QUIMIOMETRIA

Professora Márcia M. C. Ferreira
Laboratório de Quimiometria Teórica e Aplicada
Instituto de Química UNICAMP
Campinas, SP, 13083 - 970
Email: marcia@iqm.unicamp.br
URL: http://lqta.iqm.unicamp.br

UNICAMP

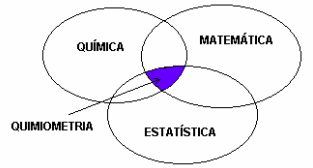
QUIMIOMETRIA

Quimiometria pode ser definida como uma área da química que usa métodos matemáticos e estatísticos para

a- planejar ou selecionar procedimentos ótimos de medidas e experimentos.
b- extrair o máximo da informação química relevante, com a análise dos dados

Outras definições:

Quimiometria é uma ciência que relaciona MEDIDAS, feitas num sistema ou processo químico, ao ESTADO do sistema utilizando métodos matemáticos e/ou estatísticos.

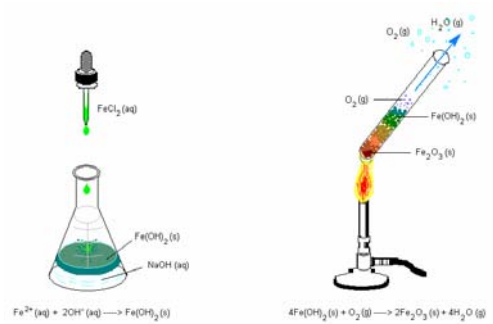


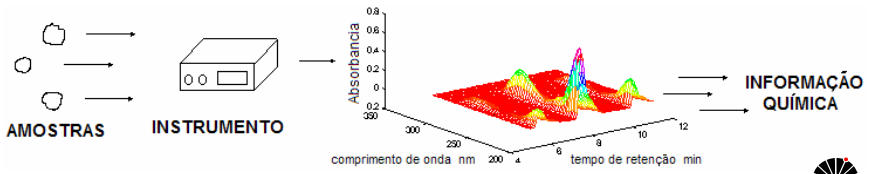
A quimiometria engloba todo um processo onde os DADOS (por exemplo, números em uma tabela) são transformados em informações usadas para tomar decisões.

Administração e processamento de informações de natureza química.

Quimiometria é o que os quimiometristas fazem.

UNICAMP





UNICAMP

REFERÊNCIAS

LIVROS

CHEMOMETRICS M. A. Sharaf, D. L. Illman and B. R. Kowalski, Wiley-Interscience (1986).

FACTOR ANALYSIS IN CHEMISTRY E. R. Malinowski, 3rd edition, John Wiley & Sons Ltd. (2002).

MULTIVARIATE CALIBRATION H. Martens and T. Naes, John Wiley & Sons Ltd. (1989)

CHEMOMETRICS A Practical Guide K. Beebe, R. Pell. M. B. Seasholtz, John Wiley & Sons (1998).

CHEMOMETRICS Data Analysis for the Laboratory and Chemical Plant Richard G. Brereton, John Wiley & Sons (2002).

HANDBOOK OF CHEMOMETRICS AND QUALIMETRICS; Data Handling In Science and Technology, Volumes 20A e B Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi P. J.; Smeyers-Verbeke. J.; Elsevier, Amsterdam, 1997.

REVISTAS ESPECIALIZADAS

Journal of Chemometrics
Chemometrics and Intelligent Laboratory Systems
Analytical Chemistry Analytica Chimica Acta Applied Spectroscopy

SOFTWARE / LINKS

MATLAB PLS_Toolbox (Eigenvector Res. Inc,) www.eigenvector.com, www.models.kvl.dk/source/
PIROUETTE (Infometrix, Inc.) www.infometrix.com
UNSCRAMBLER (CAMO Inc.) www.camo.com/
SIMCA (Umetrics) www.umetrics.com/

UNICAMP

UM ROTEIRO PARA A ANÁLISE DE DADOS MULTIVARIADOS

1- Definição do problema

2- Organização dos dados

3- Validação dos dados

4- Visualização dos dados originais

5- Transformação / Pré-processamento dos dados

6- **Análise Exploratória** dos dados

7- Construção de Modelos de

Calibração/Classificação

8- Validação dos Modelos

9- Uso dos Modelos para previsões

DEFINIÇÃO DO PROBLEMA

A primeira questão a ser colocada para um sistema em estudo:

O que queremos saber deste sistema?

Gastar um tempo em definir o problema a ser resolvido, com certeza ajuda na escolha correta das técnicas experimentais a serem usadas, e no desenvolvimento dos protocolos garantindo que as informações realmente desejadas sejam coletadas

FATORES QUE DEVEM SER CONSIDERADOS

Rever a história da origem do problema

Verificar:

- Como os dados foram gerados
- Que métodos de medida foram utilizados
- O nível de acuracidade relacionado a cada variável
- Quando os dados foram coletados, etc.

OUTROS FATORES IMPORTANTES

- Já foi feita alguma análise anterior?
- Existe alguma informação anterior que seja pertinente?
- As medidas feitas mais recentemente são diferentes das anteriores?
- A acuracidade instrumental melhorou?

ORGANIZAÇÃO DOS DADOS

É conveniente colocar os dados num único arquivo.

TENHA SEMPRE EM MENTE

- Os dados são de um único instrumento?
- Há mais de um instrumento ou tipo?
- Qual a precisão dos instrumentos?
- Há resposta de questionários ou dados coletados a mão?
- Os dados de cada amostra estão num único arquivo?
- A(s) variável(eis) dependente(s) e/ou classes estão em arquivo a parte ou serão entrados a mão?

Cada amostra corresponde a uma linha na matriz de dados, cujos elementos são os valores das variáveis medidas

$$\mathbf{x}_i^{\mathrm{T}} = [x_{i1} \quad x_{i2} \quad x_{i3} \quad \cdots \quad x_{iJ}]$$

Cada coluna $\mathbf{x}^j$ se refere a uma variável independente, ou seja, uma medida j realizada para todas as amostras

$$\mathbf{x}^j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ \vdots \\ x_{Ij} \end{bmatrix}$$

O resultado é uma matriz $\mathbf{X}(I,J)$ com um total de I linhas (amostras) e J colunas (variáveis) cujos valores x_{ij} são as respostas para as variáveis j = 1, 2, 3, ..., J, referentes à amostra i.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \vdots \\ \mathbf{x}_I^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2J} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ x_{I1} & x_{I2} & \cdots & \cdots & x_{IJ} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1 & \mathbf{x}^2 & \cdots & \cdots & \mathbf{x}^J \end{bmatrix}$$

As variáveis podem ser

☺ respostas de um instrumento multicanal (espectrômetro): as intensidades para diferentes comprimentos de onda.

