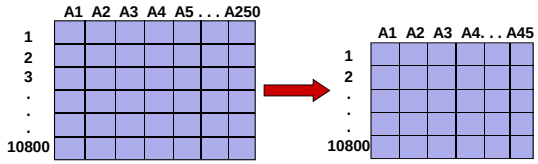
# *Métodos para Redução de Dimensionalidade*

**Stanley Robson de M. Oliveira**




# Índice da Aula

- Redução de dimensão:
  - Necessidade, motivação e aplicações.
- Principais Abordagens:
  - Extração de atributos (**não-Supervisionada**);
  - Seleção de atributos (**Supervisionada**).
- Métodos para extração de atributos:
  - Análise de Componentes Principais (**PCA**);
  - Projeção Aleatória (**Random Projection**);
  - Multidimensional Scaling (**MS**);
  - Decomposição do Valor Singular (**SVD**);
  - Latent Semantic Indexing (**LSI**).



# Por que redução de dimensão?

- Muitas técnicas de **aprendizado de máquina** e **mineração de dados** podem **não** ser **eficientes** para dados com **alta dimensionalidade**:
  - A maldição da dimensionalidade.
  - A **precisão** e a **eficiência** de uma consulta **degradam** rapidamente à medida em que a **dimensão aumenta**.
- A dimensão intrínseca pode ser menor.
  - Muitos **atributos** são **irrelevantes.**
  - **Exemplo**: o número de genes responsáveis por um certo tipo de doença pode ser menor.

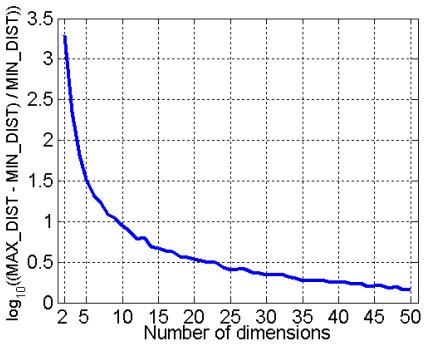


# Por que redução de dimensão? ...

- **Visualização**: projeção de dados com alta dimensionalidade em 2D ou 3D.
- **Compressão de dados**: eficiência no armazenamento e recuperação.
- **Remoção de ruído**: efeito positivo na acurácia de modelos e de consultas.

# Motivação

- Quando a **dimensionalidade aumenta**, os **dados** tornam-se progressivamente **esparsos** no espaço em que ocupam.
- **Definição de distância** entre pontos, que é critica para agrupamento e detecção de outliers, torna-se **menos significativa**.
- A análise de dados pode ficar **muito cara** se todos os atributos forem considerados.




- **500 pontos gerados aleatoriamente.**
- **Cálculo da diferença entre a distância max e min para os pares de pontos.**



# Motivação ...

- Os **alvos principais** do processo de **redução de dimensionalidade** são:
  - Melhorar a **performance** dos algoritmos de aprendizado de máquina.
  - Simplificar os **modelos de predição** e reduzir o **custo computacional** para "rodar" esses modelos.
  - Fornecer um **melhor entendimento** sobre os resultados encontrados, uma vez que existe um estudo prévio sobre o **relacionamento entre os atributos**.

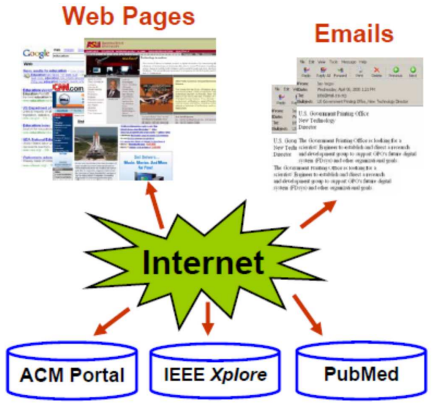


# Aplicações

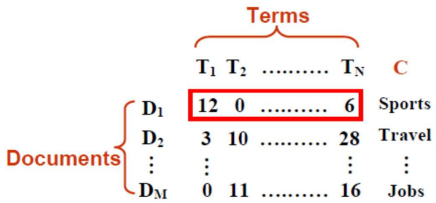
- Relacionamento com clientes (**CRM**).
- Mineração e/ou processamento de textos.
- Recuperação de informação em banco de imagens.
- Análise de dados de microarrays.
- Classificação de proteínas.
- Reconhecimento de face.
- Aplicações com dados meteorológicos.
- Química combinatorial.
- etc.



