Descritores de textura

PROF. CESAR HENRIQUE COMIN

Descritores de textura

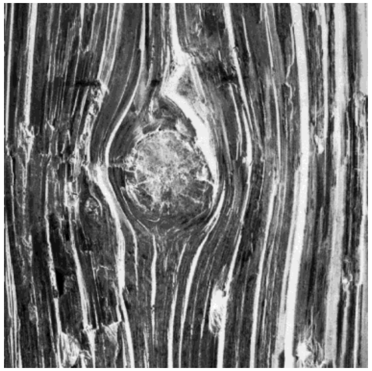
- Vimos em outras aulas o conceito de limiarização, que consiste em segmentar objetos em imagens de acordo com o nível de intensidade dos pixels.
- Mas muitas vezes nós reconhecemos objetos não por causa da intensidade dos pixels, mas por causa da **textura** dos mesmos.

Descritores de textura

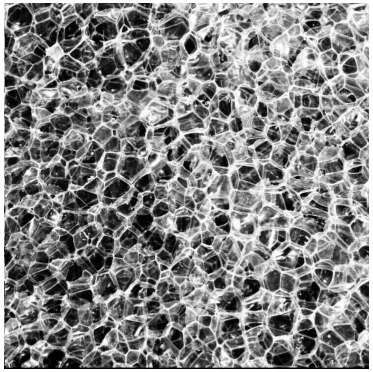
- No caso de texturas, utilizamos a informação de como o pixel se relaciona com os seus vizinhos para definirmos se ele pertence a determinado objeto.

Descritores de textura

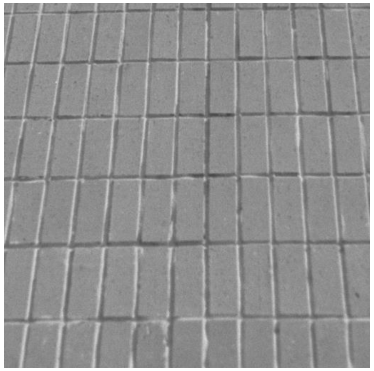
No caso abaixo, pixels próximos horizontalmente possuem fortes diferenças de intensidade. Essa diferença possui menor escala na imagem da esquerda do que na da direita



No caso abaixo, há uma escala característica na qual pixels tendem a ter altas diferenças de intensidade



No caso abaixo, pixels próximos tendem a ter pouca diferença de intensidade



Gray Level Co-Occurrence Matrix (GLCM)

- Dentre os diversos descritores de textura existentes, veremos um dos mais conhecidos;
- Esse descritor se baseia na chamada Gray Level Co-Occurrence Matrix. Essa matriz descreve como é o relacionamento entre as intensidades de pixels próximos em uma imagem;
- O principal parâmetro do método é o deslocamento $D = (r, c)$ que será utilizado na criação da matriz

Gray Level Co-Occurrence Matrix (GLCM)

- Cada linha e coluna da matriz de coocorrência G representa um nível de intensidade da imagem. Portanto, se a imagem possuir 256 níveis de intensidade, a matriz G possui 256 linhas e colunas.
- Dado um valor de deslocamento $D = (r, c)$, para cada pixel na posição (i, j) da imagem, verificamos a intensidade do pixel na posição $(i + r, j + c)$
- Seja $I_r = I(i, j)$ e $I_s = I(i + r, j + c)$ as intensidades dos pixels em questão, incrementamos em 1 o respectivo elemento $G(I_r, I_s)$ da matriz de coocorrência
- Após percorrermos todos os pixels da imagem, dividimos cada elemento de G pela soma de todos os valores de G, definindo uma nova matriz de coocorrência normalizada P

Gray Level Co-Occurrence Matrix (GLCM) - Exemplo

Imagem com intensidades no intervalo [0,7]

0	1	5	3	2
2	0	3	7	5
6	6	5	4	4
1	0	1	1	1
2	5	2	3	2

Deslocamento utilizado:

$$D = (1,2)$$

Matriz de co-ocorrência

	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Gray Level Co-Occurrence Matrix (GLCM) - Exemplo

Imagem com intensidades no intervalo [0,7]

0	1	5	3	2
2	0	3	7	5
6	6	5	4	4
1	0	1	1	1
2	5	2	3	2

Deslocamento utilizado:

$$D = (1,2)$$

Matriz de co-ocorrência

	0	1	2	3	4	5	6	7
0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Gray Level Co-Occurrence Matrix (GLCM) - Exemplo

Imagem com intensidades no intervalo [0,7]

0	1	5	3	2
2	0	3	7	5
6	6	5	4	4
1	0	1	1	1
2	5	2	3	2

Deslocamento utilizado:

$$D = (1,2)$$

Matriz de co-ocorrência

	0	1	2	3	4	5	6	7
0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Gray Level Co-Occurrence Matrix (GLCM) - Exemplo

Imagem com intensidades no intervalo [0,7]

0	1	5	3	2
2	0	3	7	5
6	6	5	4	4
1	0	1	1	1
2	5	2	3	2

Deslocamento utilizado:

$$D = (1,2)$$

Matriz de co-ocorrência

	0	1	2	3	4	5	6	7
0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Gray Level Co-Occurrence Matrix (GLCM) - Exemplo

Imagem com intensidades no intervalo [0,7]

0	1	5	3	2
2	0	3	7	5
6	6	5	4	4
1	0	1	1	1
2	5	2	3	2

Deslocamento utilizado:

$$D = (1,2)$$

Matriz de co-ocorrência

	0	1	2	3	4	5	6	7
0	0	0	0	2	1	0	0	0
1	0	0	2	0	0	0	0	1
2	0	0	0	0	0	1	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	0	0	0	0
5	0	1	0	0	0	1	0	0
6	0	2	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Gray Level Co-Occurrence Matrix (GLCM) - Exemplo

Imagem com intensidades no intervalo [0,7]

0	1	5	3	2
2	0	3	7	5
6	6	5	4	4
1	0	1	1	1
2	5	2	3	2

Deslocamento utilizado:

$$D = (1,2)$$

Matriz de co-ocorrência

	0	1	2	3	4	5	6	7
0	0	0	0	2	1	0	0	0
1	0	0	2	0	0	0	0	1
2	0	0	0	0	0	1	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	0	0	0	0
5	0	1	0	0	0	1	0	0
6	0	2	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Soma da matriz: 12

Gray Level Co-Occurrence Matrix (GLCM) - Exemplo

Imagem com intensidades no intervalo [0,7]

0	1	5	3	2
2	0	3	7	5
6	6	5	4	4
1	0	1	1	1
2	5	2	3	2

Deslocamento utilizado:

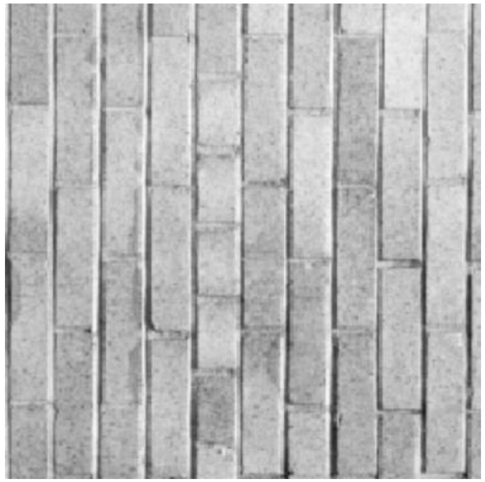
$$D = (1,2)$$

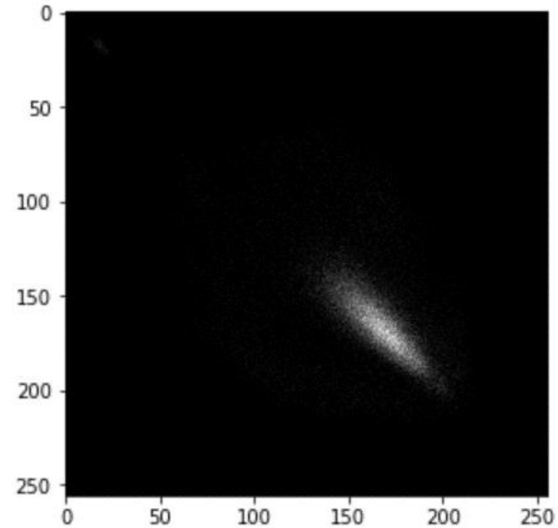
Matriz de co-ocorrência normalizada

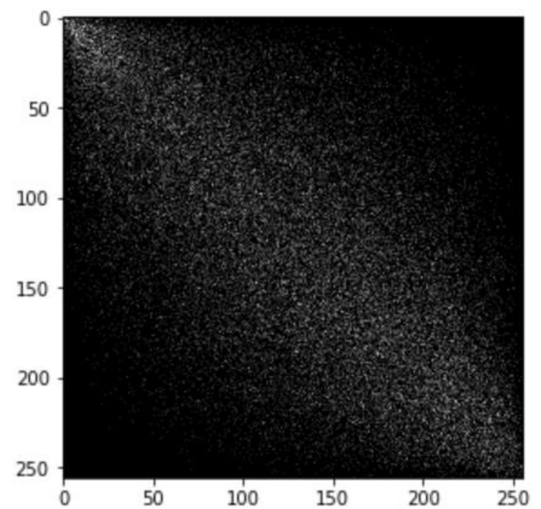
	0	1	2	3	4	5	6	7
0	0	0	0	0.17	0.08	0	0	0
1	0	0	0.17	0	0	0	0	0.08
2	0	0	0	0	0	0.08	0	0
3	0	0	0	0	0.08	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0.08	0	0	0	0.08	0	0
6	0	0.17	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Soma da matriz: 12

