

Visualização

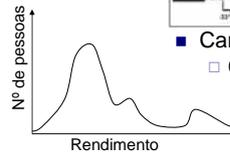
V 1.3, V.Lobo, EN/ISEGI, 2010

Armazenamento, Visualização & Representação

Victor Lobo

Mestrado em Estatística e Gestão de Informação

Uma imagem são mil palavras...



- Campanha da Rússia
- 6 variáveis diferentes !

Casos notáveis...

- Surto de cólera em Londres, em 1854
 - Gráfico da distribuição de ocorrências de casos
 - Suspeita que algo no "centro" provocava a doença
 - Provou-se que a doença tinha origem num poço de água inquinado



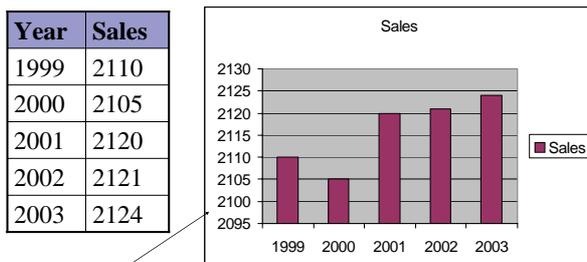
In Visual and Statistical Thinking:
Displays of evidence for making decisions

Para quê visualizar ?

- Apoiar a **exploração interactiva** dos dados
- **Analisar** os resultados
- Apresentação e **comunicação** dos resultados
- **Compreender** os dados, ter uma **perspectiva** sobre eles
- O olho humano é melhor sistema de clustering...
- Desvantagens
 - Requerem *olhos humanos*
 - É uma análise *subjectiva*
 - Podem ser *enganadores*

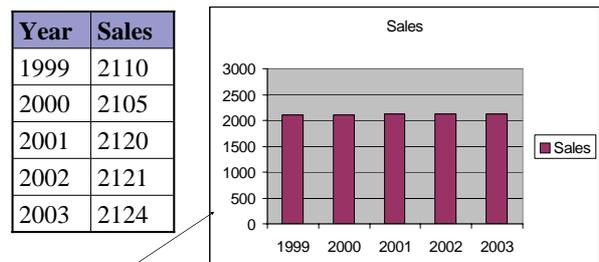
Mentir com Gráficos

Gráfico com um eixo Y "enganador"



O eixo dos Y dá uma falsa sensação de grande mudança

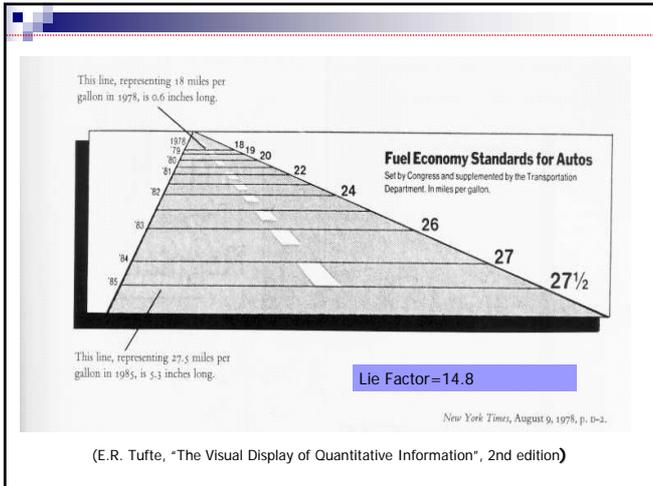
Melhor...