# Classificação de documentos



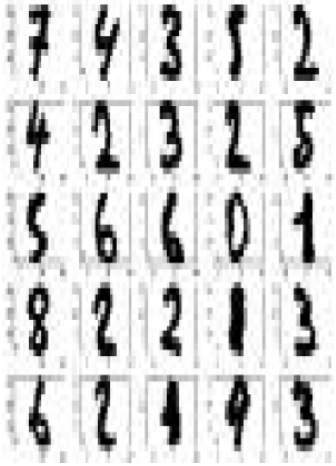

**Bibliotecas Digitais**




- **Tarefa**: classificar documentos em categorias.
- **Desafio**: milhares de termos.
- **Solução**: aplicar técnicas de redução de dimensão.

# Outros exemplos de aplicações

**Reconhecimento de face**



**Reconhecimento de dígitos manuscritos**



# Seleção de Atributos

- **IDEIA GERAL**: Processo que escolhe um **subconjunto ótimo de atributos** de acordo com uma função objetivo.
- **Objetivos**:
  - **Reduzir** dimensionalidade e **remover** ruído.
  - **Melhorar a performance** da mineração de dados:
    - Aumenta a velocidade do aprendizado.
    - Melhora a acurácia de modelos preditivos.
    - Facilita a compreensão dos resultados minerados.



# Extração de Atributos

- **IDEIA GERAL**: Ao invés de escolher um subconjunto de atributos, **define novas dimensões** em função de todos os atributos do conjunto original.
- Não considera o **atributo classe**, somente os atributos numéricos (**vetores de dados**).



# Extração de Atributos ...

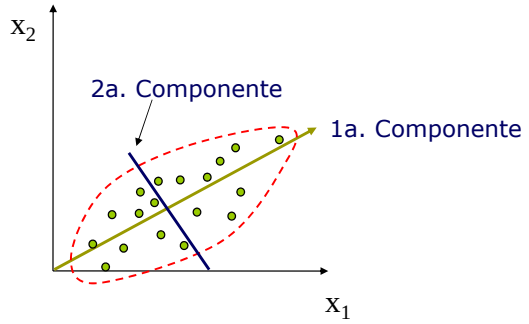
- **Ideia**:
  - Dado um conjunto de pontos no espaço *d*-dimensional,
  - Projetar esse conjunto de pontos num **espaço de menor dimensão**, preservando ao máximo as informações dos dados originais.
  - Em particular, escolher uma projeção que minimize o erro quadrático na reconstrução dos dados originais.
  - **Principais Métodos**:
    - Análise de Componentes Principais (**PCA**);
    - Projeção Aleatória (**Random Projection**);
    - Multidimensional Scaling (**MS**);
    - Decomposição do Valor Singular (**SVD**);
    - Latent semantic indexing (**LSI**).

# Seleção versus Extração

- **Extração de atributos**:
  - Todos os atributos originais são usados.
  - Os novos atributos são combinação linear dos atributos originais.
- **Seleção de atributos**:
  - Somente um subconjunto dos atributos originais são selecionados.
- **Atributos contínuos versus discretos**.



# Análise de Componentes Principais (PCA)




# Análise de Componentes Principais

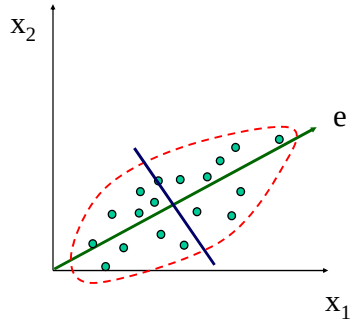
- Método para transformar **variáveis correlacionadas** em um conjunto de **variáveis não-correlacionadas** que melhor explica os **relacionamentos** entre os dados originais.
- Método para identificar as **dimensões** que exibem as maiores **variações** em um **conjunto de dados**.
- Método que possibilita encontrar a **melhor aproximação** dos dados originais usando um conjunto **menor de atributos**.



# PCA: Ideia Geral

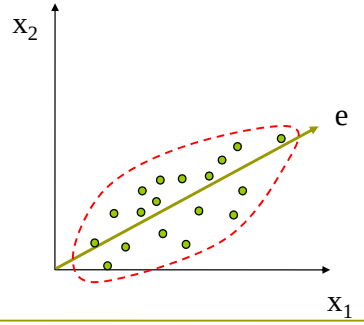
• A **linha verde** tem uma representação reduzida dos dados originais que **captura o máximo da variação** original dos dados.

