ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO (NIR)
MÓDULO II

MÉTODOS DE ANÁLISE E REGRESSÃO MULTIVARIADA: Teoria e Prática

Celio Pasquini

Julho 2011

QP-812

UNICAMP

GIA

Bibliografia

1) K. R. Beebe, R.J. Pell e M.B. Seasholtz, *"Chemometrics – A Pratical Guide"*, Wiley, N. York *1998.*
2) H. Martens e T. Naes, *"Multivariate Calibration"*, Wiley, N. York, 2002.
3) D. L. Massart, B.G.M. Vandeginste, S.N.Deming, Y. Michotte e L. Kaufman, *"Chemometrics: A Textboook"*, Elsevier, Amsterdam, 1988.
4) Matias Otto, *"Chemometrics – Statistics and Computer Application in Analytical Chemistry"*, 2nd ed, Wiley, N. York, 2007.
5) CAMO A/S, *"The Unscrambler User's Guide 7.5"*, Trondhein, 1998.
6) Kim H. Esbensen, Multivariate Data Analysis-in Practice, 5th ed, CAMO, 2006.
7) R. G. Brereton, *"Chemometrics: Data Analysis for the Laboratory and Chemical Plant"*, University of Bristol, UK, 2003.
8) Tormod Næs, Tomas Isaksson, Tom Fearn and Tony Davies, *A User-Friendly Guide to Multivariate Calib. and Classification,NIR Publications,* UK, 2002.
9) Standard Practices for Infrared Multivariate Quantitative Analysis, ASTM 1655-05, 2005.

Programa do curso:

Teoria

- A importância da abordagem multivariada no tratamento de dados.
- Pré-processamento de dados (técnicas e fatores determinantes).
- Análise hierárquica
- Interpretação de modelos multivariados baseados em PCA (scores e loadings).
- Regressão Multivariada: Regressão Linear Múltipla (MLR), Regressão em Componentes Principais (PCR), Regressão de Quadrados Mínimos Parciais (PLS).
- Requisitos práticos para construção de modelos de regressão multivariada.

Programa do curso (cont.):

- Validação e manutenção de modelos de regressão.
- Interpretação de modelos de regressão (loadings, loadings weigths e scores).
- Técnicas de Classificação (LDA, SIMCA e PLS-DA).
- Seleção de variáveis e amostras.
- Programas Quimiométricos (Ênfase no uso do Unscrambler)

Programa do curso (cont.):

Prática (Unscrambler)

Pré-tratamento de dados. Filtros digitais (Savitz-Golay), Derivadas, MSC, SNV.

Elaboração e interpretação de modelos PCA/SIMCA.

Elaboração, interpretação e validação de modelos PLS. Detecção de amostras anômalas (outliers). Aplicação em processo.

Elaboração de modelos PLS-DA

Seleção de amostras e de variáveis

O mundo é multivariado

- Todos os processos reais são multivariados, até prova em contrário e devem ser observados por meio de métodos multivariados
- 1850-1975: "Abordagem científica"
  - Uma variável (específica) por vez
  - Poucas medidas disponíveis
- 1975-2010: Métodos multivariados
  - Medidas de menor custo, experimentos de alto custo
  - Desenvolvimento de métodos adequados

➢ Como extrair informações de todos os dados?

QUIMIOMETRIA

Ramo da ciência cujo objetivo é o de utilizar técnicas matemáticas (principalmente estatísticas) no tratamento e interpretação de dados químicos."

Quimiometria – outras definições

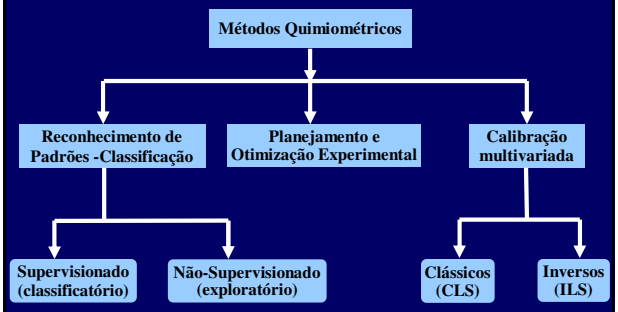
***"Disciplina da química que usa métodos matemáticos e estatísticos para planejar ou selecionar experimentos de forma otimizada e para fornecer o máximo de informação na análise de dados de natureza multivariada"* (Kowalski)*.**

***"Utilização de técnicas estatísticas e matemáticas para analisar dados químicos"* (Beeb et al).**

***"O processo completo no qual dados são transformados em informação usada para tomada de decisão"* (Beeb et al).**

***"Veículos que auxiliam os químicos a atingirem de forma mais eficiente na direção do maior conhecimento"* (Kowalski)*.**

Classificação dos Métodos Quimiométricos



O que é Quimiometria?

- 40 % Conhecimento da aplicação
- 30 % Senso comum
- 20 % Estatística
- 10 % Matemática

Gerar informação multivariada

ESPECTROSCOPIA NIR

Sinergismo

QUIMIOMETRIA

Extrair e empregar a informação

Seis Hábitos de um Quimiométrico Eficiente

"We are what we repeatedly do. Excellence, then, is not act, but a habit" (Aristóteles).

1) **Exame dos dados;**
2) **Pré-processamento (quando necessário!!!);**
3) **Estimar o modelo;**
4) **Exame dos resultados/Validação do Modelo;**
5) **Uso do modelo para previsão;**
6) **Validação da previsão.**

Hábito 1 - Exame dos Dados

Observar os dados para identificar através de gráficos e tabelas erros óbvios, características ou ocorrências que se destacam.

Hábito 2 - Pré-Processamento

Pode haver fontes de variação sistemática ou aleatória que mascarem uma variação de interesse e, assim, uma técnica de pré-processamento deve ser empregada.

Observação

1) Um pré-processamento inadequado pode remover informação útil e, por isso, um re-exame deve sempre ser realizado após o pré-processamento.

Hábito 3 - Estimar o Modelo

Determinar, calcular, estimar ou gerar os parâmetros de um modelo a partir dos dados.

Hábito 4 - Exame dos Resultados/Validação do Modelo

Ferramentas de diagnósticos (gráficos, tabelas) e conhecimento dos fenômenos químicos ou físicos do sistema são empregados nos exame dos resultados e validação dos modelos (modelo é aceitável ou não?).

Observação

1) As ferramentas de diagnósticos e de validação do modelo variam dependendo do pacote de software quimiométrico utilizado.

Hábito 5 - Uso do Modelo para Previsão

Aplicação do modelo para previsão de parâmetros em amostra desconhecidas, se o método quimiométrico gera modelos para previsão.

Hábito 6 - Validação da Previsão

Com ferramentas de diagnósticos de previsão é possível determinar se o modelo não é aplicável devido à falhas instrumentais ou à presença de amostras anômalas (outliers).

Dados em Quimiometria

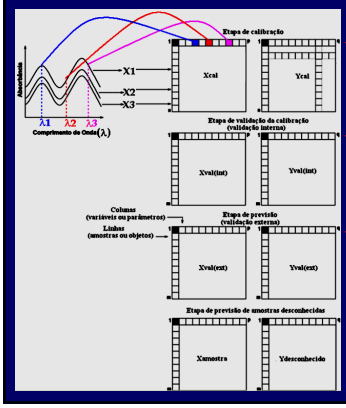
São arranjados em forma de matrizes de modo que:

Amostras → são dispostas em linhas e podem ser misturas de calibração, de testes e amostras em análise.

Variáveis → são dispostos em colunas e podem ser comprimentos de ondas, concentração de analitos, pH, pressão, etc.

Parâmetros → são propriedades derivadas da relação objetos/variáveis (valores de absorbância, pH, concentração, etc). Representam os valores dos elementos das matrizes.

Construindo um Modelo Robusto de Previsão



← Calibração
- As matrizes X_{cal} e Y_{cal} são usadas na construção dos modelos de calibração.

← Validação da Calibração (validação interna)
- As matrizes $X_{val(int)}$ e $Y_{val(int)}$ são usadas na validação dos modelos construídos na calibração.

← Validação da Previsão (validação externa)
- As matrizes $X_{val(ext)}$ e $Y_{val(ext)}$ são usadas para avaliar a robustez do modelo validado para a previsão.

← Etapa de Previsão
- A matriz $X_{amostra}$ é usada na previsão dos valores $Y_{desconhecido}$, empregando o modelo robusto de previsão.

Pré-procesamento dos dados

É definido como qualquer manipulação nos dados antes do tratamento quimiométrico.

Finalidade

Reduzir fontes de variação não informativas e tratar as matrizes de dados para tornar os cálculos melhor condicionados antes da modelagem.

Requisitos para um Bom Pré-Processamento

1) Interação entre pré-processamento e exames dos dados;

2) Conhecimento das características dos dados

Exemplos: precisão das medidas, resolução espectral

Pré-Processamento

- Das amostras → opera em uma amostra por vez e sob todas as variáveis
- Das variáveis → opera em uma variável por vez e sob todas as amostras

Tipos de Pré-Processamento das Amostras

- Normalização → remove variações sistemáticas das amostras.
- Ponderação (weighting) → enfatiza uma amostra sobre outra.
- Suaviazação (smoothing) → remove ruído aleatório
- Correção da linha de base → corrige variação sistemática da linha base

Observação

1) Alisamento ou suavização (smoothing) de sinais ruidosos é usado principalmente para remover variações aleatórias;

2) Normalização, Ponderação e Correção de Linha de Base são usados para remover variações sistemáticas.

Normalização

Norma – função que associa cada vetor de um espaço vetorial a um número real.

É efetuada dividindo cada variável por uma constante. Três constantes podem ser usadas:

1) **Normalização por unidade de área;**
2) **Normalização por unidade de comprimento;**
3) **Normalização fazendo a máxima intensidade igual a 1**

Normalização por Unidade de Área

Divide cada elemento do vetor amostra pela sua norma-1 dada por:

$$norma-1 = \sum_{j=1}^{n\,\mathrm{var}\,s} |x_j|$$

Normalização por unidade de comprimento

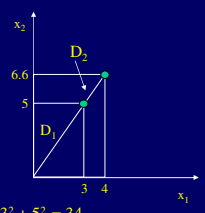
Divide cada elemento do vetor amostra pela norma-2 dada por:

$$norma-2 = \sqrt{\sum_{j=1}^{n\,\mathrm{var}\,s} |x_j^2|}$$

Normalização fazendo a máxima intensidade igual a 1

Divide cada elemento do vetor pela norma infinita, definida como o máximo valor (absoluto) do vetor.

Exemplo de normalização pela distância (comprimento do vetor)



Medida 1 = (3, 5)
Medida 2 = (4, 6.6)

Normalização pela distância

Medida 1 = (0.514, 0.858)
Medida 2 = (0.518, 0.859)

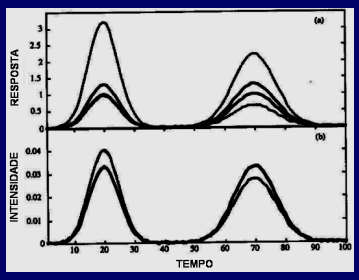
$D_1^2 = 3^2 + 5^2 = 34$
$D_1 = 5.83$

$D_2^2 = 4^2 + 6.6^2 = 59.6$
$D_2 = 7.72$

Medida 1 = (3/5.83, 5/5.83)
Medida 2 = (4/7.72, 6.6/7.72)

Variáveis que por algum motivo indesejado alteram seu valor proporcionalmente são corrigidas

Exemplo de normalização por unidade de área



Cromatogramas de 2 componentes com mesma concentração relativa mas com variações do volume injetado. Antes (a) e após (b) a normalização por unidade de área.

Observações

1) Usada neste exemplo de cromatografia para remover variação do volume injetado.
2) Os cromatogramas da mesma amostra aparecem sobrepostos após a normalização.

Ponderação

Atribui-se as amostras consideradas mais importantes pesos proporcionais a sua importância para o modelo, multiplicando cada elemento do vetor amostra pelo seu peso.

Exemplo: - dar um peso maior aos dados obtidos por um analista mais experiente.