☺ resultados de instrumentação de separação, (cromatógrafo): a altura ou área de pico em tempos de retenção específicos ou relativos correspondendo a constituintes específicos.

☺ resultados de instrumentação eletroquímica (voltametria, potenciometria): de medidas de potenciais em eletrodos seletivos para íons (potenciometria) ou intensidades de correntes para diferentes potenciais (voltametria).

☺ ensaios múltiplos específicos de instrumentação univariada

☺ testes Físicos/Químicos/ Biológicos

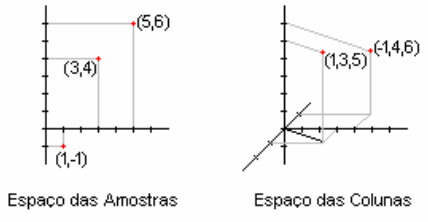
☺ resposta não instrumental: resposta de uma análise sensorial

A matriz **X**(*I*,*J*) pode ser interpretada de duas maneiras diferentes:

☺Como um arranjo de *I* pontos, num espaço de dimensão *J* onde cada ponto tem *J* coordenadas. Este é chamado de espaço das linhas. Cada amostra corresponde a um ponto neste espaço.

☺Como um arranjo de *J* pontos num espaço de dimensão *I*. Este é o espaço das colunas, onde cada coluna é representada por um vetor com *I* coordenadas.

$$\mathbf{X}=\begin{bmatrix}1 & -1\\ 3 & 4\\ 5 & 6\end{bmatrix}$$



No gráfico da esquerda (espaço linha) podemos ver a relação entre as amostras (similaridade/dissimilaridade entre elas). No gráfico da direita pode-se ver a relação entre as duas variáveis.

VALIDAÇÃO DOS DADOS

O valor x_{ij} está ausente na matriz de dados.

O QUE FAZER???

-Exclua as linhas ou colunas
-Complete com valores (as linhas ou as colunas) ZERO?
nem sempre é possível ou aconselhável

Três métodos usados para estimar x_{ij} a partir do restante dos dados:

1- x_{ij} = valor médio obtido para a variável *j* com os valores das amostras restantes.

2- x_{ij} = média ponderada dos valores das *k* amostras que não tem dados faltantes, x_{1j}, ..., x_{kj} e que estão mais próximas da amostra *i*. A contribuição de cada amostra é ponderada com base na similaridade do seu perfil com o da amostra *i*.

3- O terceiro método faz uso da análise de componentes principais (ou decomposição de valores singulares, SVD)

Em linhas gerais:

☺Seleciona-se um valor inicial para x_{ij} (o valor médio por exemplo)
☺Faz-se a decomposição de valores singulares dos dados completos
☺Selecionam-se as *A* componentes principais significativas
☺Reconstrói-se os dados com *A*PCs, produzindo uma estimativa para x_{ij}.
☺Repete-se o processo (nova SVD) até que a diferença entre duas matrizes reconstruídas convirja para um vapor arbitrário suficientemente pequeno.

O uso do valor médio é o mais rápido mas de acuracidade mais baixa.

O método dos *k*-ésimos vizinhos mais próximos em geral apresenta melhor desempenho especialmente com o acréscimo na quantidade de dados faltantes.

Os dois últimos métodos funcionam bem quando a quantidade de dados faltantes é relativamente alta.

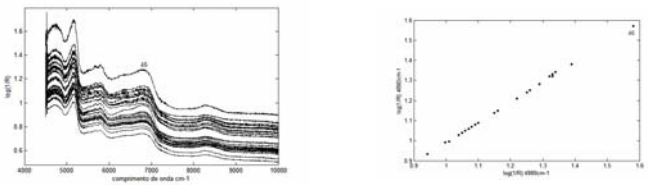
Matrizes com 50 – 100 amostras podem ser analisadas caso tenham de 10 a 20% de dados faltantes, desde que não estejam faltando segundo algum padrão sistemático. Quanto maior a matriz de dados, maior esta proporção.

VISUALIZAÇÃO DOS DADOS ORIGINAIS

A maneira mais interessante de visualizar os dados é por meio de gráficos. Os dados abaixo se referem a espectros refletância difusa na região do infravermelho próximo de 26 amostras. As variáveis são os comprimentos de onda na faixa de 4500 – 10000 cm^{-1}, com resolução de 4 cm^{-1} (incremento de 2 cm^{-1}).

Os resultados são expressos em log(1/R). Temos mais de duas mil variáveis medidas por amostra, **X** = (26, 2750).

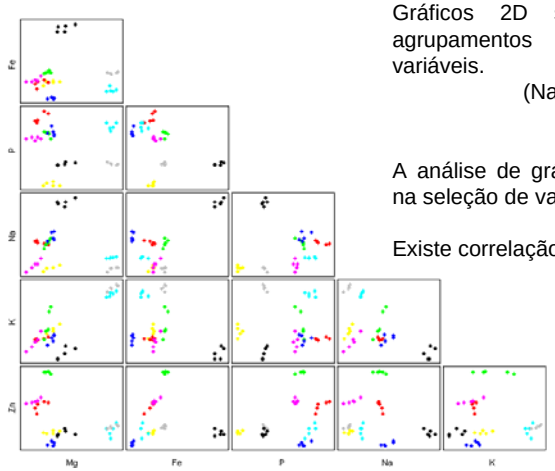
Os espectros apresentam deslocamentos e inclinação na linha de base consideráveis, característicos de espectros de reflectância difusa.



Existe alta correlação entre as variáveis indicando que elas contêm essencialmente a mesma informação (uma variável é aproximadamente função linear da outra). Isto ocorre quando temos matrizes com *J* >> *I*.

Atenção: A amostra 46 tem um comportamento atípico. Seria um "*outlier*"?