Exemplos de GLCMs







Gray Level Co-Occurrence Matrix - Propriedades

Tendo obtido a matriz de coocorrência, podemos calcular diversas propriedades dessa matriz. Por exemplo:

Nome	Fórmula
Máxima probabilidade	$\max_{i,j} p_{ij}$
Contraste	$\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} (i-j)^2 p_{ij}$
Uniformidade	$\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} p_{ij}^2$
Homogeneidade	$\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} \frac{p_{ij}}{1+\lvert i-j \rvert}$
Entropia	$-\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} p_{ij} \log_2(p_{ij})$
Correlação	$\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} \frac{(i-m_r)(j-m_c)p_{ij}}{\sigma_r \sigma_c}$

O valor p_{ij} é o elemento (i, j) da matriz de coocorrência normalizado pela soma da matriz. Ele pode ser interpretado como uma probabilidade

Gray Level Co-Occurrence Matrix - Propriedades

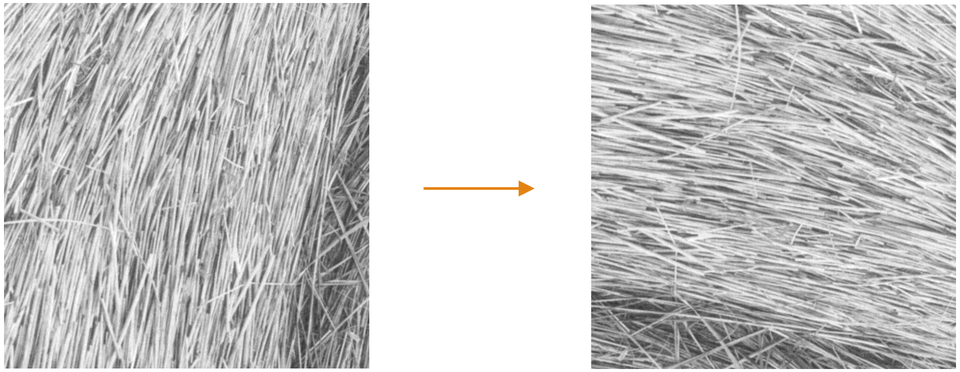
- Para uma dada imagem, as propriedades da matriz de coocorrência correspondem a diferentes características da textura presente na imagem
- Tais propriedades são globais, isto é, cada propriedade define um único valor associado à imagem

Gray Level Co-Occurrence Matrix - Propriedades

- As propriedades calculadas são, aproximadamente, invariantes à translação da imagem

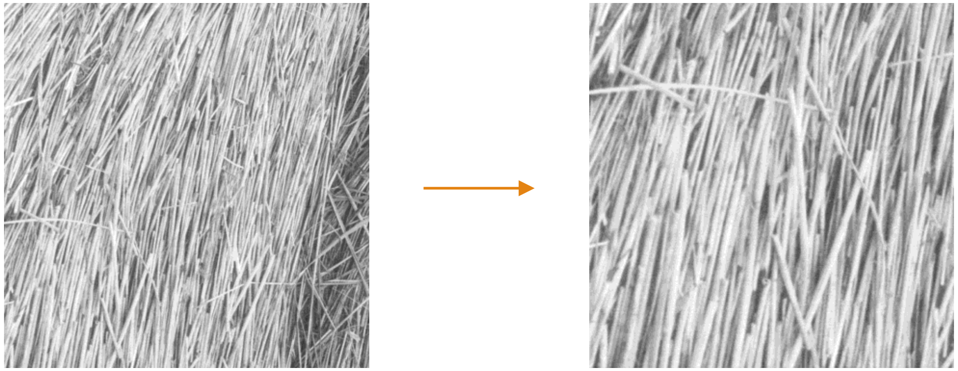
Gray Level Co-Occurrence Matrix - Propriedades

- As propriedades calculadas são, aproximadamente, invariantes à translação da imagem
- As propriedades não são invariantes à rotação
  - Possível estratégia: utilizar deslocamentos em diversos ângulos e somar os valores



Gray Level Co-Occurrence Matrix - Propriedades

- As propriedades calculadas são, aproximadamente, invariantes à translação da imagem
- As propriedades não são invariantes à rotação
    - Possível estratégia: utilizar deslocamentos em diversos ângulos e somar os valores
- As propriedades não são invariantes à escala
    - Possível estratégia: mesmo que acima, mas utilizando diferentes distâncias de deslocamento (ex: (0, 1), (0, 5), (0, 10))

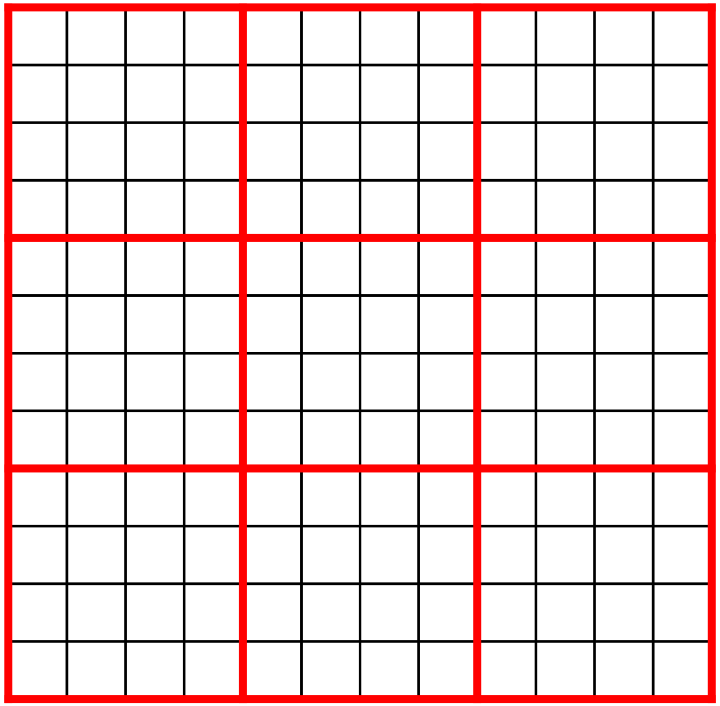


Gray Level Co-Occurrence Matrix - Propriedades

Notebook “**Graylevel Co-Ocurrence Matrix**”, seção 1 e 2

Análise local de textura usando GLCM

- Podemos analisar a textura em diferentes regiões de uma imagem utilizando GLCM
- Para isso, uma estratégia é dividirmos a imagem em regiões, e calcularmos a GLCM e suas propriedades para cada região



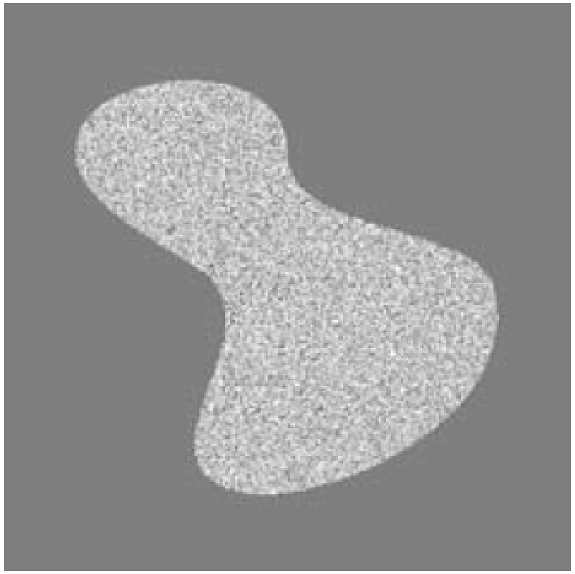
Análise local de textura usando GLCM

- A análise local de textura pode ser utilizada para segmentação de imagens
- Por exemplo:
    1. Podemos subdividir uma imagem em pequenas regiões de tamanho $R \times R$
    2. A GLCM e suas propriedades são calculadas para cada região
    3. Para cada propriedade podemos criar uma nova imagem onde a intensidade de cada pixel é o valor da propriedade.
    4. Podemos então aplicar uma limiarização na imagem criada no passo 3.