• A segunda linha (**azul**), perpendicular à primeira (**verde**), captura menos variação nos dados originais.



# PCA: Redução de Dimensão

- O alvo é encontrar uma **projeção** que capture a **maior variância** possível nos dados.
- De uma forma geral:
  - Encontrar os **autovetores da matriz de covariância** dos dados. Os **autovetores** definem o novo espaço.






# Autovalores e Autovetores

- Dado um operador linear T: V → V, estamos interessados em um vetor v ∈ V e um escalar $\lambda \in \Re$ tais que T(v) = $\lambda$v.
- Neste caso T(v) será um vetor de mesma "**direção**" que v, ou melhor, T(v) e v estão sobre a mesma reta suporte.
- Um autovalor de uma matriz $A_{n \times n}$ é um escalar $\lambda$ tal que existe um vetor *v* (**não-nulo**), com $Av = \lambda v$, onde *v* é chamado de autovetor de A associado a $\lambda$.
- Podemos encontrar os **autovalores** $\lambda$ e **autovetores** *v* pela função característica definida como:

  $p(\lambda) = det(A - \lambda I)$ onde:

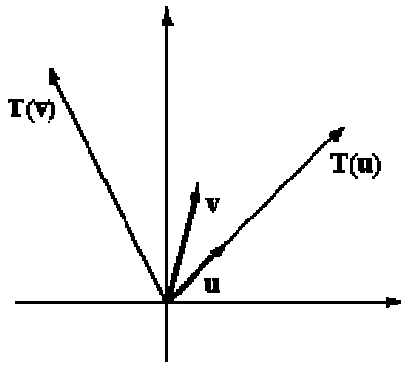
  — $p(\lambda)$ é chamado de **polinômio característico** de *A*;

  — *I* é a **matriz identidade**.



# Interpretação geométrica em $\Re^2$

- u é **autovetor** de T pois $\exists \lambda \in \Re$ / T(u) = $\lambda$u.
- v **não é autovetor** de T pois $\nexists \lambda \in \Re$ / T(v) = $\lambda$v.






# Exemplo: Autovalores e Autovetores

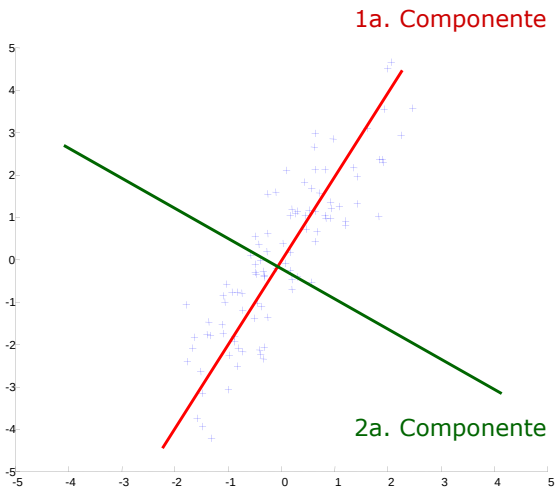
- Calcular os **autovalores** e **autovetores** da matriz: $A = \begin{pmatrix} 4 & 5 \\ 2 & 1 \end{pmatrix}$
- T: $\Re^2 \rightarrow \Re^2$ $\quad (x, y) \rightarrow (4x + 5y, 2x + y)$
- **Cálculo dos autovalores**: det (A – $\lambda$I) = 0

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} 4 & 5 \\ 2 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \det\begin{pmatrix} 4-\lambda & 5 \\ 2 & 1-\lambda \end{pmatrix}$$

- det (A – $\lambda$I) = 0 $\Leftrightarrow$ (4 – $\lambda$ )(1 – $\lambda$ ) – 10 = 0 $\Leftrightarrow \lambda^2 - 5\lambda - 6 = 0$
- Os **autovalores** são $\lambda_1 = -1$ e $\lambda_2 = 6$.
- Para cada **autovalor encontrado**, resolvemos o sistema linear (A – $\lambda$I)v = 0. Os respectivos **autovetores** são: $v_1 = (-1, 1)$ e $v_2 = (5/2, 1)$.

# Redução de Dimensão: PCA …

- As componentes principais são vetores **ortogonais**.
- Minimizar o erro quadrático (**Root Mean Square**).
- **RMS** representa a diferença entre os pontos originais e os novos pontos calculados pela transformação.