O eixo entre o 0 e os 2000 dá uma leitura correcta de pequenas alterações

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010



Lie Factor

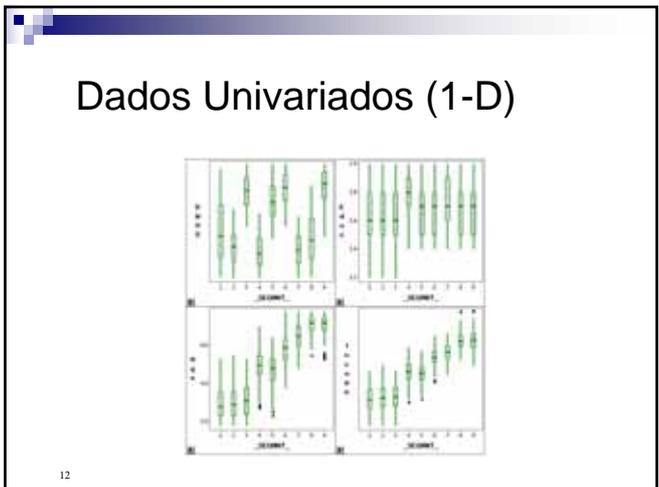
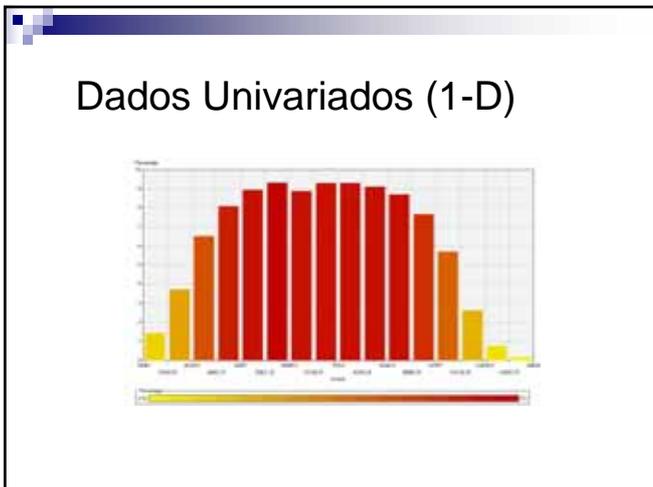
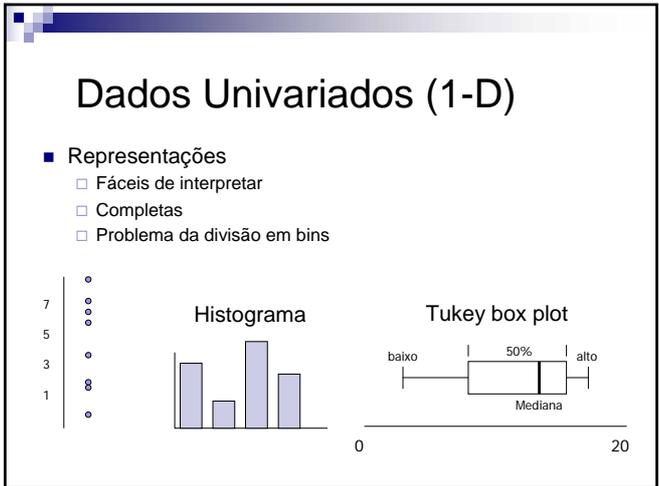
$$Lie\ Factor = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}} =$$

$$= \frac{(5.3 - 0.6)}{\frac{0.6}{(27.5 - 18.0)}} = \frac{7.833}{0.528} = 14.8$$

Tufte requirement: $0.95 < Lie\ Factor < 1.05$

(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)

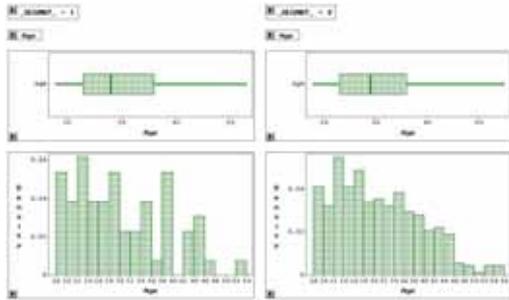
- ## Visualização de dados e dimensões
- 1 dimensão – Trivial
 - Listas, Histogramas
 - 2 dimensões – Fácil
 - Tabelas de contingência, scatterplots,
 - 3 dimensões – Complicado
 - Gráficos 3D, waterfall, contourplots
 - Multidimensionais
 - Projecções para dimensões menores
 - Coordenadas paralelas, radarplots, caras de chernoff, stick figs.
 - Dados "com interesse" são quase sempre multidimensionais !!!



Visualização

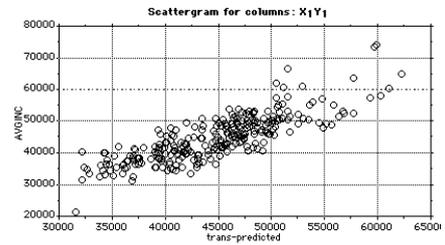
V 1.3, V.Lobo, EN/ISEGI, 2010

Dados Univariados (1-D)



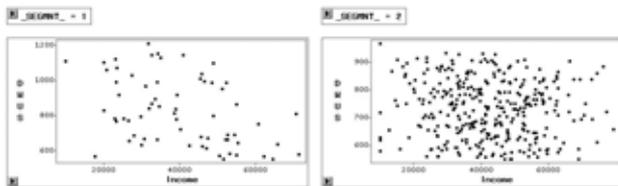
Dados Bivariados (2-D)