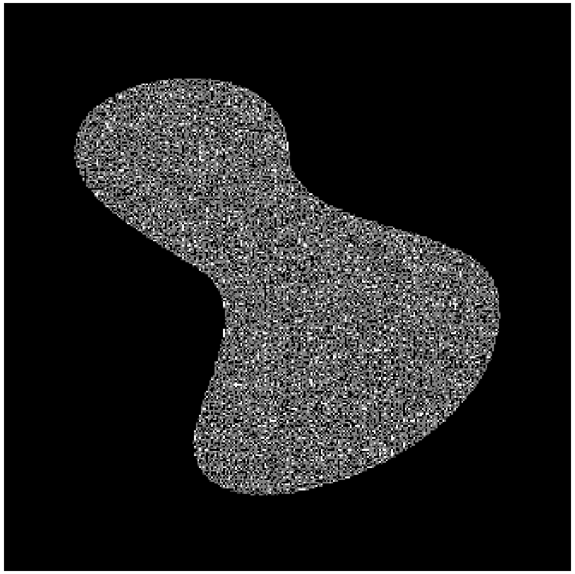
Segmentação de imagens utilizando textura

- Segmentação por textura pode levar a resultados muito melhores do que simplesmente utilizar um limiar

Original



Limiarização de intensidade



Limiarização de propriedade de textura

Segmentação de imagens utilizando GLCM

Notebook “**Graylevel Co-Ocurrence Matrix**”, seção 3

Local Binary patterns (LBP)

- Outra técnica utilizada para descrever a textura de objetos
- Essa técnica possui como objetivo a classificação de imagens. Isto é, a ideia dela é definir para uma dada imagem um grande conjunto de propriedades que representem a imagem
- Tais propriedades são então utilizadas para classificação

Local Binary patterns (LBP)

- Dada uma imagem de entrada, analisamos cada pixel da imagem (exceto pixels de borda)
- Em seguida comparamos o valor do pixel central com cada vizinho. Se o pixel central for maior do que o vizinho, associamos o valor 0 com o vizinho. Caso contrário, o valor 1 é associado com o vizinho

0	1	5	3	2
2	4	3	7	5
6	7	5	4	4
1	0	1	1	1
2	5	2	3	2

→

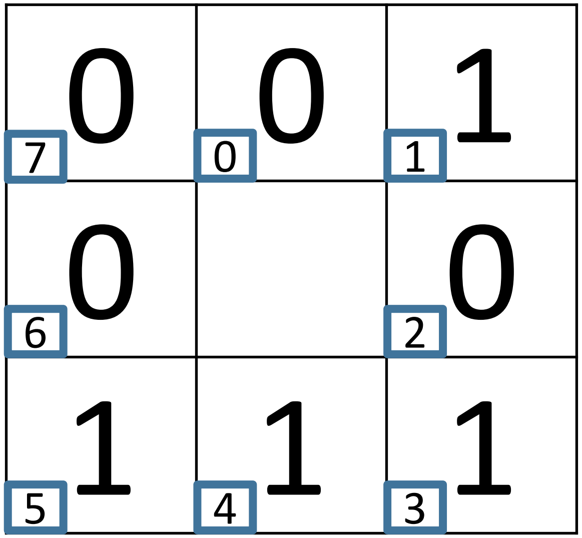
0	1	5
2	4	3
6	7	5

→

0	0	1
0		0
1	1	1

Local Binary patterns (LBP)

- O padrão obtido representa uma sequência de 8 bits
- Vamos definir que o bit 0 é o vizinho de cima do pixel referência, e que a sequência é percorrida no sentido horário
- Com isso, podemos considerar que a sequência obtida representa um número inteiro no intervalo [0,255]



0	0	1	1	1	0	1	0

$$2^6 + 2^4 + 2^3 + 2^2 = 92$$

Local Binary patterns (LBP)

- Em um novo array, armazenamos o valor obtido para cada pixel

Imagem

0	1	5	3	2
2	4	3	7	5
6	7	5	4	4
1	0	1	1	1
2	5	2	3	2

Valores LBP

	92	254	0	
	0	66	226	
	255	253	255	

Local Binary patterns (LBP)

- O histograma do array obtido é calculado, utilizando 256 caixas
- Esse histograma representa 256 propriedades associadas com a imagem
- Por exemplo, o valor na caixa 5 representa o número de pixels na imagem possuindo a vizinhança 00000101. Em outras palavras, o número de pixels para os quais o pixel é menor do que o pixel ao norte e à direita, e maior do que os demais

	92	254	0	
	0	66	226	
	255	253	255	

- Os 256 valores associados a cada imagem são utilizados para classificação

Análise de componentes Principais

Análise de componentes Principais (PCA)

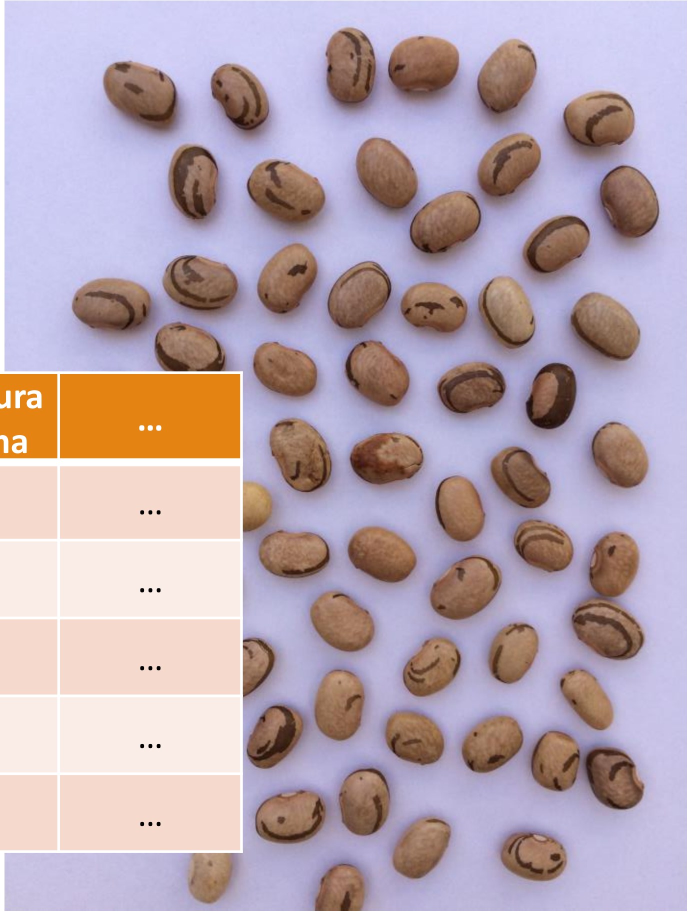
- É comum obtermos um grande número de propriedades associadas com uma dada imagem (por exemplo, área, perímetro, curvatura, elongação, contraste, correlação do GLCM, etc)
- Esse grande número de informações dificulta a interpretação dos dados obtidos
- A análise de componentes principais, do inglês Principal Component Analysis (PCA), é uma técnica estatística utilizada para reduzir o número de variáveis contidas em um conjunto de dados
- Essa redução é feita de forma que a maior parte da "informação" sobre os dados seja mantida
- Como definir informação?

PCA - Motivação

- Suponha que temos uma imagem contendo feijões
- Calculamos uma série de propriedades sobre os feijões, e organizamos o resultado em uma tabela

PCA - Motivação

- Suponha que temos uma imagem contendo feijões
- Calculamos uma série de propriedades sobre os feijões, e organizamos o resultado em uma tabela

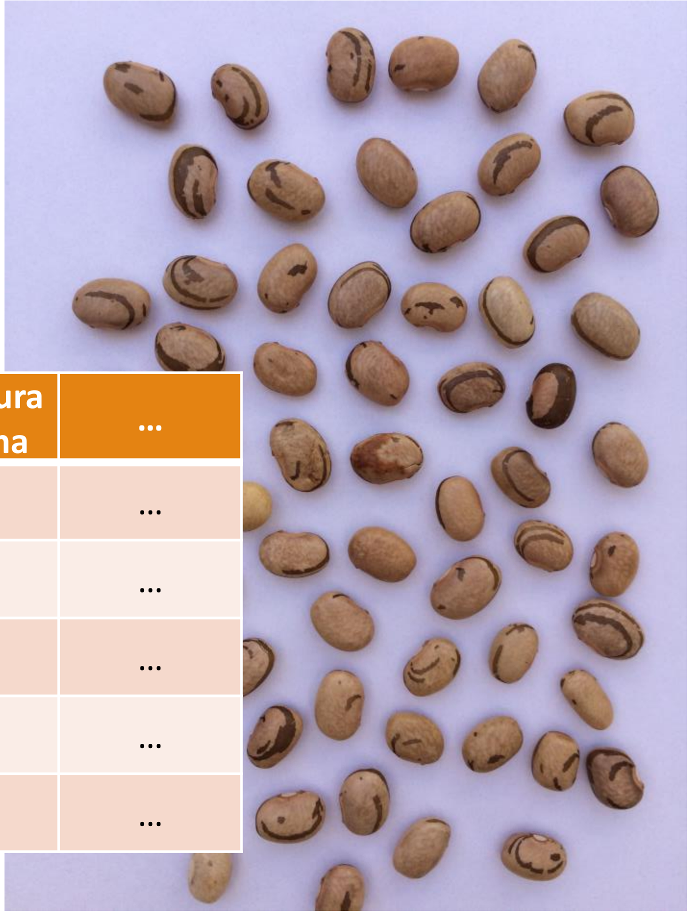


Id feijão	Área	Diâmetro	Perímetro	Elongação	Curvatura máxima	...
1	120	23	45	0.82	1.23	...
2	105	18	47	0.89	1.34	...
3	125	27	53	0.92	1.03	...
...	...	...	...	...	...	...
N	98	16	42	0.75	1.12	...

