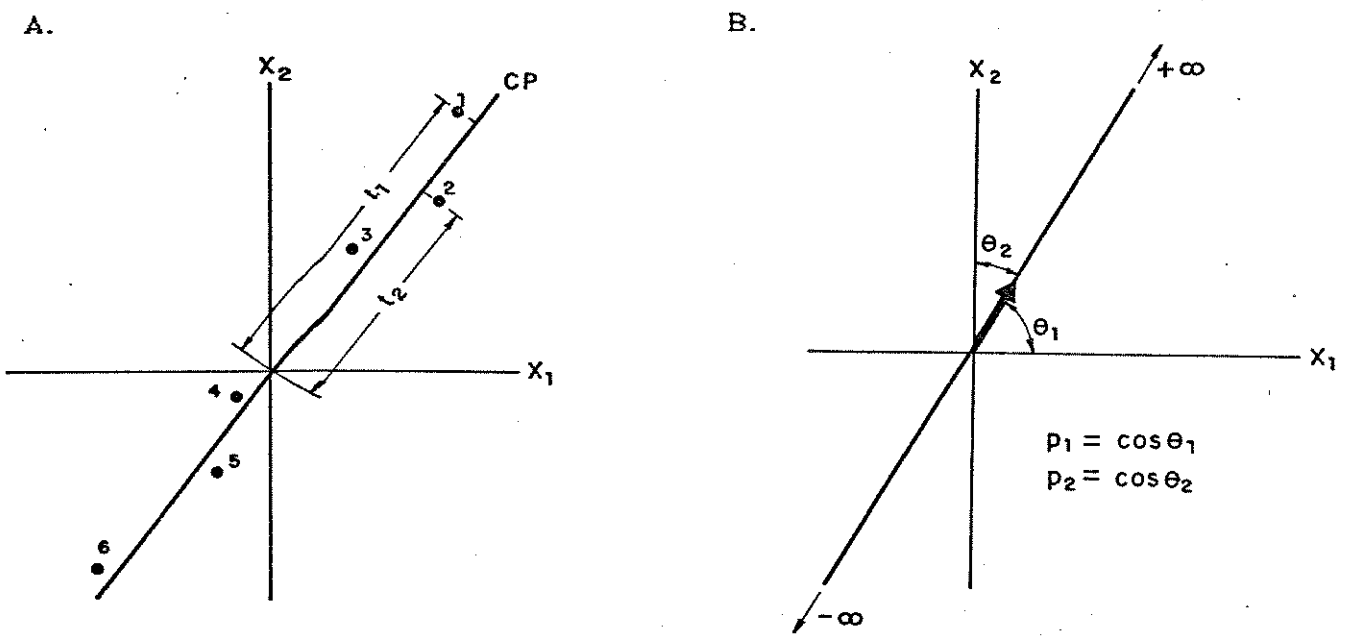


Figura 6 - O Primeiro Componente Principal no caso de 2 variáveis.

A. Os *scores* (t_1-t_6) são as projeções dos pontos (1-6) na direção do eixo do componente principal.

B. Os *loadings* (p_1 e p_2) são os cossenos dos ângulos entre cada variável e o eixo do componente principal.

OBS : Notar que os dados estão centrados na média.

Os resíduos podem ser novamente modelados, quando o primeiro componente principal não for suficiente para descrever toda a variância das amostras (a matriz de erros E permanece com valores muito grandes).

Pode-se obter desta forma um segundo componente principal, que é uma reta perpendicular ou ortogonal ao primeiro componente e que melhor se ajuste, novamente no sentido dos mínimos quadrados, aos resíduos a serem modelados. Entre todas as retas perpendiculares à do primeiro componente principal, a do segundo é a que contém a maior variância possível.

A figura 7 mostra uma representação gráfica de uma matriz 30x2, novamente com 2 variáveis, e com os 2 componentes principais. Dois novos vetores, p_2^T e t_2, são obtidos. Estes são os *loadings* e os *scores* calculados da mesma maneira anterior, só que agora em relação ao segundo componente principal.

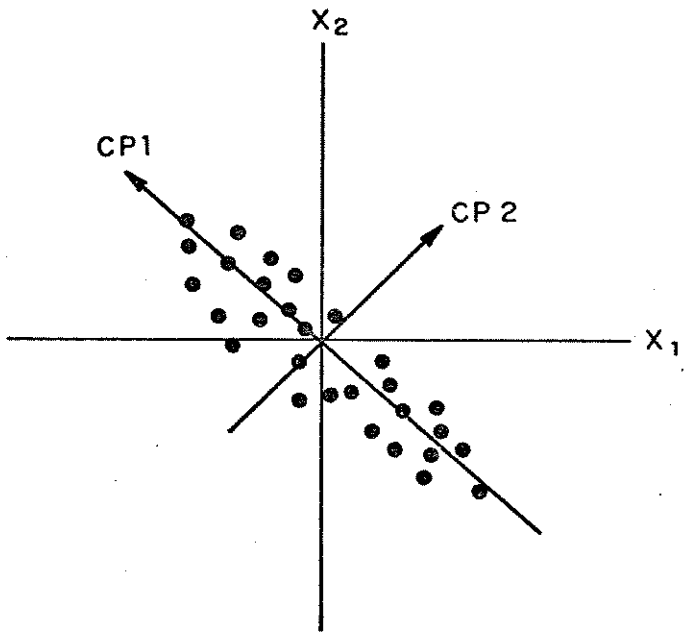


Figura 7 - Os dois primeiros componentes principais.

O modelo de dois componentes principais é representado por um plano no espaço p-dimensional :

$$X = t_1 p_1^T + t_2 p_2^T \qquad (16)$$

Novamente os resíduos podem ser modelados obtendo-se um terceiro componente principal ortogonal aos outros, e assim sucessivamente até que o valor dos resíduos seja zero ou possa ser desprezado quando comparado com o erro experimental.

Pode-se notar que os componentes principais não são calculados todos de uma única vez. Os *scores* e *loadings* são calculados aos pares para cada componente principal por um processo iterativo utilizando-se na maioria das vezes um algorítmo denominado NIPALS - *nonlinear iterative partial least squares*, desenvolvido por H. Wold [41].

No modelo desenvolvido para "A" componentes principais (A>1) tem-se a formação de uma matriz de *scores*, T, e uma matriz de *loadings*, P :

$X = t_1 p_1^T + t_2 p_2^T + \ldots + t_A p_A^T + E$, o que equivale à :

$$X = TP^T + E \qquad (17)$$

Os componentes principais podem ser usados como um novo sistema de eixos. Cada ponto (ou objeto) tem um novo conjunto de coordenadas a partir de suas projeções nestes novos eixos (os *scores*). Cada componente principal torna-se uma nova variável no

novo sistema.

A dimensionalidade do espaço original é igual ao número de colunas em **X**, ou seja, ao número de variáveis originais. A dimensionalidade do modelo descrito pelos componentes principais corresponde ao número de colunas em T (ou linhas em P), ou seja, ao número de componentes principais. Assim, se for possível descrever o sistema em estudo com um pequeno número desses componentes, o que se estará fazendo é diminuir a dimensão, praticamente sem perder informação estatística.

Utilizando-se dois componentes (geralmente os dois primeiros) é possível construir uma "janela" no espaço p-dimensional, e então projetar nela os pontos (ver figura 8). Isto fornece uma figura que mostra como devem estar dispostos os pontos no espaço original, que para espaços com dimensão maior que três é impossível de visualizar. Estes gráficos são feitos plotando-se as colunas de T umas contra as outras, que nada mais são do que os valores dos *scores* para cada componente principal.

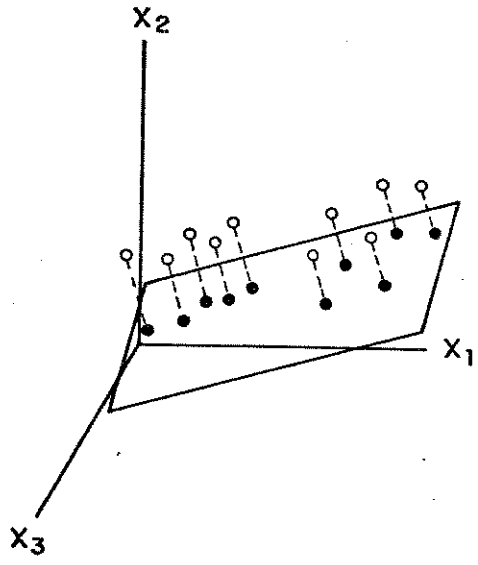


Figura 8 - A "janela" no espaço p-dimensional. Os pontos num espaço de três dimensões (p=3) sendo projetados num plano constituindo uma "janela" bidimensional. Esta "janela" é formada pelos dois primeiros componentes principais.

3.2 - A Regressão de Componentes Principais :

Este método para Calibração Multivariada pode ser dividido em duas fases. A primeira consiste em descrever o bloco das variáveis independentes (matriz **X**) pela modelagem de componentes principais, resultando na representação de **X** como sua matriz de *scores* **T**.

Desta forma um novo conjunto de variáveis independentes é gerado, com a propriedade de descrever os dados com um mínimo de perda de informação, a partir de um número reduzido de variáveis, com a vantagem de serem ortogonais entre si, não havendo,

portanto, correlação entre elas.

A transformação é :

$$T = XP \quad (= TP^{T}P = TI)$$

onde :

T é a matriz dos *scores*

X é a matriz das variáveis independentes originais

P é a matriz dos *loadings*

I é a matriz identidade

Na segunda fase do PCR, utiliza-se a Regressão Linear Múltipla para estabelecer a relação entre o bloco das variáveis dependentes (matriz Y) e a matriz dos *scores* T (o novo bloco das variáveis independentes).

Então a fórmula para a Regressão Linear Múltipla pode ser escrita como :

$$Y = TB + E \qquad (19)$$

e a solução para os coeficientes de regressão é :

$$B = (T^{T}T)^{-1}T^{T}Y \qquad (20)$$

Neste caso, a inversão de $T^{T}T$ não irá causar problemas devido à ortogonalidade mútua entre os *scores*. O PCR resolve desta forma o problema da colinearidade (garantindo que a matriz possa ser

invertida no cálculo de B).

Na utilização do PCR corre-se o risco de perder informação na escolha do número de componentes principais a serem utilizados, assim uma etapa fundamental é a escolha correta desse número.

O PCR também ignora toda a informação contida na matriz Y quando a modelagem de componentes principais é feita na matriz X. O bloco das variáveis dependentes só é utilizado na segunda fase, quando os componentes principais já foram determinados.

Maiores informações a respeito do PCR podem ser encontrados na referência [32].

3.3 - O Método dos Mínimos Quadrados Parciais - (PLS) :

O método dos Mínimos Quadrados Parciais, comumente chamado PLS, que é uma abreviação de *Partial Least Squares* proposto por H. Wold [18] tem sido utilizado na química em problemas de Calibração Multivariada.

O PLS é baseado numa extensão do algorítmo NIPALS [41], que decompõe a matriz de dados em uma soma do produto de dois vetores (os *scores* e os *loadings*). Um algorítmo para o PLS pode ser encontrado na referência [16].

Como mencionado anteriormente, é possível representar uma matriz de dados, sem perda de informação estatística útil, pela sua matriz dos *scores* com a vantagem de não haver correlação entre as variáveis.

No PLS tanto a matriz de variáveis independentes (X) como a das variáveis dependentes (Y) são representadas por seus *scores*,

pela modelagem de componentes principais. As equações para os dois blocos fica sendo :

$$X = TP^T + E = \sum_{h=1}^{a} t_h p_h^T + E \qquad (21)$$

$$Y = UQ^T + F = \sum_{h=1}^{a} u_h q_h^T + F \qquad (22)$$

com "a" componentes principais na modelagem.

Uma relação entre os dois blocos pode ser feita correlacionando-se os *scores* do bloco Y (u), com os *scores* do bloco X (t), para cada componente de cada vez. Um modelo linear é utilizado para esta relação :

$$u_h = b_h t_h \qquad (23)$$

$$\text{onde : } b_h = u_h^T t_h / t_h^T t_h \qquad (24)$$

para cada h = 1,2,...a, com "a" componentes principais.

Esse modelo, entretanto não é o melhor possível. Isto porque os componentes principais seriam calculados para os dois blocos separadamente, podendo resultar numa relação não muito satisfatória (não linear) entre os *scores* dos dois blocos. Seria mais interessante manipular a informação desses dois blocos simultaneamente para que se obtivesse a melhor correlação possível.

No PLS isto é feito por uma leve rotação dos eixos dos componentes principais, com consequente mudança dos valores dos *scores*, que são projeções dos pontos nestes eixos, de forma a produzir a melhor relação linear entre os *scores* dos dois blocos.

Como pode ser notado, no PLS há um compromisso entre a habilidade dos componentes principais em descrever as amostras nos espaços individuais (modelagem dos blocos X e Y), e o aumento na correlação entre t e u.

Pode ser produzida uma relação mista entre os dois blocos :

$$Y = TBQ^T + F \qquad (25)$$

onde o módulo de $|F|$ é minimizado com a condição que $|E|$ na equação 21 seja reduzido.

A análise de dados com o PLS pode ser sintetizada como a determinação dos componentes principais em X e Y, utilizando-se toda a informação disponível. O modelo final consiste das matrizes dos *scores* T e U que são linearmente relacionadas por um coeficiente B (B é diagonal).

Uma ilustração geométrica do PLS é mostrado na figura 9.

Previsão :

A parte fundamental de uma regressão é sua utilização na previsão do bloco dependente a partir do bloco independente conhecido.

Após o modelo ter sido estabelecido na fase de treinamento, a composição de uma nova amostra é estimada da seguinte forma :

▸ O cromatograma da amostra desconhecida (seguindo o exemplo da cromatografia já citado) é digitalizado, escalonado e centrado na média da mesma forma que na fase de calibração, produzindo um vetor das respostas instrumentais x_{teste}.

▸ Os parâmetros t_{teste} (vetor dos *scores*) e os resíduos "e" são então calculados pelo ajuste de x_{teste} para p (*loadings* do bloco X calculados na fase de treinamento):

$$x_{teste} = \sum_{a=1}^{A} t_a p_a^T + e \quad , \text{ para "A" componentes principais na modelagem.}$$

▸ A partir dos coeficientes de regressão b (equação 23), também calculados na fase de treinamento, os *scores* t_{teste} produzem uma estimativa do vetor das concentrações, já que :

$$y = \sum_{a=1}^{A} t_a b_a q_a^T \quad , \text{ para "A" componentes principais na modelagem.}$$

onde q_a (os *loadings* do bloco Y) foi calculado na fase de treinamento

▸ Os valores previstos de concentração estão escalonados e centrados na média, podendo ser transformados para as coordenadas originais pela simples aplicação das operações inversas

utilizadas.

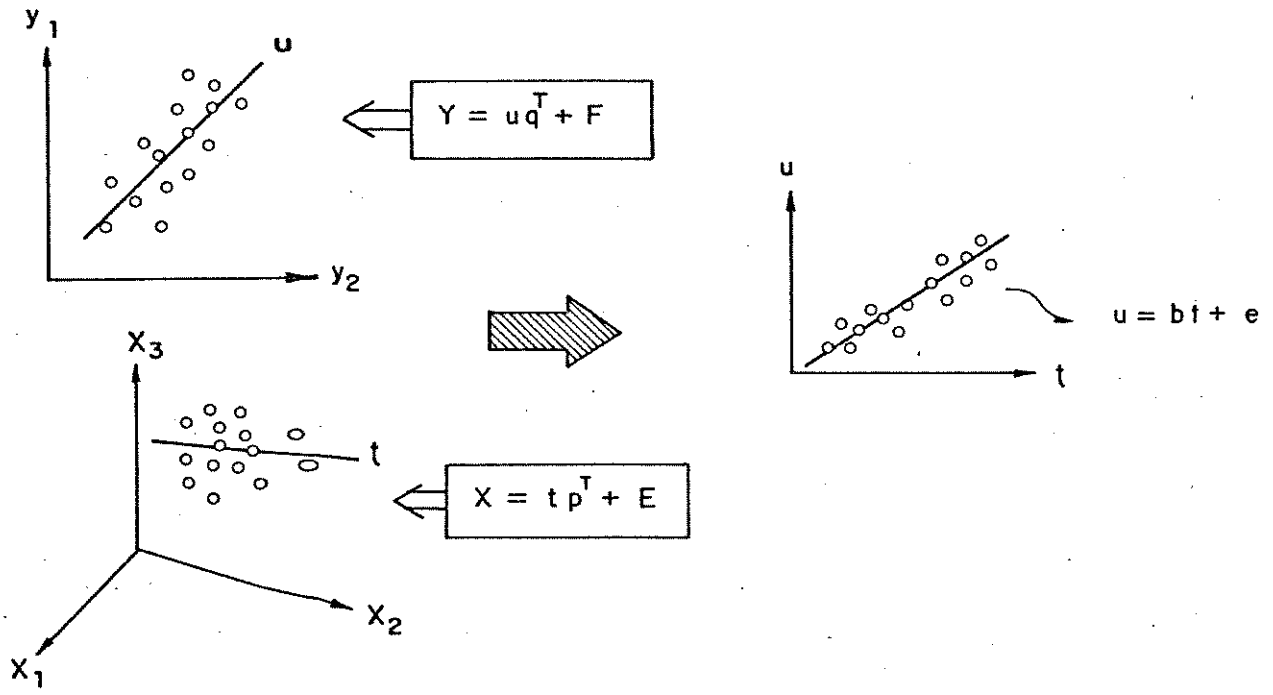


Figura 9 - Representação Geométrica do PLS com 1 Componente Principal modelando cada bloco. O bloco X contém 3 variáveis (x_1-x_3) e o bloco Y, 2 constituintes (y_1-y_2).

O Número de Componentes Principais :

Se a relação entre X e Y pode ser representada por modelos lineares, o número de componentes principais necessários para a descrição do sistema deve ser igual ao número de constituintes químicos presentes na amostra. Modelos não lineares requerem componentes principais extras, para modelarem exatamente essas não linearidades. Também a presença de interferentes, ou ainda a interação entre os constituintes pode fazer com que sejam necessários esses componentes extras.

O número de componentes principais a ser utilizado é de fundamental importância nos resultados a serem obtidos. A utilização de um número menor que o necessário fornecerá resultados não satisfatórios, uma vez que não se estará utilizando toda a informação possível a partir dos dados originais. Por outro lado, a utilização de um número de componentes principais superior ao necessário não faz mais do que modelar ruídos, aumentando desnecessariamente o número de variáveis e a complexidade do problema.

Existem muitas maneiras para a determinação do número ideal de componentes principais, podendo ser destacado um método denominado "CROSS VALIDATION" [16,42].

O número ótimo de componentes principais, também tem sido encontrado pela construção de um gráfico do desvio padrão residual (RSD) no conjunto teste contra o número de componentes principais [28], sendo o desvio padrão residual dado por :

$$RSD = \left[\frac{\sum (y - y_{prev})^2}{n} \right]^{1/2} \qquad (26)$$

onde "n" é o número de amostras no conjunto teste, "y" é o valor real e "y_{prev}" é o valor previsto pelo PLS.

O mínimo deste gráfico é tomado como tendo o número de componentes principais ideais.

Da mesma maneira tem sido feitos gráficos da soma dos quadrados dos erros de previsão (SS) para as amostras utilizadas

na fase de calibração [43], sendo essa soma dada por :

$$SS = \sum_{i=1}^{n} (y - y_{prev})^2 \qquad (27)$$

onde "n" é o número de amostras da calibração, "y" é o valor real, e "y_{prev}" é o valor previsto pelo PLS.