- Gráfico de dispersão, ou scatterplots

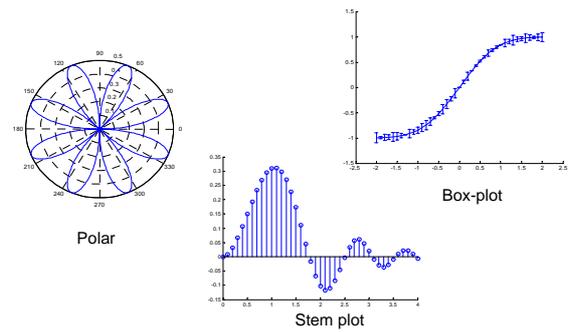


Dados Bivariados (2-D)

- Múltiplos scatterplots



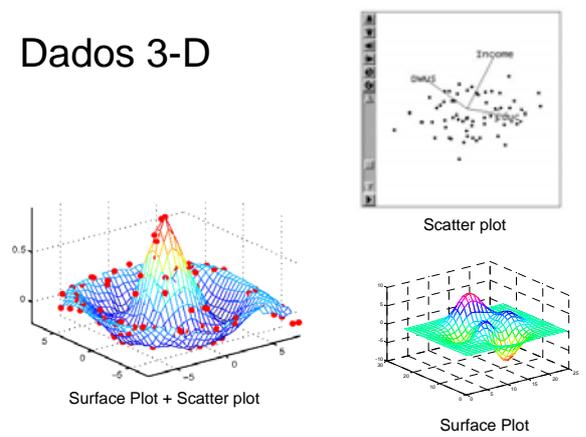
Dados Uni ou Bivariados (2-D)



Histograma a 2 dimensões (Tabela de contingência a 3D)

- Patch graph

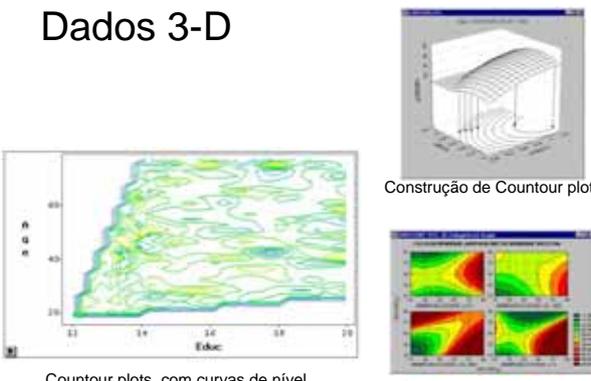
Dados 3-D



Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Dados 3-D



Construção de Contour plots

Contour plots, com curvas de nível

Contour plots, com cores

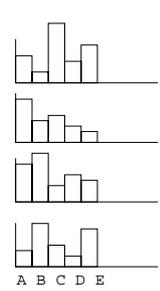
Dados multidimensionais

- Visualizações directas são impossíveis
- Alternativas:
 - Múltiplos gráficos
 - Coordenadas alternativas
 - Características não espaciais
 - Múltiplos eixos espaciais
 - Projecções sobre dimensões mais reduzidas

Múltiplos Gráficos

Dar a cada variável a seu gráfico

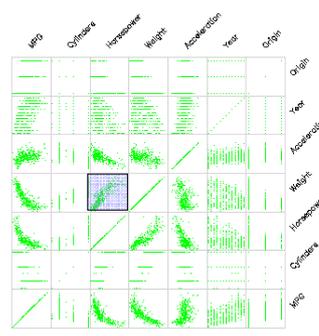
	A	B	C	D	E
1	4	1	8	3	5
2	6	3	4	2	1
3	5	7	2	4	3
4	2	6	3	1	5



Problema: não mostra as correlações

Matriz de gráficos de dispersão

Representar cada um dos possíveis pares de variáveis com o diagrama de dispersão correspondente

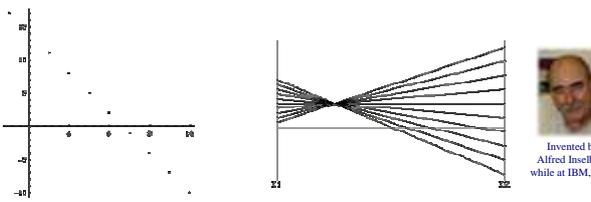


Q: Utilidade?
A: Correlações lineares

Q: Ponto fraco?
A: efeitos multivariados

Coordenadas Paralelas

- Codificar as variáveis ao longo de um eixo horizontal
- As linhas verticais especificam os valores



Dados em coordenada Cartesianas

Os mesmos dados em coordenadas paralelas

Invented by Alfred Inselberg while at IBM, 1985

Exemplo: visualizar o "iris dataset"