PCA - Motivação

- É esperado que a área, diâmetro e perímetro estejam fortemente relacionadas, pois um feijão de área maior tende a ter maior diâmetro e perímetro

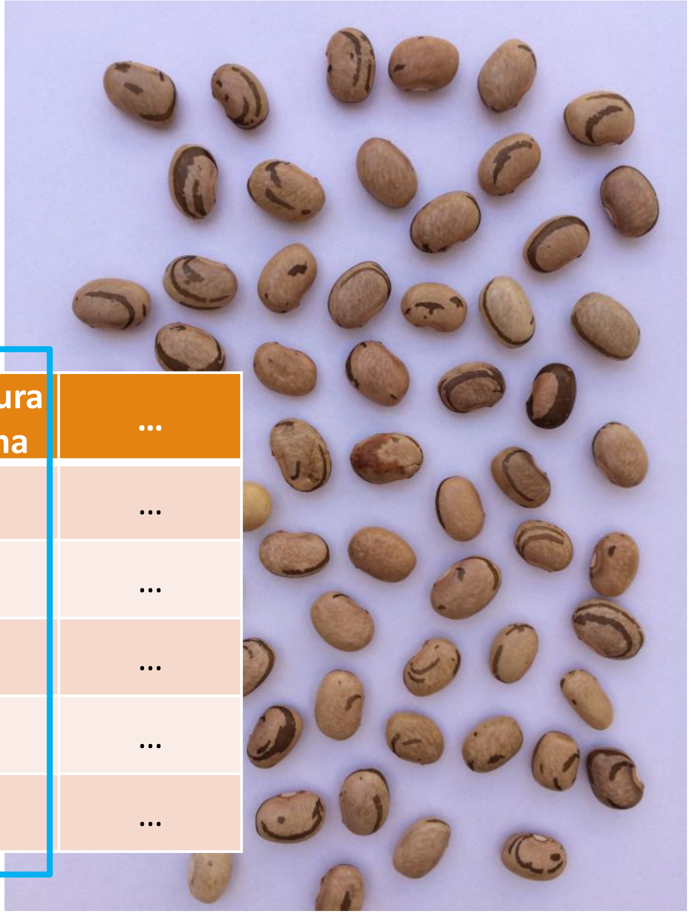
Id feijão	Área	Diâmetro	Perímetro	Elongação	Curvatura máxima	...
1	120	23	45	0.82	1.23	...
2	105	18	47	0.89	1.34	...
3	125	27	53	0.92	1.03	...
...	...	...	...	...	...	...
N	98	16	42	0.75	1.12	...



PCA - Motivação

- Dependendo de como a curvatura foi calculada, ela também terá uma forte relação com a área. Maior área tende a levar a menores valores de curvatura

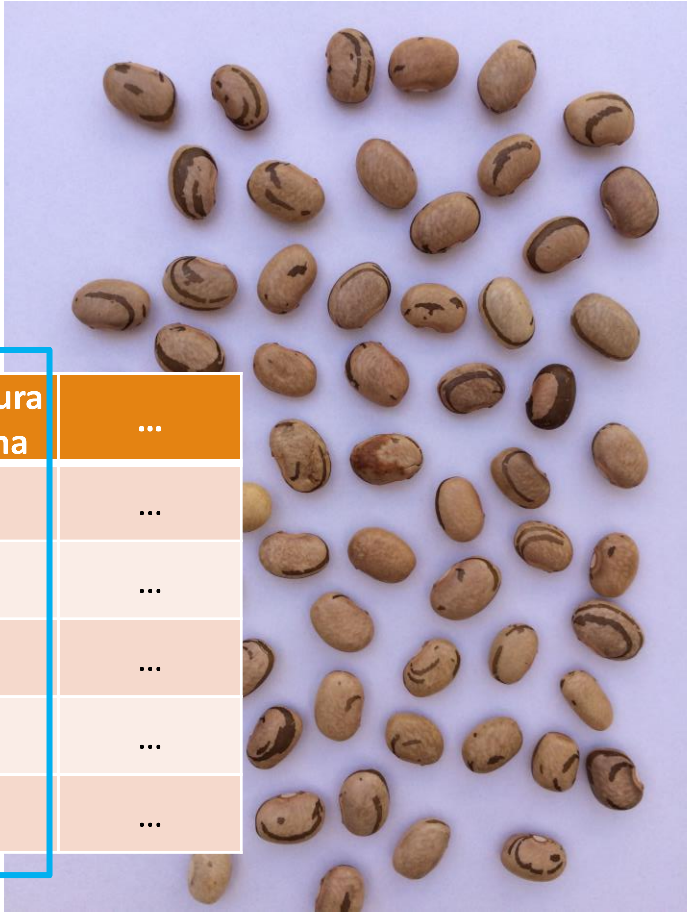
Id feijão	Área	Diâmetro	Perímetro	Elongação	Curvatura máxima	…
1	120	23	45	0.82	1.23	…
2	105	18	47	0.89	1.34	…
3	125	27	53	0.92	1.03	…
…	…	…	…	…	…	…
N	98	16	42	0.75	1.12	…



PCA - Motivação

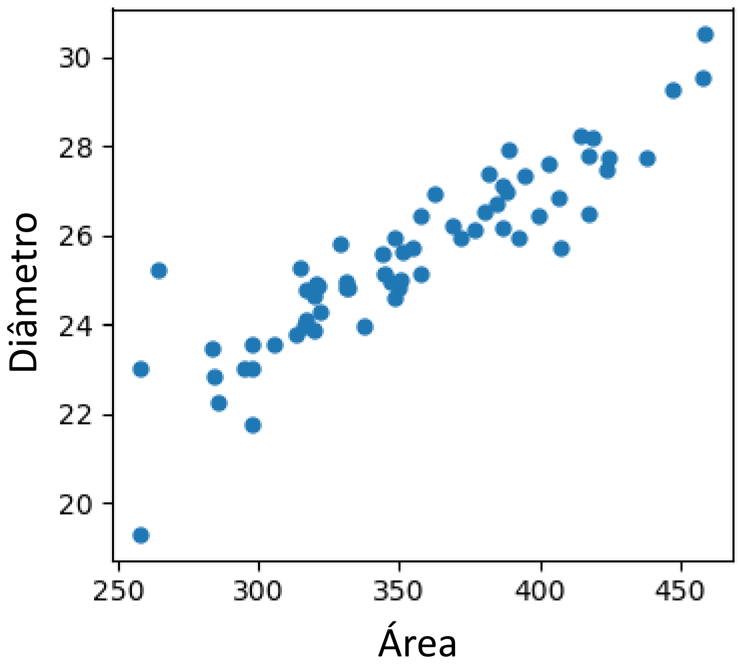
- Talvez elongação também esteja associado com a área? Por exemplo, talvez feijões maiores tendem a ser mais redondos

Id feijão	Área	Diâmetro	Perímetro	Elongação	Curvatura máxima	...
1	120	23	45	0.82	1.23	...
2	105	18	47	0.89	1.34	...
3	125	27	53	0.92	1.03	...
...	...	...	...	...	...	...
N	98	16	42	0.75	1.12	...