Observação

Usando ponderação, a influência que uma amostra tem no modelo matemático pode ser manipulada

Suavização do ruído aleatório (smoothing)

Usada para reduzir matematicamente o ruído aleatório e aumentar a relação sinal/ruído, pois os sinais instrumentais são sempre compostos pelo sinal verdadeiro e ruído aleatório.

Observação

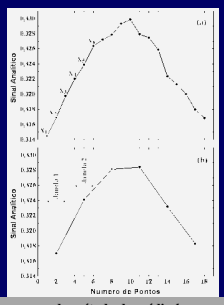
Deve-se ter muito cuidado pois a suavização pode remover informação útil.

Tipos de suavização (smoothing)

1. **por média (Boxcar);**
2. **por média móvel**
3. **por mediana (mediana) móvel**
4. **por polinômio móvel**
5. **por transformada de Fourier**
6. **por transformada Wavelet**

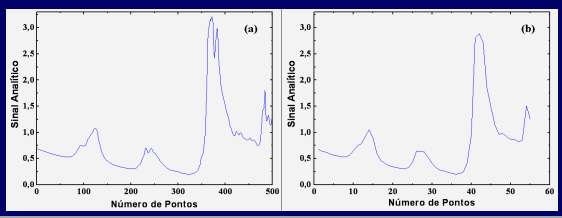
Suavização por Média (*Boxcar*)

Sua implementação consiste em se dividir o sinal analítico em uma série de janelas de comprimento n e efetuar a média em cada janela.



Exemplo ilustrativo do emprego do método da média *boxcar* na suavização de ruídos. Sinal original (a) e suavizado com uma janela n = 3 (b).

Exemplo do Emprego do método Boxcar a Dados Reais



Exemplo do emprego do método da média boxcar na suavização de ruídos do espectro NIR de um óleo vegetal. Espectro original (a) e suavizado pelo método da média *boxcar* com uma janela n = 9 (b).

Observação

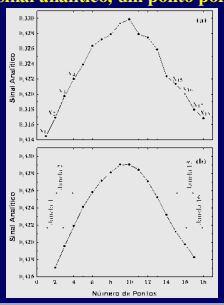
1) no método *boxcar*, o nº de pontos do sinal original é reduzido por um fator de "n";

2) Esta redução de pontos pode implicar, em alguns casos, em perda de informação analítica importante (por exemplo, picos em espectros), como ilustrado na Figura;

3) Este problema pode ser minimizado usando o método da média móvel.

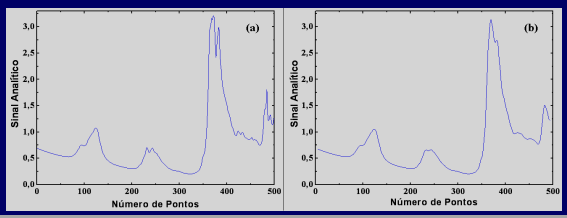
Suavização por Média Móvel

Consiste em se calcular as médias de janelas de comprimento "n" que se movem ao longo do sinal analítico, um ponto por vez.



Exemplo ilustrativo do emprego do método da média móvel na suavização de ruídos. Sinal original (a) e suavizado pelo método da média móvel com uma janela n = 3 (b).

Exemplo do Emprego do método da média móvel a Dados Reais



Exemplo do emprego do método da média móvel na suavização de ruídos do espectro NIR de um óleo vegetal. Espectro original (a) e suavizado pelo método da média *boxcar* com uma janela n = 9 (b).

Observação

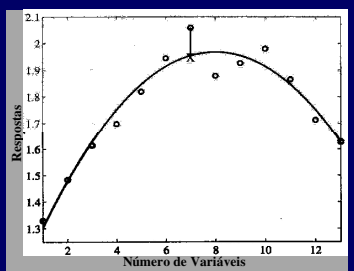
1) No método *boxcar* move-se uma janela inteira por vez, enquanto no método da média móvel move-se a janela de um ponto por vez;

2) O 1º e o último ponto não são suavizados, se fossem a janela ultrapassaria os limites do sinal original, conduzindo ao problema denominado de "efeito de borda";

3) Este problema pode ser minimizados usando o método Savitzky-Golay-Gorry.

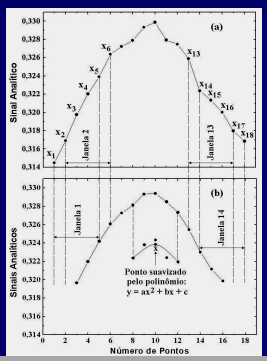
Suavização por Polinômio Móvel (Método de Savitzky-Golay)

Difere da suavização por média ou mediana móvel por usar um polinômio de baixa ordem a ser ajustado, por mínimos quadrados, aos pontos da janela.



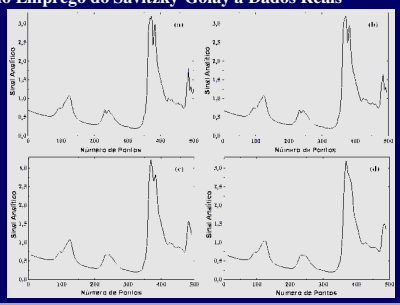
Exemplo da aplicação do método de Savitzky-Golay usando uma janela com 13 pontos. X é o valor suavizado do ponto 7 do conjunto de dados.

Exemplo ilustrativo do método de Savitzky-Golay.



Sinal original (a) e suavizado pelo método Savitzky-Golay com uma janela n = 5 (b).

Exemplo do Emprego do Savitzky-Golay a Dados Reais



Exemplo do emprego do método de Savitzky-Golay em um espectro NIR de um óleo vegetal. Espectro original (a) e suavizado pelo método de Savitzky-Golay com uma janela n = 9 (b), n = 15 (c) e n = 21 (d).

Observação

1) O problema"efeito de borda" é minimizado usando o método Savitzky-Golay-Gorry.

Suavização pelo Método de Savitzky-Golay-Gorry

Usando o método de Savitzky-Golay[1] resulta na eliminação de (tamanho da janela - 1)/2 pontos em cada extremidade do vetor (efeito de borda).

Gorry desenvolveu um método que não elimina pontos, preservando o número original de variáveis.

Observação

1) O método de Gorry produz, as vezes, perfis aberrantes nas extremidades do vetor

1) (a) A. Savitzky e M.J.E. Golay, *Anal. Chem*., 36, 1964,1627-1639

2) P.A. Gorry, *Anal. Chem*., 62, 1990, 570-573 e 63, 1991, 534-536.

3) K. Y. Hui and M. Gratzl, *Anal. Chem*.,68, 1996, 1054-1057.

Aspectos Críticos da suavização pelo Método de Savitzky-Golay

1) a escolha da largura ideal da janela;
2) a escolha do polinômio ideal;

Escolha do polinômio ideal

O polinômio ideal depende da natureza dos dados. Usa-se tipicamente polinômios de 2ª ou 3ª ordem

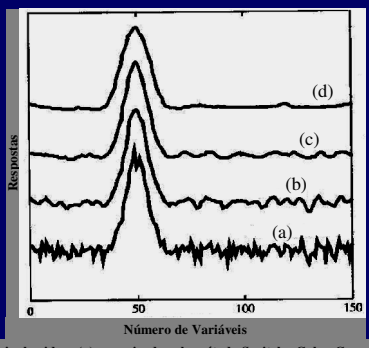
Escolha do tamanho ideal da janela

Com o aumento da janela, o ruído tende a ser removido, mas se a janela é muito grande, picos são removidos e os remanescentes são distorcidos;

Observação

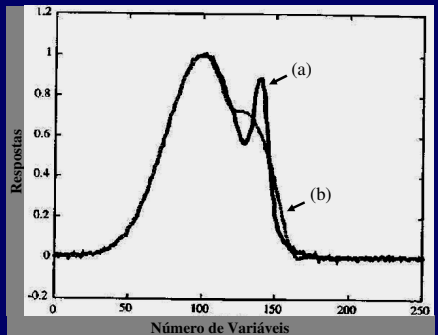
Uma janela 10 vezes maior que a largura de um pico, irá normalmente distorcê-lo ou eliminá-lo.

Efeito da escolha do tamanho da janela-Método de Savitzky-Golay



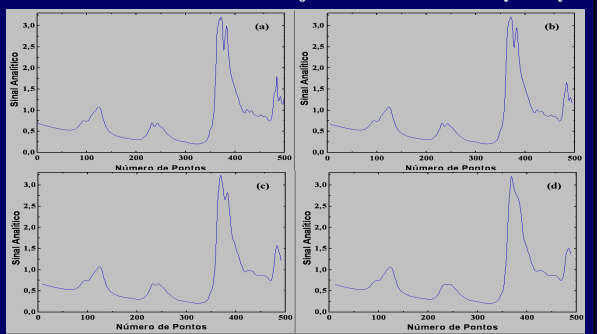
Sinal original ruidoso (a) e suavizado pelo método Savitzky-Golay-Gorry usando um polinômio de 2ª ordem e uma janela de 7 (b), 13 (c) e 21 (d) pontos.

Efeito da escolha do tamanho da janela-Método de Savitzky-Golay



Sinal original ruidoso (a) e suavizado pelo método Savitzky-Golay-Gorry usando um polinômio de 2ª ordem e uma janela de 49 (b) pontos.

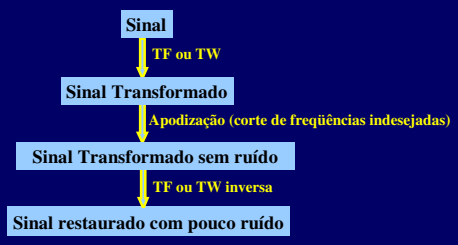
Efeito da escolha do tamanho da janela-Método de Savitzky-Golay



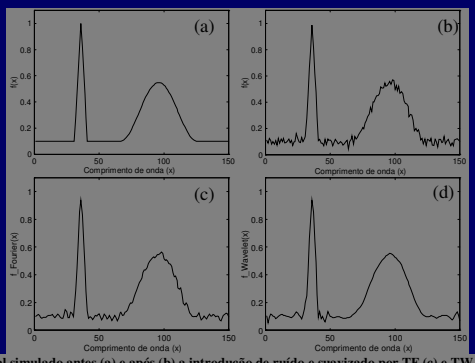
Espectro original (a) e suavizado pelo método de Savitzky-Golay com uma janela n = 9 (b), n = 15 (c) e n = 21 (d).

Suavização usando transformada de Fourier ou Wavelet

As técnicas de suavização discutidas até aqui atuam diretamente sobre o sinal e não sobre as freqüências que compõe o sinal, tal qual as técnicas que usam transformada de Fourier (TF) ou Wavelet (TW).

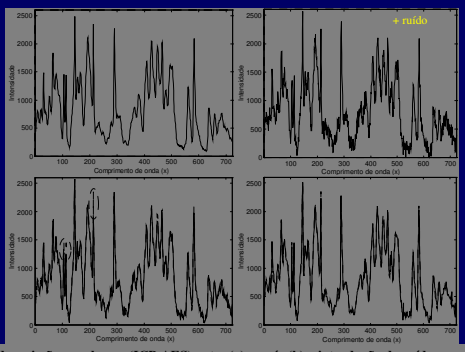


Exemplo Simulado da Suavização com TF e TW



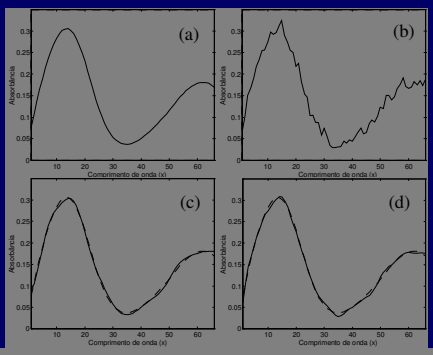
Sinal simulado antes (a) e após (b) a introdução de ruído e suavizado por TF (c) e TW (d).

Exemplo Simulado da Suavização com TF e TW



Espectro de emissão em plasma (ICP-AES) antes (a) e após (b) a introdução de ruído e suavizado (linha cheia) por TF (c) e TW (d) com espectro original sobreposto (linha tracejada).