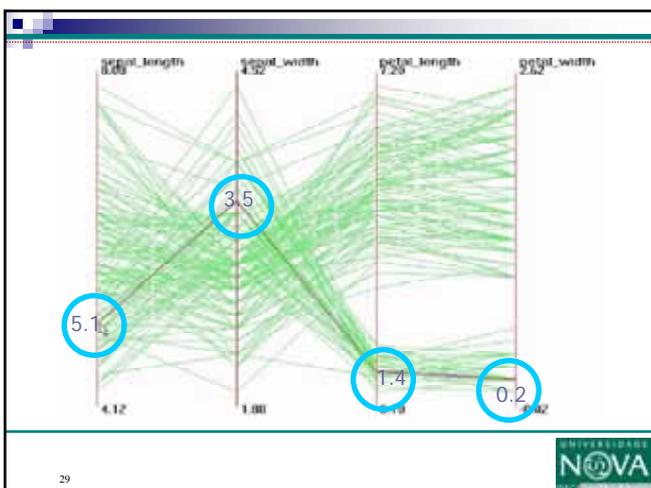
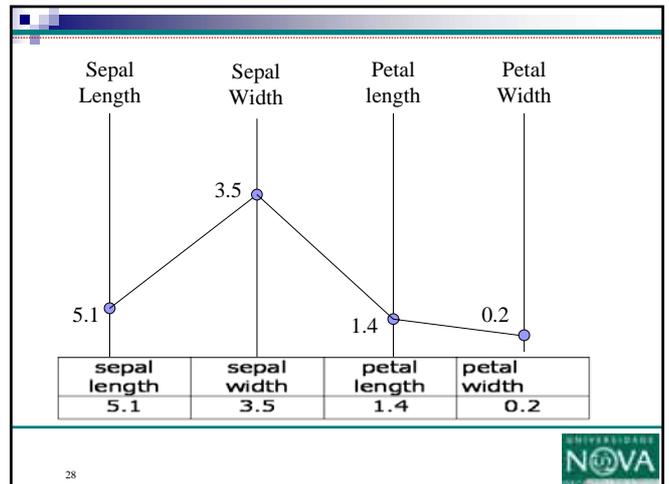
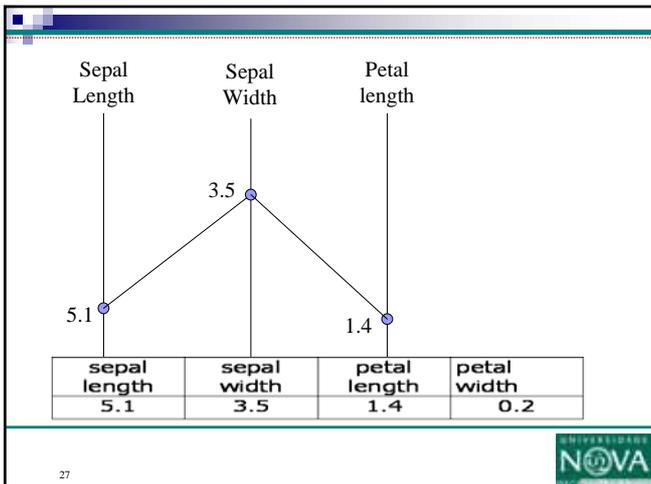
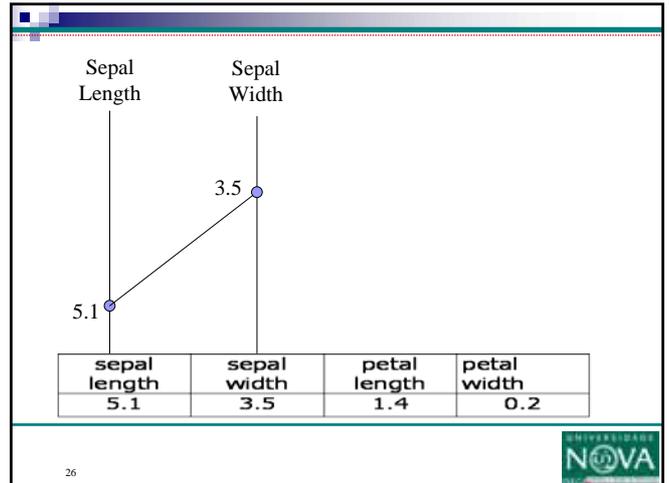
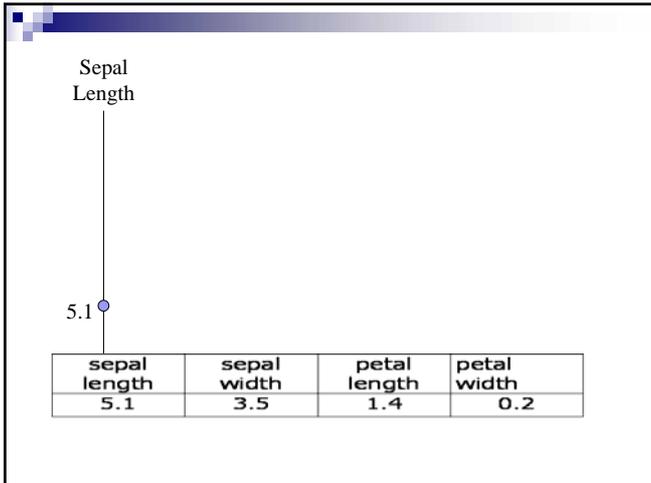
- A flor Iris tem várias variantes, 3 das quais são:
 - 1 -Iris Setosa
 - 2 -Iris Versicolour
 - 3 -Iris Virginica
- Para 50 flores de cada uma das variantes foram medidas 4 características (medidas em cm)
 - Largura da pétala
 - Comprimento da pétala
 - Largura da Sépala
 - Comprimento da Sépala
- (Questão típica)
 - É possível determinar a variante a partir desses 4 parâmetros ?