Ainda têm sido feitos gráficos da média dos erros de previsão no conjunto teste [24] contra o número de componentes principais, sendo novamente tomado o mínimo do gráfico como tendo o número ideal de componentes para a modelagem.

Identificação de amostras que não pertencem à modelagem :

Quanto mais próximos estejam dois pontos num espaço p-dimensional, mais semelhantes serão as amostras relativas a esses pontos.

Isso leva a uma abordagem particular na análise e comparação dos dados multivariados. Os pontos podem apresentar estruturas (ou sub-estruturas) que mostram agrupamentos dos objetos mais semelhantes, sendo que cada um desses agrupamentos é denominado de uma *classe* ou *categoria*. A figura 10 mostra dois agrupamentos de objetos similares no espaço, dando origem a duas classes distintas.

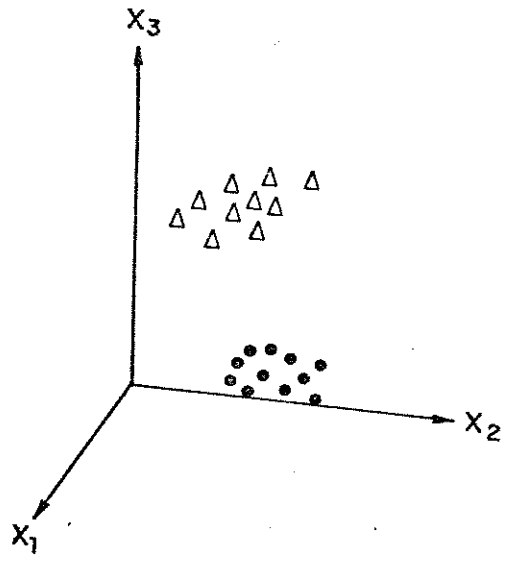


Figura 10 - Objetos em um espaço tridimensional formando duas classes distintas (triângulos e círculos).

Durante a modelagem pelo PLS, assume-se inicialmente que as amostras pertencem a uma única classe no bloco das variáveis independentes, e também a uma única classe no bloco das variáveis dependentes.

Desta forma devem ser criados mecanismos para determinar se uma amostra é similar ou não às outras pertencentes a esta *classe*.

Na equação (17) os elementos da matriz E representam a parte aleatória de **X**. Esta parte consiste de (i) erros nas medidas e outras imprecisões e (ii) erros de modelagem, ou seja, imperfeições nas aproximações. Os resíduos e_{ki}, elementos da matriz E, tem o desvio padrão, So :

$$So^2 = \sum_{k=1}^{n} \sum_{i=1}^{p} e_{ki}^2 \; / \; [(p-A)(n-A-1)] \qquad (28)$$

onde o denominador desta equação representa o número de graus de liberdade, e_{ki} representa o elemento da k-ésima linha e da i-ésima coluna da matriz de resíduos E (equação 17) após terem sido utilizados "A" componentes principais, sendo "p" o número de variáveis e "n" o número de amostras.

O desvio padrão residual "S_o" pode ser entendido como uma distância "típica" entre o modelo da classe (os componentes principais) e os objetos pertencentes a esta classe.

Os dados de um conjunto de treinamento, onde se conhece antecipadamente a classe à qual cada amostra deve pertencer, são utilizados para determinar o modelo da classe.

Um novo objeto na fase de previsão, do qual se conhece apenas as variáveis independentes (x_{teste}), é classificado como pertencente à classe pré-estabelecida pela comparação entre o desvio padrão residual no bloco X da fase de treinamento, "S_o" (equação 28), com o desvio padrão residual para este objeto que é dado por :

$$S_T^2 = \sum_{k=1}^{p} e_k^2 \ /(p-A) \qquad (29)$$

onde "p" é o número de variáveis, "A" o número de componentes principais; e os valores de e_k são calculados a partir dos valores de x_{teste}, t_{teste} (os *scores*) e b (os *loadings*) :

$$e_k = x_{teste} - \sum_{a=1}^{A} t_a b_a^T \qquad (30)$$

com "A" componentes principais.

Este desvio padrão residual para uma amostra no conjunto teste "S_T" é proporcional à distância entre o ponto que representa esta amostra e o modelo da classe no espaço p-dimensional, conforme mostrado na figura 11.

Se um objeto faz parte do conjunto de calibração, o desvio padrão residual para este objeto pode ser calculado da mesma maneira que para os do conjunto teste. É preciso multiplicá-lo por um fator de correção ϕ, devido aos diferentes números de graus de liberdade para as amostras no conjunto teste e no conjunto de calibração. Desta forma : $\phi=n/(n-A-1)$, e o desvio padrão residual para as amostras do conjunto de calibração fica sendo :

$$S_T = \sum_{k=1}^{P} e_k^2 \; (n/(P-A)(n-A-1)) \qquad (31)$$

Uma amostra pode agora ser classificada como provavelmente pertencente à classe definida se o valor de "S_T" não é muito maior que o valor "típico" para o desvio padrão residual desta classe "S_o", definido na equação 28.

Amostras no conjunto teste que são classificadas como não pertencentes à classe definida não podem ser analisadas e os valores de previsão não devem ser confiáveis. Isto significa que o modelo de calibração desenvolvido não descreve as propriedades dessa amostra, e seria necessário uma nova modelagem, com outros objetos.

Amostras no conjunto de calibração que não pertencem à classe pré-estabelecida devem ser retiradas da modelagem por não preencherem o requisito de uma única classe para cada bloco.

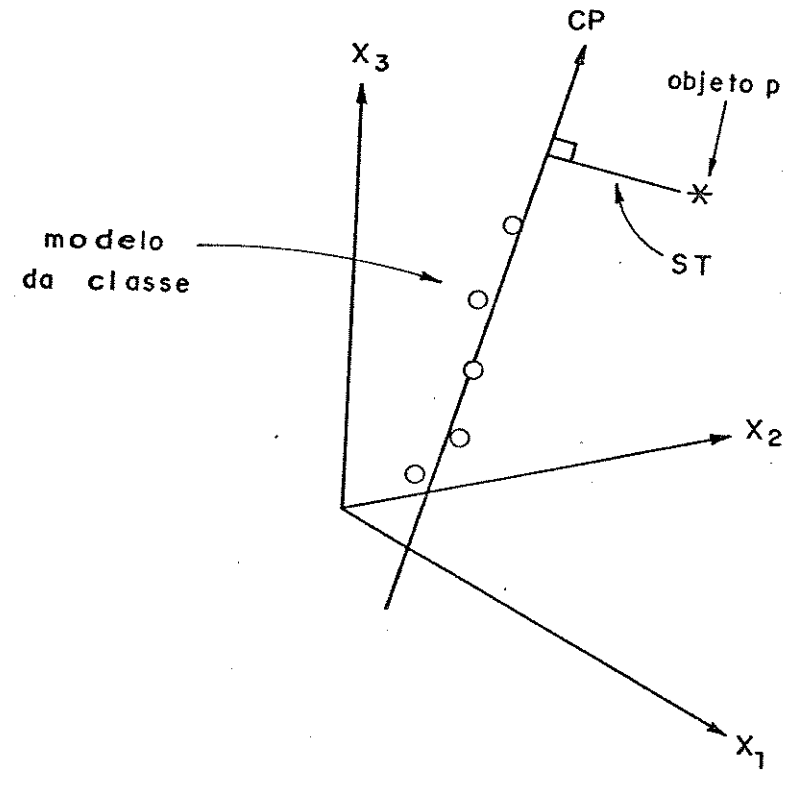


Figura 11 - A distância entre um ponto do teste (objeto "p") e o modelo da classe é proporcional ao desvio padrão residual "S_T" (equação 29).

CAPÍTULO III

ESTUDOS COM DADOS SIMULADOS

No capítulo anterior foram mostrados três métodos (PLS, PCR, Regressão Linear Múltipla) que podem ser utilizados em problemas de Calibração Multivariada. Para que se pudesse ter alguma indicação da aplicabilidade desses métodos na quantificação de picos superpostos, foram realizados inicialmente alguns estudos com dados simulados.

Essa fase da abordagem do problema tem por objetivo avaliar o desempenho dos métodos a serem utilizados, em condições onde se possa garantir a ausência de erros experimentais. Dessa forma, os erros de previsão de resultados que forem observados podem ser atribuídos diretamente à modelagem utilizada e à aplicabilidade do método para resolver este tipo de problema.

Os estudos foram realizados para o caso da superposição de sinais de dois constituintes, ou seja, a superposição de dois picos.

A função de Fraser-Suzuki [11] foi escolhida como modelo matemático para simular cada pico, sendo uma boa aproximação para os sinais obtidos na prática.

Esta função foi obtida a partir da combinação linear das funções Gaussiana e de Cauchy e pode ser expressa como :

$$f(t) = H\exp\{(-\ln 2/A^2)[\ln\{1+[A(t-t_r)/\sigma(2\ln 2)^{1/2}]\}]^2\} \quad (32)$$

onde :

"H" é a altura do pico

"σ" é o desvio padrão

"t_r" é o tempo de retenção

"A" é um fator de assimetria

A função é contínua para $A \neq 0$, e :

- se $A > 0$ então para $t > t_r - (2\ln 2)^{1/2}\sigma/A$,
- se $A < 0$ então para $t < t_r - (2\ln 2)^{1/2}\sigma/A$.

Como mencionado anteriormente, a resposta cromatográfica "R" pode ser definida como : $R = \sum_{i=1}^{L} f_i(t)$, para "L" picos (equação 2). Desta maneira um cromatograma simulado com dois picos é obtido pela soma de duas das funções escolhidas como modelo.

Na simulação assumiu-se que eram constantes os parâmetros "σ" e "A" da equação 32 em cada pico, sendo alterado o parâmetro "H", a altura do pico, para simular os diferentes cromatogramas.

Foram estudados quatro casos diferentes de cromatogramas simulados, onde a superposição dos picos variou até a condição de recobrimento total. Obtém-se os diferentes graus de superposição alterando-se o valor de "t_r" da equação 32, que corresponde ao tempo de retenção de cada um dos picos.

Os dois picos foram então definidos como :

$$\text{PICO 1 : } f(t) = H_1 \exp\{-4.3[\ln\{1+[0.4(t-t_r)/2]\}]^2\} \quad (33)$$

$$\text{PICO 2 : } f(t) = H_2 \exp\{-17.3[\ln\{1+[0.2(t-t_r)/2]\}]^2\} \quad (34)$$

A tabela 1 mostra os quatro diferentes tempos de retenção "t_r" utilizados, e um perfil de um cromatograma para cada um dos quatro casos é mostrado na figura 12.

Tabela 1 - Os tempos de retenção "t_r" utilizados.

	"t_r" do PICO 1 (u. a.)	"t_r" do PICO 2 (u. a.)
SUPERPOSIÇÃO 1	4	10
SUPERPOSIÇÃO 2	5	10
SUPERPOSIÇÃO 3	5	9
SUPERPOSIÇÃO 4	5	5

Na Calibração Multivariada a relação entre sinais (cromatogramas) e composição deve ser estabelecida. Assim é preciso encontrar uma maneira para expressar a quantidade de cada constituinte presente nos cromatogramas simulados.

Uma vez que a área de cada pico cromatográfico é proporcional à sua massa, na simulação tomou-se a área de cada pico gerado pela função de Fraser-Suzuki como sendo proporcional à quantidade presente desse constituinte.

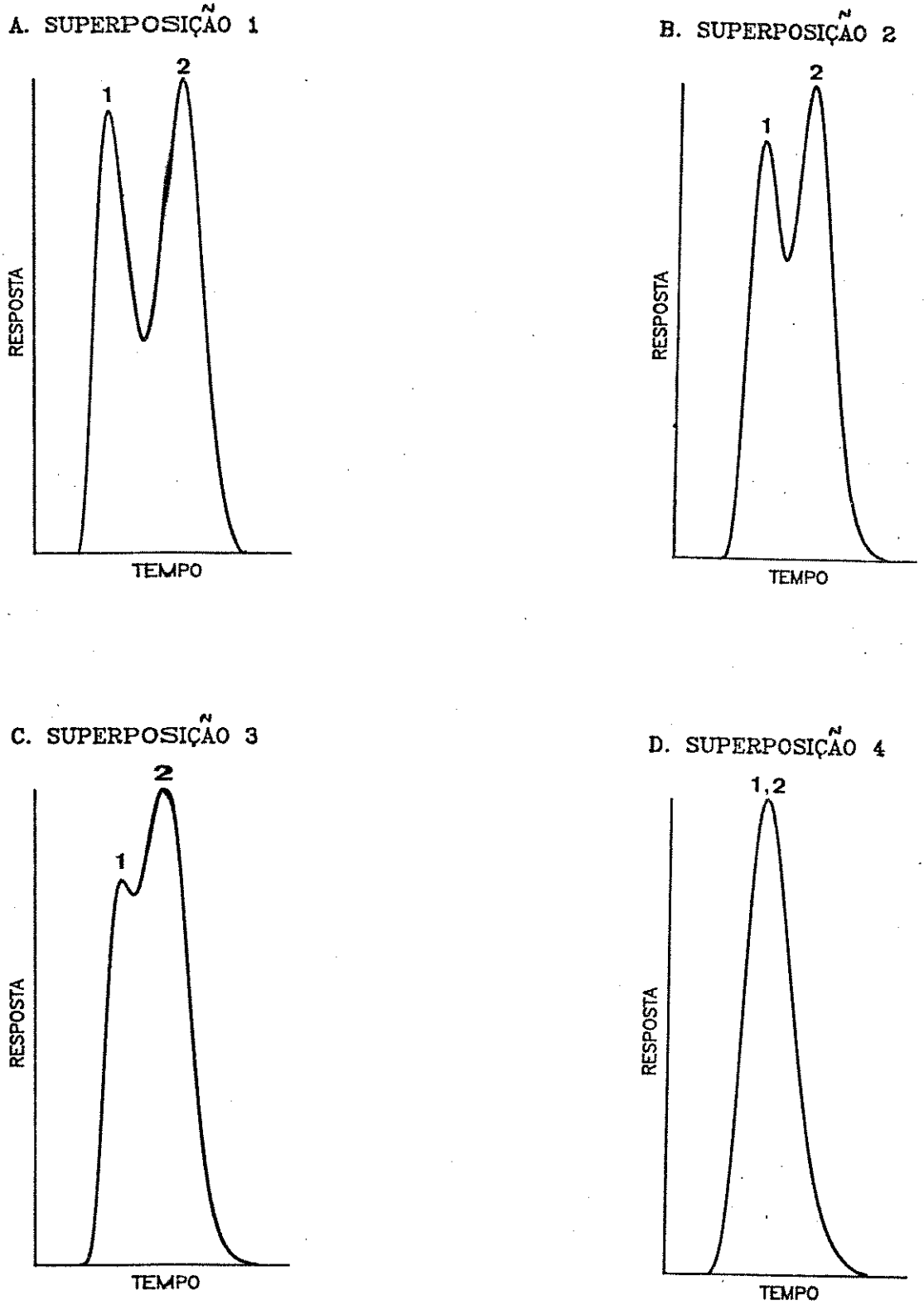


Figura 12 - O perfil de um cromatograma para os quatro casos de diferentes superposições estudadas.

Para o cálculo das áreas dessas funções foi utilizado um programa escrito em FORTRAN 77 que utiliza a regra de Simpson [44] para a integração numérica.

Assim os dados das curvas gerados pela soma das duas funções de Fraser-Suzuki podem ser correlacionados com os valores das áreas de cada uma dessas duas funções.

Foram gerados dez cromatogramas com diferentes valores de "H_1" e "H_2" (equações 33 e 34), para cada um dos quatro diferentes tipos de superposição. Em cada um dos cromatogramas simulados foram tomados 39 valores de "R", a resposta cromatográfica (equação 2), correspondentes a valores de "t" igualmente espaçados, o que implica em digitalizar os cromatogramas.

A tabela 2 mostra os valores das alturas "H_1" e "H_2" dos dois picos e as respectivas áreas produzidas.

Em cada cromatograma é produzido um vetor 1x39, que é uma linha da matriz das variáveis independentes da fase de calibração e que tem dimensão 10x39 para cada conjunto de dados. Outra matriz das variáveis dependentes da fase de calibração é formada pelos valores das áreas de cada um dos dois picos em cada cromatograma, tendo dimensão 10x2.

Para testar o método, outras três curvas foram geradas e não entraram na fase de calibração, constituindo o conjunto teste.

Desta forma foram gerados 13 cromatogramas simulados para cada um dos quatro conjuntos de dados. As áreas relativas aos picos são as mesmas para todos os conjuntos. O que varia de um conjunto para o outro é o valor do tempo de retenção dos picos.

Tabela 2 - Os valores das alturas dos picos e as respectivas áreas produzidas para os cromatogramas da fase de calibração.

AMOSTRA	PICO 1		PICO 2	
No.	H1(u.a.)	Área(u.a.)	H2(u.a.)	Área(u.a.)
1	8	36,086	8	34,558
2	10	45,108	8	34,558
3	10	45,108	6	25,918
4	9	40,597	8	34,558
5	8	36,086	6	25,918
6	9	40,597	7	30,238
7	9	40,597	10	43,197
8	6	27,065	9	38,878
9	7	31,576	8	34,558
10	6	27,086	8	34,558

Nos cálculos, tanto para o PCR como para o PLS, foram utilizados dois componentes principais, que para os quatro conjuntos de dados conseguem explicar mais de 99,5% da variância dos dados.

Os resultados obtidos na análise dos três cromatogramas do conjunto teste são mostrados na tabela 3 para o PCR, e na tabela 4 para o PLS.

Devido à alta correlação dos dados na matriz das variáveis independentes, a Regressão Linear Múltipla não consegue chegar a fornecer os valores de previsão, uma vez que sempre ocorrem

Tabela 3 - Previsões do PCR para os dados simulados (resultados em unidades arbitrárias).

AMOSTRA No.	PICO 1 VALOR REAL	PICO 1 PREV. PCR	PICO 1 ERRO[a] REL.(%)	PICO 2 VALOR REAL	PICO 2 PREV. PCR	PICO 2 ERRO REL.(%)
A. SUPERPOSIÇÃO 1						
1	36,086	36,085	0,00	30,238	30,238	0,00
2	31,576	31,576	0,00	38,878	38,878	0,00
3	31,576	31,576	0,00	43,197	43,197	0,00
média[b]			0,00			0,00
B. SUPERPOSIÇÃO 2						
1	36,086	36,092	0,02	30,238	30,251	0,04
2	31,576	31,577	0,00	38,878	38,897	0,05
3	31,576	31,581	0,02	43,197	43,212	0,04
média			0,01			0,04
C. SUPERPOSIÇÃO 3						
1	36,086	36,355	0,75	30,238	30,162	-0,25
2	31,576	31,333	-0,77	38,878	38,948	0,18
3	31,576	31,356	-0,70	43,197	43,264	0,16
média			0,74			0,20
D. SUPERPOSIÇÃO 4						
1	36,086	36,052	-0,09	30,238	30,284	0,15
2	31,576	31,420	-0,49	38,878	39,096	0,56
3	31,576	31,391	-0,59	43,197	43,457	0,60
média			0,39			0,44

a. ERRO REL.(%) $= \frac{(y_{prev} - y_{real})}{y_{real}} \times 100$

b. MÉDIA $= \frac{\sum |\text{ERRO REL.}|}{3}$

singularidades, ou seja, não é possível calcular a inversa de X^TX.