Exemplo Simulado da Suavização com TF e TW



Espectro UV-VIS antes (a) e após (b) a introdução de ruído e suavizado (linha cheia) por TF (c) e TW (d) com espectro UV-VIS original sobreposto (linha tracejada).

Correções da linha de base

Além de ruído aleatório, as medidas podem conter variações sistemáticas não relacionadas com a investigação química. Denominadas de características da linha de base (baseline features), elas podem dominar a análise se não forem removidas.

Tipos de correções da linha de base

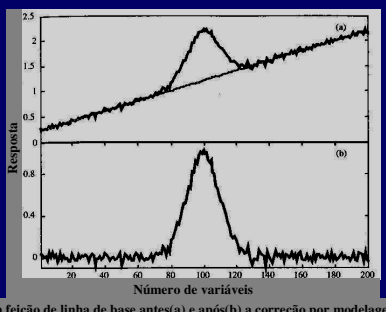
1. Por modelagem explícita
2. Por derivação
3. Por MSC
4. Por SNV

Correções da linha de base por modelagem explicita

A resposta verdadeira é obtida subtraindo os dados de uma função de linha base das respostas medidas.

$$R_{medido} = R_{verd} + \text{função de linha base } (= \beta X + \delta x^2 + \gamma x^3 + \ldots)$$

Correção da linha de base por modelagem explicita



Dados com feição de linha de base antes(a) e após(b) a correção por modelagem explicita.

Observação:

A aplicação da correção da linha de base por modelagem explicita removeu a função linear da linha base

Correções da linha de base por derivação

Este método é muito útil em casos onde a linha de base é difícil de ser identificada.

Para ver como os métodos derivativos são capazes de remover efeitos sobre a linha de base, considere a derivada da equação abaixo.

- vetor resposta: $Rmed = Rverd + (\alpha + \beta X + \delta x^2 + \xi x^3 + ...)$
- 1ª Derivada: $R'med = R'verd + 0 + \beta + 2\delta x + 3\xi x^2 + ...)$
- 2ª Derivada: $R''med = R''verd + 0 + 0 + 2\delta + 6\xi x + ...)$

Observação:

1) a 1ª derivada removerá efeitos relacionados com o offset (α).
2) a 2ª derivada removerá efeitos lineares (α + βX) e assim por diante.

Métodos de derivação

1) por simples diferença móvel;
2) por diferença de média móvel;
3) pelo Método de Savitzki-Golay.

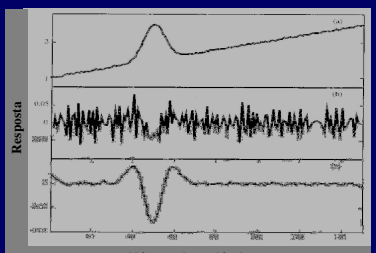
Derivação pelo método de Savitzky-Golay

- Na suavização ajusta-se um polinômio a uma janela móvel do sinal resposta e o ponto central da janela é estimado pelo polinômio.
- Na derivação por esse método é usado o valor estimado pela derivada do polinômio da janela.

Observação:

1) é matematicamente simples calcular a derivada de um polinômio.

Aplicação dos métodos de derivação em curva com função linear e ruído

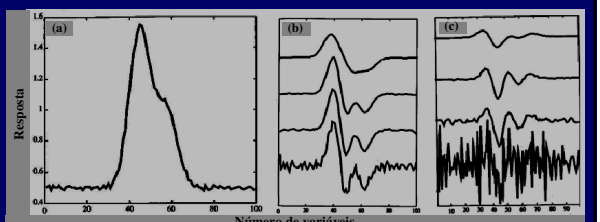


Vetor resposta com ruído e feição linear na linha de base (a) sob o qual foi aplicado a 2ª derivada por simples diferença (b) e pelo método de Savtzki-Golay com uma janela de 11 (d).

Observação:

1) Em derivadas de mais alta ordem (2ª derivada), o método Savtzki-Golay (embora com efeito de borda) torna-se mais importante se comparado ao método da derivação por simples diferença.

Influência da escolha da janela na derivação Savitzki-Golay



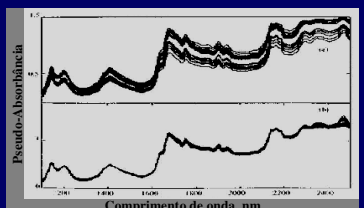
Espectro (a) usado no cálculo da 1ª (b) e 2ª (c) derivada pelo método de Savtzki-Golay-Gorry com a largura da janela variável.

Observação

1) Janelas muito pequenas resultam em derivadas com pobre relação sinal/ruído, janelas muito largas levam à perda de características úteis (informação).
2) Na janela = 3 a relação sinal/ruído é pior do que nos dados originais, principalmente no cálculo da 2ª derivada, mas na janela = 21 perde-se informação útil (pico).

Correção de sinais de espalhamento multiplicativo - MSC

Multiplicative Scatter(ou Signal) Correction (MSC) é uma ferramenta desenvolvida para corrigir espalhamento de luz em espectroscopia NIR de reflectância difusa (Martens and Naes).



Espectros de reflectância difusa NIR de polímeros sólidos antes (a) e após (b) pré-processamento com MSC.

Observação

1) o espectro pré-processado assemelha-se ao original, o que ajuda na interpretação, mas ele não corrige alterações de linha base;
2) performance comparável aos métodos derivativos.

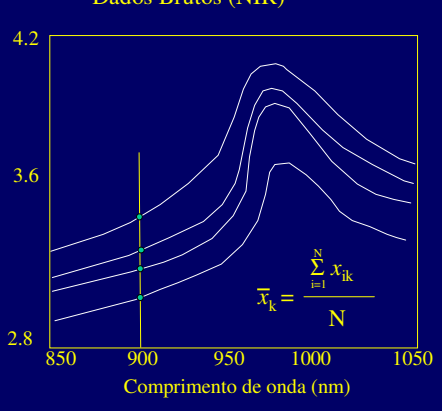
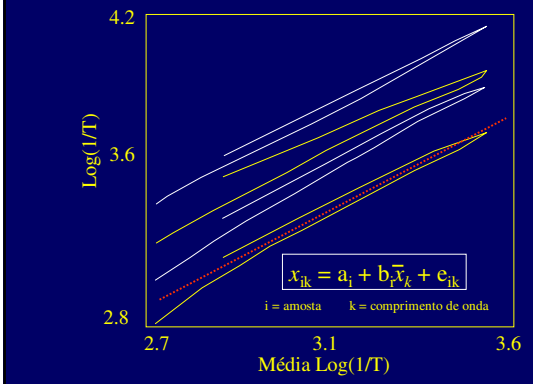
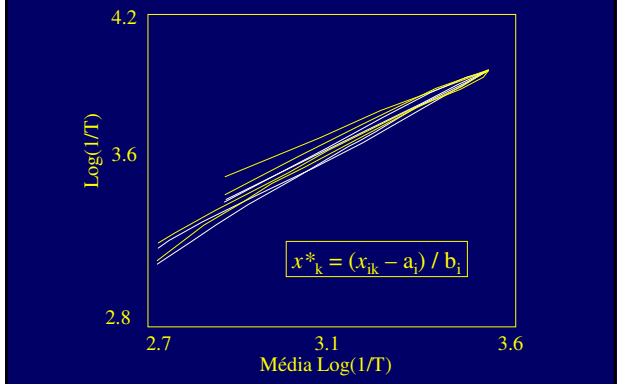
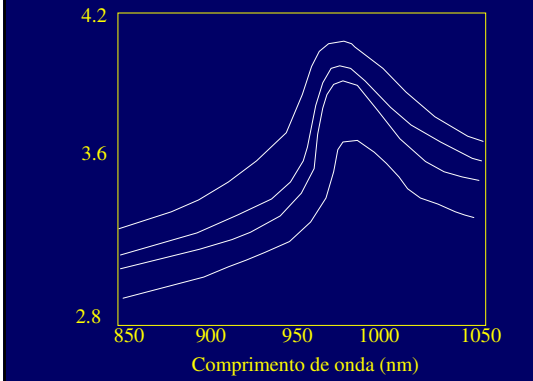


Gráfico: valores individuais x valores médios das variáveis

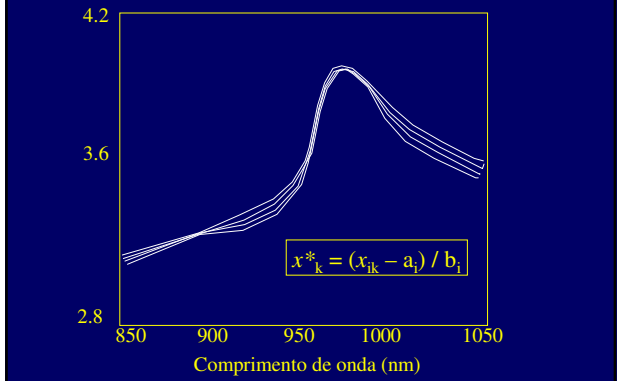




Dados Brutos (NIR)

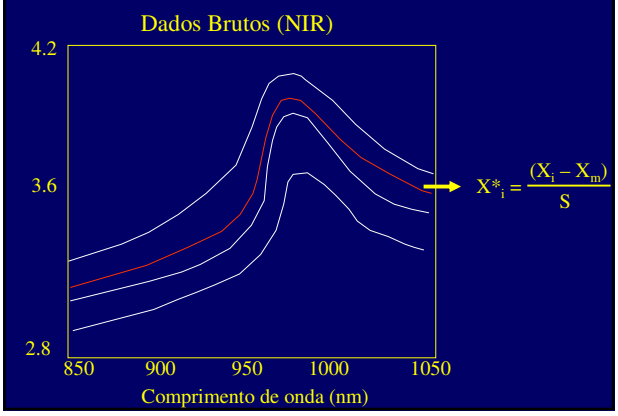


Dados Corrigidos por MSC (NIR)

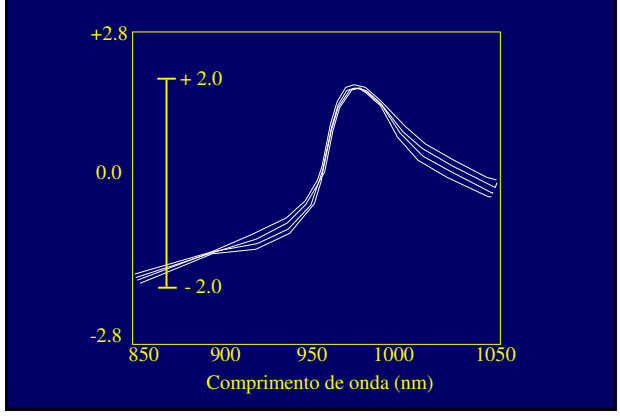


SNV – Standard Normal Variate

Dados Brutos (NIR)



Dados Corrigidos por SNV (NIR)



Técnicas de pré-processamento nas variáveis

1) Centrar na Média

2) Ponderação (weighting):
   1) Ponderação por Informação a Priori
   2) Escalonamento da Variança;
   3) Autoescalonamento
   4) Seleção de Variáveis

Observação

1) Ao se utilizar uma técnica de pré-processamento, um rigoroso re-exame dos dados deve sempre ser realizado para evitar a remoção de informações relevantes.

Centrar na média

Calcula-se o valor médio para cada variável e subtrai este valor para cada elemento de uma variável.

1) cada variável passa a ter média zero;

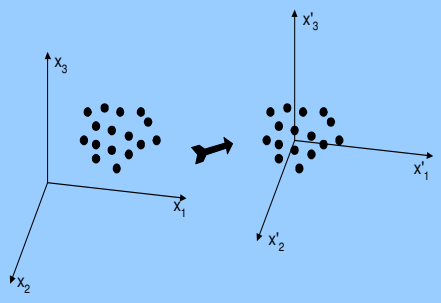
2) move-se as coordenadas para o centro dos dados.

3) diferenças nas intensidades relativas das variáveis são mais fáceis de perceber.