Iris Setosa

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010



Star plots (ou radar, ou spider)

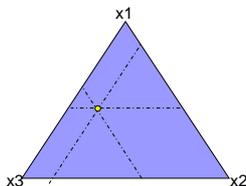
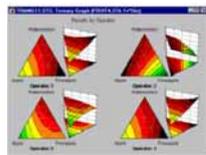
- Por os diversos eixos numa "roda"

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

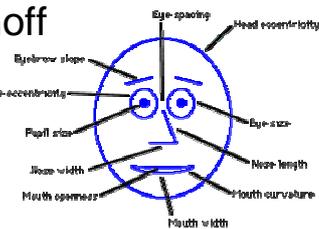
Trilinear Graphs

- Quando a soma de 3 variáveis é constante



Caras de Chernoff

- As dimensões correspondem a características da face
 - Até 11 dimensões facilmente reconhecíveis.
 - A posição da cara num gráfico 2 ou 3D acrescenta ainda mais dimensões.
 - A escolha das características pode ser polémica...



Chernoff-Faces [Che 73, Tuf 83]



Exemplos de visualizações com caras de Chernoff

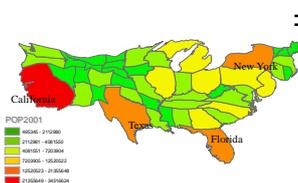
- Dados demográficos sobre Portugal
 - **Largura da face:** taxa de fecundidade de nados-vivos por 1 000 mulheres em idade fecunda: 15-49anos)
 - **Largura do nariz:** índice de envelhecimento (n.º de residentes com 65 e mais anos por 100 residentes com menos de 15 anos)
 - **Comprimento do nariz:** taxa de mortalidade (numero de óbitos por 1 000 habitantes)
 - **Curvatura da boca:** taxa de natalidade (numera de nados-vivos por 1 000 habitantes)
 - **Comprimento da boca:** nados-vivos fora do casamento (nados-vivos fora do casamento por 100 nados-vivos)
 - **Tamanho das orelhas:** taxa de nupcialidade (numero de casamentos por 1 000 habitantes)
 - **Ângulo das sobrancelhas:** taxa de divorcio (numero de divórcios por 1 000 habitantes)



[Silva 06]

Cartogramas

- Quando se quer realçar uma característica sobre um mapa geográfico



Outros...

- Andrew's curves
 - Cada variável corresponde a uma frequência [Andrew 72]
- Wireframe, contour, circular, bubble graph, high-low-close graph, Vector, surface, pictograms....

Software para visualização

- Genéricos – Excel, Matlab, Mathcad, SPSS, etc
- Dedicados
 - Tableau Software
 - www.tableausoftware.com tem demos, trials, e videos
- Applets disponíveis na net
 - <http://www.hesketh.com/schampeo/projects/Faces/interactive.html>



Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Bibliografia

- Edward R.Tufte, Visual Explanations, Graphics Press, 1997
- Edward R.Tufte, The Visual Display of Quantitative Information, Graphics Press, 1983
- Robert L. Harris, Information Graphics – A comprehensive illustrated reference, Oxford University Press, 1999
- Gene Zelazny, Say it with charts- The executive's guide to Visual Communication, McGraw-Hill, 2000
- Fayyad, Usama; Grinstein, Georges; Wierse, Andreas; Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2002
- Ana Alexandrino da Silva, Gráficos e Mapas, Lidel, 2006
- Statsoft Textbooks
 - <http://www.statsoft.com/textbook/stathome.html>