Os resultados contidos nas tabelas 3 e 4 indicam que parece não haver muita diferença entre os resultados do PCR e do PLS sendo os valores previstos praticamente idênticos em cada um dos quatro conjuntos de dados. Apenas para o conjunto "SUPERPOSIÇÃO 2" os valores previstos pelo PCR para o PICO 1 são melhores, se for considerado que uma diferença entre os erros relativos de 0,01% para 0,1% é realmente significativa.

Seria de se esperar que com o aumento na superposição entre os dois picos os valores previstos tivessem um erro cada vez maior. Isto é observado quando se parte do conjunto "SUPERPOSIÇÃO 1" até o conjunto "SUPERPOSIÇÃO 3". Já quando há a superposição total (conjunto SUPERPOSIÇÃO 4) isso não ocorre. Esse resultado deve-se ao fato de que os cromatogramas simulados são muito mais semelhantes uns com os outros nesse conjunto (há apenas um único pico), e assim esta classe é muito bem definida no espaço, o que não ocorre com o conjunto "SUPERPOSIÇÃO 3" onde se obtém cromatogramas com perfis bastante diferentes entre si.

O mais importante nesse ponto é destacar que tanto o PCR como o PLS permitem fazer excelentes previsões de resultados para o caso de sinais superpostos na ausência de erros experimentais nos dados de origem. Os erros relativos apresentados nas tabelas 3 e 4 são desprezíveis se comparados aos erros experimentais que se obtém no caso real, mesmo quando os picos se encontram completamente resolvidos.

Tabela 4 - Previsões do PLS para os dados simulados (resultados em unidades arbitrárias).

AMOSTRA No.	PICO 1 VALOR REAL	PICO 1 PREV. PLS	PICO 1 ERRO[a] REL. (%)	PICO 2 VALOR REAL	PICO 2 PREV. PLS	PICO 2 ERRO REL. (%)
A. SUPERPOSIÇÃO 1						
1	36,086	36,086	0,00	30,238	30,238	0,00
2	31,576	31,576	0,00	38,878	38,876	-0,01
3	31,576	31,576	0,00	43,197	43,197	0,00
média[b]			0,00			0,00
B. SUPERPOSIÇÃO 2						
1	36,086	36,119	0,09	30,238	30,245	0,02
2	31,576	31,614	0,12	38,878	38,888	0,03
3	31,576	31,632	0,18	43,197	43,200	0,01
média			0,13			0,02
C. SUPERPOSIÇÃO 3						
1	36,086	36,351	0,73	30,238	30,163	-0,25
2	31,576	31,337	-0,76	38,878	38,947	0,18
3	31,576	31,359	-0,69	43,197	43,263	0,15
média			0,73			0,19
D. SUPERPOSIÇÃO 4						
1	36,086	36,053	-0,09	30,238	30,284	0,15
2	31,576	31,422	-0,49	38,878	38,093	0,55
3	31,576	31,393	-0,58	43,197	43,454	0,60
média			0,39			0,43

a,b ver tabela 3.

CAPÍTULO IV

ESTUDOS COM DADOS EXPERIMENTAIS

1. Parte Experimental :

Reagentes :

Os experimentos foram realizados com misturas de Tolueno (Carlo Erba – p.a.), Isoctano (Carlo Erba – p/ cromatografia) e Etanol (Merck – p.a.) tratado previamente com peneira molecular [45] para eliminar água.

Misturas :

As amostras a serem utilizadas foram preparadas por pesagem direta do Tolueno, Isoctano e Etanol puros, e as respectivas massas injetadas no cromatógrafo foram calculadas a partir dos dados da densidade [46] à temperatura ambiente na injeção. Isto foi feito calculando-se inicialmente o volume total da solução a partir das massas pesadas de cada um dos constituintes. Sabendo-se que o volume injetado foi sempre mantido constante em 2μl, pode-se calcular as massas injetadas de cada constituinte individualmente. Foram preparadas vinte soluções a partir de diferentes massas de Tolueno, Isoctano e Etanol.

Cromatogramas :

Os dados cromatográficos foram obtidos a partir de um cromatógrafo VARIAN modelo 920, com detector de condutividade térmica e uma coluna de aço inoxidável de 2,0 metros de comprimento e 1/8 de polegada de diâmetro interno.

A coluna foi recheada utilizando-se como fase estacionária SE-30 a 5,2%, tendo como suporte Chromossorb W de 80-100 mesh.

O gás de arraste utilizado foi o hidrogênio com uma vazão de 30 ml/min., sendo a temperatura do injetor mantida a 160°C e do detector a 180°C.

Os cromatogramas foram obtidos a três diferentes temperaturas da coluna : 105°C, 120°C e 130°C, sempre sem alteração os outros parâmetros já citados acima. Desta forma, foi possível obter cromatogramas com diferentes graus de superposição entre os picos. A figura 13 mostra o exemplo de um cromatograma para cada uma das três diferentes temperaturas da coluna.

Assim, para cada uma das três temperaturas da coluna são obtidos vinte cromatogramas (perfazendo 60 no total) que formam três conjuntos de dados distintos, os quais serão posteriormente analisados em separado para verificar a adequação dos métodos matemáticos para o tratamento de casos com diferentes superposições entre os picos.

Digitalização

Todos os cromatogramas foram digitalizados manualmente.

medindo-se as alturas dos picos a 41 tempos de eluição diferentes e igualmente espaçados, de forma que cobrissem todo o cromatograma.

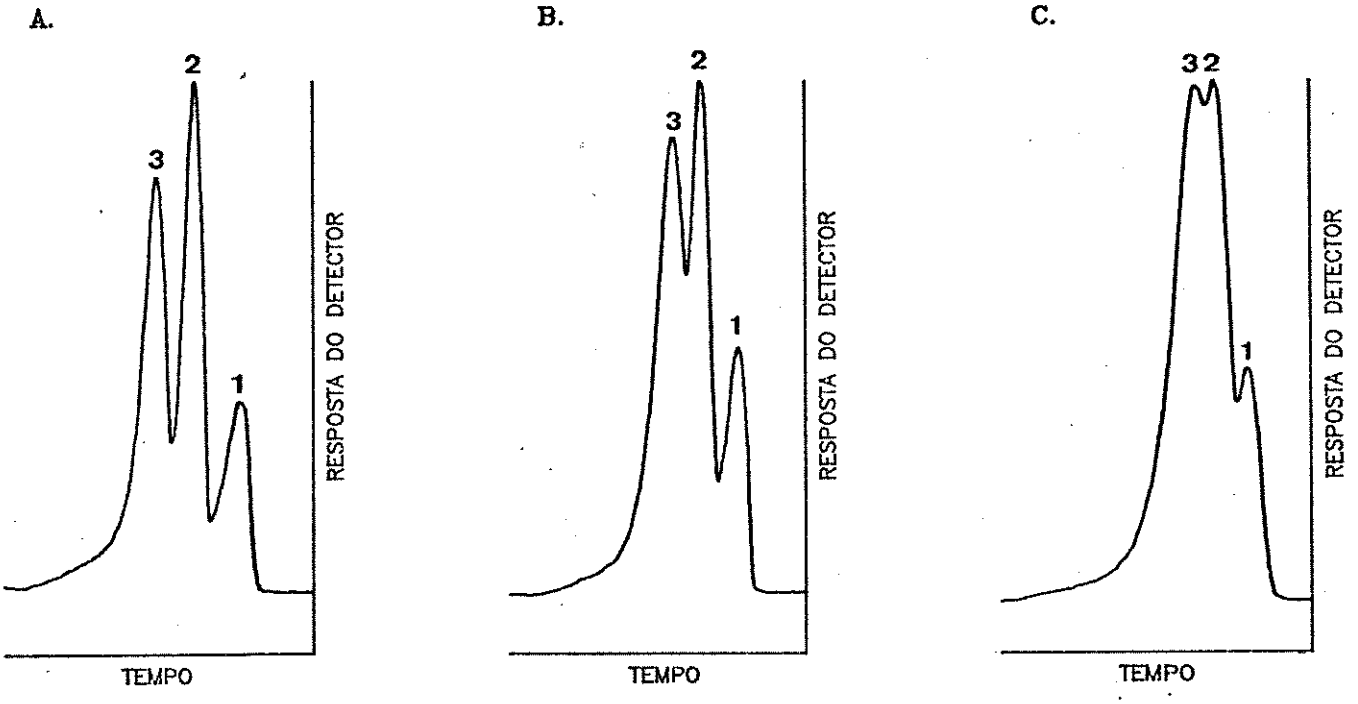


Figura 13 - Exemplo de um cromatograma para cada temperatura da coluna.

Condições :

Instrumento : VARIAN modelo 920

Detector : Condutividade térmica

Gás de arraste : H_2 a 30 ml/min.

Coluna : 5,2 % de SE-30 em Chromossorb W de 80-100 mesh

2,0 metros de comprimento e 1/8" de diâmetro interno

Temperatura do injetor : 160°C

temperaturas da coluna : A. 105°C, B. 120°C, C. 130°C

Amostras : 1. Etanol, 2. Isoctano, 3. Tolueno

Cada uma dessas alturas será uma variável independente na matriz de dados, e para assegurar que cada variável tenha o mesmo significado em cada um dos cromatogramas de um mesmo conjunto (mesma temperatura da coluna), foi necessário estabelecer critérios para o início e o fim da digitalização.

O início da digitalização foi estabelecido de tal forma que o máximo do primeiro pico (Etanol) correspondesse à sexta variável. O final da digitalização (variável 41) foi dado pelo ponto onde a tangente à curva de descida do último pico (Tolueno) cruza o eixo dos tempos (linha base). A figura 14 mostra o critério utilizado na digitalização de um cromatograma e os dados correspondentes a esta digitalização são apresentados na tabela 5.

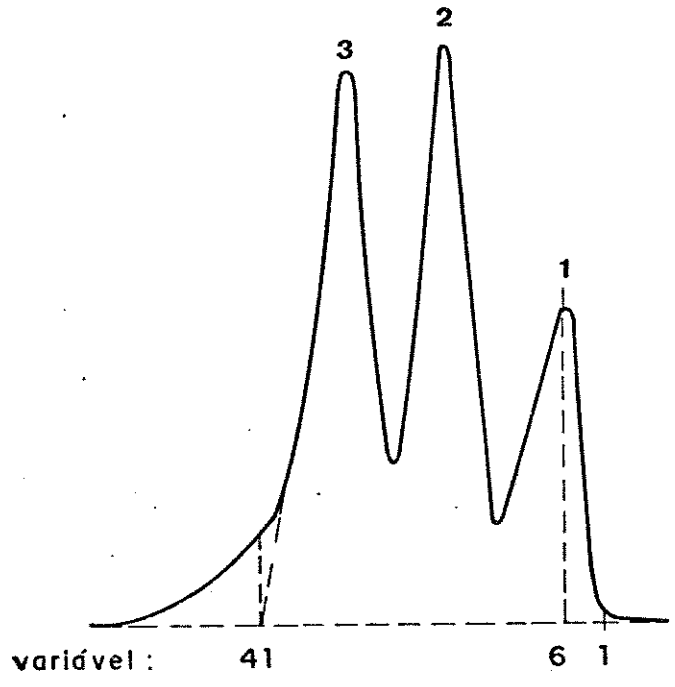


Figura 14 - O critério da digitalização. Amostras : 1. Etanol, 2. Isoctano, 3. Tolueno

Tabela 5 - Dados correspondentes à digitalização do cromatograma apresentado na figura 14.

variável :	1	2	3	4	5	6	7	8	9
valor :	1,0	2,0	16,0	45,0	65,5	73,0	64,0	55,0	47,5
variável :	10	11	12	13	14	15	16	17	18
valor :	40,5	33,5	28,0	23,5	25,0	58,0	100,0	125,0	123,0
variável :	19	20	21	22	23	24	25	26	27
valor :	106,5	85,0	61,5	45,0	35,5	43,5	78,5	109,0	120,0
variável :	28	29	30	31	32	33	34	35	36
valor :	114,5	101,5	80,5	64,5	50,0	35,5	30,0	24,0	22,5
variável :	37	38	39	40	41				
valor :	15,5	13,0	11,5	10,0	9,0				

Um disquete com todos os dados utilizados neste trabalho pode ser adquirido do autor, ou do orientador na UNICAMP. Neste disquete irão conter dados simulados e experimentais.

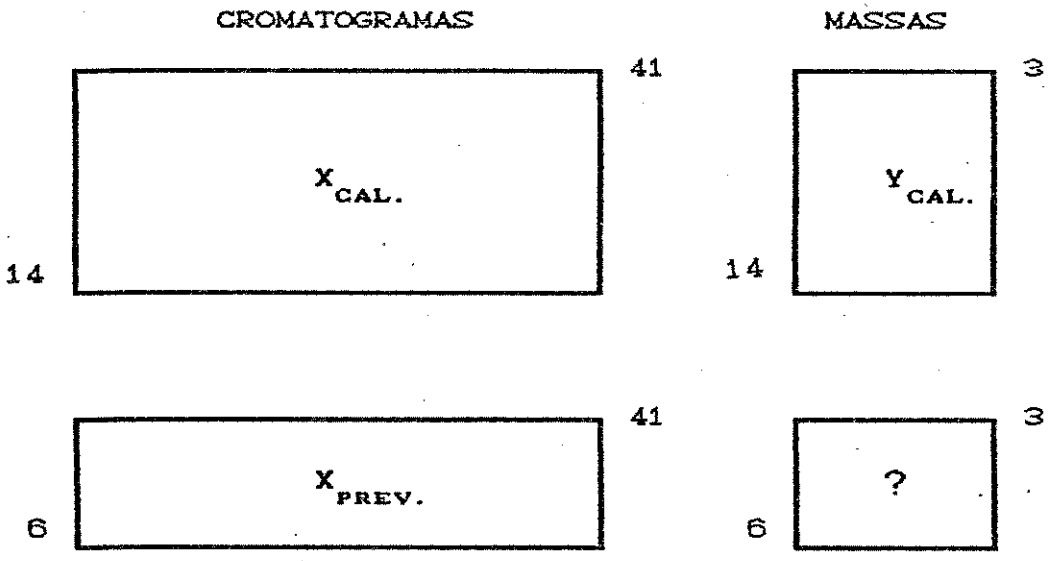
Organização dos Dados para a Análise Multivariada :

Dos vinte cromatogramas para cada conjunto, quatorze deles foram escolhidos para construir a modelagem. Tem-se a formação, após a digitalização, de uma matriz 14x41 para as variáveis independentes. Também é formada, com os valores das respectivas massas injetadas dos três constituintes, uma matriz 14x3, que corresponde ao bloco das variáveis dependentes. Essas duas

matrizes fazem parte da fase de calibração.

Os outros seis cromatogramas para cada uma das três temperaturas da coluna, que não entraram na fase de calibração são utilizados para testar os métodos de Calibração Multivariada. É formada uma matriz 6x41 para as variáveis independentes, que será utilizada na fase de previsão.

Um esquema das matrizes utilizadas nas duas fases é :



2. - Resultados e Discussões da Análise Multivariada :

2.1 - Método dos Mínimos Quadrados Parciais (PLS) :

Na previsão das massas pelo método PLS, todas as variáveis foram inicialmente escalonadas de forma que ficassem com desvio padrão unitário. As variáveis também foram inicialmente centradas na média, ou seja, todas as variáveis ficaram com média zero.

A partir das 41 variáveis independentes (cromatograma digitalizado) e das 3 dependentes (massas) nas 14 amostras da fase

de calibração e para os 3 conjuntos de dados, foram construídos modelos com o número de componentes principais variando de um até seis.

As massas correspondentes às outras 6 amostras, não utilizadas para a construção dos modelos, foram então previstas em todos os casos. Como as amostras utilizadas, embora tratadas como incógnitas neste estudo, são de fato conhecidas, é possível calcular o erro relativo cometido na estimativa da massa de cada constituinte. Esse erro relativo é utilizado como um critério prático para a seleção do melhor modelo a ser utilizado, no que diz respeito ao número de componentes principais empregados.

A partir dos valores dos erros relativos de cada amostra é possível calcular uma média, quando o número de componentes principais na modelagem varia de um até seis. A figura 15 mostra os três gráficos obtidos (um para cada conjunto de dados), quando são plotados a média dos erros relativos contra o número de componentes principais.

Para facilitar a apresentação dos resultados, os três conjuntos de dados passarão a ser chamados de :

- CRO 105 : para os dados com a temperatura da coluna em 105°C.

- CRO 120 : para os dados com a temperatura da coluna em 120°C.

- CRO 130 : para os dados com a temperatura da coluna em 130°C.

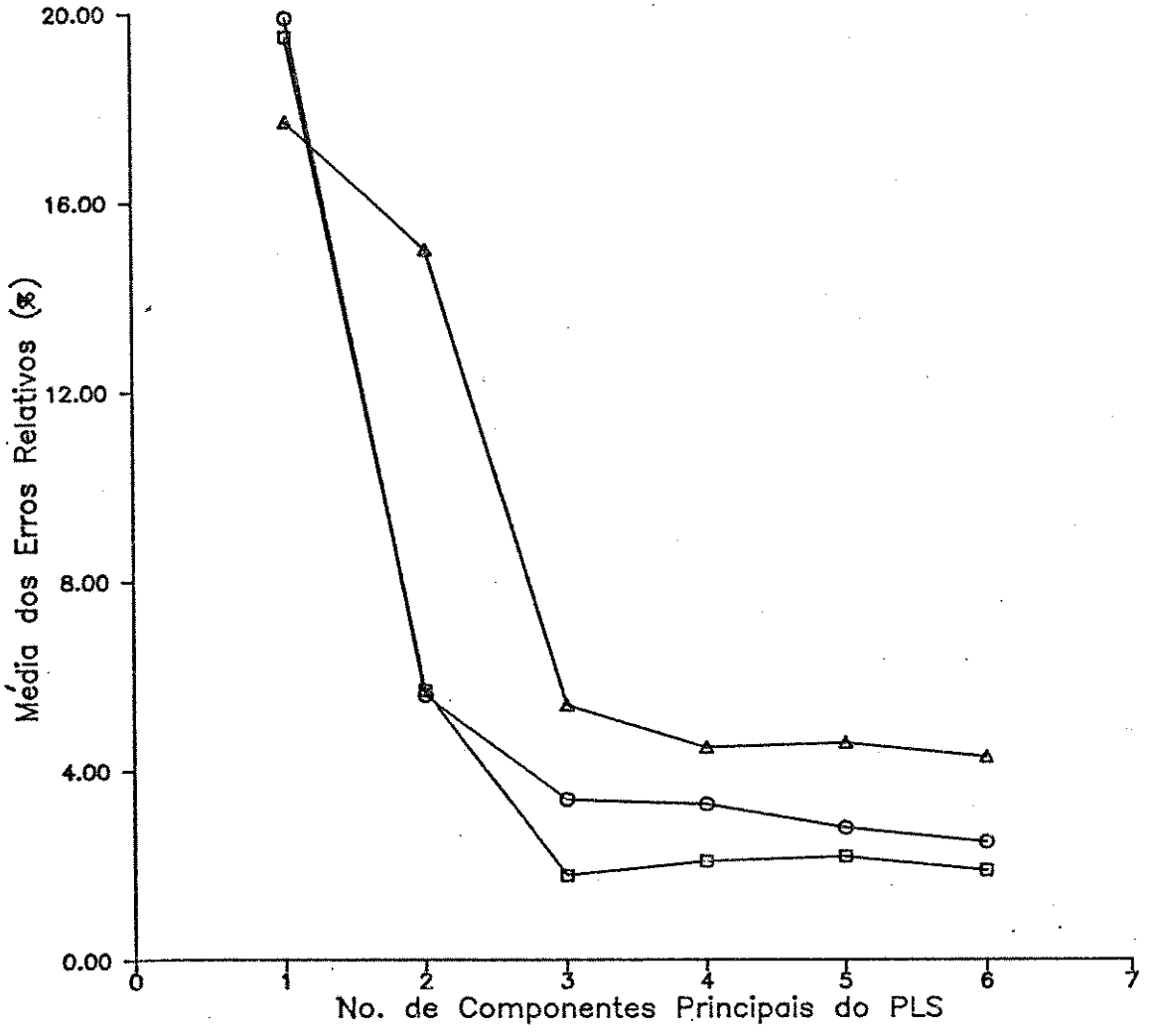


Figura 15 - Média dos Erros Relativos para diferentes números de componentes principais. □ CRO 105, Δ CRO 120, o CRO 130.