# PCA: Algoritmo

- **Algoritmo PCA**:
  - X ← Matriz de dados (N x d), em que cada linha é um vetor $x_n$.
  - X ← Em cada linha, subtrair a média *x* de cada elemento *do* vetor $x_n$ em X.
  - Σ ← matriz de covariância de X.
  - Encontrar os autovalores e autovetores de Σ.
  - PC's ← os *K* autovetores com os maiores autovalores.

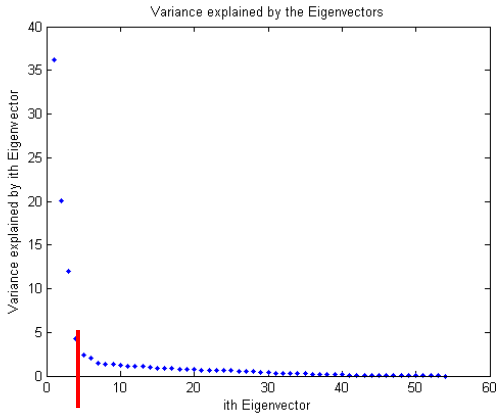


# Algoritmo PCA no Matlab

```
% generate data
Data = mvnrnd([5, 5],[1 1.5; 1.5 3], 100);
figure(1); plot(Data(:,1), Data(:,2), '+');

%center the data
for i = 1:size(Data,1)
  Data(i, :) = Data(i, :) - mean(Data);
end

DataCov = cov(Data); %covariance matrix
[PC, variances, explained] = pcacov(DataCov); %eigen

% plot principal components
figure(2); clf; hold on;
plot(Data(:,1), Data(:,2), '+b');
plot(PC(1,1)*[-5 5], PC(2,1)*[-5 5], '-r')
plot(PC(1,2)*[-5 5], PC(2,2)*[-5 5], '-b'); hold off

% project down to 1 dimension
PcaPos = Data * PC(:, 1);
```

# Qual é o número ideal de componentes?

- Verifique a **distribuição dos autovalores**.
- Selecione um **número de autovetores** que **cubra 80-90% da variância**.



**Exemplo**: Dados sobre a eficiência de cana-de-açúcar para 20 municípios em SP, em 2002.

| Municipio | Chuva | Tmax | Rad_Sol | Def_Hid | Prod | Eficiencia |
|---|---|---|---|---|---|---|
| Barretos | 1347,8 | 30,2 | 18,8 | 391,6 | 91,5 | 0,02 |
| Campinas | 1335,2 | 28,1 | 18,2 | 354,6 | 89,9 | 0,72 |
| Campos do Jordão | 1494,6 | 26,0 | 18,2 | 339,2 | 84,0 | 0,01 |
| Caraguatatuba | 1558,0 | 28,2 | 17,9 | 327,4 | 87,0 | 0,40 |
| Cubatão | 1460,7 | 28,2 | 17,8 | 328,4 | 75,4 | 0,52 |
| Franca | 1415,3 | 28,6 | 18,7 | 391,4 | 90,3 | 0,82 |
| Ilha Solteira | 1198,7 | 31,0 | 18,8 | 382,9 | 101,2 | 0,50 |
| Itanhaem | 1456,0 | 28,3 | 17,9 | 325,2 | 84,5 | 0,72 |
| Leme | 1329,1 | 28,3 | 18,4 | 373,7 | 78,9 | 0,47 |
| Limeira | 1320,6 | 28,2 | 18,3 | 362,5 | 79,2 | 0,54 |
| Ourinhos | 1348,7 | 29,5 | 19,0 | 332,7 | 99,5 | 0,70 |
| Paulínia | 1329,3 | 28,1 | 18,2 | 358,6 | 78,5 | 0,35 |
| Piracicaba | 1318,3 | 28,3 | 18,3 | 356,2 | 84,9 | 0,75 |
| Presidente Prudente | 1349,7 | 29,9 | 18,8 | 333,8 | 109,4 | 0,03 |
| Ribeirão Preto | 1368,5 | 29,2 | 18,7 | 388,0 | 88,8 | 0,67 |
| Rio Claro | 1323,6 | 28,3 | 18,4 | 366,4 | 88,3 | 0,28 |
| São João da Boa Vista | 1352,6 | 27,8 | 18,4 | 387,2 | 86,1 | 0,68 |
| São José do Rio Preto | 1255,3 | 30,2 | 18,7 | 380,8 | 92,0 | 0,67 |
| São Paulo | 1417,2 | 27,8 | 17,9 | 334,5 | 79,1 | 0,02 |
| Ubatuba | 1614,2 | 28,0 | 17,9 | 336,2 | 80,5 | 0,03 |