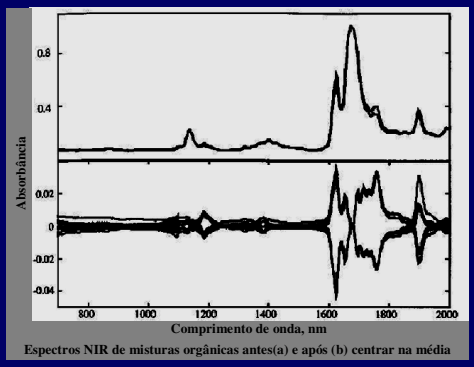
4) recomenda-se sempre centrar na média ao se usar métodos baseados componentes principais.

5) centrar na média geralmente não prejudica a análise e sempre ajuda e por isso é sempre usada como "default" nos pacotes de softwares.

Centralização dos dados



Exemplo de espectros centrados na média



Espectros NIR de misturas orgânicas antes(a) e após (b) centrar na média

Ponderação (weight)

Atribui-se às variáveis consideradas mais importantes pesos proporcionais a sua importância para o modelo, multiplicando cada elemento do vetor variável pelo seu peso.

Técnicas de ponderação

Quatro tipos de pré-processamento das variáveis podem ser considerados como casos particular de ponderação:

1) Ponderação por Informação a Priori;
2) Escalonamento da Variância;
3) Auto-escalonamento;
4) Seleção de Variáveis.

Observação

1) A seleção de variáveis e a ponderação por informação a priori enfatizam umas variáveis sobre as outras;
2) O escalonamento da variância e o auto-escalonamento colocam todas variáveis em pé de igualdade.

Ponderação por informação a priori

Informações teóricas ou experiências anteriores podem fazer com que o analista dê pesos maiores as variáveis mais úteis e pesos menores aquelas com baixa relação sinal/ruído.

Sabendo-se, a priori, que certas variáveis são mais importantes que outras para a modelagem, atribui-se as variáveis mais importantes pesos proporcionais a sua importância.

Se não há informação prévia sobre quais variáveis são mais importantes em relação a outras, recomenda-se usar o escalamento da variância.

Escalamento da variância

Divide-se cada elemento de uma variável pelo desvio padrão global desta variável. Isto faz com que:

1) a variância de cada variável torne-se unitária (= 1);

2) as variáveis passem a ser expressas em unidades de desvios padrão;

3) as influências relativas das diferentes variáveis nos cálculos tornem-se independentes das suas unidades;

4) as variáveis com maior magnitude tenham uma menor influência na modelagem;

Exemplo: variáveis em unidades de kilograma passam a ter a mesma influência que as variáveis em grama.

Auto-escalamento

Na prática, o escalamento pela variância é sempre realizada em conjunto com a centralização na média. Estes dois processos realizados conjuntamente em qualquer ordem é chamado de auto-escalamento.

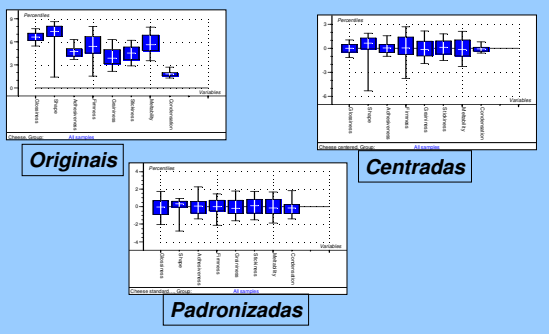
1) interpretação de gráficos de dados auto-escalados é difícil porque as unidades foram removidas.

2) graficamente parece que o auto-escalamento remove informações porque padrões característicos comuns (de espectros, por exemplo) desaparecem.

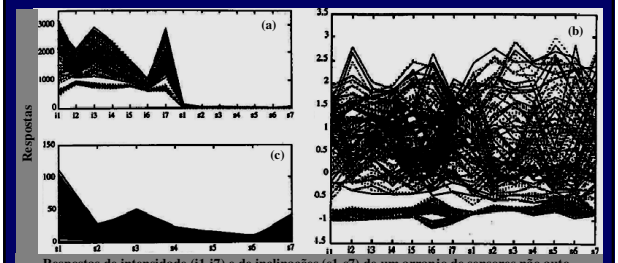
3) em geral, o auto-escalamento melhora os resultados das análises multivariadas. Principalmente quando variáveis de natureza distinta estão sendo avaliadas.

Escalamento das Variáveis

Padronização: Leva todas as variáveis para a mesma escala.



Exemplo da aplicação do auto-escalamento



Respostas de intensidade (i1-i7) e de inclinações (s1-s7) de um arranjo de sensores não auto-escalonado (a) auto-escalonado (b) e zoom das respostas de inclinação (c).

Seleção de variáveis

É um caso extremo de ponderação onde é atribuído peso zero as variáveis não informativas.

1) reduz o ruído;

2) minimiza a propagação de erros;

3) melhora a previsão.

Seleção de variáveis por informação a priori

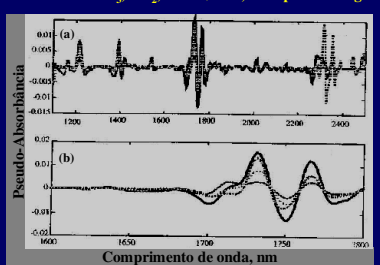
Conhecimento físico-químico do sistema em investigação permite que o analista selecione faixas de variáveis úteis para modelagem, eliminando as regiões não informativas.

Exemplo de seleção de variáveis por informação a priori

- Seleção de regiões de espectros NIR de reflectância difusa para:

1) reduzir variações no caminho óptico pela eliminação da região 2250 a 2500nm;

2) melhorar a seleção dos tipos de plásticos pela escolha da região 1600-1800nm que inclui os 1os sobretons de CH_3, CH_2, CH e ø-CH, úteis para distinguir plásticos.



Espectros NIR de reflectância antes (a) e após(b) a seleção de variáveis por informações a priori.

Seleção de variáveis por técnicas de otimização

Métodos matemáticos/estatísticos (algoritmos de otimização) são desenvolvidas e utilizadas na seleção de variáveis que portam informações úteis e não são correlacionadas, descartando as variáveis não relevantes, redundantes ou não informativas para a modelagem.

Otimização

Determinar a combinação de fatores (variáveis) que maximiza ou minimiza um dado índice (função objetivo)

1) Cada combinação de fatores: Solução

2) Melhor combinação: Solução Ótima

Exemplo de Otimização Multivariada em Processo Industrial

Variáveis → temperatura, pressão, pH, catalisador, etc;
Objetivo → selecionar a melhor combinação destas variáveis;
Função-objetivo → maximizar rendimento ou minimizar tempo, etc.

Exemplo de otimização em seleção de variáveis espectrométricas

Variáveis: comprimentos de onda empregados;
Objetivo: selecionar variáveis que maximize capacidade preditiva;
Função-Objetivo: minimizar (*Root Mean Square error of prediction - RMSEP*);

Técnicas de otimização aplicáveis à seleção de variáveis

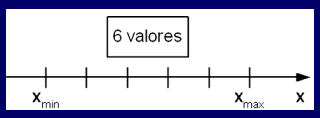
1) Busca Exaustiva;
2) Cálculo Matemático;
3) Técnicas de gradiente: steepest ascendent (hill-climbing) ou descent;
4) *Simulated Annealing* ("Têmpera Simulada");
5) *Simplex* ("Poliedros Flexíveis");
6) Algoritmo da Projeções Sucessivas (APS);
7) Algoritmo Genético (AG), etc;

Técnica da busca exaustiva univariável

Calcula-se $f(x)$ para todos os valores de x entre x_{min} e x_{max}.

Variável "x" inteira

Seis variáveis inteiras implicam no cálculo de 6 valores para f(x).



Variável x real

Faz-se uma discretização dos valores entre x_{min} e x_{max}.

discretização mais fina → - maior precisão, - maior nº de cálculos.

Busca exaustiva multivariada

$x = (x_1, x_2, ..., x_N)$

Cada variável → M possíveis valores

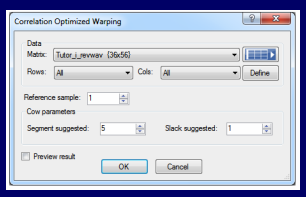
M^N cálculos para f(x)

10 variáveis, 5 valores para cada variável

$5^{10} \approx 10$ milhões de valores de f(x)

Explosão em Busca Exaustiva

Alinhamento (COW – Correlation Optimised Warping)



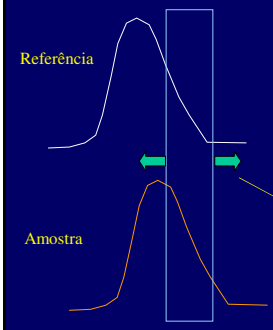
No. mínimo de variáveis = 20

1. Eleger uma amostra típica de referência (com todos os picos) Guardá-la em separado

2. Definir tamanho do segmento (< N / 4)

3. Definir "slack" = shift máximo a ser investigado (≤ segmento)

Alinhamento (COW – Correlation Optimised Warping)

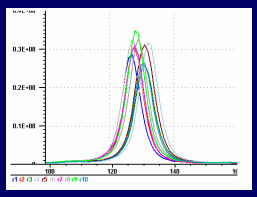


Otimizar a correlação entre a amostra e a referência.

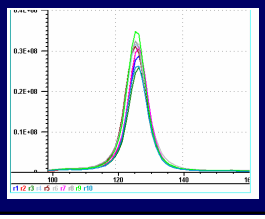
Qual "shift" faz isso?

Slack (alteração permitida na posição para ser encontrada)

Alinhamento (COW – Correlation Optimised Warping)



Cromatogramas originais



Cromatogramas após pré-tratamento com COW
Segmento = 100
Slack = 20

PRÉ-PROCESSAMENTO DE DADOS

- Diversos tipos de pré-processamentos de dados podem ser aplicados aos dados originais antes do início de uma análise multivariada propriamente dita ou do desenvolvimento de um modelo multivariado de calibração.

- Qualquer tipo de pré-processamento pode ser empregado se estes levarem a produção de um modelo que ofereça uma melhor precisão e que atenda as especificações de validação.

Exemplos:

Derivadas (1ª., 2ª.), Centrar na média, Filtros digitais MSC, SNV, OSC, etc.

Pré-tratamentos disponíveis no UNSCRAMBLER 9.8

Transformação

- Média
- Suavização
- Normalização
- MSC
- Derivada
- SNV
- Correção de linha base
- Ruído
- Espectroscópica

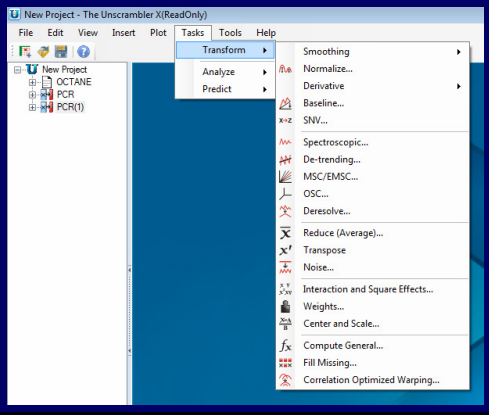
(Absorbância para Reflectância, Reflectância para Absorbância, Reflectância para Kubelka-Munck)

- UDT (Transformação definida pelo usuário)
  - Pacote de acessórios para Espectroscopia
  - Transformações programadas em Matlab, C++,…

Computação

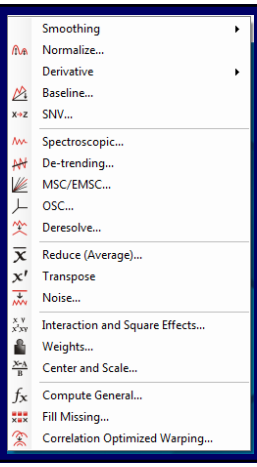
- Operações numéricas ou matriciais
- Cálculo de funções (Raíz quadrada, logaritmo, inversa, …)