Este exemplo contém uma série de gráficos bi-variados de amostras de oito tipos diferentes de embutidos de peru (salsicha, hambúrguer, almôndega, banque, role, presunto, presunto defumado e peito de peru defumado), que foram analisados por espectrometria de emissão óptica com plasma indutivamente acoplado (ICP-OES). As variáveis são as concentrações de seis elementos minerais: Na, K, Mg, Fe, Zn e P.



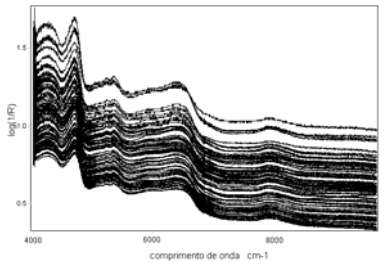
Gráficos 2D são bons para visualizar agrupamentos de amostras em algumas variáveis.

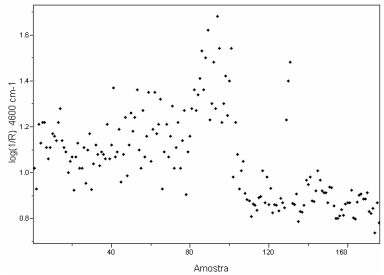
(Na x Zn); (Na x P)

A análise de gráficos bi-variados é muito útil na seleção de variáveis.

Existe correlação entre as variáveis?

Os dados abaixo se referem aos espectros no infravermelho próximo das amostras de café cru.





O gráfico acima mostra uma tendência sistemática da variável correspondente ao número de onda 4800 cm^{-1} com o número crescente da amostra (problema??)

O deslocamento da linha de base varia com o tempo. A distribuição das absorbâncias não está centrada no valor 1,1 como esperado.

Gráficos desta natureza podem indicar um acréscimo monotônico da variável ou uma flutuação *"drift"* não calibrada da medida, originada de variações ambientais (temperatura, umidade), vibrações, variações na fonte, etc.

PRÉ-TRATAMENTO DOS DADOS

OS DADOS EXPERIMENTAIS SÃO PREPARADOS PARA ANÁLISE

Objetivo: remover matematicamente fontes de variação indesejáveis que não serão removidas naturalmente durante a análise dos dados.

TRANSFORMAÇÃO	PRÉ-PROCESSAMENTO
↓	↓
Aplicado às amostras	Aplicado às variáveis
Linhas da matriz **X**	Colunas da matriz **X**

TRANSFORMAÇÃO DOS DADOS

Os sinais medidos consistem de:

SINAL MEDIDO =

CONTRIBUIÇÃO DETERMINÍSTICA + CONTRIBUIÇÃO ESTOCÁSTICA

↓ ↓

Sinal verdadeiro Ruído aleatório

Informação relevante + Informação indesejada

Variações sistemáticas ou aleatórias devem ser removidas.

Variações aleatórias: (ruído experimental) são removidas por meio de técnicas de alisamento (*smoothing*) com o objetivo de aumentar a razão sinal-ruído S/R.

Variações sistemáticas: reduzidas ou eliminadas por meio de correções da linha de base.

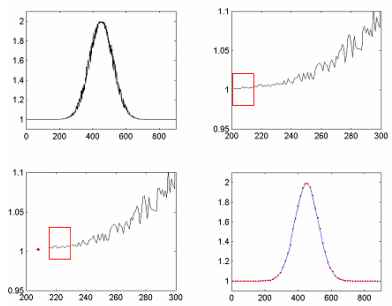
TÉCNICAS DE ALISAMENTO

☺ Aumentam a razão sinal-ruído.
☺ Em geral utilizam uma janela.
☺ Todos os pontos da janela são usados para determinar a resposta no centro da mesma.
☺ A janela percorre todo o espectro.

O alisamento com filtro de Fourier usa uma metodologia diferente.

ALISAMENTO PELA MÉDIA

É usado quando se deseja diminuir o número de variáveis (*J*).



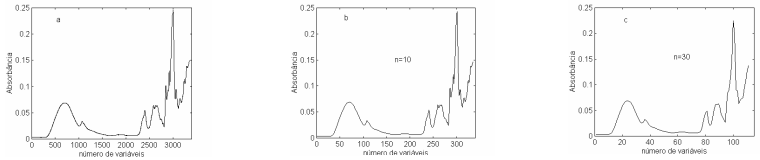
Seleciona-se uma janela de abertura = n+1, onde n é um número inteiro par. (as n+1 primeiras variáveis),

Calcula-se a média das respostas, que será a primeira variável do espectro alisado, com comprimento de onda igual ao do centro da janela (n/2 + 1).

Faz-se o mesmo com as variáveis n+2 ate 2n+2

O resultado é um espectro com um número de variáveis igual ao inteiro mais próximo de $J/(n+1)$.

CUIDADO, POIS O TAMANHO DA JANELA PODE ELIMINAR INFORMAÇÕES RELEVANTES.



(a) espectro original. (b) espectro alisado com janela n = 10. (c) espectro alisado com janela n = 30.

No eixo das ordenadas é mostrado o número das variáveis e não o número de onda.

Note a perda de resolução no último espectro alisado com janela n = 30.

Alisamento pela média na linguagem do MATLAB

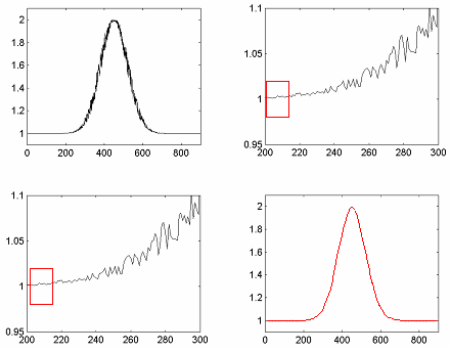
```
% Esta rotina, faz o alisamento pela média
%  X é a matriz a ser alisada (amostras nas linhas)
%  Xalis é a matriz alisada e "janela" é o tamanho da janela
X=X';
Xalis=[];
[J,I]=size(X);
  for j=1: round(J/(janela+1)-.5)
       Xalis=[Xalis ; mean(X((j-1)*janela+j: j*janela+j,:))];
  end
%para a última janela
  if (J-(j*janela+j)) > janela/2
       Xalis=[Xalis ; mean(X(j*janela+j:J,:))];
  end
Xalis= Xalis';
```

ALISAMENTO PELA MÉDIA MÓVEL

Idêntico ao alisamento pela média, só que neste caso, a janela move de elemento em elemento ao invés de janela em janela.