Projeções para 2 dimensões

Projeções sobre espaços visualizáveis

- Ideia geral:
 - Mapear os dados para um espaço de 1 ou 2 dimensões
- Mapear para espaços de 1 dimensão
 - Permite definir uma ordenação
- Mapear para espaços de 2 dimensões
 - Permite visualizar a “distribuição” dos dados (semelhanças, diferenças, clusters)

Exemplos

Problemas com as projeções

- Perdem informação
 - Podem perder MUITA informação e dar uma imagem errada
- Medidas para saber “o que não estamos a ver”
 - Variância explicada
 - Stress
 - Outros erros (erro de quantização, topológico, etc)

Dimensão *intrínseca*

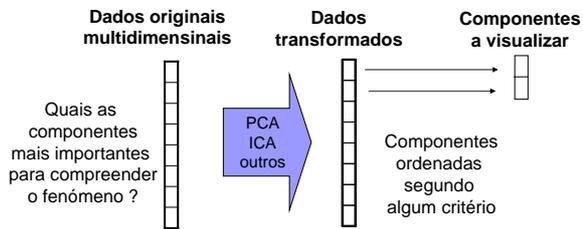
- Dimensão do sub-espaço dos dados
 - Pode ou não haver um mapeamento linear
- Estimativas da dimensão intrínseca
 - Com PCA – Verificar a diminuição dos V.P.
 - Basicamente, medir a variância explicada
 - Com medidas de stress (em MDS)
 - Com medidas de erro

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Seleccionar componentes mais “relevantes” para visualização

- Será sempre uma “boa” escolha ?

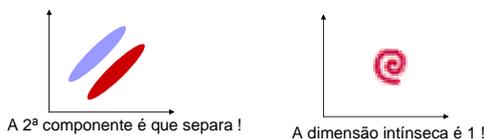


PCA – Principal Component Analysis

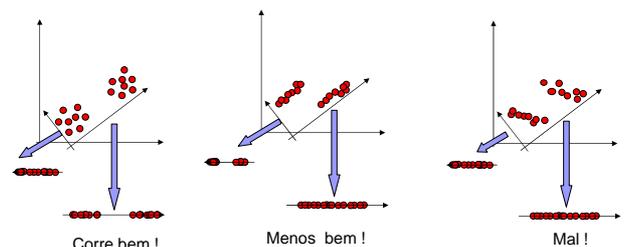
- Principal Component Analysis
 - Análise de componente principais
 - Transformada (discreta) de Karhunen-Loève
 - Transformada linear para o espaço definido pelos **vectores próprios** da matriz de **covariância dos dados**.
 - Não é mais que uma **mudança de coordenadas** (eixos)
 - Eixos ordenados pelos valores próprios
 - Utiliza-se normalmente SVD

Componentes principais

- Mudança de eixos
 - Os novos eixos estão “alinhados” com as direcções de maior de variação
 - Continuam a ser eixos perpendiculares
 - Podem “esconder aspectos importantes”

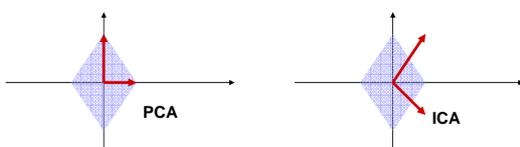


Problemas com ACP



Componentes Independentes

- ICA – Independent Component Analysis
 - Maximizam a independência estatística (minimizam a informação mútua)
- Diferenças em relação a PCA



Componentes Independentes

- Bom comportamento para clustering
 - Muitas vezes melhor que PCA por “espalhar” melhor os dados
- Bom para “blind source separation”
 - Separar causas independentes que se manifestam no mesmo fenómeno
- Disponibilidade
 - Técnica recente... ainda pouco divulgada
 - Boas implementações em Matlab e C
 - Livro de referencia (embora não a ref.original):
 - Hyvärinen, A., J. Karhunen, et al. (2001). Independent Component Analysis. Wiley-Interscience.

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Referências sobre ICA

- Primeiras referências
 - B.Ans, J.Herault, C.Jutten, "Adaptive Neural architectures: Detection of primitives", COGNITIVA'85, Paris, France, 1985
 - P.Comon, "Independent Component Analysis, a new concept?", Signal Processing, vol36,n3,pp278-283, July 1994
- Algoritmo mais usado, FastICA
 - Hyvärinen, A., J. Karhunen, et al. (2001). Independent Component Analysis, Wiley-Interscience.
 - V.Zarzos, P.Comon, "How Fast is FastICA?", Proc.European Signal Processing Conf., Florence, Italy, Setember 2006
- Recensão recente
 - A.Kachenoura et al., "ICA: A Potential Tool for BCI Systems", IEEE Signal processing Magazine, vol25, n.1, pp 57-68, January 2008
- Código freeware e material de apoio
 - FastICA para Matlab, R, C++, Python, e muitos apontadores para informação
 - <http://www.cis.hut.fi/projects/ica/fastica/>