Pre-tratamentos disponíneis no Unscrambler X

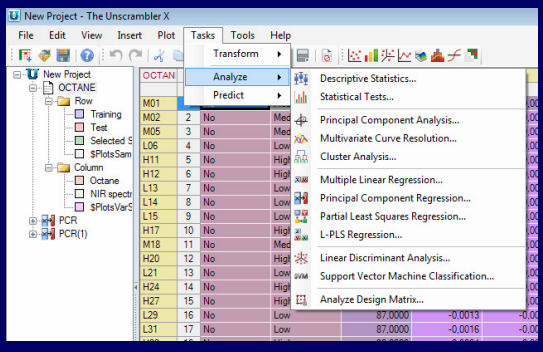


Pré-tratamentos

Unscrambler X



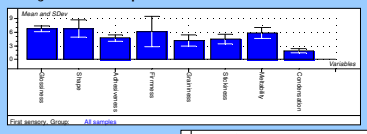
Análise de Dados – Unscrambler X



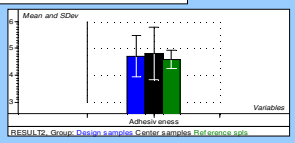
Média e Desvio Padrão

- Média = valor médio das respostas
- DP (*Desvio Padrão*) = dispersão ao redor da média

➨ Comparação de respostas:



➨ Comparação de grupos de amostras:

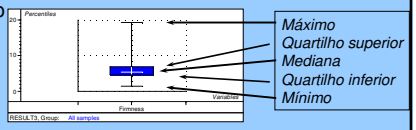


Percentilhos

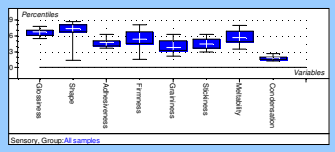
- Gráfico: divide os valores de resposta em 4 fatias iguais

➨ Faixa de Variação

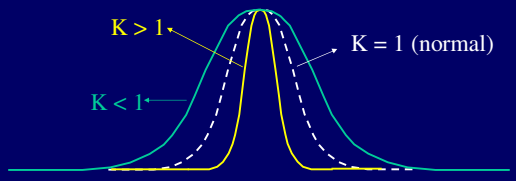
➨ Simetria



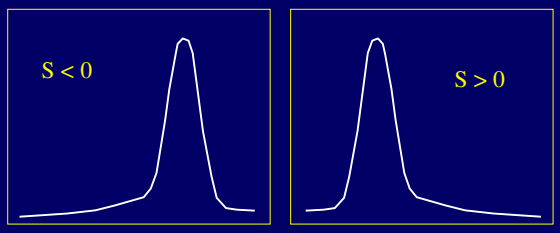
➨ Comparação das distribuições de várias respostas



Avaliação da Distribuição Normal dos Dados (Kurtosis)



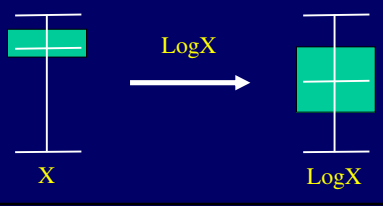
A medida da assimetria da distribuição dos valores das variáveis pode ser observada também pelo "skewness" (S)



S = 0 ⟹ Distribuição Simétrica

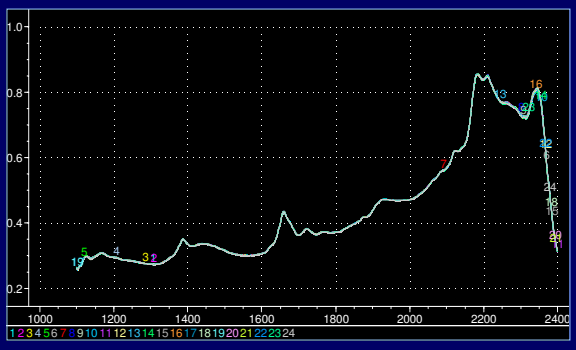
Quando uma distribuição muito assimétrica de uma variável for detectada, deve-se tentar corrigir essa distribuição transformando-se essa variável individualmente empregando-se, por exemplo, o Log da variável.

A nova distribuição deve ser verificada com um gráfico de percentilhos (quartilhos).

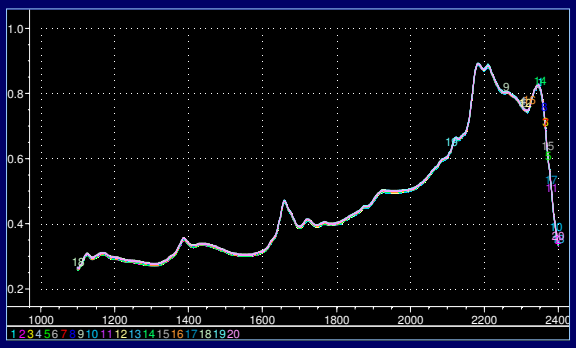


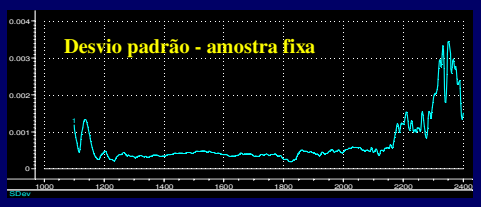
Exemplo Prático de Aplicação Correta de um Pré-Tratamento

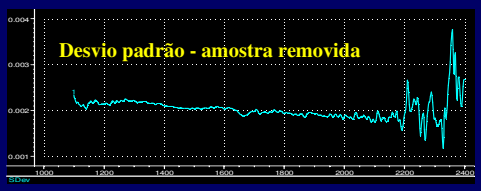
24 espectros de refletância de Aspirina 500 mg
(mesmo comprimido, sem retirar do suporte)

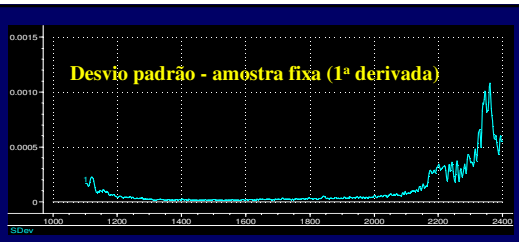


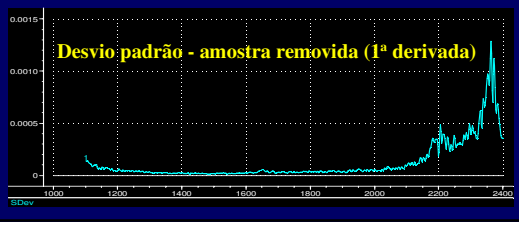
20 espectros de refletância de Aspirina 500 mg
(mesmo comprimido, retirando do suporte)





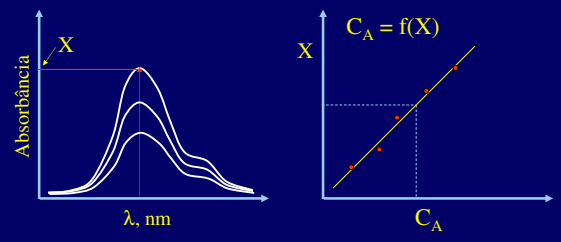




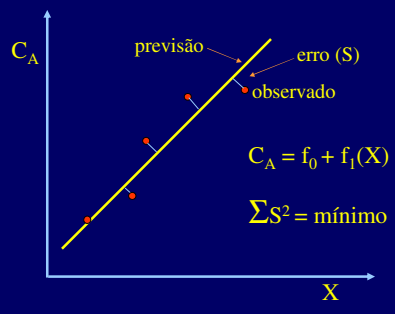


Técnicas Multivariadas

Técnicas Univariadas



Regressão Linear (mínimos quadrados)



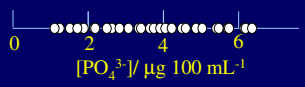
Técnicas Multivariadas

Empregam a interpretação e avaliação de toda a informação produzida e armazenada (nos computadores) pelas técnicas analíticas modernas.

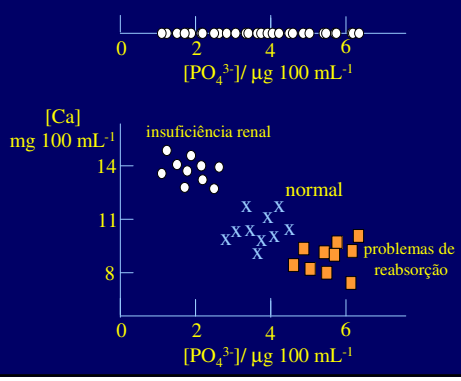
Objetivos:

1. Agrupar amostras de acordo com a suas características físicas e/ou químicas similares.
2. Classificar amostras com base em classes conhecidas.
3. Calibração de um único constituinte através do uso de todo o espectro.

Poder de discriminação de uma variável (teor de fosfato) em relação aos diferentes grupos de pacientes: normais, com insuficiência renal e com problemas de reabsorção

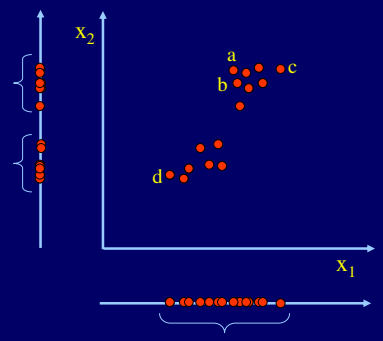


Poder de discriminação de duas variáveis

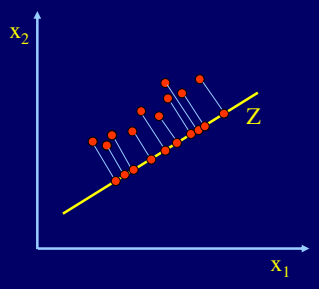


Distribuição projetada em uma dimensão

x_1 e x_2 = variáveis

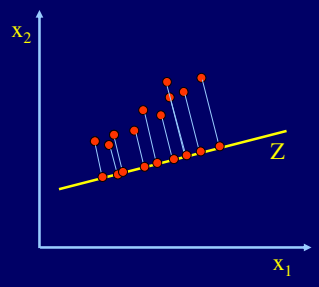


$Z = c_1x_1 + c_2x_2$



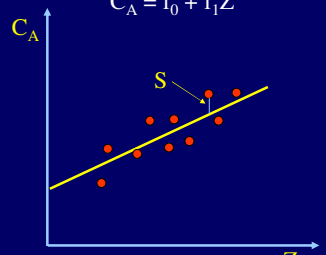
Projeção de duas variáveis em uma

$Z = c_1x_1 + c_2x_2$



Projeção de duas variáveis em uma

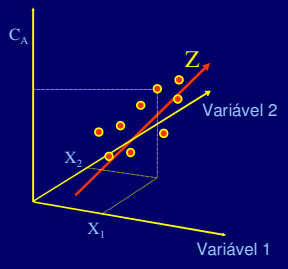
$C_A = f_0 + f_1Z$



$Z = c_1x_1 + c_2x_2$ $\qquad$ $C_A = f_0 + f_1Z$

$$C_A = f_0 + (f_1c_1)x_1 + (f_1c_2)x_2$$

Regressão Linear Múltipla (MLR)
Encontrar o Z que minimiza S^2



$$C_A = f_0 + (f_1c_1)x_1 + ... \ (f_1c_p)x_p$$

Calibração Multivariada – Regressão Linear Múltipla - MLR

Critério dos Mínimos Quadrados

- O vetor b dos coeficientes lineares é estimado, na etapa de calibração, utilizando o critério dos "Mínimos Quadrados" que minimiza o vetor dos resíduos (e = Y – Xb) e é calculado por:

$$b = (X^t X)^{-1} X^t Y$$

- O vetor concentração Y_{am} de uma amostra desconhecida é obtida por:

$$Y = X_{am} b$$

Problemas da MLR

- colinearidade → Correlação alta em X, a inversa de (X^T X) pode não existir ou dar erros grandes em b;
- infinitas soluções → número de variáveis não pode ser maior do que o nº de misturas de calibração.

Colinearidade

- Em MRL, para calcular b e prever Y (as concentrações dos analitos) é necessário calcular a inversa de $X_{cal}{}^T X_{cal}$ e este cálculo só pode ser realizado se o determinante $X_{cal}{}^T X_{cal}$ for diferente de zero.