Por esses gráficos é possível verificar que praticamente não há mais diferença na média dos Erros Relativos porcentuais após o terceiro componente principal, ou seja, a utilização de 4, 5 ou 6 componentes não altera significativamente os resultados. O número ótimo de componentes principais a ser utilizado para os três conjuntos de dados deve ser três, e a variância explicada através

de cada componente principal para os blocos das variáveis independentes e dependentes, é mostrada para os três conjuntos de dados nas tabelas 6.a para o conjunto CRO 105, 6.b para o conjunto CRO 120 e 6.c para o conjunto CRO 130.

Tabela 6.a - Variância porcentual descrita pela modelagem com o PLS para o conjunto CRO 105.

No. de componentes principais do PLS	variáveis independentes		variáveis dependentes	
	cada	total	cada	total
1	43,15	43,15	43,55	43,55
2	40,40	87,55	53,55	97,01
3	4,25	91,80	2,45	99,46
4	3,32	95,12	0,14	99,60
5	1,10	96,22	0,22	99,82
6	1,14	97,36	0,09	99,91

Tabela 6.b - Variância porcentual descrita pela modelagem com o PLS para o conjunto CRO 120.

No. de componentes principais do PLS	variáveis independentes		variáveis dependentes	
	cada	total	cada	total
1	49,66	49,66	41,36	41,36
2	34,37	84,03	51,41	92,77
3	3,17	87,20	4,68	97,45
4	4,11	91,31	1,36	98,81
5	2,51	93,82	0,19	99,00
6	1,00	94,82	0,10	99,10

Tabela 6.c - Variância porcentual descrita pela modelagem com o PLS para o conjunto CRO 130.

No. de componentes principais do PLS	variáveis independentes		variáveis dependentes	
	cada	total	cada	total
1	60,73	60,73	40,80	40,80
2	27,88	88,61	52,37	93,17
3	3,41	92,02	5,58	98,75
4	2,45	94,47	0,78	99,53
5	1,60	96,07	0,16	99,69
6	0,64	96,71	0,13	99,82

Pelas tabelas 6.a, 6.b e 6.c é possível verificar que nos três conjuntos de dados os dois primeiros componentes principais explicam mais de 80% da variância dos dados, tanto no bloco das variáveis independentes como no das dependentes. Também é possível observar que a variância explicada quando são utilizados três componentes principais é superior a 90% para os conjuntos CRO 105 e CRO 130, enquanto para o conjunto CRO 120 esse valor é menor, ficando em torno de 87% no bloco das variáveis independentes.

Os resultados obtidos para as previsões das massas quando são utilizados três componentes principais são mostrados na tabela 7.

Uma análise dos resultados obtidos mostra que para os conjuntos CRO 120 e CRO 130, os erros de previsão para as massas de Tolueno e Etanol são maiores que para o Isoctano. Este fato parece indicar que, principalmente nos casos onde a superposição dos sinais é maior, a qualidade da previsão das massas fica

comprometida para os picos que ocupam as extremidades do cromatograma.

Tabela 7 - Previsões pelo método PLS para a modelagem com 3 componentes principais (massas em mg).

amostra no.	TOLUENO			ISOCTANO			ETANOL		
	VALOR REAL	PREV. PLS	ERRO[a] REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)
A. CRO 105									
1	0,631	0,645	2,2	0,403	0,405	0,5	0,529	0,513	-3,0
2 *[b]	0,202	0,262	29,7	0,895	0,819	-8,5	0,359	0,389	8,4
3	0,396	0,409	3,5	0,615	0,599	-2,6	0,501	0,508	1,4
4	0,415	0,413	-0,5	0,303	0,294	-3,0	0,839	0,852	1,6
5	0,613	0,591	-3,6	0,608	0,605	-0,5	0,311	0,334	7,4
6	0,521	0,523	0,4	0,496	0,515	3,8	0,523	0,500	4,4
média[c]			2,0			2,1			3,6
B. CRO 120									
1	0,629	0,652	3,7	0,402	0,416	3,5	0,528	0,493	-6,6
2 *	0,201	0,195	-3,0	0,893	0,885	-0,9	0,358	0,374	4,5
3	0,395	0,365	-7,6	0,614	0,609	-0,8	0,500	0,534	6,8
4	0,415	0,436	5,1	0,303	0,298	-1,7	0,838	0,824	-1,7
5 *	0,612	0,546	10,8	0,607	0,605	-0,3	0,310	0,374	20,6
6	0,520	0,475	-8,7	0,495	0,516	4,2	0,522	0,558	6,9
média			6,3			2,6			5,5
C. CRO 130									
1	0,629	0,580	-7,8	0,402	0,440	9,5	0,527	0,483	-8,4
2 *	0,201	0,251	24,6	0,892	0,863	-3,3	0,358	0,346	-3,4
3	0,395	0,365	-7,6	0,613	0,617	0,7	0,500	0,522	4,4
4 *	0,414	0,400	-3,2	0,302	0,256	-15,2	0,837	0,903	7,9
5 *	0,611	0,582	-4,8	0,606	0,633	4,5	0,310	0,306	-1,3
6	0,520	0,520	0,0	0,494	0,492	-0,4	0,522	0,523	0,2
média			5,1			3,5			4,3

a. ERRO REL. (%) = [(PREVISTO - REAL)/REAL] X 100

b. Amostras com "*" não fazem parte do modelo.

c. Nas médias somente foram consideradas as amostras que pertenciam ao modelo, e os erros foram tomados em módulo.

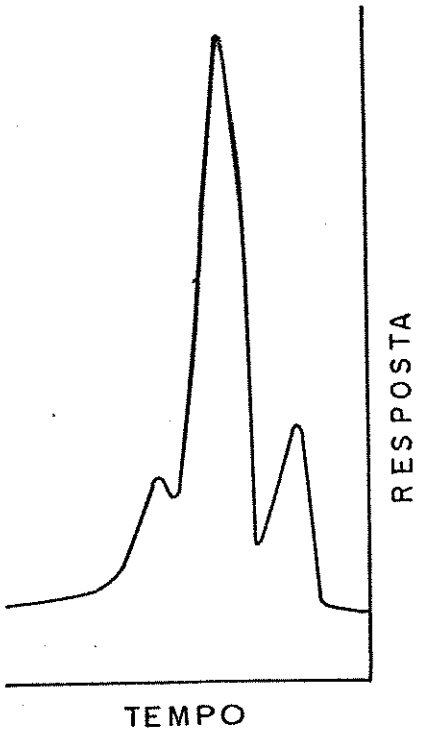
Identificação das amostras que não pertencem à modelagem :

Na apresentação dos resultados na tabela 7, algumas amostras foram assinaladas com um asterisco e não entraram no cálculo das médias dos erros relativos por serem consideradas como não pertencentes ao modelo.

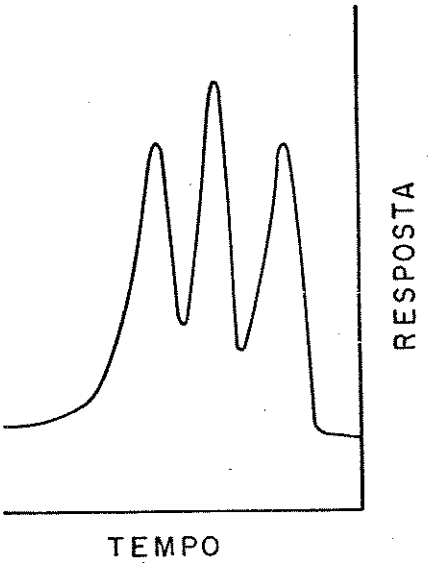
Isto significa que estas amostras não devem fazer parte da classe definida na fase de calibração, ficando o ponto referente ao cromatograma dessa amostra mais distante dos demais, em suma : seu cromatograma deve possuir um perfil bastante diferente daqueles utilizados para a calibração. A figura 16 mostra os perfis de alguns cromatogramas do conjunto CRO 105, onde se pode verificar que aquele referente à "amostra 2" do conjunto teste, considerado como não pertencente ao modelo, realmente é diferente dos demais.

Para facilitar essa comparação dos cromatogramas, pode-se construir um gráfico dos *scores* correspondentes aos três primeiros componentes principais do PLS para o bloco das variáveis independentes. Nesse gráfico de três dimensões, mostrado na figura 17, cada ponto representa um cromatograma, e desta forma é possível ter uma visão das similaridades, pois quanto mais próximos estejam os pontos, mais similares serão os cromatogramas representados por eles. Assim um espaço que tinha 41 dimensões foi reduzido para apenas três, com perda de apenas 8% de informação (os 3 componentes principais em conjunto explicam 92% da variância), havendo assim uma enorme redução no número de variáveis com pouca perda de informação.

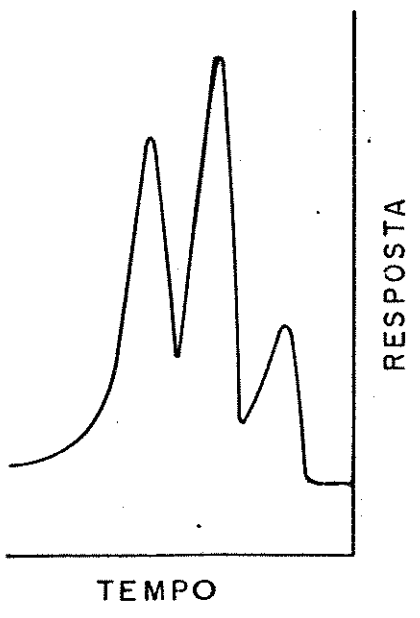
A. "AMOSTRA 2" DO TESTE



B. "AMOSTRA 4" DA CALIBRAÇÃO



C. "AMOSTRA 10" DA CALIBRAÇÃO



D. "AMOSTRA 13" DA CALIBRAÇÃO

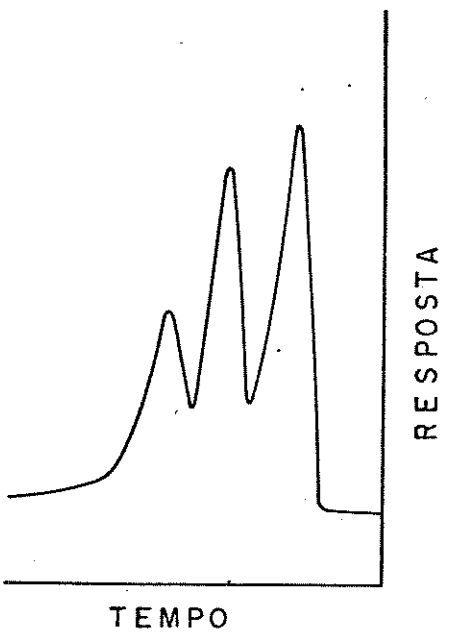


Figura 16 - Alguns cromatogramas do conjunto CRO 105. Notar que o cromatograma correspondente à "amostra 2" do teste tem o perfil diferente dos demais.

Ainda na figura 17, pode-se notar que o ponto correspondente ao cromatograma da "amostra 2" da fase de previsão está mais distante dos demais.

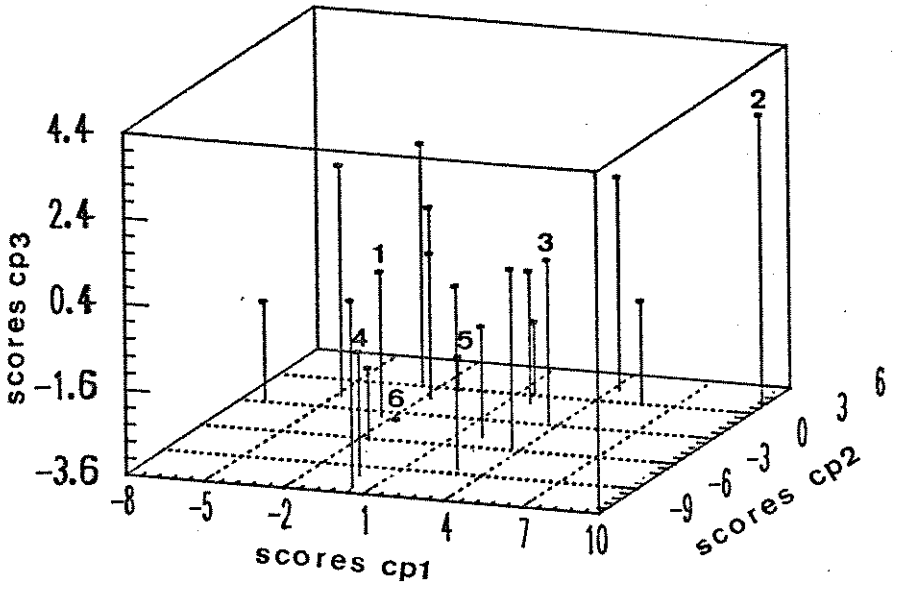


Figura 17 - Os *scores* para o bloco das variáveis independentes para os 3 primeiros componentes principais. Cada ponto representa um cromatograma, sendo assinaladas de 1 até 6 as amostras do teste. Os demais pontos são referentes aos cromatogramas utilizados na fase de calibração.
Notar a "amostra 2" mais distante das demais.

É possível classificar a "amostra 2" do conjunto CRO 105 como não pertencente ao modelo, pela análise dos desvios padrão residuais das amostras da fase de teste, S_T (equação 29, capítulo 2), que dá uma distância do ponto aos eixos dos componentes, e compará-los com o desvio padrão da calibração, S_o (equação 28 do capítulo 2), que fornece uma distância "típica" dos pontos

utilizados na calibração, aos eixos dos componentes.

A tabela 8 mostra estes desvios padrão para o conjunto CRO 105, quando da utilização de 3 componentes principais.

Tabela 8 - Desvios padrão residuais das amostras do teste e para a calibração no conjunto CRO 105 utilizando-se 3 componentes principais.

DESVIO PADRÃO DA CALIBRAÇÃO : $S_o = 0,2847$

DESVIOS PADRÃO RESIDUAIS DAS AMOSTRAS DO TESTE :	AMOSTRA No.	"S_r"
	1	0,2821
	2	0,4991
	3	0,3305
	4	0,3239
	5	0,3011
	6	0,2905

Pode-se notar analisando-se esta tabela que a "amostra 2"da fase de teste tem um desvio padrão residual muito maior que as demais. Além disso esta amostra tem também um desvio padrão residual muito maior que o desvio padrão residual da calibração, o que não ocorre com as demais amostras. É possível assim afirmar que a "amostra 2" não deve fazer parte da categoria definida na fase de calibração.

Da mesma maneira como mostrado para o conjunto CRO 105, as

amostras do teste nos conjuntos CRO 120 e CRO 130 que possuíam desvios padrão residuais (S_r) grandes quando comparados com o desvio padrão residual da calibração (S_o) não foram consideradas nos cálculos. Ademais, os pontos relativos aos cromatogramas dessas amostras nos gráficos dos *scores* para o bloco das variáveis independentes dos 3 primeiros componentes principais ficarem deslocados em relação aos demais, o que confirma a análise numérica dos desvios padrão.

Esses pontos, por não pertencerem à categoria definida, geralmente não apresentam bons resultados para a previsão, sendo que só foram incluídos para efeito de demonstração da capacidade do método em discriminar as amostras que não podem ser analisadas a partir da calibração construída.

No conjunto CRO 130, a "amostra 5" do teste (tabela 7), apesar de ter sido considerada como não pertencente ao modelo, apresenta resultados satisfatórios na previsão. Isto deve-se ao fato que no PLS a previsão é feita a partir das curvas dos *scores* para o bloco das variáveis independentes contra os *scores* do bloco das dependentes (como mostrado na figura 9, capítulo 2).

Assim, pode ocorrer que um ponto, mesmo distante do modelo, apresente um valor de *score* que permita uma previsão aceitável de resultados. Como não é possível prever quando isto ocorre, os resultados para essa amostra não podem ser considerados como de total confiança.

Também as amostras que fizeram parte da fase de calibração devem pertencer à categoria definida, e isto é verificado novamente pela comparação dos desvios padrão residuais das

amostras da calibração (equação 31, capítulo 2), com o desvio padrão residual da calibração (equação 28, capítulo 2). Assim, amostras cujos desvios padrão residuais são muito maiores que o desvio padrão da calibração não devem fazer parte desta categoria, e devem ser retiradas da fase de calibração, pois podem distorcer toda a modelágem.

Todas as amostras, nos três conjuntos de dados, mostraram pertencer à categoria definida na fase de calibração.

2.1.1 - Normalização dos cromatogramas :

Além de escalonar as variáveis de forma que ficassem com o desvio padrão unitário e centrar as variáveis na média, cada cromatograma foi normalizado de maneira que a soma das intensidades consideradas fosse unitária. Este procedimento equivale a normalizar as linhas da matriz de dados, como já feito anteriormente com as colunas. Isto deve fazer com que os dados fiquem melhor condicionados, pois cada linha da matriz passa a representar a mesma quantidade, evitando-se uma discriminação falsa dos objetos.

Com os dados normalizados, foram utilizadas para os três conjuntos de dados as mesmas amostras das fases de calibração e previsão citadas anteriormente, e novamente foram estimadas as massas presentes quando o número de componentes principais variava de um até seis na modelagem.

As amostras consideradas como não pertencentes ao modelo foram descartadas, e as médias dos erros relativos obtidas de

acordo com o número de componentes principais na modelagem foram calculadas, sendo os gráficos obtidos para os três conjuntos de dados mostrados na figura 18.