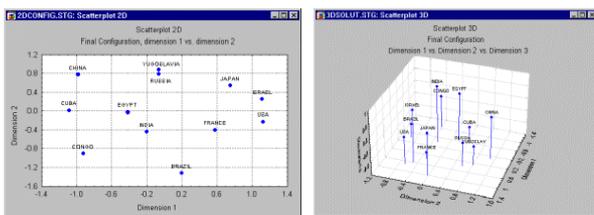
MDS – MultiDimensional Scaling

- Objectivo
 - Representação gráfica a 2D que preserva as distâncias originais entre objectos
- Vários algoritmos (e por vezes nomes diferentes)
 - Sammon Mapping (1968)
 - Também conhecido como Perceptual Mapping
 - É um processo iterativo
 - Não é, rigorosamente, um mapeamento...
- Stress
 - Mede a distorção que não foi possível eliminar

$$Stress = \sqrt{\frac{(d_{ij} - \hat{d}_{ij})^2}{(d_{ij} - \bar{d})^2}}$$

d_{ij} = distância verdadeira
 \hat{d} = distância no grafico 2d
 \bar{d} = média das distâncias

Exemplos de MDS



Exemplo com países do mundo caracterizados por indicadores socio-económicos

- Nota:
 - Ao acrescentar mais um dado é necessário recalculer tudo !

Transformações tempo/frequência

- Transformada de Fourier
 - É uma mudança de referencial !
 - Projecta um espaço sobre outro
- Transformadas tempo/frequência
 - Wavelets
 - Wigner-Ville
 - Identificam a ocorrência (localizada no tempo) de fenómenos que se vêm melhor na frequência...

Transformada de Fourier

- Aplicações
 - Análise de séries temporais
 - Análise de imagens
 - Análise de dados com dependências "periódicas" entre eles
- Permite:
 - Invariância a "tempo"
 - Invariância a "posição"
- O que é:
 - Um decomposição em senos e cosenos
 - Uma projecção do espaço original sobre um espaço de funções

Transformada de Fourier

- O que é a "decomposição" ?

$$x(t) = \text{[Complex Wave]} = \text{[Cosine]} + \text{[Sine]} + \text{[Higher Frequency]} + \dots$$

- Com o que é que fico ? Com o que quiser...
 - Com as amplitudes de cada frequência...
 - Com os valores das 2 frequências mais "fortes"...
- Notas:
 - Para não perder informação N-pontos geram N-pontos
 - Posso calcular a transformada mesmo que faltem valores

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Curvas principais, SOM, etc

- Curvas principais
 - Hastie 1989
 - Define-se parametricamente a família de curvas sobre o qual os dados são projectados
- SOM
 - Kohonen 1982
 - Serão discutidas mais tarde

Bibliografia

- Sammon, J. W., Jr (1969). "A Nonlinear Mapping for Data Structure Analysis." IEEE Transactions on Computers **C-18**(5)
- Hastie, T. and W. Stuetzle (1989). "Principal curves." Journal of the American Statistical Association **84**(406): 502-516.
- Hyvarinen, A. and E. Oja (2000). "Independent component analysis: algorithms and applications." Neural Networks **13**: 411-430
- Hyvärinen, A., J. Karhunen, et al. (2001). Independent Component Analysis, Wiley-Interscience.

Exemplo prático (TPC opcional 1)

- Numa escola universitária são realizados inquéritos aos alunos sobre as características dos professores.
- É necessário promover um dos professores auxiliares a associado.
- Os profs catedráticos gostariam de conhecer o mais possível as características dos professores auxiliares para escolher o "melhor". Gostariam de contar com o "input" dos alunos sobre o desempenho pedagógico.
- Usando os dados disponibilizados pelos inquéritos, prepare uma apresentação 1 minuto (60segundos) para esses professores, deixando-lhes depois uma folha A4 com o que fôr mais importante.

Pré-Processamento dos dados

Porquê pré-processar os dados

- Valores omissos (missing values)
- Factores de escala
- Invariância a factores irrelevantes
- Eliminar dados contraditórios
- Eliminar dados redundantes
- Discretizar ou tornar contínuo
- Introduzir conhecimento "à priori"
- Reduzir a "praga da dimensionalidade"
- Facilitar o processamento posterior

Crucial !