Exemplo:

$$X_{cal} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix} \Rightarrow X_{cal}{}^T X_{cal} = \begin{bmatrix} 14 & 28 \\ 28 & 56 \end{bmatrix}$$

$$\det\left(X_{cal}{}^T X_{cal}\right) = 14 \times 56 - 28 \times 28 = 784 - 784 = 0$$

Problema:

- Quando as colunas de X_{cal} são linearmente dependentes, as variáveis das colunas x_1 e x_2 são colineares, ou seja, quando x_1 aumenta, x_2 aumenta na mesma proporção. Portanto, ou as variáveis x_1 ou as variáveis x_2 portam informação redundante !

Colinearidade, Instabilidade Numérica e Propagação de Erros

- Se duas variáveis forem aproximadamente colineares, o determinante de $X_{cal}{}^T X_{cal}$ poderá ser muito pequeno (muito próximo de zero) e isto causa:
  - matriz $X_{cal}{}^T X_{cal}$ não pode ser invertida, dando um erro matriz singular (det ≅ 0);
  - problemas de instabilidade numérica devido ao ruído das variáveis;
  - propagação de erros na determinação das concentrações Y_{cal}

Possível Solução:

- Retirar do modelo as variáveis que causam problema de colinearidade, pois sua informação é redundante.
- Para selecionar as variáveis não colineares recomenda-se usar os métodos tais como: stepwise MLR, GA-MLR e APS-MLR

Análise de Componentes Principais(PCA)

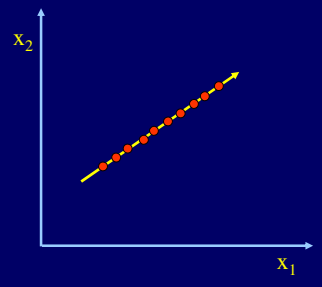
Princípio da PCA:

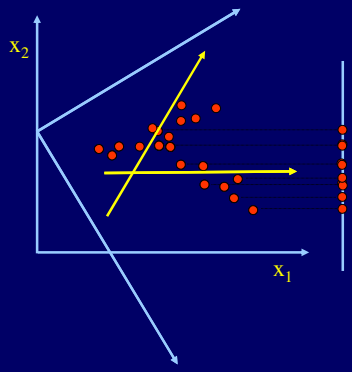
"Possibilitar a extração de informação, a partir de dados multivariados complexos, difícil de ser visualizada diretamente"

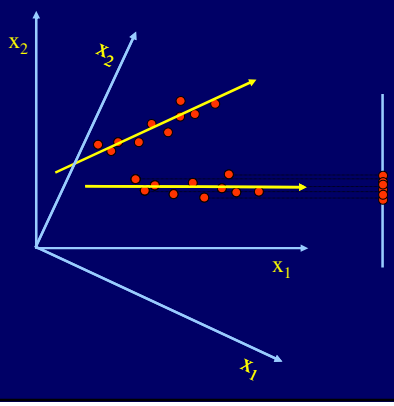
"Encontrar as direções no espaço ao longo das quais a distância entre os pontos dos dados seja a maior possível. Encontrar as combinações lineares das variáveis iniciais que fazem as amostras diferirem umas das outras"

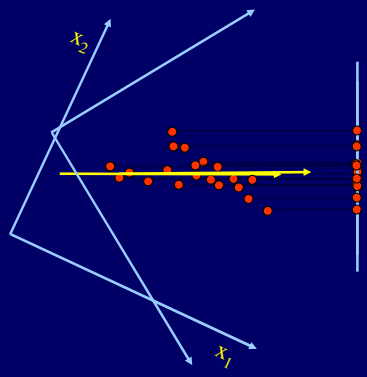
PCA (SIMCA, PCR e PLS) permite:

- Dizer a que respeito uma amostra é diferente da outra.
- Quais variáveis contribuem mais para estas diferenças.
- Quais variáveis contribuem da mesma forma (correlacionadas) ou independentemente uma da outra.
- Detectar padrão de amostras (agrupamentos).
- Diferenciar informação útil de ruído.

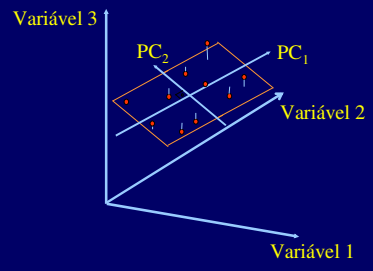


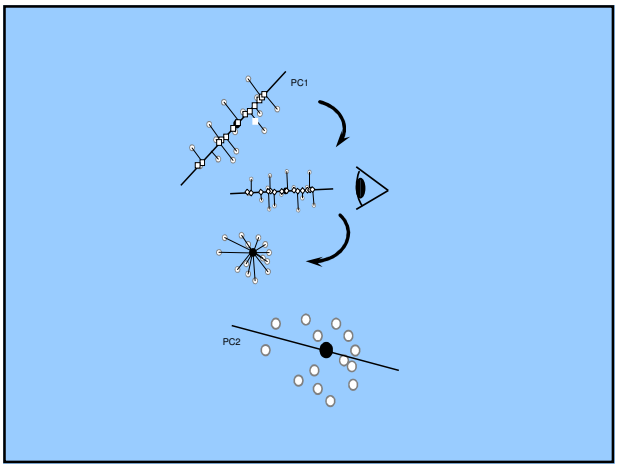






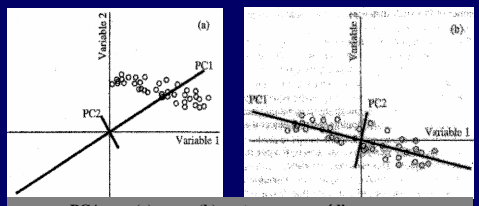
PCA





Importância de Centrar na Média em PCA

Como PC1 deve descrever a direção a partir da origem até a nuvem dos dados e PC2 é restringido ser perpendicular e também passar pela origem, logo pode-se obter as figuras:



PCA sem (a) e com (b) centragem na média e com c

Observação

1) sem centrar as 2 PCs varrem o espaço mas não estão orientadas com a nuvem no espaço linha.

2) centrando na média, as PCs descrevem efetivamente a direção da nuvem de pontos (PC1 descreve principal fonte de variação (o comprimento) e PC2 a largura.

Informação → Variabilidade

Duas amostras são similares se elas têm valores próximos para a maioria das variáveis (coordenadas próximas no espaço multi-dimensional.

Duas amostras são diferentes se elas têm valores que diferem para pelo menos algumas variáveis (coordenadas muito diferentes)

Análise de Componentes Principais (PCA - NIPALS)

$$X = TP^t$$

$$\underset{N}{\boxed{X}}^{K} = \underset{N}{\boxed{T}}^{A} \underset{A}{\boxed{P^t}}^{K} + \underset{N}{\boxed{E}}^{K}$$

X = matriz de dados originais
(N - linhas - amostras x K colunas - variáveis)

A = número de componentes principais

T = Matriz de scores

P = Matriz de loadings E =Matriz erro

Distâncias

Quando se determina a distância de um objeto a um grupo, (ou entre dois grupos) é necessário se considerar a posição do objeto comparada com a direção do eixo passando através do grupo.

Esta medida é fornecida pela distância de Mahalanobis.

Essencialmente, a distância de Mahalanobis é igual a distância Euclidiana com um fator adicional que considera a correlação entre as variáveis.

Distância Euclidiana (DE)

A distância Euclidiana entre as amostras *x* e *y* é calculada em dimensões *nvar* (nº de variáveis medidas) por:

$$DE = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_{n\,\mathrm{var}} - y_{n\,\mathrm{var}})^2}$$

onde x_i e y_i são as coordenadas das amostras *x* e *y* na i-ésima dimensão (onde *i* varia de 1 a *nvar*).

Distância de Mahalanobis (DM)

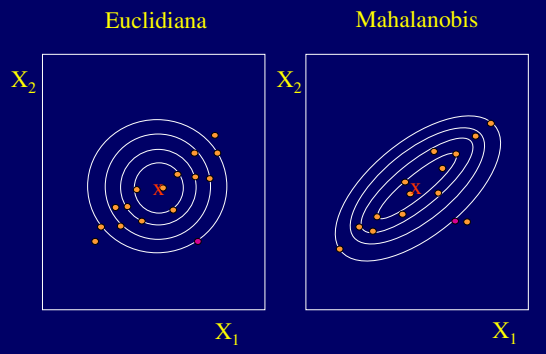
Para uma distribuição bivariada (x_1 e x_2), a distância de Mahalanobis é dada por:

$$DM^2 = \left[\frac{(x_1 - \mu_1)}{\sigma_1}\right]^2 + \left[\frac{\frac{(x_2 - \mu_2)}{\sigma_2} - \rho\frac{(x_1 - \mu_1)}{\sigma_1}}{\sqrt{1-\rho^2}}\right]^2$$

Observações

1) **se coeficiente de correlação ρ = 0, a distância de Mahalanobis é igual a distância Euclidiana;**

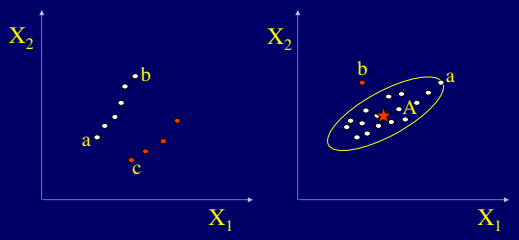
2) **a distância de Mahalanobis é igual a distância Euclidiana com um fator adicional que considera a correlação entre as variáveis.**



Pontos nos círculos têm distâncias Euclidianas iguais

Pontos nas elipses têm distâncias de Mahalanobis iguais

Similaridade e Distância



"a" é mais similar a "b" do que a "c"

"a" está mais próxima do grupo "A" do que "b"

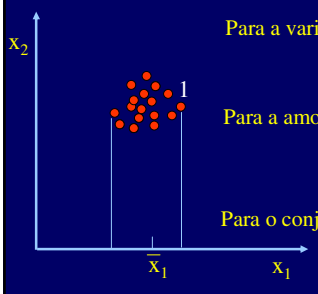
Resultados da PCA

Variâncias: medidas de erros; elas dizem quanto da informação está sendo considerada pelos sucessivos PCs. Variância - de uma variável, de uma amostra ou de todo o conjunto de dados. Variação quadrática média corrigida pelos graus de liberdade restantes.

Variância Residual: expressa o quanto da variação nos dados resta para ser explicada uma vez que o PC corrente seja considerado.

Variância Explicada: Expressa como a porcentagem da variância total nos dados; é uma medida da proporção da variação nos dados considerada pelo PC corrente.

Variâncias



Para a variável $X_1 = \dfrac{\Sigma\,(X_{1i} - \bar{X}_1)^2}{(N-1)}$

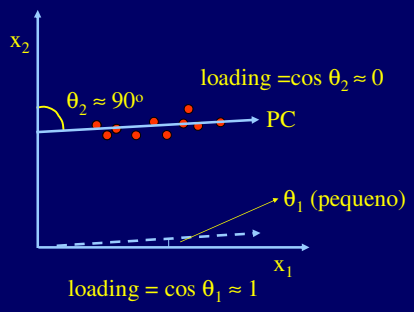
Para a amostra 1 = $(X_{11} - \bar{X}_1)^2 + (X_{21} - \bar{X}_2)^2$

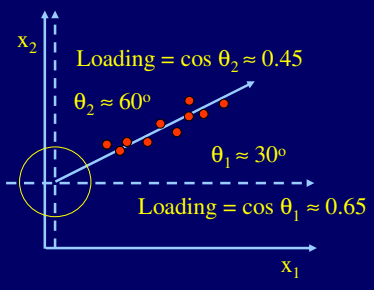
Para o conjunto = $\dfrac{\Sigma\,(X_{1i} - \bar{X}_1)^2 + (X_{2i} - \bar{X}_2)^2}{(N-1)}$

Resultados - PCA

Loadings: descrevem as relações entre as variáveis. Cada variável tem um loading em cada PC. Ele reflete o quanto aquela variável contribui para aquele PC e com que qualidade aquele PC considera a sua variação sobre os dados.

Geometricamente o loading é o coseno do ângulo entre a variável e o componente principal. Quanto menor o ângulo, maior o loading, refletindo uma maior importância da variável naquele PC.