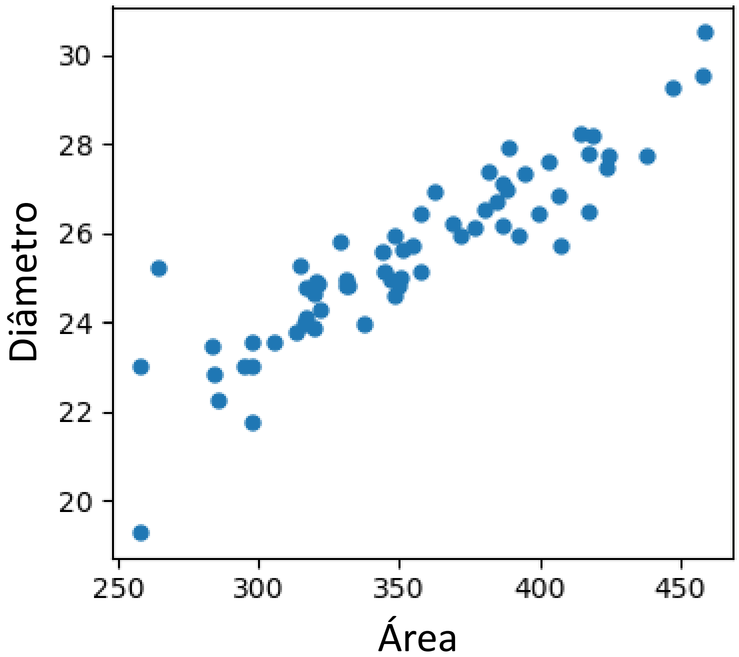
PCA - Motivação

- Por exemplo, se plotarmos a área dos feijões em função do diâmetro obteremos o seguinte gráfico:



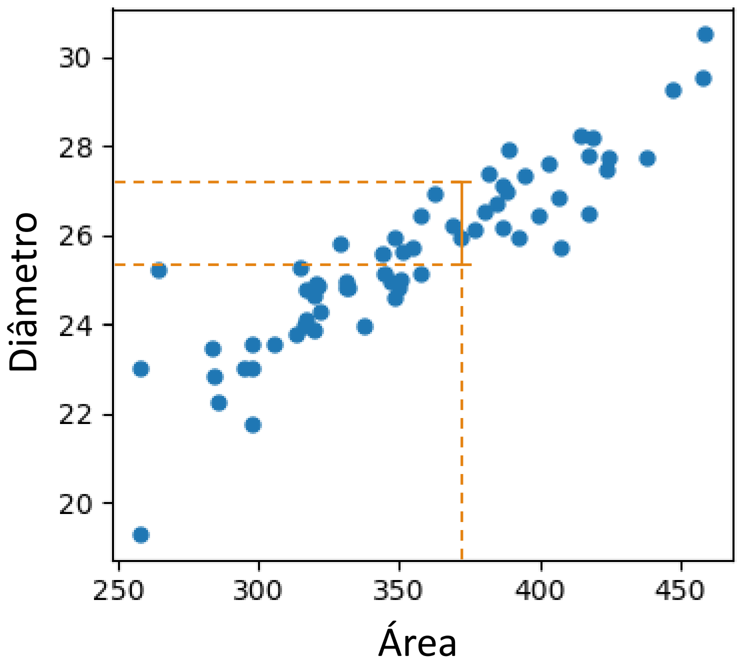
PCA - Motivação

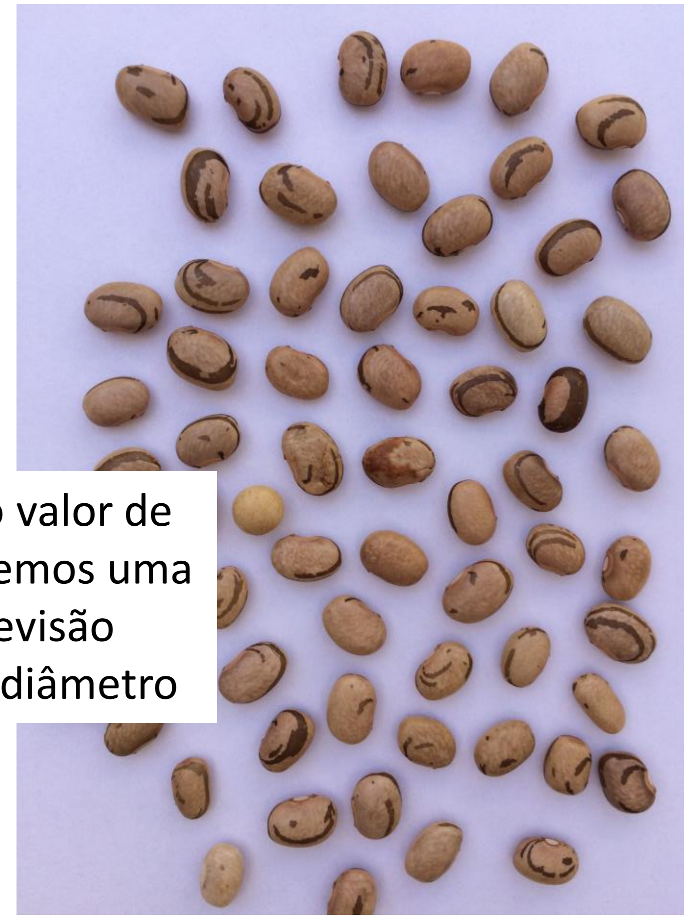
- Vemos que utilizar a área conjuntamente com o diâmetro para caracterizar os feijões é altamente redundante, pois uma propriedade já fornece quase toda a informação contida nas duas propriedades



PCA - Motivação

- Vemos que utilizar a área conjuntamente com o diâmetro para caracterizar os feijões é altamente redundante, pois uma propriedade já fornece quase toda a informação contida nas duas propriedades

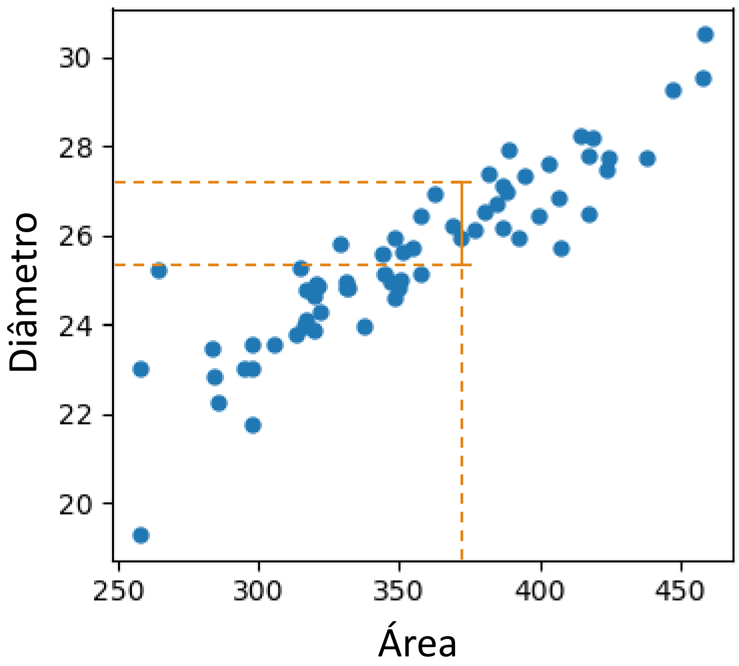




Dado o valor de área, temos uma boa previsão para o diâmetro

PCA - Motivação

- Como quantificar o grau de relacionamento entre duas propriedades?



Coeficiente de correlação de Pearson

- Existem diversas formas de quantificarmos o grau de relacionamento entre duas propriedades
- Uma das mais utilizadas é o coeficiente de correlação de Pearson

Coeficiente de correlação de Pearson

- Existem diversas formas de quantificarmos o grau de relacionamento entre duas propriedades
- Uma das mais utilizadas é o coeficiente de correlação de Pearson

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

X e Y: Representam duas propriedades
x_i e $\mathrm{y_i}$: O valor de cada propriedade medido para o objeto ou imagem i
$\mathrm{Cov}(X,Y)$: Covariância entre as propriedades X e Y
μ_X: Valor médio de X
σ_X: Desvio padrão de X
N: Número de objetos ou imagens
ρ: Coeficiente de correlação de Pearson

Covariância

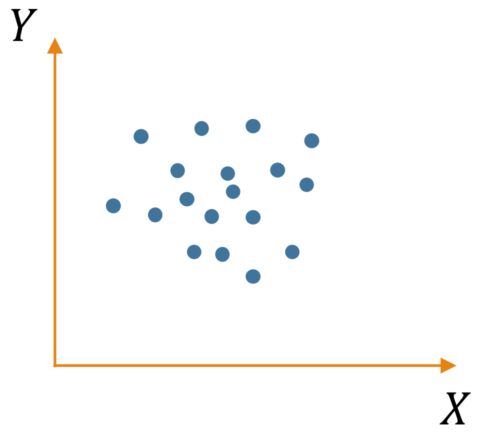
$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

A covariância quantifica o grau de variação de uma variável causada pela mudança de uma outra variável

Covariância

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

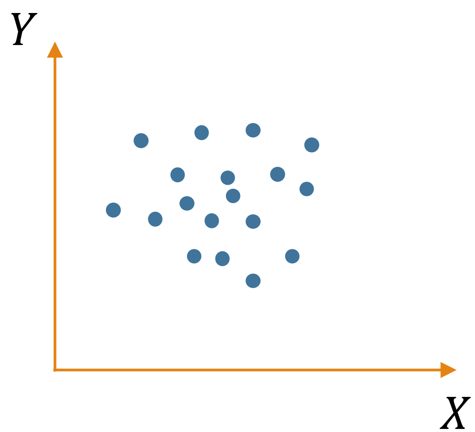
- Suponha que duas variáveis possuem os seguintes valores para um conjunto de imagens
- Intuitivamente, vemos que as duas variáveis fornecem informações distintas