O espectro alisado contendo basicamente o mesmo número de variáveis que o original.

Em lugar da média, pode-se usar a mediana da janela ajustar um polinômio aos pontos da janela e substituir o ponto central da janela pelo valor estimado pelo polinômio.

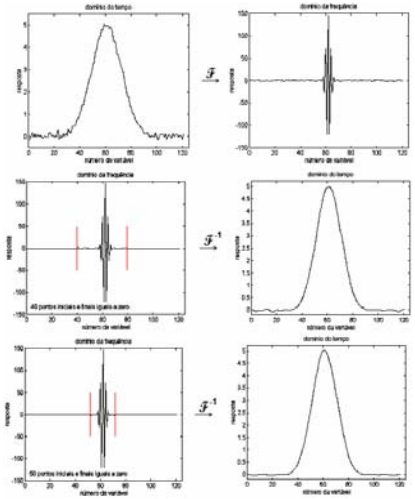


(a) a banda simulada é uma gaussiana com ruído adicionado.
(b) parte da banda com a janela móvel em vermelho.
(c) primeira janela alisada e a indicação da segunda janela.
(d) banda alisada sobreposta à banda original sem ruído.

ALISAMENTO COM FILTROS DE FOURIER

Os espectros que estão no domínio do tempo são transformados para o domínio da frequência (interferograma) pela transformada de Fourier.

As componentes de alta frequência no início e final do interferograma são removidas fazendo os coeficientes de Fourier iguais a zero.



O resultado é então transformado de volta ao domínio do tempo.

Foram retidas as componentes de baixa frequência, este filtro é denominado "*low pass*".

Cuidado para evitar um super alisamento que pode distorcer ou eliminar características importantes do espectro.

Com o alisamento há uma perda na resolução e o espectro alisado contém menos informação que o original.

Este é o preço pago pelo acréscimo na razão S/R.

CORREÇÕES DA LINHA DE BASE

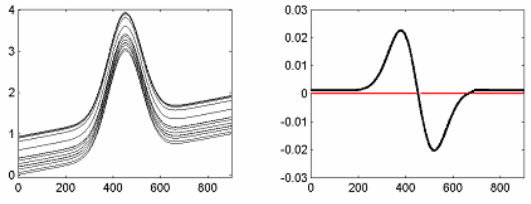
DERIVADAS

Problemas de linha de base podem ser corrigidos tomando-se as derivadas do espectro.

O algoritmo mais utilizado é o de Savitsky-Golay.

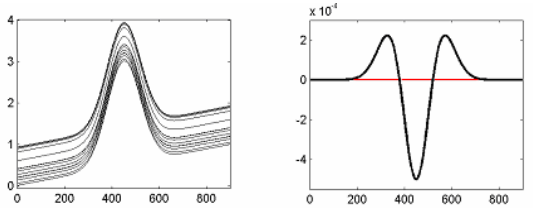
Primeira Derivada $$2\delta \frac{dA}{d\lambda}(\lambda_j) \cong \Delta A(\lambda_j) = A_{\lambda_j+\delta} - A_{\lambda_j-\delta}$$

O espectro inteiro pode estar deslocado de uma quantidade constante (*Offset* na linha de base), que pode ser corrigido tomando-se a primeira derivada.



Espectro que apresenta um problema de inclinação na linha de base (*bias*), "subindo um morro" à medida que decresce o número de onda, pode ser corrigido tomando a segunda derivada.

Segunda Derivada $(2\delta)^2 \frac{d^2A}{d\lambda^2}(\lambda_j) \cong \Delta^2 A(\lambda_j) = A_{\lambda_j+\delta} + A_{\lambda_j-\delta} - 2A_{\lambda_j}$



ALERTA: alguns algoritmos usados para calcular derivadas introduzem mais ruído nos resultados.

O analista deve decidir se o *offset* e *bias* eliminados usando derivadas compensam o ruído introduzido.

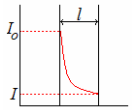
A segunda derivada mede a concavidade de uma curva. Esta característica é muito útil para identificar picos especialmente quando eles estão sobrepostos.

Outras Transformações Importantes

LOGARÍTMO

O logaritmo (log 10) pode ser aplicado com o objetivo de linearizar os dados e a escolha da base logarítmica não afeta a interpretação dos resultados.

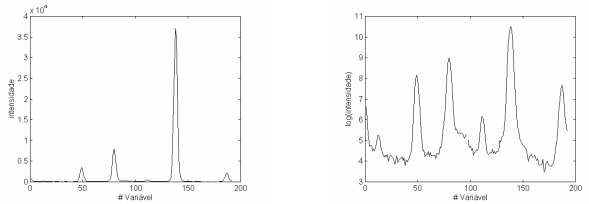
☺ Espectros de transmitância ou refletância não são lineares com a concentração e devem ser transformados para absorbância.



$A_\lambda = -\log T_\lambda = -\log I/I_o$

$T_\lambda = 10^{-\alpha_\lambda lc}$

☺ Esta transformação também pode ser usada para enfatizar intensidades baixas.



Transformação logarítmica de um espectro de fluorescência de raios-X com intensidades variando de 0,1x10⁴ a 3,8 x10⁴.

☺ Em estudos de QSAR se deseja obter uma relação funcional *f* entre uma série de descritores estruturais e a atividade biológica da forma:

atividade = *f*(*descritores*)

Na construção dos modelos, as atividades são transformadas para a forma logarítmica.

NORMALIZAÇÃO

Divide-se cada uma das variáveis de uma dada amostra *i* por um fator de normalização: pela norma da amostra *i*, representada por $||\mathbf{x}_i||$.

O resultado é que todas as amostras estarão numa mesma escala.

$$x_{ij(norm)} = \frac{x_{ij}}{\|\mathbf{x}_i\|} \quad , \quad j = 1, 2, ..., J$$

As normas mais utilizadas são:

$\|\mathbf{x}_i\|_\infty = \max_{1 \le j \le J} |x_{ij}|$ norma sup, ou l_∞ $\qquad$ $\|\mathbf{x}_i\|_1 = \sum_{j=1}^{J} |x_{ij}|$ norma l_1

$\|\mathbf{x}_i\|_2 = \sqrt{\sum_{j=1}^{J} x_{ij}^2}$ norma Euclideana ou norma l_2

Normalização pela norma sup: a resposta máxima de cada uma das amostras se torna igual a 1.