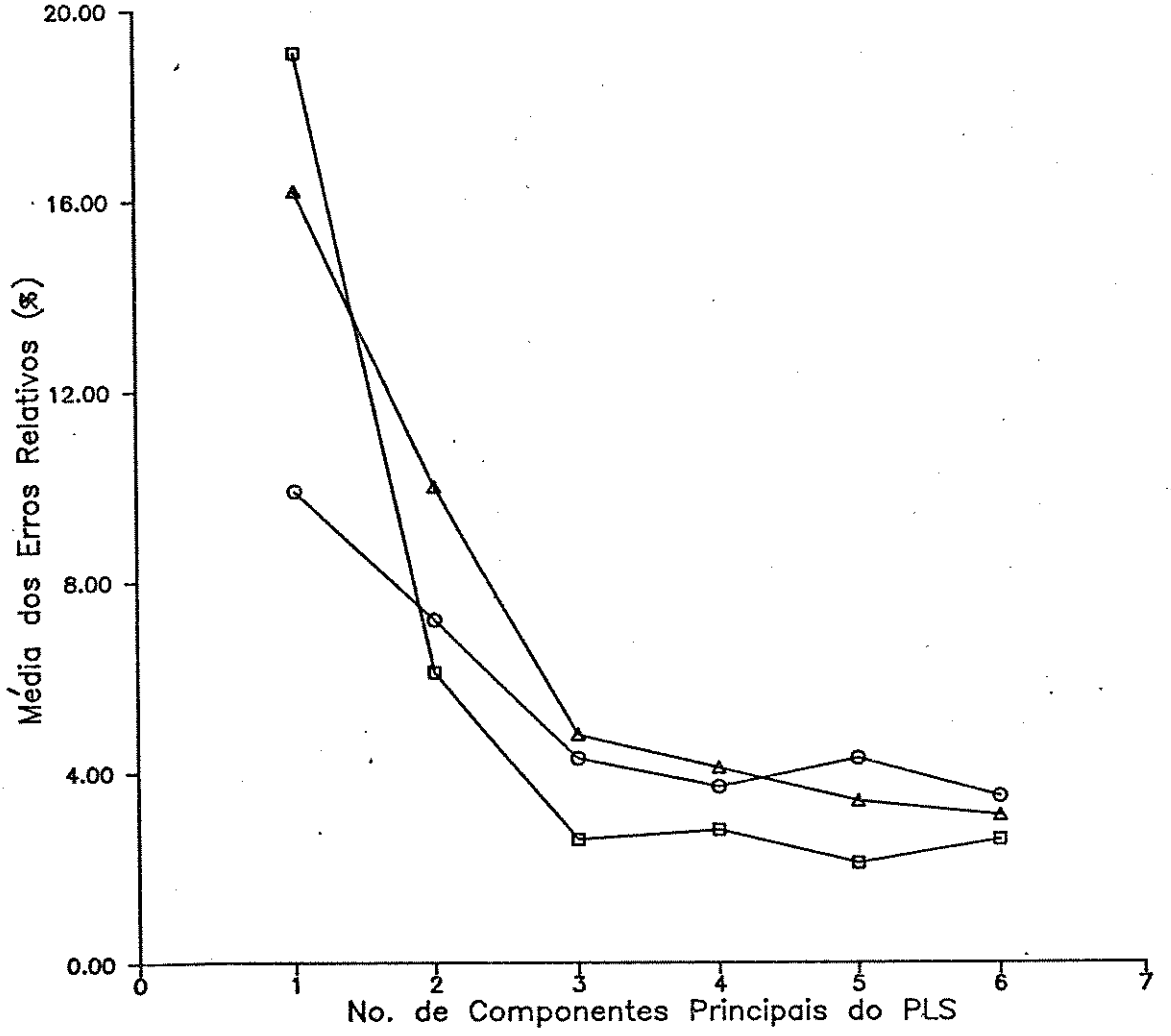


Figura 18 - Média dos Erros Relativos para diferentes números de componentes Principais, para os cromatogramas normalizados à unidade. □ CRO 105, Δ CRO 120, o CRO 130.

Novamente pode-se notar que para os três conjuntos de dados, a partir do terceiro Componente Principal praticamente não há mais alteração na média dos erros relativos. Assim o número de componentes principais a ser utilizado mais uma vez deve ser igual a três.

Os resultados para as previsões das massas injetadas com o uso de três componentes principais na modelagem para os cromatogramas normalizados à unidade são mostrados na tabela 9.

A utilização de três componentes principais faz com que se consiga explicar 92,2% da variância no bloco das variáveis independentes e 99,5% no das dependentes para o conjunto CRO 108. No conjunto CRO 120, com esse mesmo número de componentes é explicada 88,7% da variância no bloco das variáveis independentes e 98,4% no das dependentes. Finalmente para o conjunto CRO 130, também com a utilização de três componentes principais, é explicada 93,3% da variância para o bloco das variáveis independentes e 97,5% para o bloco das dependentes.

A comparação desses números com os resultados das tabelas 6.a, 6.b e 6.c, mostra que o aumento na variância descrita em cada conjunto pelo uso de três componentes principais não é significativa quando se normaliza os cromatogramas à unidade.

Comparando-se a tabela 7 (resultados sem normalização dos cromatogramas) com a tabela 9 (resultados para dados já normalizados), é possível verificar que para o conjunto CRO 130 há uma significativa mudança nas amostras que não pertencem à modelagem. Enquanto que para os dados sem normalização havia três amostras não pertencentes ao modelo (amostras "2", "4" e "5"),

após a normalização apenas a "amostra 1", que antes pertencia à modelagem mas tinha resultados para a previsão não muito bons, passa a ser considerada como não pertencente ao modelo.

Tabela 9 - Previsões do PLS utilizando-se três componentes principais, para os cromatogramas normalizados (massas em mg).

amostra no.	TOLUENO			ISOCTANO			ETANOL		
	VALOR REAL	PREV. PLS	ERRO[a] REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)
A. CRO 105									
1	0,631	0,632	0,2	0,403	0,406	0,7	0,529	0,524	-1,0
2 *[b]	0,202	0,235	16,3	0,895	0,842	-5,9	0,359	0,388	8,1
3	0,396	0,399	0,8	0,615	0,591	-3,9	0,501	0,526	5,0
4	0,415	0,416	0,2	0,303	0,300	-1,0	0,839	0,842	0,4
5	0,613	0,617	0,7	0,608	0,595	-2,1	0,311	0,322	3,5
6	0,521	0,533	2,3	0,496	0,503	1,4	0,523	0,504	-3,6
média[c]			0,8			1,8			2,7
B. CRO 120									
1	0,629	0,690	9,7	0,402	0,385	4,4	0,528	0,494	-6,4
2 *	0,201	0,198	-1,5	0,893	0,906	-1,4	0,358	0,347	-3,1
3	0,395	0,399	1,0	0,614	0,593	3,5	0,500	0,521	4.2
4	0,415	0,454	9,4	0,303	0,283	-6,6	0,838	0,824	-1,7
5 *	0,612	0,541	-11,6	0,607	0,590	-2,8	0,310	0,396	27,7
6	0,520	0,475	-8,7	0,495	0,511	3,2	0,522	0,554	6,1
média			7,2			4,4			4,6
C. CRO 130									
1 *	0,629	0,665	5,7	0,421	0,421	4,7	0,527	0,472	-10,4
2	0,201	0,207	3,0	0,892	0,907	1,7	0,358	0,335	-6,4
3	0,395	0,376	-4,8	0,613	0,599	-2,3	0,500	0,533	6,6
4	0,414	0,419	1,2	0,302	0,291	-3,6	0,837	0,846	1,1
5	0,611	0,602	-1,5	0,606	0,652	7,6	0,310	0,280	-9,7
6	0,520	0,517	-0,6	0,494	0,496	0,4	0,522	0,522	0,0
média			2,2			3,1			4,8

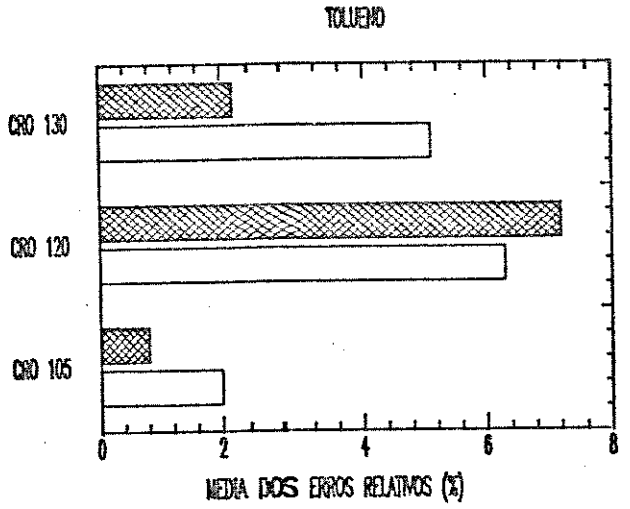
a,b,c idem tabela 8.

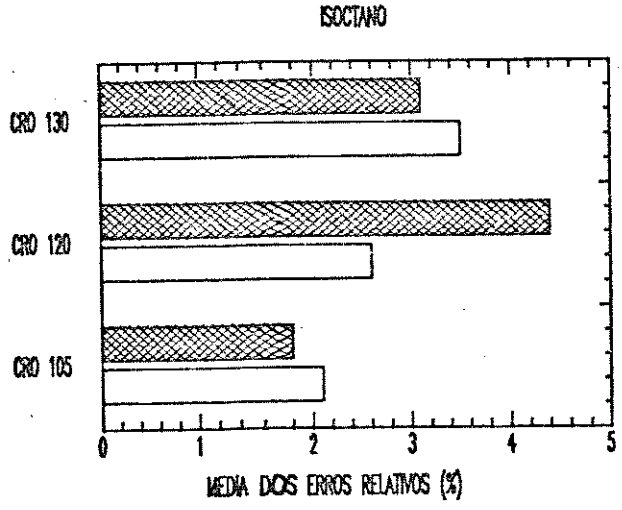
Quanto aos resultados apresentados na tabela 9, novamente é possível verificar que os melhores resultados são para o conjunto CRO 105, onde há menor grau de superposição entre os picos. Também os piores resultados aparecem para o conjunto CRO 120, principalmente para o Tolueno.

Os resultados para as previsões quando os cromatogramas são normalizados, podem ser analisados para cada um dos três constituintes (Tolueno, Isoctano e Etanol) individualmente, através de gráficos de barras onde são comparadas as médias dos erros relativos para cada um desses constituintes, antes e após a normalização, para os três conjuntos de dados, como mostrado na figura 19.

Pela análise da figura 19, apenas os resultados para a previsão do Tolueno nos conjuntos CRO 105 e CRO 130 podem ser considerados significativamente melhores após a normalização dos cromatogramas, enquanto que pioram para o Isoctano no conjunto CRO 120. Quanto aos demais resultados fica difícil tomar alguma decisão à respeito.

Embora a normalização dos cromatogramas imponha algumas alterações importantes no comportamento dos dados analisados, como a redução das amostras consideradas como pontos deslocados, a conveniência do seu emprego não fica evidente nos casos estudados nesse trabalho, em termos do resultado final da previsão das massas.





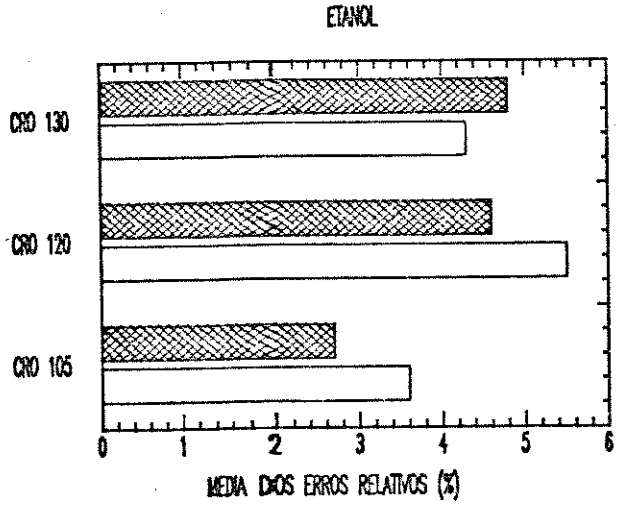


Figura 19 - Média dos Erros Relativos Porcentuais para as previsões antes e após a normalização dos cromatogramas.
Barras claras : sem normalização.
Barras sombreadas : cromatogramas normalizados.

2.1.2 - Redução das Variáveis :

Pelos resultados obtidos até o momento, os piores valores para previsão (maiores erros relativos), principalmente no conjunto CRO 120, foram para o Etanol e Tolueno.

Verificándo-se os cromatogramas, notou-se que para uma mesma temperatura da coluna (mesmo conjunto de dados) praticamente sempre o início do primeiro pico (Etanol) e o final do último (Tolueno), ocorriam com os mesmos tempos de eluição, ou seja, as regiões inicial e final de todos cromatogramas para aquela temperatura eram idênticas entre si. Assim, as intensidades medidas nessas regiões não adicionam muita informação ao sistema, contribuindo entretanto para aumentar o ruído (erros de digitalização, erros instrumentais).

Por essa razão decidiu-se investigar a aplicação do método PLS a conjuntos de cromatogramas em que a parte inicial e final fossem consideradas irrelevantes para o problema em estudo.

Desta maneira foram eliminadas as 4 primeiras variáveis e as 11 últimas, que correspondiam à região onde sempre os cromatogramas eram semelhantes. Na figura 20 são mostrados dois cromatogramas de um mesmo conjunto e as variáveis que foram utilizadas.

Assim, os cálculos passaram a ser realizados utilizando-se 26 variáveis independentes (variáveis "5" até "30") para os três conjuntos da dados, com as mesmas amostras usadas anteriormente na calibração e previsão.

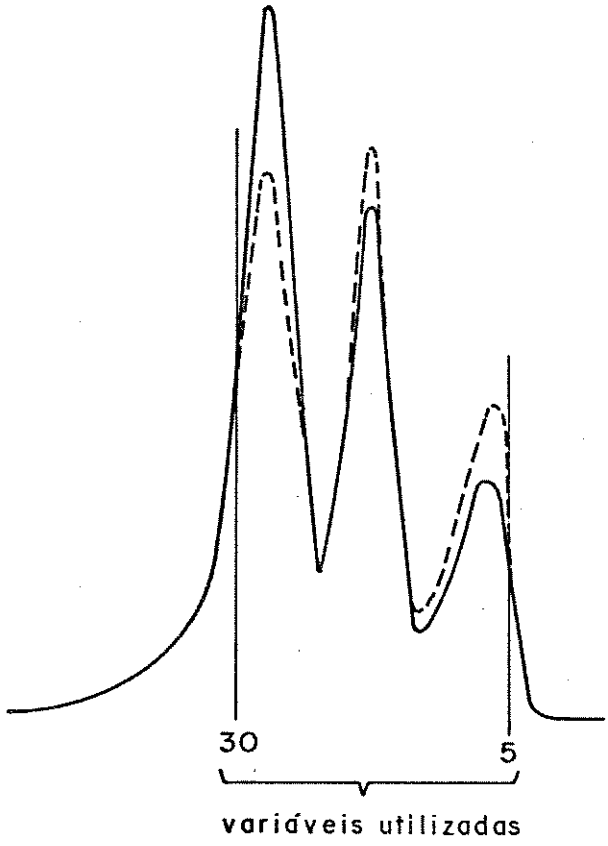


Figura 20 - Dois cromatogramas de um mesmo conjunto, com a faixa das variáveis utilizadas.

Novamente todas as variáveis foram inicialmente escalonadas para desvio padrão unitário e centradas na média. As previsões foram feitas com o número de componentes principais variando de um até seis, sendo descartadas as amostras consideradas como não pertencentes ao modelo.

As médias dos erros relativos de acordo com o número de componentes principais foram calculadas e os gráficos desses erros contra o número de componentes principais, para cada conjunto de dados, são mostrados na figura 21.

Pelos gráficos obtidos (fig. 21), nota-se que para os conjuntos CRO 105 e CRO 130, após o segundo componente principal

já não há mais mudança significativa dos valores das médias dos erros relativos, ou seja, o número mínimo de componentes passou a ser dois para esses dois conjuntos de dados. Já para o conjunto CRO 120 o número de componentes principais continua sendo igual a três.

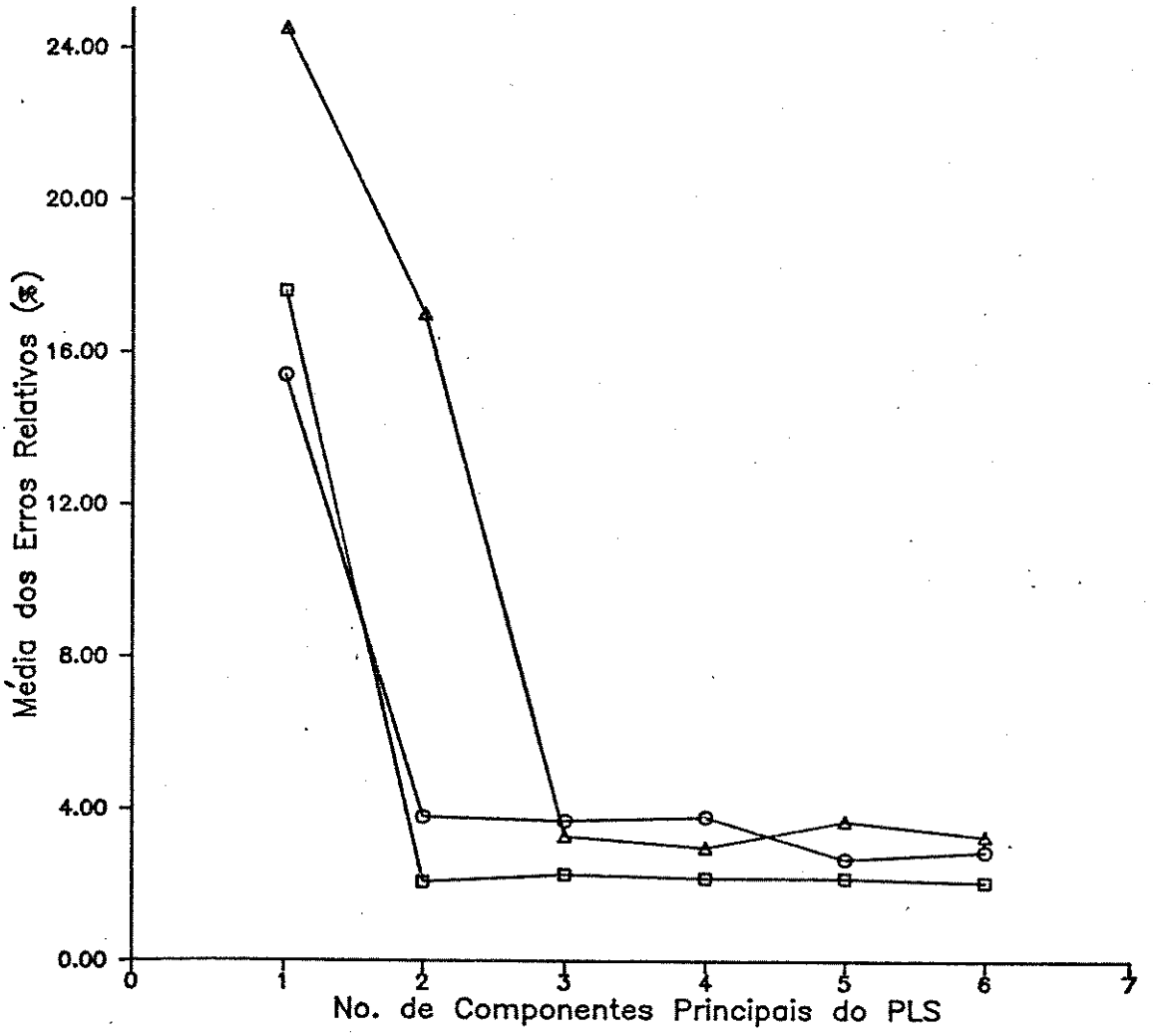


Figura 21 - Média dos Erros Relativos para diferentes números de componentes principais, após a redução das variáveis. □ CRO 105 , Δ CRO 120, o CRO 130.

As variâncias retidas em cada componente principal para cada um dos blocos (das variáveis independentes e dependentes), após a redução das variáveis, para os três conjuntos de dados, são mostrados nas tabelas 10.a para o conjunto CRO 105, 10.b para o conjunto CRO 120 e 10.c para o conjunto CRO 130.