Covariância

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

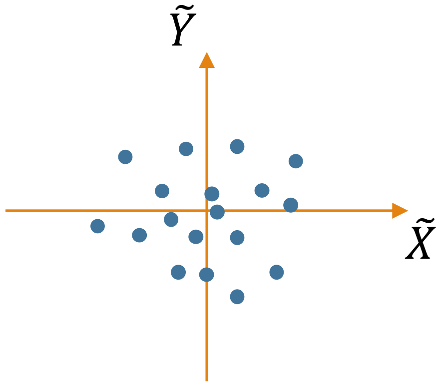
- Para calcular a covariância, primeiro transladamos os dados de forma que a média dos valores esteja na origem



Covariância

$$\text{Cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_i \tilde{y}_i$$

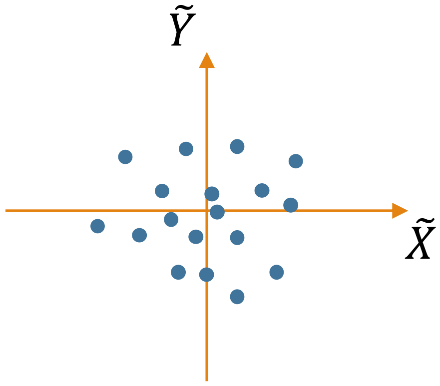
- Para calcular a covariância, primeiro transladamos os dados de forma que a média dos valores esteja na origem



Covariância

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\tilde{x}_i\tilde{y}_i$$

- Em seguida, a covariância é calculada através da multiplicação de todos os valores de $\tilde{x}_i$ e $\tilde{y}_i$



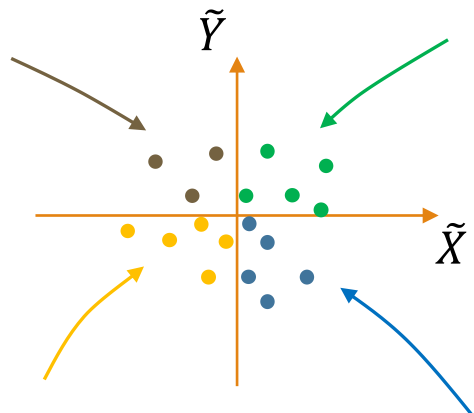
Covariância

$$\mathrm{Cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_i \tilde{y}_i$$

- Em seguida, a covariância é calculada através da multiplicação de todos os valores de $\tilde{x}_i$ e $\tilde{y}_i$

A multiplicação $\tilde{x}_i\tilde{y}_i$ resultará em valores negativos aqui ($x_i < 0$ e $y_i > 0$)

A multiplicação $\tilde{x}_i\tilde{y}_i$ resultará em valores positivos aqui ($x_i > 0$ e $y_i > 0$)



A multiplicação $\tilde{x}_i\tilde{y}_i$ resultará em valores positivos aqui ($x_i < 0$ e $y_i < 0$)

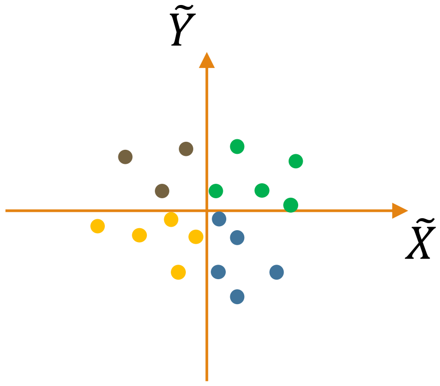
A multiplicação $\tilde{x}_i\tilde{y}_i$ resultará em valores negativos aqui ($x_i > 0$ e $y_i < 0$)

Covariância

$$\mathrm{Cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_i \tilde{y}_i$$

- Em seguida, a covariância é calculada através da multiplicação de todos os valores de $\tilde{x}_i$ e $\tilde{y}_i$

A soma de todas as contribuições resultará em $\mathrm{Cov}(X, Y) \approx 0$ para esses dados

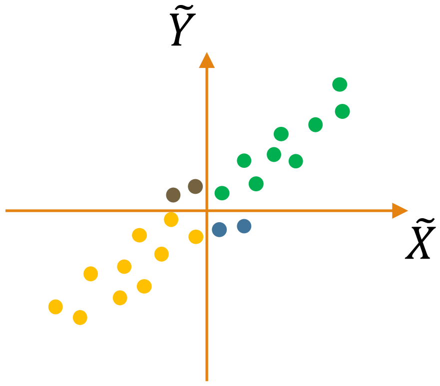


Covariância

$$\text{Cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_i \tilde{y}_i$$

- Em seguida, a covariância é calculada através da multiplicação de todos os valores de $\tilde{x}_i$ e $\tilde{y}_i$

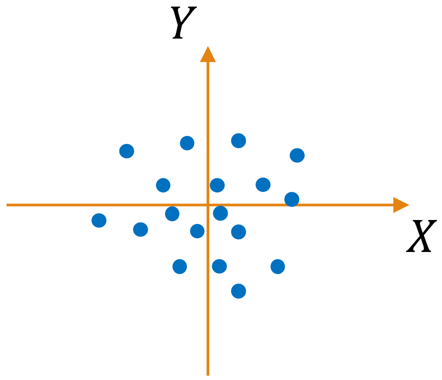
A soma de todas as contribuições resultará em $\text{Cov}(X, Y)$ alto para esses dados



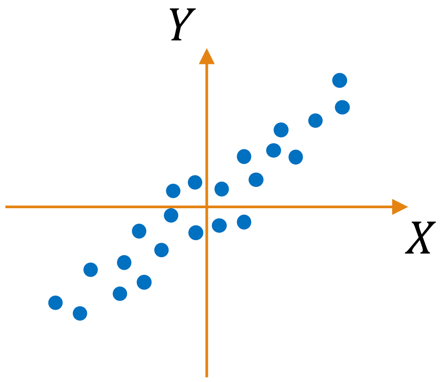
Covariância

Portanto, nós temos que

Covariância baixa



Covariância alta



De Covariância para Correlação de Pearson

- Um problema da covariância é que ela não é normalizada
- Propriedades possuindo valores altos tendem a resultar em altos valores de covariância
- O coeficiente de correlação de Pearson é dado pela covariância normalizada

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- A covariância é dividida pelo desvio padrão das propriedades, definindo um valor que não possui escala (o que é bom)

Standard score

- A transformação de uma propriedade

$$\hat{X} = \frac{X - \mu_X}{\sigma_X}$$

- É conhecida como escore padrão (standard score) ou z-score

Standard score

- Suponha que medimos a altura e peso de um conjunto de pessoas
- Temos que tomar cuidado ao comparar as duas variáveis, pois elas estão expressas em unidades diferentes (metros e quilograma)

Id pessoa	Altura (metros)	Peso (kg)
1	1.74	80
2	1.68	73
3	1.82	78
4	1.79	77
5	1.80	82
6	1.75	68
…	…	…

Standard score

- Suponha que medimos a altura e peso de um conjunto de pessoas
- Temos que tomar cuidado ao comparar as duas variáveis, pois elas estão expressas em unidades diferentes (metros e quilograma)

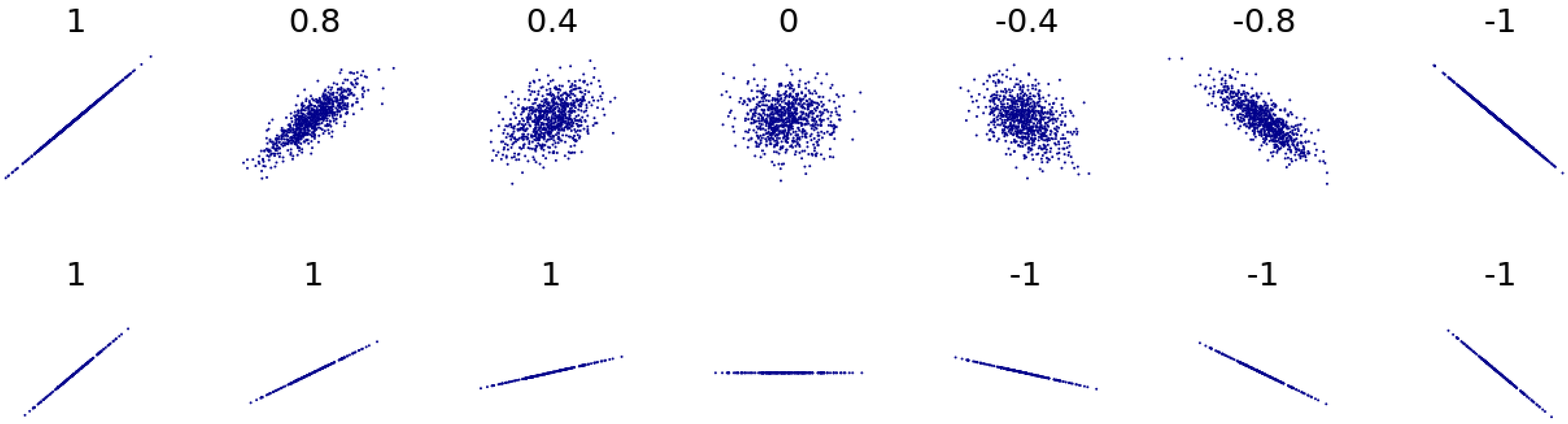
Id pessoa	Altura (metros)	Peso (kg)
1	1.74	80
2	1.68	73
3	1.82	78
4	1.79	77
5	1.80	82
6	1.75	68
…	…	…