Normalização pela norma l_1: a área sob cada um dos espectros é unitária.

Normalização pela norma l_2: cada espectro terá comprimento igual a 1.

NOTAS:

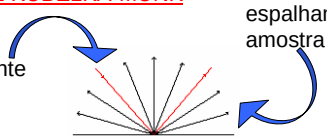
☺ A normalização é usada principalmente para remover variação sistemática, em geral associada com o tamanho da amostra.

☺ Corrige-se o efeito da variação no volume de injeção em cromatografia normalizando cada cromatograma para área unitária.

☺ Normaliza-se o pico com maior *m/e* em espectrometria de massa utilizando-se a norma sup.

TRANSFORMAÇÃO DE KUBELKA-MUNK

Reflexão da luz incidente numa amostra perfeitamente lisa: __



espalhamento de uma amostra rugosa: __

O mais comum para linearização de espectros de reflectância difusa é o uso da transformação mencionada anteriormente que é –log R_l.

A equação original de Kubelka-Munk relaciona a reflectância difusa absoluta com os coeficientes de espalhamento, *s*, e absorção molar, *k*.

Na prática a reflectância relativa foi substituída pela reflectância relativa (em relação a um padrão).

Esta equação, que define uma relação linear entre a intensidade espectral relativa (em relação a um padrão) e a concentração, é mais sofisticada que a simples transformação logarítmica, -log R_l.

$$\left[\frac{(1 - R_{\lambda_1})^2}{2R_{\lambda_1}}\right] = \frac{k}{s}$$

CORREÇÃO MULTIPLICATIVA DE SINAL – MSC

Usada para corrigir efeitos de espalhamento de luz em espectroscopia por refletância, causados por diferenças no tamanho e na forma das partículas.

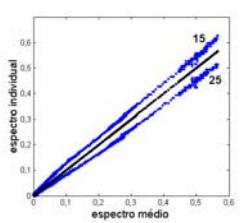
Estes efeitos são eliminados fazendo-se a regressão linear das variáveis espectrais ($\mathbf{X}^T$) nas variáveis do espectro médio ($\mathbf{x}_m$).

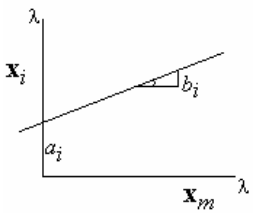
$$\mathbf{x}_i = a_i\mathbf{1} + b_i\,\mathbf{x}_m.$$

Os coeficientes a_i e b_i da amostra *i* são calculados por quadrados mínimos fazendo-se a regressão de cada espectro no espectro médio (um conjunto *a*, *b* para cada amostra),

O espectro corrigido $\mathbf{x}_{imsc}$ é obtido subtraindo-se a absorbância de cada comprimento de onda do espectro original, $\mathbf{x}_i$, de a_i e dividindo-a por b_i.

O resultado final da regressão é uma matriz de coeficientes (2 x *I*).

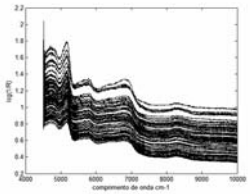


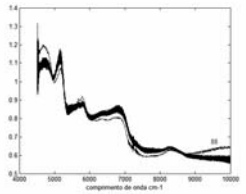


$$\begin{bmatrix}\vdots\\ \mathbf{x}_i\\ \vdots\end{bmatrix} = \begin{bmatrix}1 & \vdots\\ \vdots & \mathbf{x}_m\\ 1 & \vdots\end{bmatrix}\begin{bmatrix}a_i\\ b_i\end{bmatrix} \qquad \begin{bmatrix}a_i\\ b_i\end{bmatrix} = \left(\mathbf{X}_m^T\mathbf{X}_m\right)^{-1}\mathbf{X}_m^T\begin{bmatrix}\vdots\\ \mathbf{x}_i\\ \vdots\end{bmatrix} \quad (9), \qquad \begin{bmatrix}\vdots\\ \mathbf{x}_{imsc}\\ \vdots\end{bmatrix} = \frac{1}{b_i}\left(\begin{bmatrix}\vdots\\ \mathbf{x}_i\\ \vdots\end{bmatrix} - \begin{bmatrix}a_i\\ \vdots\\ a_i\end{bmatrix}\right)$$

Rotina em linguagem do MATLAB para o cálculo do espectro corrigido

```
[I,J]=size(X);
Xm=[ones(J,1) mean(X)'];
% Xm é uma matriz de duas colunas, apresentando uma
% coluna à esquerda com todas as suas entradas unitárias (1's),
coef=inv(Xm'*Xm)*Xm'*X';
Xmsc=(X'-ones(J,1)*coef(1,:))./(ones(J,1)*coef(2,:));
Xmsc=Xmsc';
```





A vantagem deste tratamento em relação às derivadas, é que o espectro corrigido se assemelha ao espectro original, o que auxilia na interpretação.

TRANSFORMAÇÃO DE WAVELET:

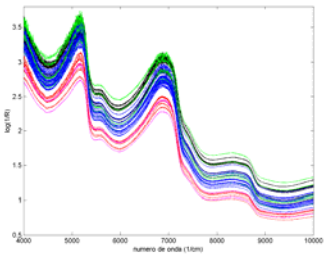
Tal como a transformada de Fourier, esta é uma transformação que fornece informações nos domínios do tempo e da frequência. Na transformada de Fourier faz-se uma combinação linear de senos e cosenos enquanto que na transformada de wavelet as funções usadas são as funções de wavelet.
Esta transformação é bastante utilizada para minimizar efeitos de ruídos e desvios tendenciosos "*drifts*" da linha de base. Para uma descrição mais detalhada, o leitor deve consultar a literatura.

O EXEMPLO A SEGUIR ILUSTRA O EFEITO DE ALGUMAS DAS TRANSFORMAÇÕES VISTAS

O conjunto de dados contém os espectros originais, registrados na região do infravermelho próximo, NIR, na faixa de 4000 cm^{-1} a 10000 cm^{-1} com resolução de 4 cm^{-1} de 42 amostras de produtos de tomate.

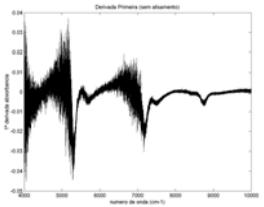
♣ Os espectros apresentam deslocamento e uma inclinação na linha de base.