Tabela 10.a – Variância porcentual descrita pela modelagem com o PLS para o conjunto CRO 105, após a redução das variáveis.

No. de componentes principais do PLS	variáveis independentes		variáveis dependentes	
	cada	total	cada	total
1	46,39	46,39	45,04	45,04
2	47,09	93,48	50,76	95,80
3	1,02	94,50	1,15	98,95
4	1,91	96,41	0,67	99,62
5	2,38	98,79	0,25	99,87
6	0,18	98,97	0,06	99,93

Tabela 10.b – Variância porcentual descrita pela modelagem com o PLS para o conjunto CRO 120, após a redução das variáveis.

No. de componentes principais do PLS	variáveis independentes		variáveis dependentes	
	cada	total	cada	total
1	43,51	43,51	40,52	40,52
2	42,00	85,51	52,33	92,85
3	5,04	90,55	5,00	97,85
4	5,04	95,59	0,22	98,07
5	0,39	95,98	1,05	99,12
6	2,90	98,88	0,04	99,16

Tabela 10.c – Variância porcentual descrita pela modelagem com o PLS para o conjunto CRO 130, após a redução das variáveis.

No. de componentes principais do PLS	variáveis independentes		variáveis dependentes	
	cada	total	cada	total
1	64,93	64,93	40,87	40,87
2	25,29	90,22	52,65	93,52
3	5,09	95,31	4,54	98,06
4	2,55	97,86	0,97	99,03
5	0,91	98,77	0,54	99,57
6	0,29	99,06	0,16	99,73

Pode-se verificar pelas tabelas 10.a, e 10.c que a variância descrita pelos dois primeiros componentes principais para o bloco das variáveis independentes nos conjuntos CRO 105 e CRO 130 é superior a 90% (93,5% e 90,2% respectivamente) e isso confirma ainda mais a possibilidade de utilizar-se apenas dois componentes principais nos cálculos. Já para o conjunto CRO 120 são necessários três componentes principais para que a variância descrita ultrapasse os 90% no bloco das variáveis independentes.

Os resultados obtidos para a previsão das massas com a utilização de dois componentes principais nos cálculos para os conjuntos CRO 105 e CRO 130, e três componentes principais para o conjunto CRO 120 são apresentados na tabela 11.

Tabela 11- Previsões pelo Método PLS com a redução das variáveis. Foram utilizados 2 componentes principais nos cálculos para os conjuntos CRO 105 e CRO 130, e 3 componentes principais no conjunto CRO 120 (massas em miligramas).

amostra no.	TOLUENO			ISOCTANO			ETANOL		
	VALOR REAL	PREV. PLS	ERRO[a] REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)
A. CRO 105									
1	0,631	0,642	1,7	0,403	0,402	-0,3	0,529	0,520	-1,7
2 *[b]	0,202	0,245	21,3	0,895	0,840	-6,2	0,359	0,382	6,4
3	0,396	0,401	1,3	0,615	0,595	-3,3	0,501	0,519	3,6
4	0,415	0,419	1,0	0,303	0,288	-5,0	0,839	0,853	1,7
5	0,613	0,601	-2,0	0,608	0,609	0,2	0,311	0,320	2,9
6	0,521	0,509	-2,3	0,496	0,513	3,4	0,523	0,514	-1,7
média[c]			1,7			2,4			2,3
B. CRO 120									
1	0,629	0,605	-3,8	0,402	0,434	8,0	0,528	0,498	-5,7
2 *	0,201	0,265	31,8	0,893	0,870	-2,6	0,358	0,327	-8,7
3	0,395	0,402	1,8	0,614	0,599	-2,4	0,500	0,519	3,8
4	0,415	0,402	-3,1	0,303	0,314	3,6	0,838	0,825	-1,6
5	0,612	0,640	4,6	0,607	0,588	-3,1	0,310	0,320	3,2
6	0,520	0,514	-1,2	0,495	0,492	-0,6	0,522	0,539	3,3
média			2,9			3,5			3,5
C. CRO 130									
1	0,629	0,559	-11,1	0,421	0,425	5,7	0,527	0,517	-1,9
2 *	0,201	0,226	12,4	0,892	0,858	-3,8	0,358	0,374	4,5
3	0,395	0,388	-1,8	0,613	0,617	0,7	0,500	0,501	0,2
4	0,414	0,404	-2,4	0,302	0,284	-6,0	0,837	0,867	3,6
5 *	0,611	0,600	-1,8	0,606	0,670	10,6	0,310	0,247	-20,3
6	0,520	0,512	-1,5	0,494	0,522	5,7	0,522	0,496	-5,0
média			4,2			4,5			2,7

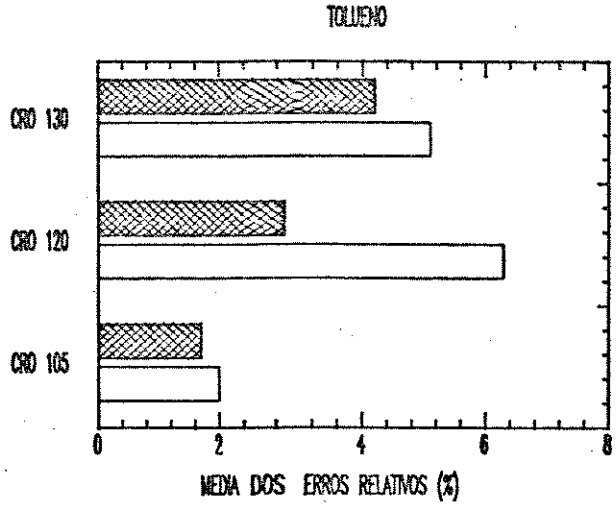
a,b,c idem tabela 8.

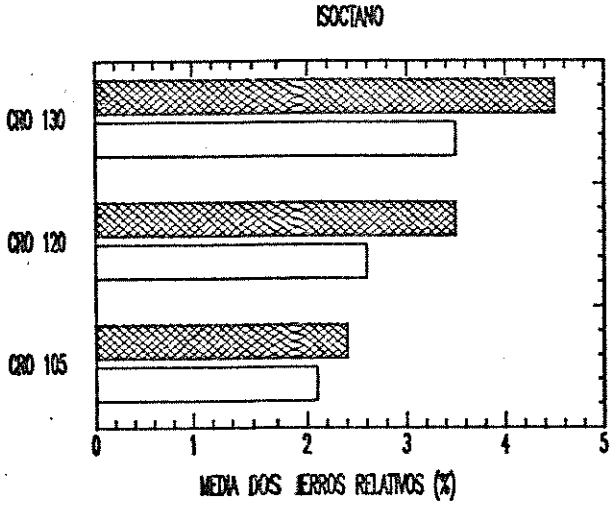
Pelos resultados obtidos com a redução das variáveis (tabela 11), pode-se notar que os erros de previsão no conjunto CRO 120 diminuíram bastante. Ainda, os resultados para o conjunto CRO 105 são os melhores de todos, como seria o esperado.

Também os resultados obtidos antes e após a redução das variáveis, para cada um dos três constituintes químicos presentes, podem ser comparados por gráficos de barras onde são mostradas as médias dos erros relativos de cada um desses constituintes, para os três conjuntos de dados, conforme mostrado na figura 22.

Pela análise da figura 22, pode-se dizer que para o conjunto CRO 105 praticamente não há alteração nos resultados, apenas para o Etanol há uma pequena diminuição dos erros de previsão com a redução das variáveis. Entretanto, para o conjunto CRO 120 há uma sensível melhora nos valores de previsão para o Etanol e principalmente para o Tolueno. Também no conjunto CRO 130 os resultados para Etanol e Tolueno apresentam uma diminuição nos valores dos erros de previsão. Para o Isoctano, esses erros acabam por ter um pequeno aumento.

Uma vez que as variáveis descartadas descreviam os sinais relativos a Etanol e Tolueno, é razoável que a previsão dessas massas tenha sido a mais afetada em decorrência da redução de variáveis imposta ao conjunto de dados.





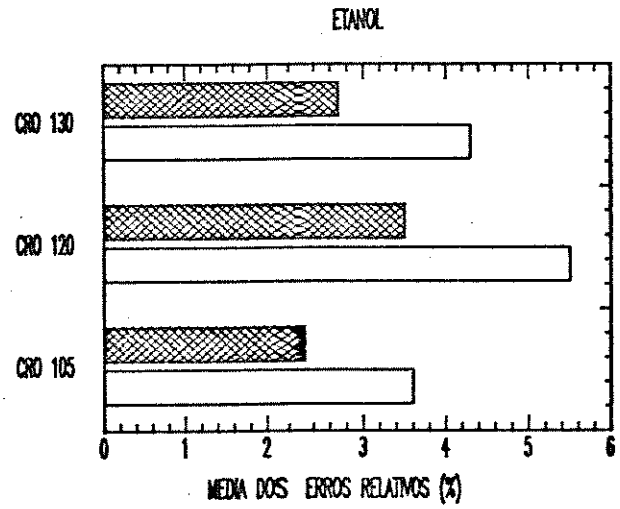


Figura 22 - Média dos Erros Relativos Porcentuais para as previsões antes e após a redução das variáveis.
Barras claras : sem redução das variáveis.
Barras sombreadas : com a redução.

Um fato importante que deve ser destacado com a redução das variáveis é a enorme redução dos erros de previsão no conjunto CRO 120 para o Tolueno. No conjunto CRO 120 os cromatogramas relativos às várias amostras são muito diferentes entre si, o que não ocorre para os outros dois conjuntos. Isso acaba fazendo com que uma mesma variável, em cada cromatograma desse conjunto, não tenha exatamente o mesmo significado, e por consequência, os pontos no hiperespaço que representam os cromatogramas digitalizados (cada cromatograma é representado por um ponto num espaço de 26 dimensões) ficam muito espalhados, dificultando a modelagem com os componentes principais.

Deve-se ter sempre em mente que quanto mais semelhantes sejam dois cromatogramas, mais próximos devem estar os pontos referentes a eles no hiperespaço considerado. Na figura 23 são apresentados dois cromatogramas do conjunto CRO 120, mostrando a diferença no significado de uma mesma variável entre eles.

Como o início da digitalização foi dado pelo pico referente ao Etanol, terminando portanto pelo pico do Tolueno, essas diferenças das variáveis de um cromatograma para o outro acabam ficando muito críticas para o Tolueno, pois vão sendo acumuladas. Com o corte das últimas variáveis, acabaram sendo descartadas exatamente aquelas que possuíam mais erros.

Também para o conjunto CRO 130 deve acontecer algo semelhante, só que em uma intensidade muito menor. Já para o conjunto CRO 105 isto não deve ocorrer, pois os resultados praticamente não se alteram com a redução das variáveis. Este fato dos resultados não se alterarem indica que realmente estas

variáveis não contém nenhuma informação imprescindível à solução do problema.

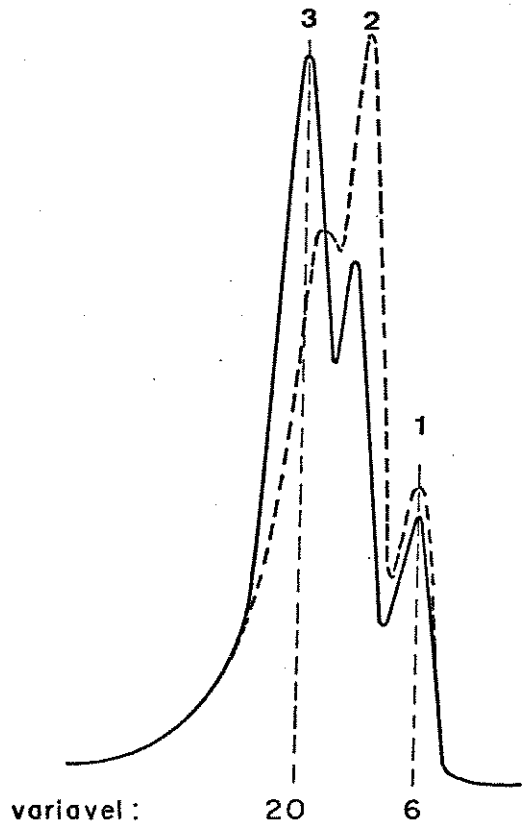


Figura 23 - A diferença que uma mesma variável significa em dois cromatogramas do conjunto CRO 120. Amostras : 1. Etanol , 2. Isoctano, 3. Tolueno.

Número de Componentes Principais :

Outro fato importante que deve ser destacado com a redução das variáveis é que para os conjuntos CRO 105 e CRO 130, o número de componentes principais mínimo necessário para os cálculos passou de três, antes da redução, para dois com a redução. Já para o conjunto CRO 120 esse número mínimo continuou sendo igual a

três.

Isso indica que para os conjuntos CRO 105 e CRO 130, um terceiro componente principal estava sendo necessário para modelar o efeito dos ruídos causados pelas variáveis que posteriormente foram eliminadas. Para o conjunto CRO 120, um terceiro componente principal deve estar sendo necessário devido à estrutura dos pontos no hiperespaço, já que, como mencionado anteriormente, os cromatogramas nesse conjunto são mais diferentes entre si.

É possível que a estrutura dos dados para este conjunto não possa ser explicada por modelos lineares, e assim um componente a mais pode estar sendo necessário para modelar essa não linearidade.

Como citado no capítulo 3, "em sistemas lineares o número de componentes principais necessários deve ser igual ao número de constituintes químicos presentes". Desta maneira, como existem três constituintes químicos : tolueno, isoctano e etanol, deveriam ser necessários pelo menos três componentes principais na modelagem, e não dois como obtido para os conjuntos CRO 105 e CRO 130.

A possibilidade de modelar os dados com apenas dois componentes principais é decorrente da maneira como os experimentos foram conduzidos, já que sempre foram injetadas misturas constantes em 2 μl. Desta forma, há um compromisso entre os valores das massas, pois basta determinar a massa de dois constituintes químicos para que, sendo o volume constante, a massa do terceiro esteja automaticamente determinada.

2.1.3 - Normalização dos Cromatogramas com redução das variáveis :

Após a redução das variáveis foram realizados cálculos que envolveram o escalonamento para desvio padrão unitário, a centralização das variáveis em torno da média e a normalização de cada cromatograma à unidade.

Foram utilizadas as mesmas amostras já citadas anteriormente na calibração e previsão, e os cálculos foram feitos novamente com o número de componentes principais variando de 1 até 6, sendo descartadas as amostras consideradas como não pertencentes à modelagem. As médias dos erros relativos de acordo com o número de componentes principais utilizados foram calculadas, e o gráfico obtido quando são plotados esses erros contra o número de componentes principais, para os três conjuntos de dados, é mostrado na figura 24.

Pelos gráficos obtidos, nota-se mais uma vez que para os conjuntos CRO 105 e CRO 130 pode-se utilizar dois componentes principais, enquanto que para o conjunto CRO 120 esse número é três ,ou seja, após a normalização dos cromatogramas à unidade, não há alteração no número de componentes principais necessários para modelar os dados.

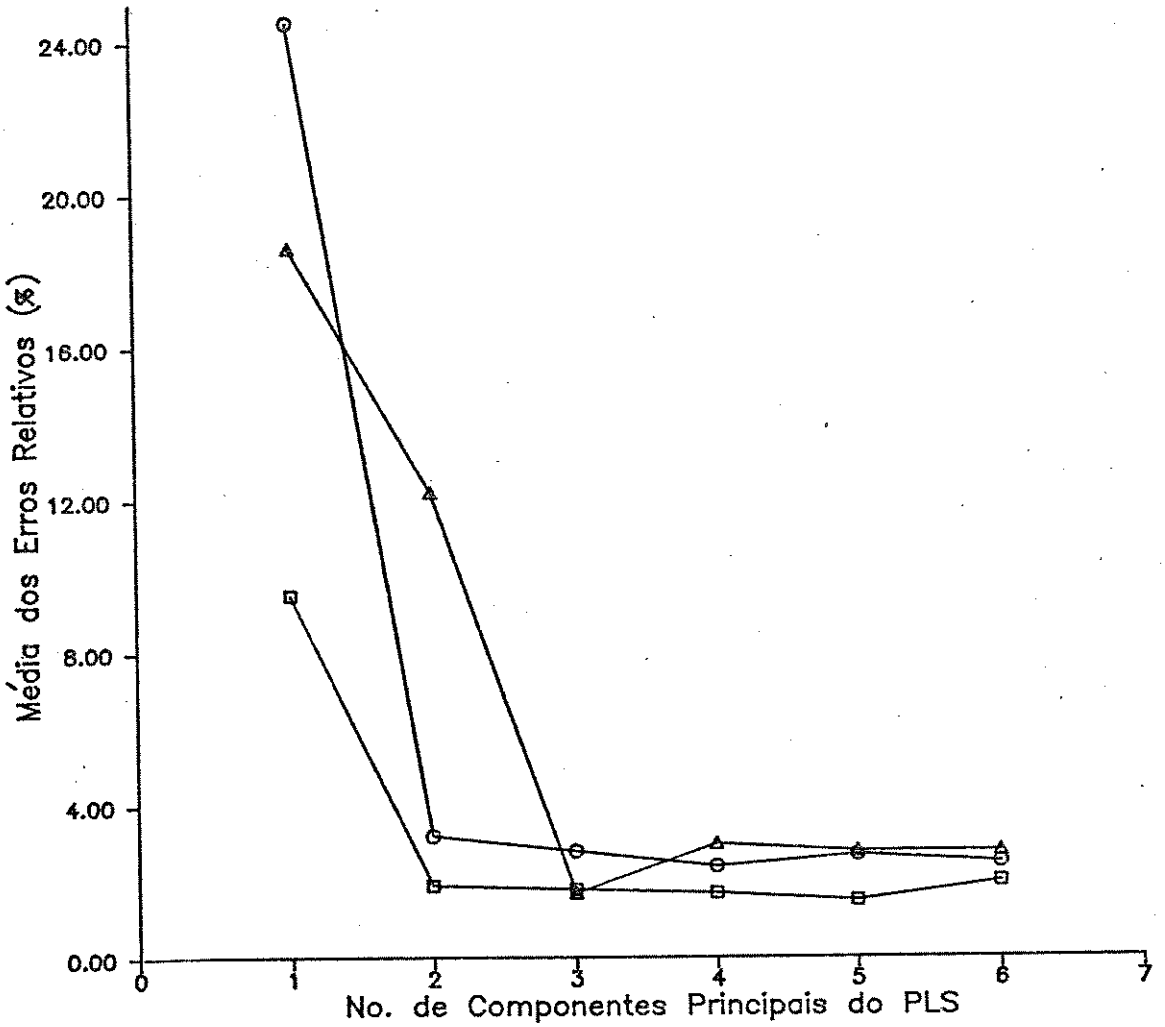


Figura 24 - Média dos Erros Relativos para diferentes números de componentes principais, para os cromatogramas normalizados após a redução das variáveis. □ CRO 105, Δ CRO 120, o CRO 130

A utilização de dois componentes principais faz com que se explique, no conjunto CRO 105, para o bloco das variáveis independentes 94,05% da variância e 99,73% para o bloco das dependentes. Para o conjunto CRO 130 a utilização de dois componentes explica 96,47% da variância no bloco das variáveis

independentes e 96,57% para o bloco das variáveis dependentes. Já para o conjunto CRO 120, o uso de três componentes principais faz com que se explique 92,88% da variância no bloco das variáveis independentes e 98,24% no das dependentes.

Os resultados obtidos para a previsão com a utilização de dois componentes principais nos cálculos para os conjuntos CRO 105 e CRO 130, e três componentes principais para o conjunto CRO 120, são apresentados na tabela 12.