Garbage in /
Garbage out

Valores omissos

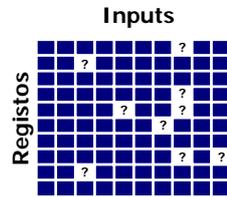
- Usar técnicas que lidem bem com eles
- Substitui-los
 - Por valores "neutros"
 - Por valores "médios" (média, mediana, moda, etc)
 - Por valores "do vizinho mais próximo"
 - K-vizinhos, parzen, etc
 - Interpolações
 - Lineares, com "splines", com Fourier, etc.
 - Com um estimador "inteligente"
 - Usar os restantes dados para fazer a previsão

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Alternativa: Eliminar valores omissos

- Eliminar registos
 - Podemos ficar com poucos dados
 - (neste caso 3 em 10)
- Eliminar variáveis
 - Podemos ficar com poucas características
 - (neste caso 4 em 9)



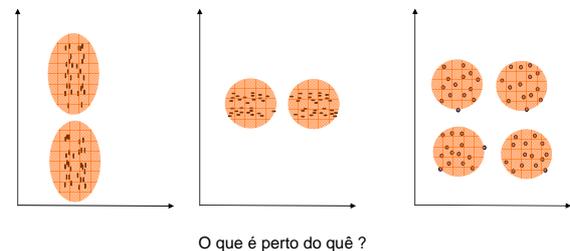
Abordagem iterativa

- Usar primeiro uma aproximação “grosseira”
 - Eliminar registos / variáveis
 - Usar simplesmente valores médios
- Observar os resultados
 - Conseguem-se boas previsões ?
 - Resultados são realistas ?
- Abordagem mais fina
 - Estimar valores para os omissos
 - Usar “clusters” para definir médias

Normalização dos dados

Nomalização

- Efeitos de mudanças de escala

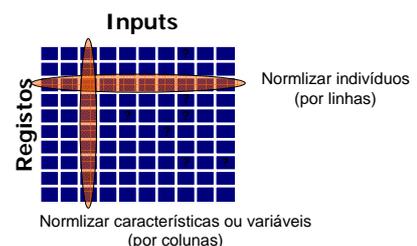


Porquê normalizar

- Para cada variável individual
 - Para não comparar “alhos com bugalhos” !
- Entre variáveis
 - Para que métodos que dependem de distâncias (logo de escala) não fiquem “trancados” numa única característica
 - Para que as diferentes características tenham importâncias proporcionais.

Porquê normalizar

- Entre indivíduos
 - Para insensibilizar a factores de escala
 - Para identificar “prefis” em vez de valores absolutos



Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Objectivos possíveis

- Aproximar a distribuição de uniforme
 - “Espalha” maximamente os dados
- Aproximar a distribuição normal
 - Identifica bem os extremos e deixa que estes sejam muito diferentes
- Ter maior resolução na “zona de interesse”

Pré-processamento

- Algumas normalizações mais comuns

- Min-Max
 - $y' \in [0,1]$
$$y' = \left(\frac{y - \min}{\max - \min} \right)$$

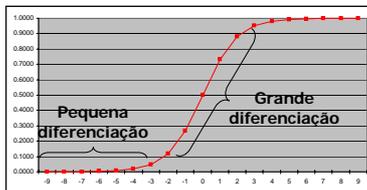
- Z-score
 - y' centrado em 0 com $\sigma=1$
$$y' = \frac{y - \text{média}}{\text{Desvio Padrão}}$$

- Percentis
 - Distribuição final sigmoidal
$$y' = n^{\circ} \text{ de ordem}$$

- Sigmoidal (logística)
 - y' com maior resolução “no centro”
$$y' = \frac{1 - e^{-\alpha}}{1 + e^{-\alpha}}$$

Normalização sigmoidal

- Diferencia a “zona de transição”



Outros problemas de pré-processamento

Eliminar outliers

- Efeito de alavanca dos outliers
- Efeito de “esmagamento” dos outliers
- Eliminar outliers
 - Estatística (baseado em σ)
 - Problema dos “inliers”
 - Métodos “detectores” de outliers
 - Com k-médias
 - Com SOM

Conversões entre tipos de dados

- Nominal / Binário
 - 1 bit para cada valor possível
- Ordinal / Numérico
 - Respeitar ou não a escala ?
- Numérico / Ordinal
 - Como discretizar ?

Visualização

V 1.3, V.Lobo, EN/ISEGI, 2010

Outras transformações

- Médias para reduzir ruído
- Ratios para insensibilizar a escala
- Combinar dados
 - É introdução de conhecimento “à priori”

Quanto pré-processamento ?

- Mais pré-processamento
 - Maior incorporação de conhecimento à priori
 - Mais trabalho inicial, tarefas mais fáceis e fiáveis mais tarde
- Menos pré-processamento
 - Maior esforço mais tarde
 - Maior “pressão” sobre sistema de classificação/ previsão / clustering
 - Princípio: “garbage in – garbage out”