♣ As regiões de 4000 cm^{-1} a 5500cm^{-1} e de 6300 cm^{-1} a 7300 cm^{-1}apresentam um nível mais alto de ruído.

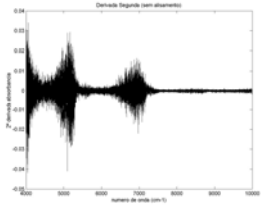


Usar o método das derivadas para resolver os problemas de linha de base; a segunda derivada, em princípio, deveria resolver ambos os problemas.

Espetros das amostras de produtos de tomate após a derivação



primeira derivada

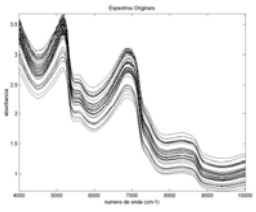


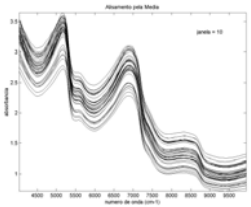
segunda derivada.

É visível a introdução do ruído originado com este método. A segunda derivada dos espectros apresenta um nível bem maior de ruído que a primeira derivada.

A eliminação do deslocamento e inclinação da linha feita desta maneira compensa o ruído introduzido?

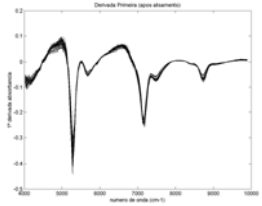
Estes resultados podem ser melhorados aumentando a razão S/R com um alisamento pela média nos dados originais.
Para isto foi usada uma janela de tamanho *n* = 10 para o alisamento.



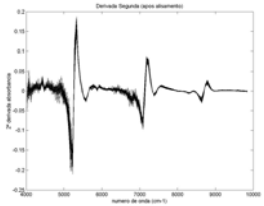


O número de variáveis foi reduzido para 273.
É visível o acréscimo na relação S/R especialmente na região de 4000 cm^{-1} – 5500 cm^{-1}. A forma geral do espectro não sofreu alteração e portanto, não houve perda de informação com esta transformação.

Após o alisamento, os espectros foram então derivados para a correção dos efeitos de linha de base.



primeira derivada

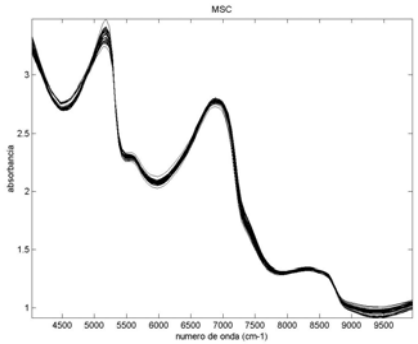


segunda derivada

Este exemplo mostra a importância do uso correto das transformações.

Aplicação da correção multiplicativa de sinal nos espectros alisados.

Espetros alisados pela média e então corrigidos pela correção multiplicativa de espalhamento MSC.



O deslocamento da linha de base foi em grande parte removido e a inclinação da linha de base permanece.

Isto se deve ao fato de que a MSC utiliza a projeção dos espectros no espectro médio que possui esta mesma tendência.

PRÉ-PROCESSAMENTO DOS DADOS

Métodos aplicados às variáveis

CENTRAGEM DOS DADOS NA MÉDIA (apenas uma translação de eixos)

$$x_{ij(cm)} = x_{ij} - \bar{x}_j$$

onde $\bar{x}_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij}$ é a média da *j*-ésima coluna dos dados.

ESCALAMENTO PELA VARIÂNCIA $x_{ij(sv)} = \frac{x_{ij}}{s_j}$

onde $s_j^2 = \frac{1}{I-1}\sum_{i=1}^{I}\left(x_{ij} - \bar{x}_j\right)^2$ é o quadrado do desvio padrão da *j*-ésima ariável.

AUTOESCALAMENTO $x_{ij(as)} = \frac{x_{ij} - \bar{x}_j}{s_j}$

ESCALAMENTO PELA AMPLITUDE*: Neste pré-processamento, os dados são escalados como mostra a equação abaixo

$$x_{ij(sr)} = \frac{x_{ij} - x_{j(\min)}}{x_{j(\max)} - x_{j(\min)}} \quad \text{onde} \quad x_{j(\min)} = \min_{1\le i\le I}\left|x_{ij}\right| \qquad x_{j(\max)} = \max_{1\le i\le I}\left|x_{ij}\right|$$

[*] Escalamento pelo *Range*

NOTAS

A escolha adequada do pré-tratamento é essencial para o sucesso de qualquer análise.

♣ Quando as variáveis têm diferentes unidades ou quando a faixa de variação dos dados é grande, recomenda-se o autoescalamento das variáveis.

♣ Todos estes métodos de escalamento são sensíveis à presença de amostras anômalas, que têm um comportamento diferenciado do restante do conjunto, "*outliers*". O escalamento pela amplitude é método mais sensível, porque uma amostra com comportamento distinto aumenta a faixa de variação e pode deslocar as demais para o lado oposto a ela.

♣ Recomenda-se centrar os dados na média para a construir modelos de calibração com dados de espectroscopia.

♣ Medidas de espectroscopia óptica têm correlação significativa entre as variáveis e, portanto, não requerem escalamento por variância ou autoescalamento.

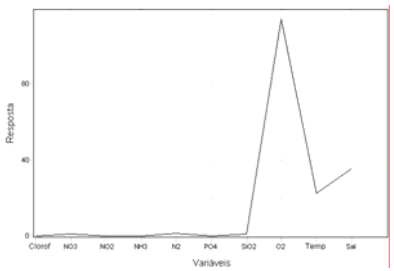
♣ Em estudos de QSAR, o autoescalamento é o procedimento universal.

EXEMPLO

A Tabela abaixo contém dados de concentrações de vários constituintes encontrados em amostras de água do mar coletadas na região de Cabo Frio no litoral norte de São Paulo durante uma expedição feita no verão de 1986, onde estão sendo consideradas também a temperatura (Temp) e a salinidade (Sal) da água.