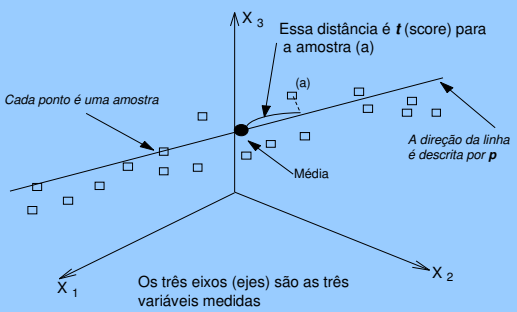
X_1 é correlacionada com X_2
(neste caso, correlação positiva)

Scores:

Descrevem as propriedades das variáveis. Apontam as diferenças e similaridades entre as amostras. Cada amostra tem um score em cada PC. O score é a coordenada da amostra no PC.

Amostras com valores de scores semelhantes no mesmo PC são semelhantes (elas têm valores similares para as variáveis correspondentes).

O que é um "Score"?



Diagnosticando um modelo PCA

Diagnosticar significa acessar a qualidade do modelo.

1. Checar Variâncias.
2. Avaliar presença de "outliers" (amostras que não pertencem ao conjunto, espurios)

Modelo PCA → Aproximação da Realidade

$$X = TP^t + E$$

Matriz de Variáveis

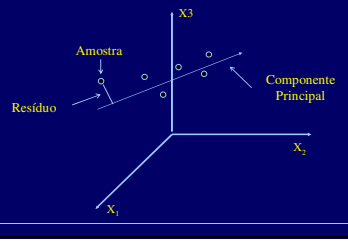
Matriz dos scores

Matriz dos loadings

Matriz dos erros

Residuais das Amostras

Cada ponto é aproximado por outro ponto que está no hiperplano gerado pelos componentes do modelo (PCs) A diferença entre o ponto original e aquele aproximado (ou projeção no modelo) é denominada de Resíduo da Amostra



Resíduos das Variáveis

Os vetores originais das variáveis são aproximados por suas projeções nos componentes principais do modelo. A diferença entre o vetor original e o projetado é o resíduo da variável.

Variação Residual

Soma dos quadrados dos resíduos da amostra para todos os componentes (PCs) do modelo.

Quadrado da distância entre a localização original da amostra e sua projeção no modelo.

Mesma definição para a variação residual das variáveis

Variância residual total

É a média da variância residual para todas as variáveis.

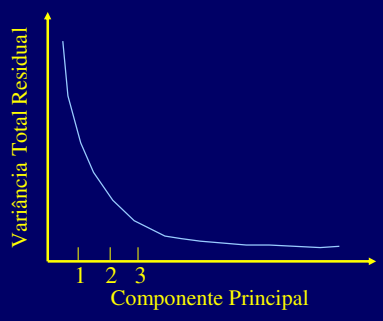
Resume o erro total de modelagem

Variância Explicada

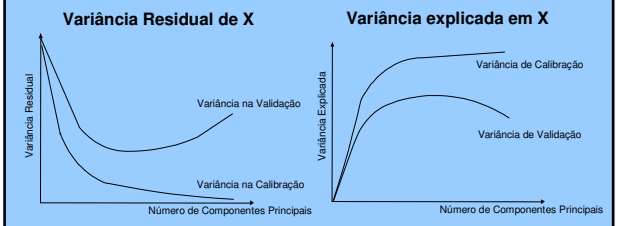
É o complemento da variância residual, expressa como porcentagem da variância total nos dados. A variância explicada de uma variável é a fração da variância global considerada pelo modelo.

Variância total explicada

Mede quanto da variação original dos dados é descrita pelo modelo. Proporção de estrutura encontrada nos dados através do modelo.



Variâncias



Como Utilizar Resíduos e Variâncias Explicadas

O resíduo total e a variância total explicada mostram quão bem o modelo se adapta aos dados. (Qualidade do modelo).

Modelos simples (variância explicada próxima a 100%) com poucos componentes principais.

Modelos complexos com muitos componentes principais refletem um grande ruído presente nos dados ou uma estrutura dos dados muito complexa para ser explicada por poucas PCs (p.e. muitos componentes químicos presentes na amostra).

Como Utilizar Resíduos e Variâncias Explicadas

Variáveis com variância residual pequena para um componente em particular são explicadas adequadamente pelo modelo.

Se algumas variáveis têm variância residual muito maior que outras variáveis para todos os PCs (ou para os primeiros 3 ou 4), tentar retirar estas variáveis e produzir um novo modelo. Este modelo poderá ser interpretado mais facilmente.

Como Utilizar Resíduos e Variâncias Explicadas

A variância da calibração é baseada na proximidade dos dados com o modelo. A variância da validação é computada testando-se o modelo em dados que não foram utilizados na construção do modelo. Se a diferença entre as duas é grande, a representatividade dos dados utilizados na calibração ou teste do modelo pode ser questionada.

A presença de Outliers pode se constituir na razão da obtenção de uma grande variação residual.

Como detectar Outliers (anomalias ou espúrios) em PCA

Outlier = Espurios

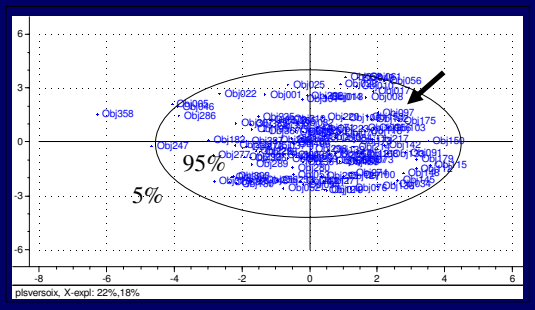
Outlier:- é uma amostra que se apresenta tão diferente das outras que ela não pode ser descrita adequadamente pelo modelo ou influencia muito o próprio modelo. Desta forma, pelo menos um dos PCs pode estar sendo utilizado somente para descrever o comportamento das variáveis associadas a esta amostra em particular, mesmo que isto seja irrelevante para a estrutura mais importante presente nas outras amostras.

Como detectar Outliers em PCA

Gráfico dos Scores: Mostra os padrões das amostras de acordo com um ou dois componentes. É fácil localizar uma amostra que se situa distante das outras. Esta amostra apresenta probabilidade de ser um outlier.

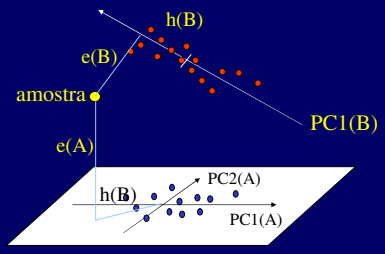
Resíduos: mede o quão bem as amostras ou variáveis são descritas pelo modelo. Uma amostra com alto resíduo é descrita de forma pobre pelo modelo. Esta amostra é candidata a ser um outlier.

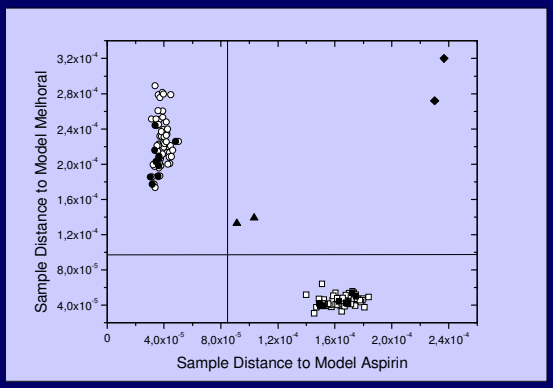
"Outliers" em PCA: Elípse de Hotelling T^2



Classificação (baseada em PCA)

SIMCA - Soft Independent Modelling of Class Analogies

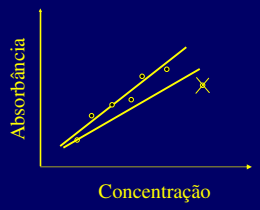




REGRESSÃO

Regressão é um termo genérico empregado para todos os métodos que procuram criar um modelo que descreva os dados observados de forma a quantificar a relação entre dois grupos de variáveis. O modelo produzido pode ser utilizado para descrever esta relação ou para prever novos valores.

Regressão Linear (univariada)
(efeito de um outlier)



Concentração = a + b(Absorbância)

Calibração Multivariada – PCR e PLS

- **PCR e PLS são os métodos de calibração multivariada mais utilizados em quimiometria.**

Modelagem implícita

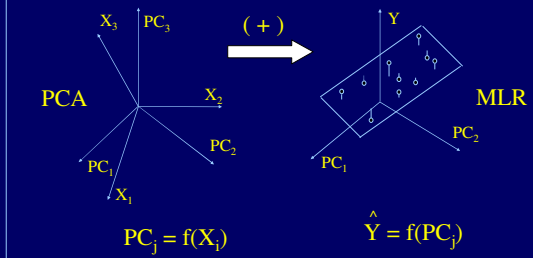
- **Como MLR, PCR e PLS também usam o processo de calibração inversa onde é possível calibrar o(s) componente(s) desejado(s), modelando (levando em conta) implicitamente todas as fontes de variação.**

Solução do problema da Inversão de Matriz

- **A calibração inversa envolve a inversão de uma matriz tipicamente instável e este problema é solucionado por PCR e PLS substituindo as variáveis originais por combinações lineares das variáveis (PCs, fatores ou variáveis latentes).**

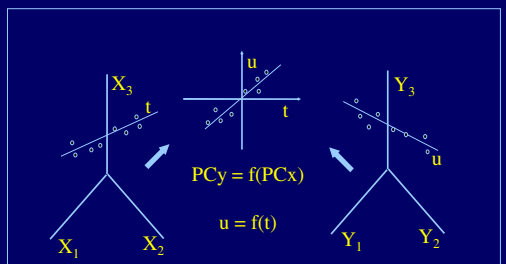
Observação:
1) este problema é solucionado em MLR usando um método de seleção de variáveis.

PCR - Regressão de Componentes Principais



$PC_j = f(X_i)$ $\quad \hat{Y} = f(PC_j)$

PLS - Regressão de Quadrados Mínimos Parciais



PLS1 trabalha com uma só variável de resposta por vez (assim como MLR e PCR).

PLS2 trabalha com várias respostas simultaneamente

Regressão: Vocabulário Básico

Contexto	X	Y
Geral	Previsores	Respostas
MLR	Variáveis Independentes	Variáveis Dependentes
Espectroscopia	Espectro	Constituintes/ Propriedades

Y = f(X)

Função do Modelo: explicar ou prever as variações nas variáveis **Y** através das variações nas variáveis **X**.

A ligação entre os valores de **X** e **Y** é encontrado com o uso de um grupo comum de amostras para as quais os valores de X e Y são conhecidos.

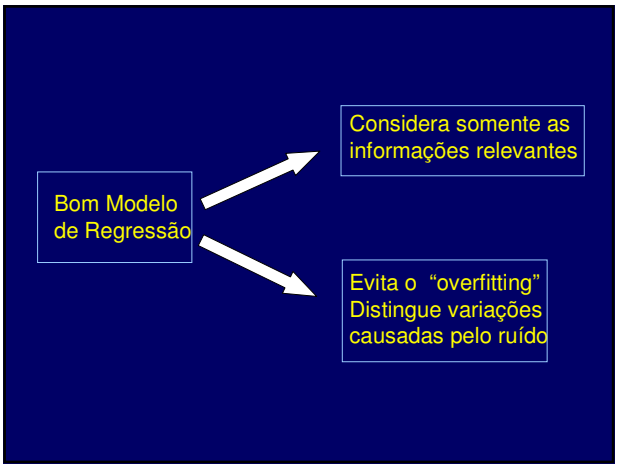
Regressão Univariada: Utiliza um só previsor, o que pode ser frequentemente insuficiente para modelar uma dada propriedade.

Regressão Multivariada: Considera muitas variáveis simultaneamente modelando a propriedade de interesse com maior exatidão.

Ruído: Variações aleatórias na resposta devido ao erro experimental.

Informação Irrelevante: está presente nos previsores que têm pouco ou nada haver com o fenômeno que está sendo modelado.