Novamente pode-se verificar, analisando a tabela 12, que para o conjunto CRO 130 há uma alteração nas amostras que são consideradas como não pertencentes ao modelo. Enquanto antes eram consideradas as amostras "2" e "5", agora somente a "amostra 1" não pertence ao modelo.

Após a redução das variáveis e a normalização dos cromatogramas à unidade, apenas uma amostra em cada conjunto de dados passa a ser considerada como não pertencente ao modelo. Isto faz supor que o fator preponderante para desclassificar uma amostra no modelo inicial, residia basicamente na escolha adequada das variáveis que se deva incluir no tratamento dos dados, e da forma como essas variáveis eram tratadas antes da aplicação do PLS.

A utilização de variáveis desnecessárias à descrição do sistema e que contém algum ruído, além de não adicionar informação relevante ao problema tem o efeito de distorcer o modelo, levando à falsa conclusão de que alguns objetos do conjunto teste não pertencem à classe definida na calibração.

Tabela 12- Previsões pelo Método PLS para os cromatogramas normalizados à unidade após a redução das variáveis. Foram utilizados 2 componentes principais nos cálculos para os conjuntos CRO 105 e CRO 130, e 3 componentes principais no conjunto CRO 120 (massas em mg).

amostra no.	TOLUENO			ISOCTANO			ETANOL		
	VALOR REAL	PREV. PLS	ERRO[a] REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)
A. CRO 105									
1	0,631	0,623	-1,3	0,403	0,405	0,5	0,529	0,534	1,0
2 *[b]	0,202	0,241	19,3	0,895	0,839	-6,3	0,359	0,387	7,8
3	0,396	0,399	0,8	0,615	0,590	-4,1	0,501	0,527	5,2
4	0,415	0,414	-0,2	0,303	0,293	-3,3	0,839	0,852	1,6
5	0,613	0,609	-0,7	0,608	0,602	-1,0	0,311	0,321	3,2
6	0,521	0,507	2,7	0,496	0,509	2,6	0,523	0,521	-0,4
média[c]			1,1			2,3			2,3
B. CRO 120									
1	0,629	0,637	1,3	0,402	0,421	4,7	0,528	0,501	-5,1
2 *	0,201	0,234	16,4	0,893	0,954	6,8	0,358	0,259	-27,7
3	0,395	0,414	4,8	0,614	0,610	-0,7	0,500	0,488	-2,4
4	0,415	0,413	-0,5	0,303	0,303	0,0	0,838	0,839	0,1
5	0,612	0,600	-2,0	0,607	0,619	2,0	0,310	0,308	-0,7
6	0,520	0,518	-0,4	0,495	0,501	1,2	0,522	0,518	-0,8
média			1,8			1,7			1,8
C. CRO 130									
1 *	0,629	0,711	13,0	0,421	0,396	-1,5	0,527	0,459	-12,9
2	0,201	0,207	3,0	0,892	0,898	0,7	0,358	0,347	-3,1
3	0,395	0,387	-2,0	0,613	0,602	-1,8	0,500	0,520	4,0
4	0,414	0,410	-1,0	0,302	0,300	-0,7	0,837	0,844	0,8
5	0,611	0,570	-6,7	0,606	0,658	8,6	0,310	0,285	-8,1
6	0,520	0,501	-3,7	0,494	0,511	3,4	0,522	0,519	-0,6
média			3,3			3,0			3,3

a,b,c idem tabela 8.

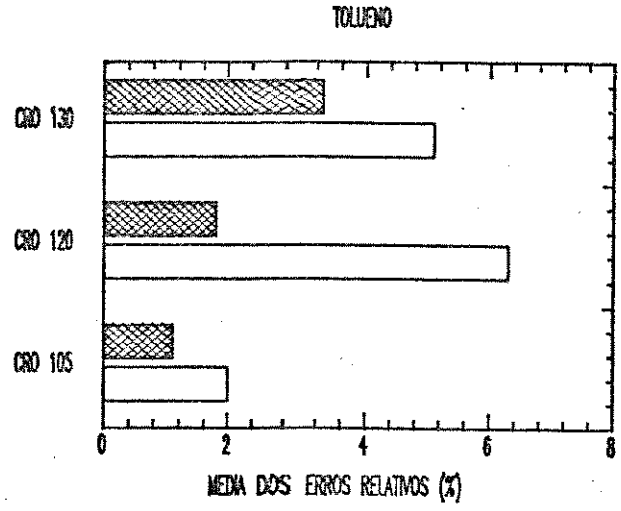
Para os conjuntos CRO 120 e CRO 130 os resultados para a previsão tiveram uma pequena melhora após a normalização, se forem comparados com aqueles obtidos somente com redução de variáveis. Nesses resultados parece que fica mais acentuado a fato que com o aumento na superposição dos picos há um aumento no erro, se forem comparados os resultados obtidos para os conjuntos CRO 120 e CRO 130.

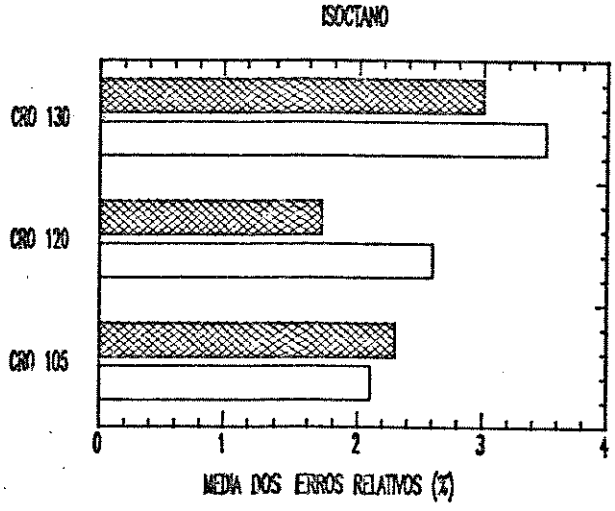
Com a normalização dos cromatogramas digitalizados à unidade após a redução das variáveis, chega-se a um procedimento geral para o pré-tratamento dos dados. Esse procedimento implica, como visto, em abandonar o início e o final do cromatograma, para fazer uso das variáveis que contém informações úteis, além de tornar a soma das intensidades medidas em cada cromatograma unitária.

A figura 25 apresenta uma comparação gráfica entre os erros de previsão com e sem o pré-tratamento dos dados.

Apesar da diminuição do erro relativo de previsão para Tolueno e Etanol observados para o conjunto CRO 105, onde a superposição dos picos é menor, a aplicação do pré-tratamento dos dados parece ser dispensável nesse caso. Isto porque o comportamento das amostras após esse pré-tratamento praticamente não se altera, no que diz respeito aos objetos que pertencem ou não à classe modelada na fase de treinamento.

Para os conjuntos CRO 120 e CRO 130, onde a superposição entre os picos é mais pronunciada, o pré-tratamento dos dados torna-se indispensável, pois além da diminuição dos erros de previsão, algumas amostras que antes se comportavam como não pertencentes ao modelo estabelecido, podem, após o pré-tratamento, serem analisadas.





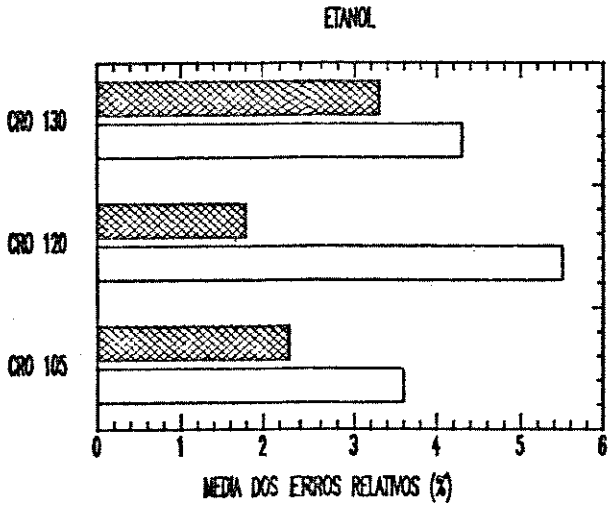


Figura 25 - Média dos Erros Relativos Porcentuais para as previsões antes e após o pré- tratamento dos dados.
Barras claras : sem pré-tratamento.
Barras sombreadas : com o pré-tratamento.

2.2 - Regressão de Componentes Principais (PCR) :

Era intenção inicial utilizar a Regressão Linear Múltipla neste trabalho, para que se pudesse comparar seus valores de previsão com os do PLS e do PCR. Entretanto, mesmo após a redução das variáveis, o número de variáveis independentes (41 originalmente e depois 26) continua maior que de amostras (foram utilizadas 14 amostras na fase de calibração), e quando isso ocorre não é possível obter uma solução discreta com o uso da Regressão Linear Múltipla.

Reduzir-se ainda mais o número de variáveis não foi possível, pois perde-se muita informação relevante, e os resultados com a utilização do PLS e do PCR pioram muito, ficando sem sentido uma comparação. Também não foi possível aumentar o número de amostras por questões técnicas. Desta maneira, não foram realizados estudos com a Regressão Linear Múltipla.

Assim, além do PLS o outro Método de Calibração Multivariada estudado foi o PCR. Este método também utiliza componentes principais, só que neste caso apenas a matriz correspondente às variáveis independentes é descrita pelos componentes principais.

Por esse método, como no PLS, é possível determinar as amostras que não pertencem à categoria definida na fase de calibração. Isso é feito como anteriormente, pela comparação entre os desvios padrão residuais das amostras do teste (equação 29, capítulo 2), com o desvio padrão para a calibração (equação 28 , capítulo 2).

As mesmas 14 amostras, nos três conjuntos de dados utilizados

no PLS para a fase de calibração (construção do modelo), foram utilizados nesta fase, assim como as 6 amostras utilizadas anteriormente como teste. Este cuidado foi tomado a fim de permitir uma comparação do desempenho dos dois métodos, em idênticas condições.

Como no caso do PLS, foi investigada a aplicação do PCR ao conjunto de 41 variáveis e ao conjunto reduzido para 26 variáveis. Em ambos os casos também foi estudado o efeito da normalização dos cromatogramas à unidade. Novamente, o conjunto de dados que se mostrou mais adequado para a utilização foi o de 26 variáveis normalizadas. Esse conjunto da dados, autoescalonado para média zero e variância unitária, forneceu os resultados que serão apresentados nesta seção.

Para se obter resultados comparáveis aos do PLS, sempre foi necessário um componente principal a mais que os utilizados no PLS. Assim, para os conjuntos CRO 105 e CRO 130 foram utilizados três componentes principais, e para o conjunto CRO 120 foram utilizados quatro componentes principais. Os resultados para os valores de previsão obtidos pelo PCR, com a utilização do número de componentes principais já citados, são mostrados na tabela 13.

A utilização de três componentes principais para o conjunto CRO 105 faz com que se consiga explicar 97,3% da variância no bloco das variáveis independentes, e para o conjunto CRO 130 98,4%. Já a utilização de quatro componentes principais para o conjunto CRO 120 faz com que se consiga explicar 98,5% da variância no bloco das variáveis independentes.

Tabela 13 — Previsões pelo Método PCR. Foram utilizados 3 componentes principais nos cálculos para os conjuntos CRO 105 e CRO 130, e 4 componentes principais no conjunto CRO 120 (massas em mg).

amostra no.	TOLUENO			ISOCTANO			ETANOL		
	VALOR REAL	PREV. PLS	ERRO[a] REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)	VALOR REAL	PREV. PLS	ERRO REL. (%)
A. CRO 105									
1	0,631	0,624	-1,1	0,403	0,404	0,3	0,529	0,534	1,0
2 *[b]	0,202	0,254	25,7	0,895	0,830	-7,3	0,359	0,385	7,2
3	0,396	0,393	-0,8	0,615	0,594	-3,4	0,501	0,527	5,2
4	0,415	0,416	0,2	0,303	0,291	-4,0	0,839	0,852	1,6
5	0,613	0,611	-0,3	0,608	0,600	-1,3	0,311	0,321	3,2
6	0,521	0,505	-3,1	0,496	0,510	2,8	0,523	0,521	-0,4
média[c]			1,1			2,4			2,3
B. CRO 120									
1	0,629	0,636	1,1	0,402	0,425	5,7	0,528	0,500	-5,3
2 *	0,201	0,236	17,4	0,893	0,941	5,4	0,358	0,279	-22,1
3	0,395	0,416	5,3	0,614	0,583	-5,0	0,500	0,498	-0,4
4	0,415	0,412	-0,7	0,303	0,310	2,3	0,838	0,837	-0,1
5	0,612	0,600	-2,0	0,607	0,610	0,5	0,310	0,311	0,3
6	0,520	0,518	-0,4	0,495	0,493	-0,4	0,522	0,521	-0,2
média			1,9			2,8			1,3
C. CRO 130									
1 *	0,629	0,646	2,7	0,421	0,434	8,0	0,527	0,478	-9,3
2	0,201	0,195	-3,0	0,892	0,904	1,4	0,358	0,350	-2,2
3	0,395	0,388	-1,8	0,613	0,600	-2,1	0,500	0,521	4,2
4	0,414	0,405	-2,2	0,302	0,299	-1,0	0,837	0,850	1,6
5	0,611	0,590	-3,4	0,606	0,668	10,2	0,310	0,288	-7,1
6	0,520	0,520	0,0	0,494	0,501	1,4	0,522	0,513	-1,7
média			2,1			3,2			3,4

a,b,c idem tabela 8.

Os resultados obtidos pelo PCR, também apresentam erros de previsão bastante aceitáveis nos três conjuntos de dados, indicando claramente a possibilidade da sua utilização para quantificar espécies em casos de picos superpostos.

As previsões realizadas pelo PCR são praticamente idênticas aos do PLS, para os três conjuntos de dados, quando se utiliza o pré-tratamento já citado (redução das variáveis e normalização dos cromatogramas). Isto pode ser verificado no gráfico de barras mostrado na figura 26, onde são comparados os erros relativos em cada constituinte químico separadamente, nos três conjuntos de dados, quando da utilização do PLS e do PCR.

Desta maneira, para os casos estudados de picos superpostos em cromatografia gasosa, não parece haver muita diferença, em termos de resultados, na utilização de um dos dois métodos estudados.

Nos cálculos utilizando o PCR, a necessidade da utilização de um componente principal a mais que no PLS, ocorre porque no PCR somente um dos blocos das variáveis (as independentes) é modelado com componentes principais, sendo assim necessário obter-se o máximo de informação a partir dos dados desta bloco. No PLS, onde se modela os dois blocos e se maximiza sua correlação por uma rotação dos componentes, cada um desses componentes deve conter mais informação sobre a previsão dos constituintes do que no caso do PCR.

A necessidade da utilização de um componente principal a mais nos cálculos com o PCR, também já foi observado por alguns autores [47].

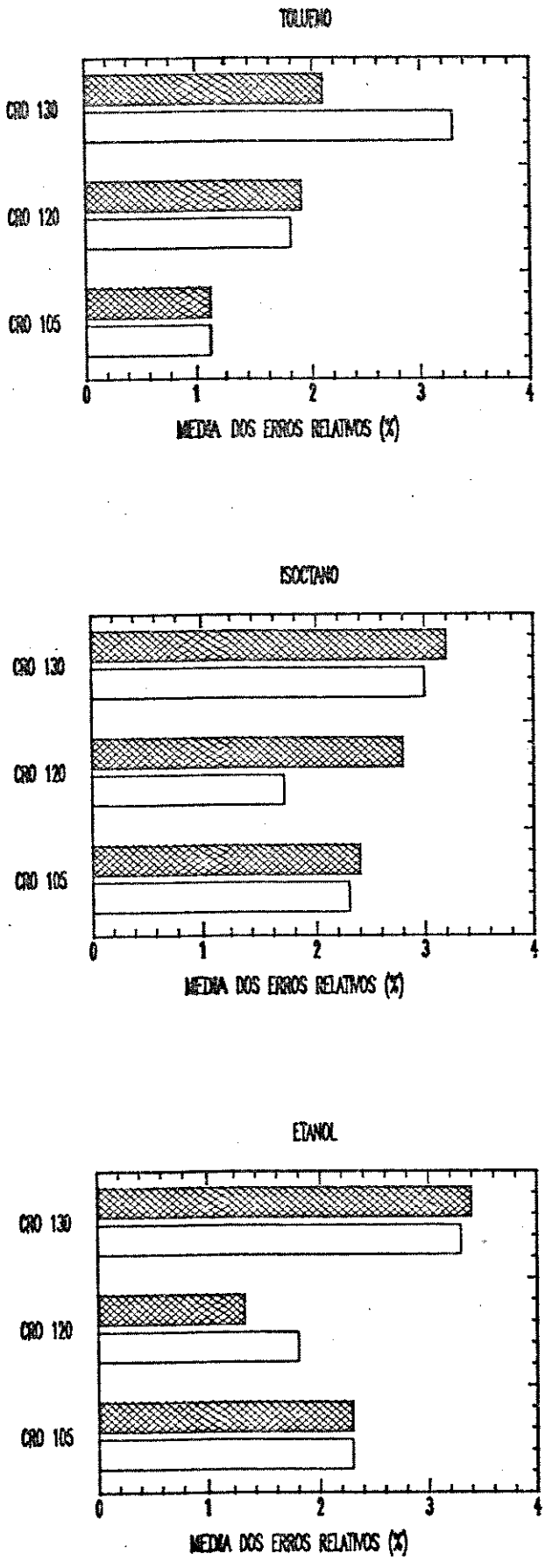


Figura 26 - Média dos Erros Relativos Porcentuais para as previsões utilizando o PLS e o PCR.

Barras claras : resultados dos PLS.

Barras sombreadas : resultados do PCR.

Os cálculos computacionais utilizando-se o PLS são mais rápidos que os do PCR, principalmente quando envolvem uma grande quantidade de amostras e variáveis. Este fato já foi observado por alguns autores [21,48], e é devido que no PLS, todo o processo para a construção da modelagem da calibração é feito em uma única etapa, enquanto que no PCR (utilizando-se por exemplo o algorítmo "NIPALS") são inicialmente calculados os *scores* e os *loadings*, para só então, numa segunda etapa, com o programa da Regressão Linear Múltipla obter-se os parâmetros da modelagem (coeficientes de regressão).

Deve ser destacada a importância da aplicação da Análise de Componentes Principais (PCA) à matriz das variáveis independentes originais . São produzidas, com o PCA, novas variáveis que pelo fato de serem ortogonais entre si e em número reduzido, podem agora ser utilizadas na Regressão Linear Múltipla (esta é a base da Regressão de Componentes Principais), o que não era possível com os dados originais.

3. - Métodos de Separação Linear :

Quando não se dispõe de métodos computacionais para tratar o caso de sinais superpostos, métodos manuais de separação linear tem sido sugeridos [6] para permitir a quantificação das espécies presentes. Nesse trabalho apenas foi empregado o método mais comumente utilizado, que é aquele onde os picos superpostos são separados por um linha perpendicular à linha de base tomada no ponto mínimo dos picos (Método da perpendicular). A figura 27

mostra o critério utilizado na aplicação deste método para um cromatograma do conjunto CRO 120.