# Resultado da Análise (Minitab)

Análise de autovalores e autovetores matriz de correlação

| | | | | | | |
|---|---|---|---|---|---|---|
| Autovalor | 3,3115 | 1,1964 | 0,7118 | 0,3769 | 0,3075 | 0,0958 |
| Proporção | 0,552 | 0,199 | 0,119 | 0,063 | 0,051 | 0,016 |
| Acumulado | 0,552 | 0,751 | 0,870 | 0,933 | 0,984 | 1,000 |

| Variável | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| Chuva | -0,449 | -0,235 | 0,126 | -0,692 | -0,492 | -0,085 |
| Tmax | 0,458 | -0,233 | 0,130 | 0,360 | -0,759 | -0,113 |
| Rad_Sol | 0,506 | -0,139 | -0,083 | -0,390 | 0,255 | -0,708 |
| Defic_hidr | 0,370 | 0,418 | -0,600 | -0,368 | -0,238 | 0,369 |
| Produtividade | 0,403 | -0,508 | 0,325 | -0,272 | 0,239 | 0,585 |
| Eficiencia | 0,184 | 0,662 | 0,703 | -0,177 | -0,055 | -0,003 |

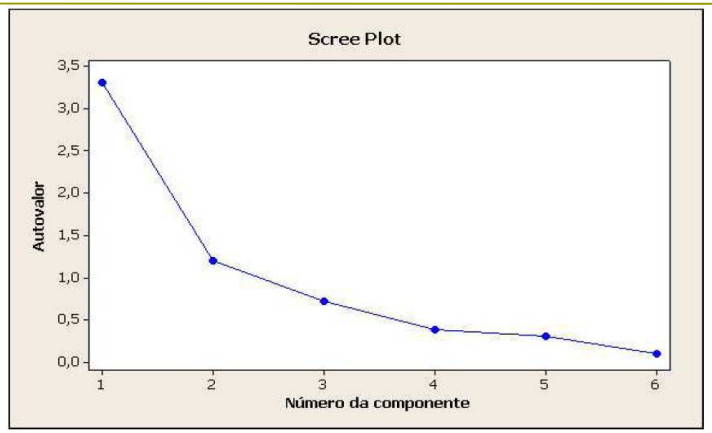
É possível explicar **aproximadamente 90%** da variabilidade total observada nos dados com apenas **três componentes principais**:

$Z_1$ = -0.45 Chuva + 0.46 Tmax + 0.51 Rad_Sol + 0.37 Def_Hidr + 0.40 Prod + 0.18 Efic

$Z_2$ = -0.24 Chuva - 0.23 Tmax - 0.14 Rad_Sol + 0.42 Def_Hidr - 0.51 Prod + 0.66 Efic

$Z_3$ = 0.13 Chuva + 0.13 Tmax - 0.09 Rad_Sol - 0.60 Def_Hidr + 0.33 Prod + 0.70 Efic

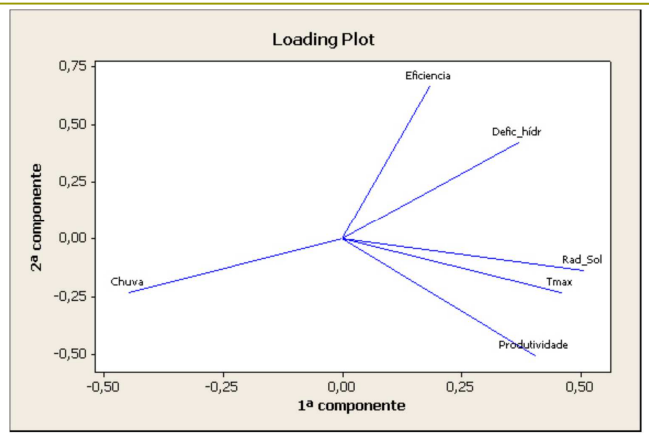
# Resultado da Análise (Minitab) ...




A Figura acima evidencia a importância das três primeiras componentes, em relação às demais (quanto **maior é o autovalor**, **maior será a porcentagem de variação explicada** pela componente correspondente).