Por exemplo: Os espectros de absorbância NIR podem conter informações relativa ao solvente e não somente ao composto ou propriedade de interesse.



Métodos de Regressão:

Regressão Linear Múltipla (MLR).

Estima os coeficientes do modelo através da equação:

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

Inversão de matriz está envolvida o que leva a problemas de colinearidade se as variáveis não forem linearmente independentes.

Requer mais amostras do que variáveis de previsão (predictors, variáveis independentes).

$$y = b_0 + b_1X_1 + ... + b_kX_k + e$$

Valores previstos de Y:
Computados para cada amostra aplicando-se o modelo.

Resíduos:
Diferença entre o valor observado (experimental) de Y e o valor determinado pelo modelo.

Variação residual de Y:
Expressa o quanto da variação permanece nas respostas observadas quando a parte modelada é retirada .

RMSEC e RMSEP:
Erro de calibração e erro de previsão, respectivamente, expressos nas mesmas unidades das variáveis de resposta .

MODELOS PCR e PLS

PCR:
Decompõe a matriz **X** por PCA e então constrói um modelo MLR usando os PCs ao invés dos dados dos previsores originais.

PLS (1 e 2):
Projeção de Estruturas Latentes. Modela **X** e **Y** simultaneamente para encontrar variáveis latentes em **X** que melhor prevêem as variáveis latentes em **Y**. Os componentes PLS são similares aos PCs.

PLS1 - uma variável de resposta por vez.
PLS2 - várias respostas

Resultados Principais do PCR:

Scores, Loadings, Resíduos :

Mesmo significado do que em PCA e MLR.

ETAPAS NA ELABORAÇÃO DE UM MODELO

. Coletar espectros amostras representativas para as quais se obteve os valores de referência

. Escolher adequadamente o pré-processamento

. Calibração

. Escolher o número de componentes

. Validação

. Diagnóstico do modelo

. Interpretar "loadings e scores"

. Previsão para novas amostras

COMO DETECTAR NÃO-LINEARIDADES

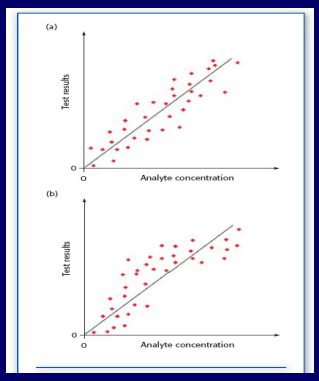
Modelo Bom ➡ Resíduos devem se distribuir aleatoriamente

Resíduos em Y vs. Y previstos

Resíduos em Y vs. scores

PLS - X-Y Relation Outliers - relacionamento de X e Y ao longo de um componente do modelo .

Distribuição dos Resíduos



COMO DETECTAR OUTLIERS

Outliers são detectados usando-se os gráficos de scores, resíduos e leverages.

1. Uma amostra pode ser um outlier em relação a variável X ou Y ou ambas. Pode não ser em relação a X e Y separadamente mas ser quando se considera a relação de X e Y. Neste caso, gráficos X-Y Relation Outliers são utilizados.

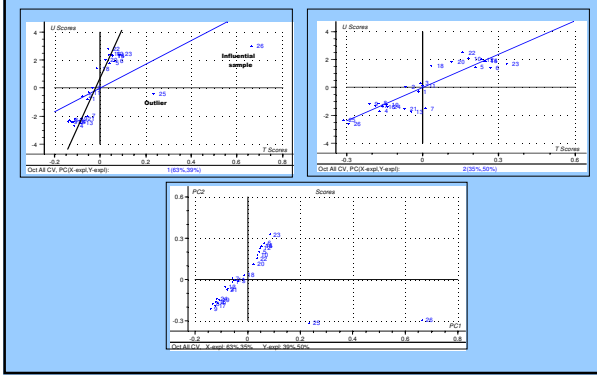
2. **Resíduos**. Gráfico da **variância residual** por amostra. Gráfico do **resíduo das variáveis** para amostras que apresentam valores altos para sua variância residual.

COMO DETECTAR OUTLIERS

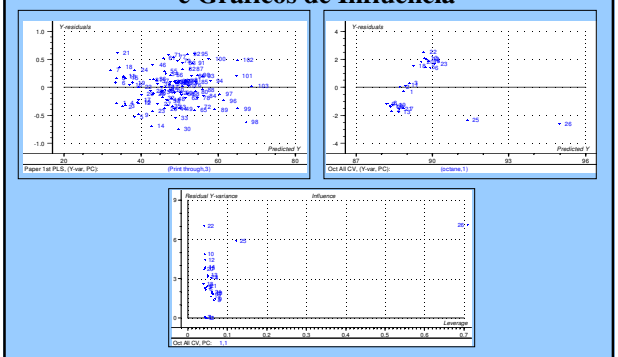
3. **Leverages**. Graficadas em função das amostras. Amostras com altos valores de leverage podem ser *outliers*.

4. Combinar os três itens anteriores com Y-resíduos vs Y-previstos.

"X-Y Relation Outliers" e Gráfico de "Score"



Resíduos vs. Previstos e Gráficos de Influência



MLR vs PCR vs PLS

MLR:

Número de variáveis deve ser menor que o número de amostras. MLR tende a incluir ruído no modelo. Deve ser usado quando o número de variáveis X é pequeno e com pequena correlação entre elas.

PCR:

Utiliza MLR na regressão. Um modelo PCR que utilize todos os PCs tem a mesma solução que um MLR. PCR e MLR modelam uma só variável Y por vez.

PLS(1-2):

Utiliza as variáveis X e Y para definir o modelo. PLS usualmente requer menos PCs para atingir o ótimo do modelo. PLS2 é melhor quando se tem mais de uma variável Y. PLS1 e PCR fornecem melhores resultados se existe a presença de fortes não-linearidades pois modela cada Y em separado, considerando as não-linearidades. Por outro lado se as variáveis Y apresentam alto ruído, sendo fortemente correlacionadas, o PLS2 é melhor.

VALIDAÇÃO DE MODELOS MULTIVARIADOS

Validar: Verificar como o modelo vai operar com novos dados.

Validação por conjunto de teste:
Conjunto de amostras não empregado na calibração.

Cross-validation (validação cruzada):
Completa, Segmentada. Comutação. Utiliza o mesmo conjunto de amostra da calibração. Deixa uma ou mais fora da calibração e prevê os valores de Y para elas e calcula o RMSEP.

PREVISÃO

Último estágio da análise multivariada.

Amostras com valores Y desconhecidos.

Modelo de Regressão (MLR, PCR e PLS) que expressam a resposta da variável ou das variáveis Y em função das variáveis X.

O modelo deve ser **calibrado** com amostras cobrindo a região na qual as novas amostras irão ser encontradas.

O modelo deve ter sido **validado** em amostras cobrindo esta mesma região.

Resultados da Previsão:

Previsão com Desvios - Os valores previstos de Y são mostrados com valores de *"deviation"*. Estes desvios expressam o quanto a amostra é similar ao conjunto de calibração. Quanto mais similar, menor o desvio. Pode ser interpretado como um limite de confiança a 95%.

Previsão vs. Referência -fornecido somente se os valores de referência estiverem disponíveis.

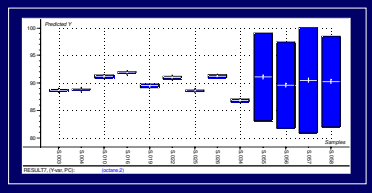
Limite de Incerteza na Previsão

(Valor Previsto ± Desvio)

Incerteza na previsão de um objeto novo específico. Baseado em similaridade com os objetos de calibração.

Similar → Previsão confiável

Diferente → Não- confiável



Cálculo do "deviation" (UNSCRAMBLER)

$$Y(\text{Deviation}) = \sqrt{\text{VarResY(validação)} \left[\frac{\text{ResXValAmPrev}}{\text{ResXValTot}} \right] + H_i + \frac{1}{\text{Ical}} \left[1 - \frac{a+1}{\text{Ical}} \right]}$$

VarResY(validação) = Variância residual de Y obtida na validação do modelo (RMSEP)

ResXValTot = Variância total residual em X apurada na validação do modelo

ResXValAmPrev = Variância residual em X da amostra prevista

Ical = Número de amostras de calibração

H_i = Leverage da amostra

a = Número de variáveis latentes

Controle de Qualidade da Calibração

A manutenção da qualidade de um método multivariado NIR/IR requer a avaliação do instrumento e do modelo.

Para efetuar esta avaliação periódica pode ser necessário se escolher uma ou mais amostras de controle de qualidade que deverão monitorar alterações de performance do instrumento ou modelo.

As amostras de controle de qualidade devem ser identificadas por ocasião da construção do modelo.

Controle de Qualidade da Calibração

As amostras de controle devem:

1. Química e fisicamente ser compatível com a amostra, não deve contaminar as amostras e não apresentar problemas de segurança.

2. Os materiais devem ser quimicamente estáveis.

3. O espectro do material de controle deve ser compatível com o modelo. O espectro do material deve ser o mais similar possível com os espectros das amostras de calibração.

Controle de Qualidade da Calibração

Procedimento para o controle de qualidade:

1. Coletar pelo menos 20 espectros do material de controle. Diferentes amostras devem ser empregadas para cada medida.

2. Os espectros são analisados pelo modelo e o valor médio (Y_{qc}) é calculado.

3. Calcula-se o desvio padrão das estimativas (σ_{qc}).

4. O valor estimado deve para o material de controle de qualidade deve estar entre: $(Yqc - t \times \sigma_{qc})$ e $(Yqc + t \times \sigma_{qc})$

Atualização do Modelo

- Algumas vezes se faz necessário o acréscimo de mais amostras de calibração a um modelo pré-existente para aumentar sua faixa de aplicação.

- Os procedimentos de identificação de *outliers* devem ser aplicados ao novo conjunto. Assim, se amostras adicionais estão sendo incluídas para aumentar a faixa de concentração pode ser necessário adicionar várias amostras do novo tipo para se evitar que estas sejam classificadas como outliers.

- O modelo deve ser revalidado. A porcentagem de amostras adicionadas no conjunto de validação deve ser no mínimo tão grande como aquela das novas amostras adicionadas no conjunto de calibração.

Questionário para Calibração Multivariada

1. A técnica matemática utilizada na calibração foi MLR, PCR ou PLS1?

2. A metodologia inclui a capacidade de detectar *outliers* de alta leverage, empregando estatística com a distância de Mahalanobis?

3. A metodologia apresenta a capacidade de detectar *outliers* através da análise dos resíduos espectrais?

4. Número de amostras no conjunto de calibração n > 6K se o modelo não é centrado na média ou 6(k +1) se o modelo for centrado na média?

n = número de amostras no conjunto de calibração
k = número de variáveis (MLR comprimentos de onda, componentes principais ou PLS variáveis latentes).

5. O número de amostras no conjunto de calibração é no mínimo 14?

6. Um conjunto separado de amostras de validação foi empregado no teste do modelo?

7. Os espectros de validação caracterizados como *outliers* devido a alta leverage ou alto resíduo foram excluídos do conjunto de validação?

8. O número de amostras de validação é maior que 4k se o modelo não é centrado na média ou maior que 4(k+1) se o modelo é centrado na média?

9. O número de amostras de validação é, no mínimo, 20?

10. As amostras de validação se distribuem pelo menos por 95% da faixa das amostras de calibração?

11. 95% dos resultados das amostras de validação estão entre

$$\pm t \times SEC \times \sqrt{1 + D^2}$$

t para n-k (não centrado) ou (n-k-1) (centrado)?

12. A validação mostra tendência (*bias*) insignificante?

13. A precisão do modelo foi foi determinada usando ($t \geq k \geq 3$) amostras de teste e ($r \geq 6$) medidas de replicatas por amostra?

14. Se a metodologia de análise e calibração inclui pré-processamento ou pós-processamento, estas operações são feitas automaticamente?

SE VOCÊ RESPONDEU **AFIRMATIVAMENTE**

TODAS AS 14 QUESTÕES, PARABÉNS VOCÊ

TEM UM MODELO MULTIVARIADO QUE

OBEDECE A NORMA ASTM 1655-05!!!

FIM