	NO2	NH3	N2	PO4	SiO2	O2	Temp	Sal
	0,00	1,20	1,48	0,02	1,64	106,00	23,50	35,92
	0,09	0,50	0,85	0,02	0,43	110,20	22,74	35,25
	0,01	0,40	0,71	0,07	0,61	111,83	20,98	35,28
	0,00	0,10	0,29	0,12	0,68	96,48	16,63	35,43
	0,06	0,10	0,29	0,36	2,44	81,00	15,35	35,49
	0,31	1,00	2,34	0,47	5,02	76,42	15,57	35,65
	0,01	0,30	0,54	0,00	1,25	104,50	22,00	35,49
	0,07	0,50	0,65	0,04	0,43	100,68	21,99	35,32
	0,06	0,50	0,70	0,08	2,92	96,48	19,74	35,82
	0,04	0,30	0,57	0,06	1,40	108,60	17,09	35,72
	0,08	0,60	0,75	0,01	1,01	103,33	23,21	35,63
	0,00	0,30	0,30	0,01	0,18	101,00	23,16	35,68
	0,06	0,50	0,68	0,14	3,55	81,92	17,17	35,75
	0,08	0,60	13,14	0,43	2,68	85,83	14,06	35,38
Média	0,06	0,49	1,66	0,13	1,73	97,45	19,51	35,56
Desvio Padrão	0,08	0,31	3,34	0,16	1,41	11,65	3,40	0,21

Quando a faixa de variação dos dados é grande, recomenda-se o autoescalamento de cada variável por um valor. Assim, minimizamos o efeito (influência) de uma variável dominante em cálculos posteriores.



As variáveis O_2, Temp e Sal têm uma resposta média muito alta em comparação com as outras variáveis, se bem que a faixa de variação da salinidade é bem estreita.

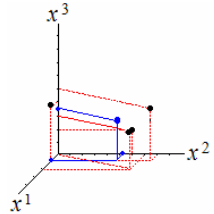
Os valores das respostas da variável N_2 são pequenos mas o seu conjunto tem um desvio padrão bastante alto (especialmente devido à concentração da última amostra), da mesma ordem de grandeza do desvio padrão da variável Temp cujo valor médio é mais do que dez vezes maior.

Estes dados devem ser autoescalados antes da análise.

Com o autoescalamento, as variáveis que tem uma faixa de variação alta serão encolhidas e aquelas com baixo desvio padrão serão alongadas, como por exemplo as variáveis 1, 2, 4 e 8.

EXEMPLO NUMÉRICO

$$\mathbf{X} = \begin{array}{c} \begin{matrix} x^1 & x^2 & x^3 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 4 \\ 2 & 7 & 3 \\ 0 & 6 & 2 \\ 1 & 8 & 5 \end{bmatrix} \end{array}$$



O ponto em azul indica o ponto médio do conjunto de dados.

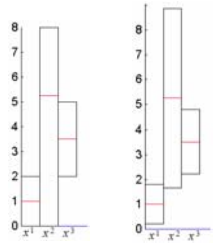
$$\bar{\mathbf{x}}^T = [1{,}00;\quad 5{,}25;\quad 3{,}50] = [\overline{x^1};\quad \overline{x^2};\quad \overline{x^3}]$$

$$\mathbf{s}^T = [0{,}82;\quad 3{,}59;\quad 1{,}29] = [s_1;\quad s_2;\quad s_3]$$

$\bar{\mathbf{x}}^T$ é um vetor linha contendo as médias das colunas de **X**

$\mathbf{s}^T$ é um vetor linha, contém o desvio padrão de cada coluna.

Estes dados podem ser representados num gráfico de barras, uma para cada variável e suas respectivas médias, ou ainda num gráfico de barras de variância.



As linhas vermelhas se referem aos valores médios de cada variável.

(a) faixa de variação de cada variável;

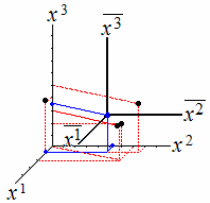
(b) faixa de variação do desvio padrão de cada variável.

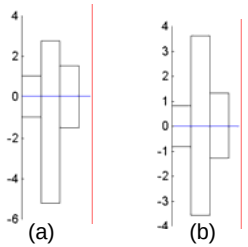
CENTRAGEM DOS DADOS NA MÉDIA

Geometricamente equivale a fazer uma translação do sistema de eixos ao longo do vetor (1,00; 5,25; 3,50), para o centro do conjunto de dados.

```
[I,J] = size(X);
xbar = mean(X);
Xcm = X – ones(I,1)*xbar;
```

$$\mathbf{X}_{cm} = \begin{bmatrix} 1 & 0 & 4 \\ 2 & 7 & 3 \\ 0 & 6 & 2 \\ 1 & 8 & 5 \end{bmatrix} - \begin{bmatrix} 1{,}0 & 5{,}25 & 3{,}5 \\ 1{,}0 & 5{,}25 & 3{,}5 \\ 1{,}0 & 5{,}25 & 3{,}5 \\ 1{,}0 & 5{,}25 & 3{,}5 \end{bmatrix} = \begin{bmatrix} 0 & -5{,}25 & 0{,}5 \\ 1{,}0 & 1{,}75 & -0{,}5 \\ -1{,}0 & 0{,}75 & -1{,}5 \\ 0 & 2{,}75 & 1{,}5 \end{bmatrix}$$





Cada uma das barras é deslocada até que os valores médios coincidam com o zero mas, mantendo intacto o tamanho de cada uma.

ESCALAMENTO PELA VARIÂNCIA

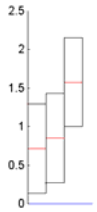
Se os dados originais são escalados pela variância, cada elemento da matriz original será dividido pelo desvio padrão da respectiva coluna,

```
s=sqrt((sum(X.^2))/(I-1));
%ou
s=std(X);
Xv=X./(ones(I,1)*s)
```

$$\bar{\mathbf{x}}_v^T = [1{,}23;\quad 1{,}46;\quad 2{,}71] \qquad \mathbf{s}_v^T = [1{,}0;\quad 1{,}0;\quad 1{,}0]$$

$$\mathbf{X}_v = \begin{bmatrix} 1 & 0 & 4 \\ 2 & 7 & 3 \\ 0 & 6 & 2 \\ 1 & 8 & 5 \end{bmatrix} \begin{bmatrix} 1/0{,}82 & 0{,}00 & 0{,}00 \\ 0{,}00 & 1/3{,}59 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 1/1{,}29 \end{bmatrix} = \begin{bmatrix} 1{,}23 & 0 & 3{,}10 \\ 2{,}45 & 1{,}95 & 2{,}32 \\ 0{,}00 & 0{,}75 & 1{,}55 \\ 1{,}23 & 2{,}75 & 3{,}87 \end{bmatrix}$$

A escala foi tomada em unidades de variância (variância unitária), tal como mostrado no vetor de desvio padrão , das colunas de $\mathbf{X}_v$.