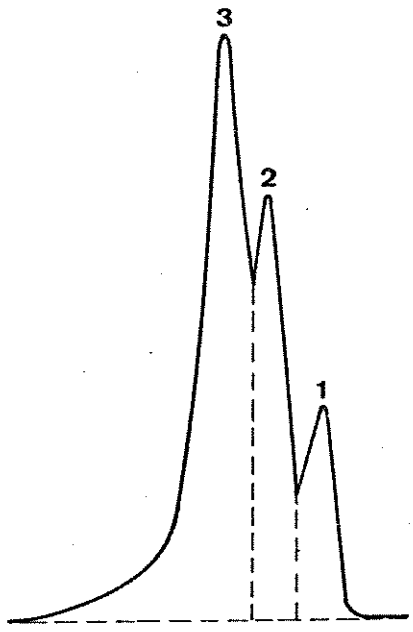


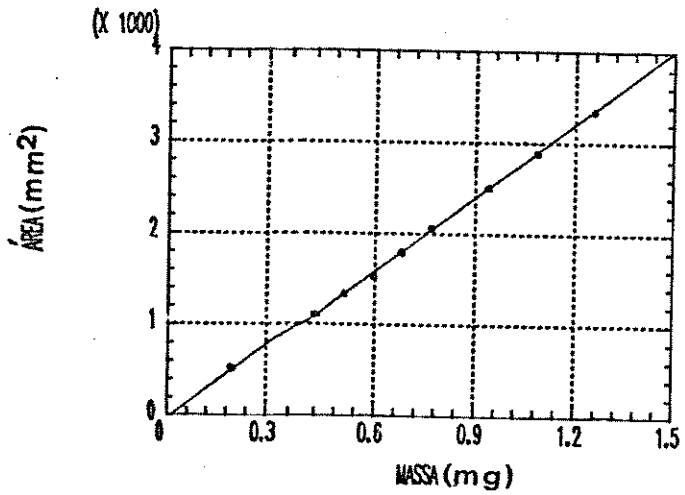
Figura 27 - Método de Separação Linear utilizado para a análise de um cromatograma do conjunto CRO 120. Amostras : 1. Etanol, 2. Isoctano, 3. Tolueno

Esse método gráfico foi aplicado apenas ao conjunto de dados obtidos quando a temperatura da coluna foi fixada em 120°C. Para os cromatogramas obtidos a 130°C, a superposição dos sinais é tão grande que a ausência de um vale entre os picos impede sua aplicação. Já para o conjunto de dados obtidos para a temperatura da coluna em 105°C, a superposição é pequena e os erros devem ser baixos.

Para a utilização desse método é necessário a construção de curvas de calibração independentes para cada um dos três constituintes presentes nas amostras.

As curvas de calibração foram então construídas mantendo a temperatura da coluna em 120°C e os demais parâmetros, conforme citados anteriormente. Foi utilizado o-xileno (Carlo Erba - p.a.) como solvente, sendo as soluções preparadas pela pesagem direta de soluto e solvente. A massa injetada foi calculada a partir dos dados de densidade à temperatura ambiente [46] e do volume injetado, que ficou constante em 2μl.

Foram preparadas nove soluções para cada um dos três constituintes e levantada a curva de calibração da área do pico (em milimetros quadrados) contra a massa injetada (em miligramas). A figura 28.a mostra a curva de calibração para o Tolueno, a 28.b para o Isoctano e a 28.c para o Etanol

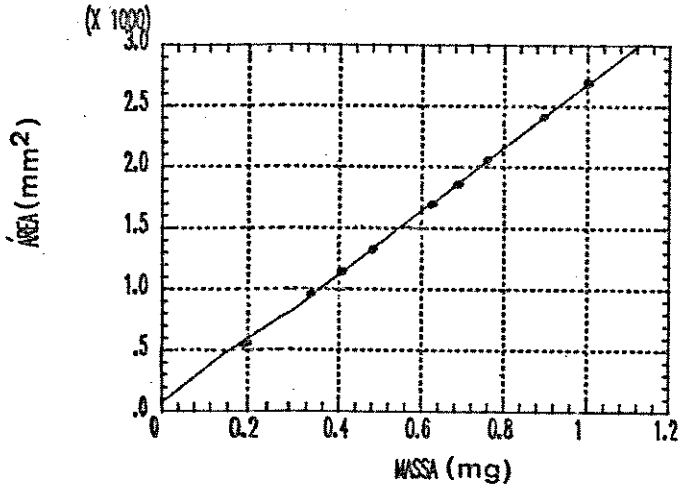


Coef. Angular : 2696,7 mm^2/mg

Coef. Linear : -32,97 mm^2

Coef. Correlação : 0,9996

Figura 28.a - Curva de Calibração para o Tolueno

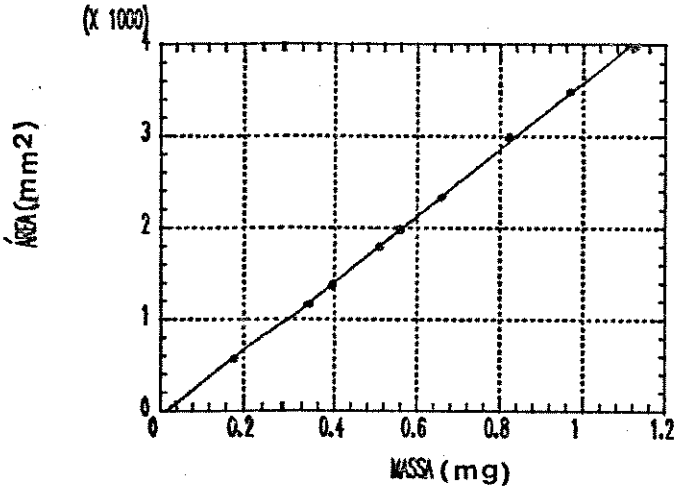


Coef. Angular : 2603,6 mm^2/mg

Coef. Linear : 42,83 mm^2

Coef. Correlação : 0,9998

figura 28.b - Curva de Calibração para o Isoctano.



Coef. Angular : 3648,6 mm^2/mg

Coef. Linear : -67,87 mm^2

Coef. Correlação : 0,9995

Figura 28.c - Curva de Calibração para o Etanol.

A partir dessas curvas de calibração construídas para cada um dos 3 constituintes individualmente, foram quantificadas as mesmas 6 amostras que foram utilizadas como conjunto teste na análise multivariada, e os resultados obtidos são apresentados na tabela 14.

Tabela 14 - Resultados da análise pelo método de Separação Linear (massas em mg).

amostra no.	TOLUENO			ISOCTANO			ETANOL		
	REAL	PREV.	ERRO REL. (%)	REAL	PREV.	ERRO REL. (%)	REAL	PREV.	ERRO REL. (%)
1	0,629	0,848	34,8	0,402	0,431	7,2	0,528	0,426	-19,3
2	0,201	——[a]	—	0,893	——	—	0,359	0,385	-15,9
3	0,395	0,481	21,8	0,614	0,652	6,2	0,501	0,527	-20,8
4	0,415	0,506	21,9	0,303	0,366	20,8	0,839	0,852	-18,0
5	0,612	0,815	33,2	0,607	0,679	11,9	0,311	0,321	-21,9
6	0,520	0,668	28,5	0,496	0,514	3,8	0,523	0,521	-21,6

a. Não é possível utilizar este método.

Pelos resultados obtidos pode-se notar que para o Etanol, as massas sempre foram menores que as reais. Os picos do Etanol possuem cauda, e com esse método faz-se um corte de uma parte do pico, e desta forma parte da sua área é perdida. Para o Isoctano, alguns resultados são bastante razoáveis, e isto talvez se deva ao fato que a perda da área a ser sofrida pelo corte da parte final do pico, inerente ao próprio método, acabe sendo compensada pelo respectivo aumento devido às influências do Etanol e do Tolueno. Pode-se notar que para a menor massa de Isoctano ("amostra 4"), o erro na determinação aumenta muito, pois nesse caso a área do pico

que é perdida com o corte, é proporcionalmente muito menor que aquela referente à influência do Tolueno.

Já para o Tolueno, as massas previstas são sempre muito maiores que as reais, devido à influência do Isoctano fazendo aumentar a área relativa ao Tolueno.

No geral, pode-se verificar que este método não é apropriado para a quantificação dos picos superpostostos estudados neste caso, devendo ser utilizado apenas quando a superposição entre os picos é pequena, e para picos que não possuem cauda.

4. - Erro na Análise sem superposição

Foi determinado o erro cometido na análise de cada um dos três constituintes separadamente, no caso de não haver superposição.

Para que se pudesse avaliar esse erro, foram preparadas outras três soluções de cada um dos constituintes, da mesma maneira que na construção das curvas de calibração. Utilizando-se agora as curvas de calibração, foram estimadas as massas injetadas nessas três soluções.

Os erros relativos em cada estimativa puderam então ser calculados, uma vez que se conhecia antecipadamente as massas. As médias desses erros para cada constituinte são apresentadas na tabela 15.

Esses erros devem expressar ruídos na instrumentação utilizada, assim como erros do operador, tanto na preparação das amostras como na injeção.

Tabela 15 - Erros relativos porcentuais na quantificação sem superposição.

CONSTITUINTE	ERRO RELATIVO (%)
tolueno	2,5
isoctano	2,1
etanol	3,0

5. - Análise Global dos Métodos utilizados :

Agora, é possível fazer uma análise global dos métodos utilizados para a quantificação dos picos cromatográficos superpostos.

O método tradicionalmente utilizado nesse caso que consiste em separar graficamente os picos com o auxílio de uma reta perpendicular à linha base traçada a partir do vale entre dois sinais, apresentou para os casos estudados, em que os picos possuem cauda e a superposição já é bastante pronunciada, erros de previsão muito altos. Isto mostra a ineficiência desse método em determinações quantitativas com sinais superpostos.

Por outro lado, os métodos baseados nos componentes principais como o PCR e o PLS apresentaram erros relativos que, em média, são praticamente iguais aos encontrados quando foram feitas determinações onde não havia nenhuma superposição entre os picos.

A comparação entre os resultados obtidos pelo PCR, PLS e a determinação gráfica, com os encontrados na análise sem superposição é mostrada na tabela 16.

Tabela 16 - Comparação entre as médias dos Erros Relativos Porcentuais do PCR, PLS, Método Gráfico e Análise sem superposição.

CONSTITUINTE	PCR			PLS			Separação Linear *	Sem superposição *
	CRO 105	CRO 120	CRO 130	CRO 105	CRO 120	CRO 130		
TOLUENO	1,1	1,9	2,1	1,1	1,8	3,3	28,0	2,5
ISOCTANO	2,4	2,8	3,2	2,3	1,7	3,0	10,0	2,1
ETANOL	2,3	1,3	3,4	2,3	1,8	3,3	19,6	3,0

* valores obtidos a 120°C.

Pela tabela 16, fica evidênciado que a aplicação dos métodos de Calibração Multivariada nos casos estudados, apresentam grandes vantagens quando comparada com os métodos clássicos. Esses resultados, muito expressivos no sentido de demonstrar o potencial dos métodos de Calibração Multivariada em resolver esse tipo de problema, ganham ainda maior significado se for lembrado que, em alguns casos, a superposição dos sinais é tão acentuada que o método gráfico não pode ser aplicado.

CAPÍTULO V

CONCLUSÕES

Os métodos de Calibração Multivariada baseados nos Componentes Principais, como o PCR e principalmente o PLS, tem sido aplicados na química apenas recentemente, e muitos dos seus aspectos ainda estão sendo investigados.

Até a pouco tempo a aplicação de Métodos de Calibração Multivariada em problemas similares aos estudados nesse trabalho, limitava-se à Regressão Linear Múltipla ou à Regressão Múltipla por passos. Somente com a utilização do PCR e do PLS começaram a surgir resultados confiáveis que permitem a aplicação prática da Calibração Multivariada para quantificar sinais superpostos.

No nosso entender, alguns aspectos do PLS devem ainda ser objeto de estudos mais aprofundados, como é o caso da determinação do número de Componentes Principais a ser utilizado, e a fixação de critérios para definir as amostras que não devem fazer parte da categoria definida na fase de calibração. Esses parâmetros são fundamentais para que se possa obter resultados corretos e confiáveis.

Quanto aos resultados obtidos, foi possível verificar que para os casos estudados, os resultados do PCR foram praticamente iguais aos do PLS. Esta é uma conclusão importante para o grupo de Quimiometria da UNICAMP, uma vez que o programa para os cálculos com o PCR foi montado a partir de pacotes computacionais abertos e bem conhecidos, enquanto que os programas para os cálculos com o

PLS são protegidos e possuem limitação no número de usuários.

Ficou ainda evidenciada a enorme dificuldade que se encontra quando da utilização do método manual de separação linear, o que justifica, nesse caso, a afirmação encontrada na referência [6], "Somente análises com picos resolvidos completamente produzirão resultados de alta precisão e exatidão". Por outro lado, os resultados da Calibração Multivariada parecem sugerir que afirmações desse tipo sejam revistas.

Na aplicação do PLS foi ainda destacada a importância do pré-tratamento dos dados, pois somente após a redução das variáveis e normalização dos cromatogramas à unidade tornou-se possível obter resultados comparáveis aos da análise sem superposição.

Os resultados obtidos pelos Métodos de Calibração Multivariada são praticamente iguais nos três casos de superposição estudadas, ou seja, pode-se quantificar picos completamente superpostos com um mínimo de erro, fato que superou as espectativas iniciais do trabalho.

Tudo isso aponta para um enorme conjunto de aplicações para esses métodos na quantificação de sinais superpostos, não só em cromatografia, mas também em muitos outros métodos instrumentais.

REFERÊNCIAS

[1]. M. A. Sharaf, D. L. Illman e B. R. Kowalski, "Chemometrics", Wiley, New York, 1986.

[2]. R. R. Meglen, *Chemometrics Intell. Lab. Systems*, 3 (1988) 17.

[3]. J. Novák, "Quantitative Analysis by Gas Chromatography" - Chromatografic Science Series - vol. 5, Marcel Dekker, 1975.

[4]. E. Proksch, H. Bruneder, V. Grauzner, *J. Chromatogr. Sci*, 7 (1969) 473.

[5]. A. W. Westerberg, *Anal. Chem.*, 41 (1969) 1770.

[6]. R. Ciola, "Fundamentos da Cromatografia a Gás", Edgard Blücher, São Paulo, 1985.

[7]. S. M. Roberts, *Anal. Chem.*, 44 (1972) 502.

[8]. H. M. Gladney, B. F. Dowden, J. D. Swalen, *Anal. Chem.*, 41 (1969) 883.

[9]. T. S. Buys, K. de Clerk, *Anal. Chem.*, 44 (1972) 1273.

[10]. E. Grushka, M. N. Myers, J. C. Giddings, *Anal. Chem.*, 42 (1970) 21.

[11]. R. D. B. Frazer, E. Suzuki, *Anal. Chem.*, 41 (1969) 37.

[12]. A. H. Anderson, T. C. Gibb, A. B. Littlewood, *Anal. Chem.*, 42 (1970) 434.

[13]. J. P. Gourlia, J. Bordet, *J. Chromatogr. Sci.*, 19 (1981) 35.

[14]. S. Wold, K. Esbensen, P. Geladi, *Chemometrics Intell. Lab. Systems*, 2 (1987) 37.

[15]. S. Wold, C. Albano, W. J. Dunn III, K. Esbensen, S. Hellberg, E. Johansson, M. Sjostrom, em "Food Research and Data Analysis", H. Martens e H. Russwurn Jr. eds., Applied Science,

London, 1983.

[16]. P. Geladi, B. R. Kowalski, *Anal. Chim. Acta*, 185 (1986) 1.

[17]. K. B. Beebe, B. R. Kowalski, *Anal. Chem.*, 59 (1987) 1007A.

[18]. H. Wold em "Sistems Under Indirect Observation", K. G. Joreskog e H. Wold eds., North-Holland, Amsterdam, 1982.

[19]. S. Wold, A. Ruhe, H. Wold e W. Dunn, *SIAM J. Sci. Stat. Comput.*, 5 (1984) 735.

[20]. M. Otto, W. Wegscheider, *Anal. Chem.*, 57 (1985) 63.

[21]. W. Lindberg, J. A. Persson, S. Wold, *Anal. Chem.*, 55 (1983) 643.

[22]. P. Geladi, D. MacDougall, H. Martens, *Appl. Spectrosc.*, 39 (1985) 491.

[23]. M. Martens, H. Martens, *Appl. Spectrosc.*, 40 (1986) 303.

[24]. T. V. Karstang, R. J. Eastgate, *Chemometrics Intell. Lab. Systems*, 2 (1987) 209.

[25]. M. Otto, J. D. R. Thomas, *Anal. Chem.*, 57 (1985) 2647.

[26]. I. Lukkari, W. Lindberg, *Anal. Chim. Acta*, 211 (1988) 1.

[27]. W. Lindberg, J. Öhman, S. Wold, H. Martens, *Anal. Chim. Acta*, 171 (1985) 1.

[28]. W. Lindberg, J. Öhman, S. Wold, H. Martens, *Anal. Chim. Acta*, 174 (1985) 41.

[29]. I. S. Scarminio, Tese de Doutorado, Universidade Estadual de Campinas - Instituto de Química, 1989.

[30]. O programa SIMCA-3B pode ser obtido junto ao Prof. S. Wold no seguinte endereço : Principal Data Components, 2505 Shepard Blvd., Columbia, MO 65201. U.S.A.

[31]. B. R. Kowalski, C. F. Bender, *J. Am. Chem. Soc.*, 94 (1972)

5632.

[32]. N. Draper, H. Smith, "Applied Regression Analysis", Wiley, New York, 1981.

[33]. T. Naes, H. Martens, *Trends Anal. Chem.*, 3 (1984) 266.

[34]. H. Martens, T. Karstang, T. Naes, *J. Chemometrics*, 1 (1987) 201.

[35]. B. R. Kowalski, *Anal. Chem.*, 52 (1980) 112R.

[36]. I. E. Frank, B. R. Kowalski, *Anal. Chem.*, 54 (1982) 232R.

[37]. M. F. Delaney, *Anal. Chem.*, 56 (1984) 261R.

[38]. L. S. Ramos, K. R. Beebe, W. P. Carey, E. Sanches M., B. C. Erickson, B. E. Wilson, L. E. Wangen, B. R. Kowalski, *Anal. Chem.*, 58 (1986) 294R.

[39]. S. D. Brown, T. Q. Barker, R. J. Lariver, S. L. Monfre, H. R. Wilk, *Anal. Chem.*, 60 (1988) 252R.

[40]. S. Wold, P. Geladi, K. Esbensen, J. Öhman, *J. Chemometrics*, 1 (1987) 41.

[41]. B. G. M. Vandeginste, C. Sielhorst, M. Gerritsen, *Trends Anal. Chem.*, 7 (1988) 286.

[42]. P. Geladi, B. R. kowalski, *Anal. Chim. Acta*, 185 (1986) 19.

[43]. W. P. Carey, K. R. Beebe, B. R. Kowalski, *Anal. Chem.*, 59 (1987) 1529.

[44]. T. L. Isenhour, P. C. Jurs, "Introduction to Computer Programming for Chemists", Allyn and Bacon Inc., Boston, 1972.

[45]. J. A. Riddick, W. B. Bunger, "Organic Solvents", 3rd. ed, Wiley - Interscience, New York, 1970.

[46]. "International Critical Tables of Numerical data - Physics, Chemistry and Technology, Editor-in-chief E. W. Washburn, McGraw -

Hill Book company, New York, 1982, vol. 3, pg. 27.

[47]. M. Sjöström, S. Wold, W. Lindberg, J. A. Persson, H. MArtens, *Anal. Chim. Acta*, 150 (1983) 61.

[48]. D. M. Haaland, E. V. Thomas, *Anal. Chem.*, 60 (1988) 1202.