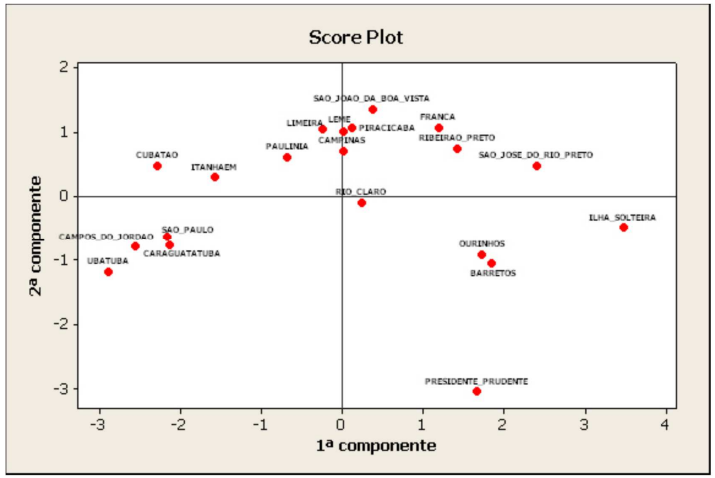


# Resultado da Análise (Minitab) ...




A Figura acima ilustra **geometricamente** como as **seis variáveis** do exemplo podem ser adequadamente representadas por **duas componentes principais** ($Z_1$ e $Z_2$).

# Resultado da Análise (Minitab) ...




As **duas componentes** descrevem, de uma forma geral, características das **cidades vizinhas** que possuem **climas** e **condições de cultivo semelhantes**.

# PCA: Descarte de Atributos

- Dados *N* vetores no espaço *n*-dimensional, encontrar $k \leq n$ vetores ortogonais (***componentes principais***) que podem ser melhor usados para representar os dados.
- **Passos**:
  - Normalizar dados originais: todos atributos ficam na mesma faixa (**intervalo**).
  - Calcular *k* vetores ortogonais, i.e., **componentes principais**.
  - Cada vetor (**original**) é uma combinação linear dos *k* vetores (**componentes principais**).
  - As componentes principais são ordenadas (**ordem decrescente**) representando a "**significância**" ou "**força**".
  - Como as componentes são ordenadas, o tamanho dos dados pode ser reduzido eliminando-se as **componentes fracas**, i.e., aquelas com baixa variância.



# PCA: Descarte de Atributos ...

- **IDEIA GERAL**:
  - Executar **PCA** sobre uma matriz de correlação com ***p*** variáveis.
  - Inicialmente, ***k*** variáveis são selecionadas (**retidas**).
  - No final, (***p*** – ***k***) variáveis serão descartadas.



# PCA: Descarte de Atributos ...

- **Algoritmo**:
  - Selecione o autovetor (**componente**) correspondente ao menor autovalor;
  - Rejeite a variável com maior coeficiente (**valor absoluto**) na componente.
  - O processo continua até que os (***p*** – ***k***) menores autovalores sejam considerados.

**Princípio para descarte de variáveis**: uma componente com baixo autovalor é menos importante e, consequentemente, a variável que domina essa componente deve ser menos importante ou redundante.

# PCA: Descarte de Atributos ...

- A escolha de ***k*** (**variáveis retidas**):
  - **Jolliffe (1972)** recomenda o *threshold* $\lambda_0 = 0.70$ depois de investigar vários conjuntos de dados;
  - Qualquer autovalor $\lambda_0 \leq 0.70$ **contribui muito pouco** para a explicação dos dados.

**Jolliffe, I. T. (1972)**. Discarding variables in principal component analysis I: artificial data. Appl. Statist., **21**, 160-173.

**Jolliffe, I. T. (1973)**. Discarding variables in principal component analysis II: real data. Appl. Statist., **22**, 21-31.



# PCA: Descarte de Atributos ...

**Dataset: IRIS**

| Variáveis | Componentes Principais (Autovetores) | | | |
|---|---|---|---|---|
| | V1 | V2 | V3 | V4 |
| sepallength | 0.5224 | -0.3723 | 0.7210 | 0.2620 |
| sepalwidth | -0.2634 | -0.9256 | -0.2420 | -0.1241 |
| petallength | 0.5813 | -0.0211 | -0.1409 | -0.8012 |
| petalwidth | 0.5656 | -0.0654 | -0.6338 | 0.5235 |
| | | | | |
| Autovalor $\lambda_i$ | 2.91082 | 0.92122 | 0.14735 | 0.02061 |
| Proporção | 0.7277 | 0.23031 | 0.03684 | 0.00515 |
| % acumulado | 0.7277 | 0.95801 | 0.99480 | 1 |

$\lambda_i < 0.70$

- ✓ **Variáveis descartadas**: petallength, sepallength.
- ✓ **Variáveis retidas**: sepalwidth, petalwidth.