Gráficos de barra para os dados escalados pela variância.
A faixa de variação do desvio está em escala de 1/(*I*-1) unidades de variância.

AUTOESCALAMENTO

Para autoescalar os dados, vamos centrá-los na média e escalar pela variância

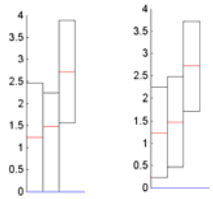
```
Xas = Xcm./(ones(I,1)*s);)
```

$$\mathbf{X}_{as} = \begin{bmatrix} 0 & -1{,}46 & 0{,}39 \\ 1{,}23 & 0{,}49 & -0{,}39 \\ -1{,}23 & 0{,}21 & -1{,}16 \\ 0 & 0{,}77 & 1{,}16 \end{bmatrix} \qquad \bar{\mathbf{x}}_{as}^T = [0{,}0;\quad 0{,}0;\quad 0{,}0] \qquad \mathbf{s}_{as}^T = [1{,}0;\quad 1{,}0;\quad 1{,}0]$$

```
Xas = Xcm./((ones(I,1)*s*sqrt(I-1));
```

$$\mathbf{s}_{as}^T = [0{,}58;\quad 0{,}58;\quad 0{,}58] \qquad \mathbf{X}_{as} = \begin{bmatrix} 0{,}0 & -0{,}84 & 0{,}22 \\ 0{,}71 & 0{,}28 & -0{,}22 \\ -0{,}71 & 0{,}12 & -0{,}67 \\ 0{,}0 & 0{,}44 & 0{,}67 \end{bmatrix}$$

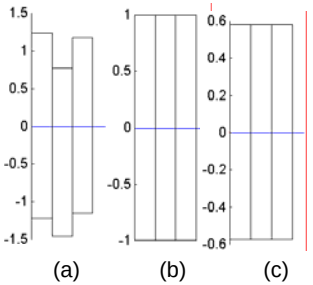
Com este pré-processamento o valor médio de cada coluna mudou e também o tamanho de cada barra.
Agora, todas as barras de desvio padrão têm o mesmo tamanho (desvio padrão $s_v = \pm 1{,}0$).



A matriz de dados também pode ser escalada para 1/(*I*-1) unidades de variância. Neste caso, $\mathbf{X}_v$ é ligeiramente diferente.

```
Xv=X./(ones(I,1)*std(X)*sqrt(I-1))
% o símbolo './' indica que os respectivos elementos
% de cada matriz serão divididos entre si.
```

$$\mathbf{X}_v = \begin{bmatrix} 0{,}71 & 0 & 1{,}79 \\ 1{,}41 & 1{,}13 & 1{,}34 \\ 0{,}00 & 0{,}96 & 0{,}89 \\ 0{,}71 & 1{,}29 & 2{,}24 \end{bmatrix} \qquad \bar{\mathbf{x}}_v^T = [0{,}71;\quad 0{,}84;\quad 1{,}57] \qquad \mathbf{s}_v^T = [0{,}58;\quad 0{,}58;\quad 0{,}58]$$



Gráficos de barra para os dados autoescalados.

(a) faixa de variação das variáveis centradas na média e escaladas para variância unitária;

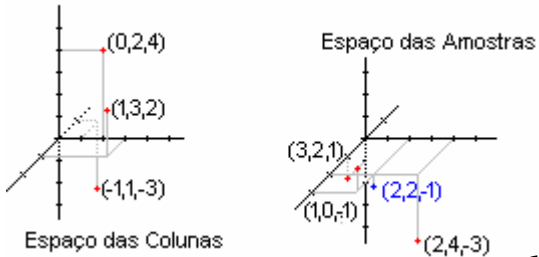
(b) faixa de variação do desvio padrão de cada variável (± 1,0);

(c) faixa de variação do desvio padrão de cada variável normalizado para 1/(*I*-1).

É possível visualizar o que acontece no espaço das linhas e no espaço das colunas quando se fazem os pré-processamentos, usando um exemplo mais simples de uma matriz **X**(3,3).

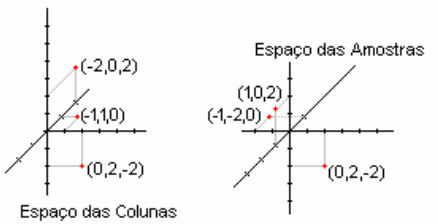
$$\mathbf{X} = \begin{bmatrix} 1 & 0 & -1 \\ 3 & 2 & 1 \\ 2 & 4 & -3 \end{bmatrix}$$

$$\bar{\mathbf{x}}^T = [2;\quad 2;\quad -1] \qquad \mathbf{s}^T = [1;\quad 2;\quad 2]$$



CENTRANDO OS DADOS NA MÉDIA:

$$\mathbf{X}_{cm} = \begin{bmatrix} -1 & -2 & 0 \\ 1 & 0 & 2 \\ 0 & 2 & -2 \end{bmatrix}$$

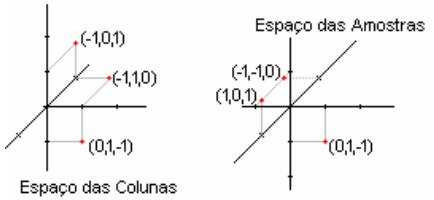


No espaço das amostras: translação da origem do sistema de eixos para o centróide. A distância entre as amostras foi conservada.

No espaço das colunas: a distância entre os pontos não foi preservada.

AUTOESCALANDO OS DADOS PARA VARIÂNCIA UNITÁRIA:

$$\mathbf{X}as = \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$



No espaço das amostras: as distâncias entre as amostras não foram preservadas.

No espaço das colunas: as distâncias de todos os pontos à origem é a mesma. Esta distância não é “1” porque os dados foram escalados para variância unitária.

Com os dados autoescalados, os pontos no espaço das colunas estão localizados numa hiperesfera centrada na origem do espaço.

UNICAMP