Standard score →

Id pesosa	Altura (metros)	Peso (kg)
1	-0.5	0.79
2	-1.79	-0.72
3	1.22	0.36
4	0.57	0.14
5	0.79	1.22
6	-0.29	-1.79
…	…	…

Coeficiente de correlação de Pearson

- O coeficiente de correlação de Pearson quantifica o grau de relacionamento linear entre duas propriedades
- Ele possui valores no interval $[-1,1]$



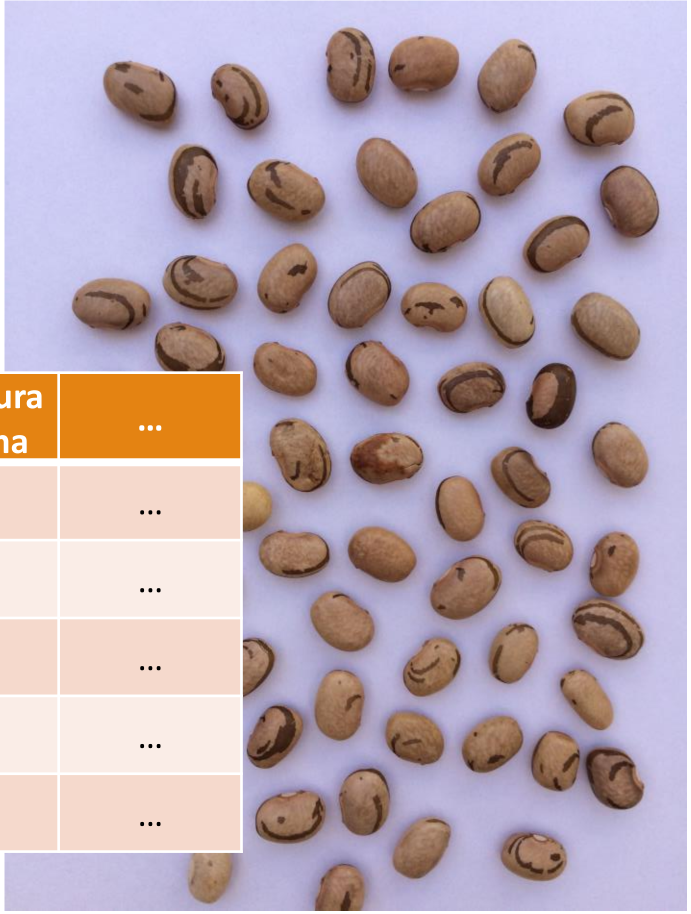
Coeficiente de correlação de Pearson

- Quando duas variáveis possuem alta correlação de Pearson, sabemos que uma delas pode ser excluída sem muita perda de informação

Coeficiente de correlação de Pearson

Estratégia:

- Calcule a correlação de Pearson entre todos os pares de propriedades;
- Para pares que possuem alta correlação, elimine uma das propriedades;
- Os atributos restantes proporcionam uma descrição "sucinta" dos feijões.



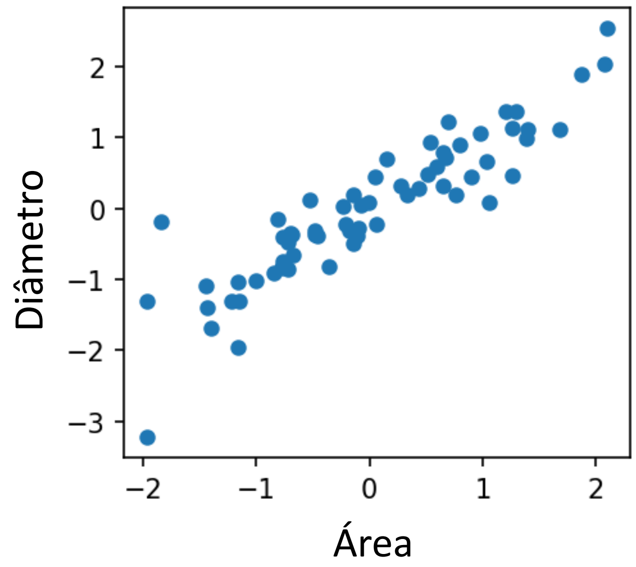
Id feijão	Área	Diâmetro	Perímetro	Elongação	Curvatura máxima	…
1	120	23	45	0.82	1.23	…
2	105	18	47	0.89	1.34	…
3	125	27	53	0.92	1.03	…
…	…	…	…	…	…	…
N	98	16	42	0.75	1.12	…

Análise de Componentes Principais (PCA)

- Análise de componentes principais é uma técnica utilizada para definir novos atributos, dados por combinações lineares entre os atributos originais
- Os novos atributos são definidos de forma que a maior parte da “informação” contida nos dados originais seja mantida

Análise de Componentes Principais (PCA)

No caso abaixo, qual novo atributo podemos definir que vai manter o máximo de informação sobre os dados?

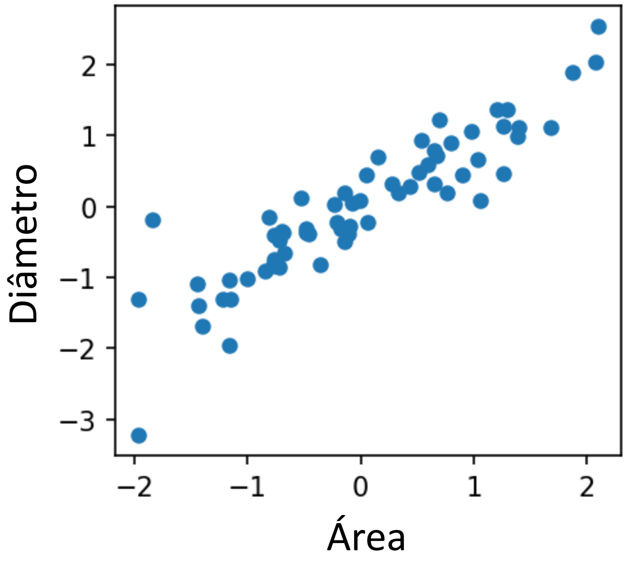


*Note que as medidas estão normalizadas (standard score) no gráfico

Análise de Componentes Principais (PCA)

Uma possibilidade:

$$f = 1 * \text{Área} + 0 * \text{Diâmetro}$$

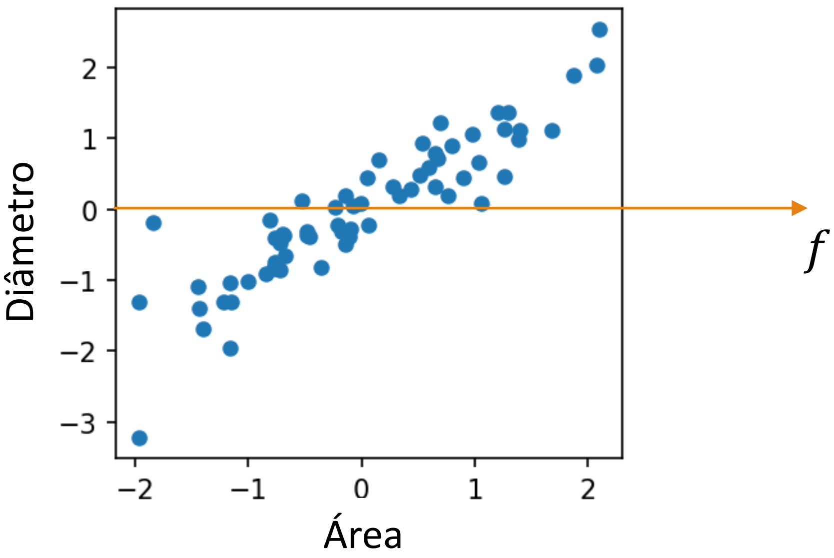


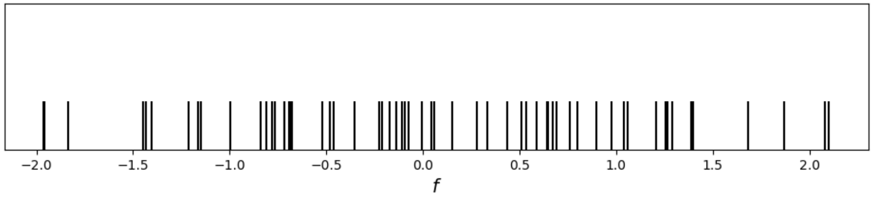
Análise de Componentes Principais (PCA)

Uma possibilidade:

$$f = 1 * \text{Área} + 0 * \text{Diâmetro}$$

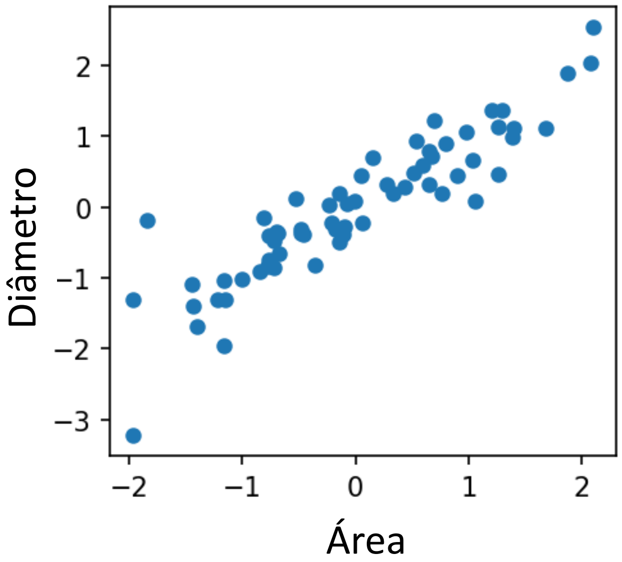
Esse novo atributo é simplesmente a área dos feijões





Análise de Componentes Principais (PCA)

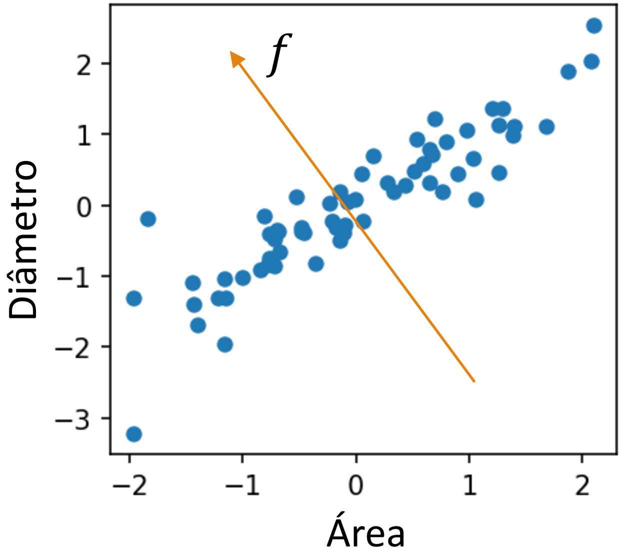
Talvez $f = -0.5 * \text{Área} + 0.5 * \text{Diâmetro}$?

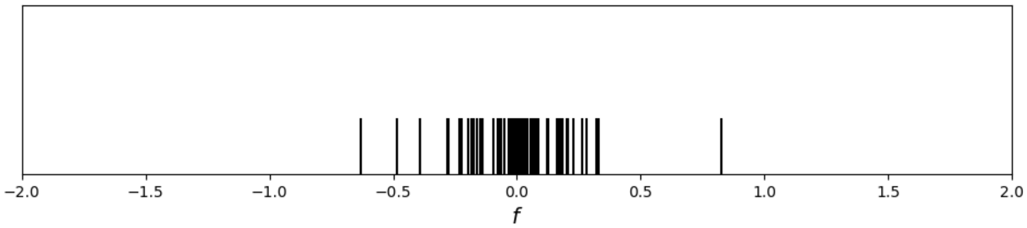


Análise de Componentes Principais (PCA)

Talvez $f = -0.5 * \text{Área} + 0.5 * \text{Diâmetro}$?

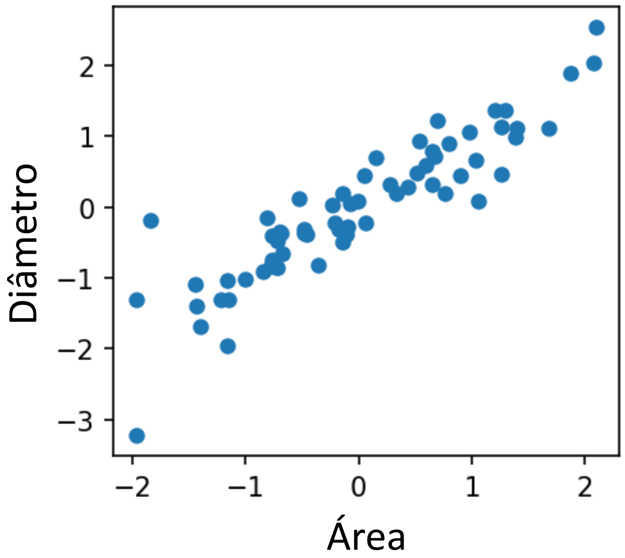
Não há muita variação da medida nessa escala, a maioria dos valores estará próxima de 0





Análise de Componentes Principais (PCA)

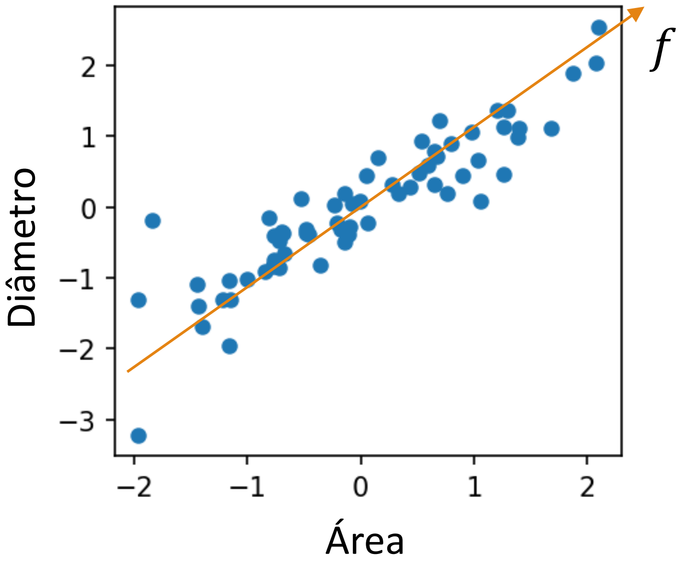
Talvez $f = 0.5 * \text{Área} + 0.5 * \text{Diâmetro}$?

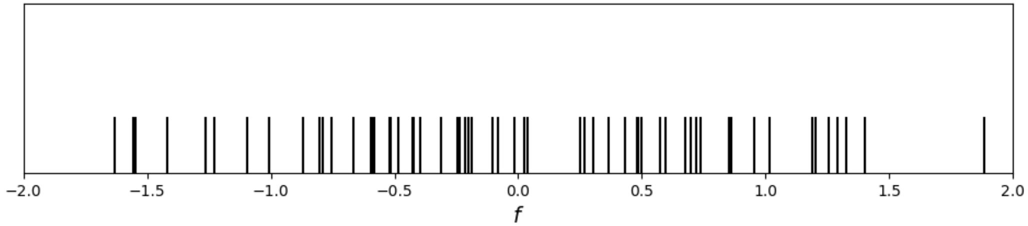


Análise de Componentes Principais (PCA)

Talvez $f = 0.5 * \text{Área} + 0.5 * \text{Diâmetro}$?

Essa parece ser uma boa propriedade. Os valores estão bem espalhados, o que indica que os feijões estarão bem caracterizados





Análise de Componentes Principais (PCA)

- Obter um novo conjunto de atributos seguindo os critérios mencionados é um problema de otimização muito custoso
- PCA é uma técnica que obtém de forma algébrica, sem otimização, novos atributos seguindo os critérios mencionados

- De forma geral, o algoritmo do PCA é o seguinte:
  1. Obtenha a matriz de correlação de Pearson
  2. Calcule os autovetores e autovalores dessa matriz
  3. Mantenha apenas os autovetores associados aos L maiores autovalores, esses autovetores definem o novo espaço de atributos, ou seja, as novas medidas a serem utilizadas
  4. Utilize os L autovetores para projetar os dados originais, obtendo assim os valores dos novos atributos

Exemplo de PCA em 4D – Dataset Iris

Iris setosa

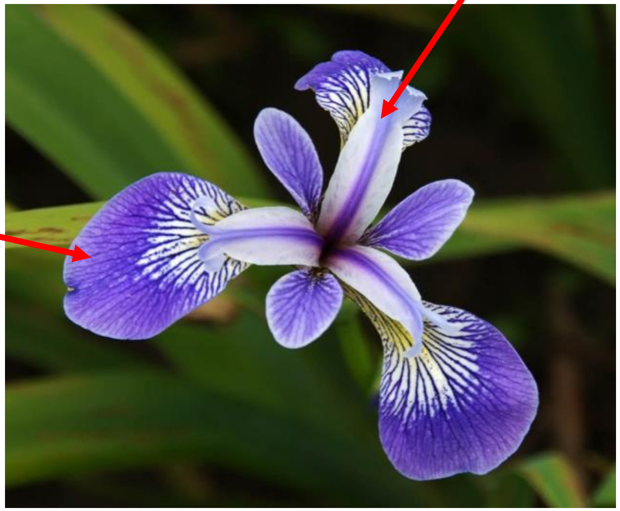
Iris virginica

Iris versicolor

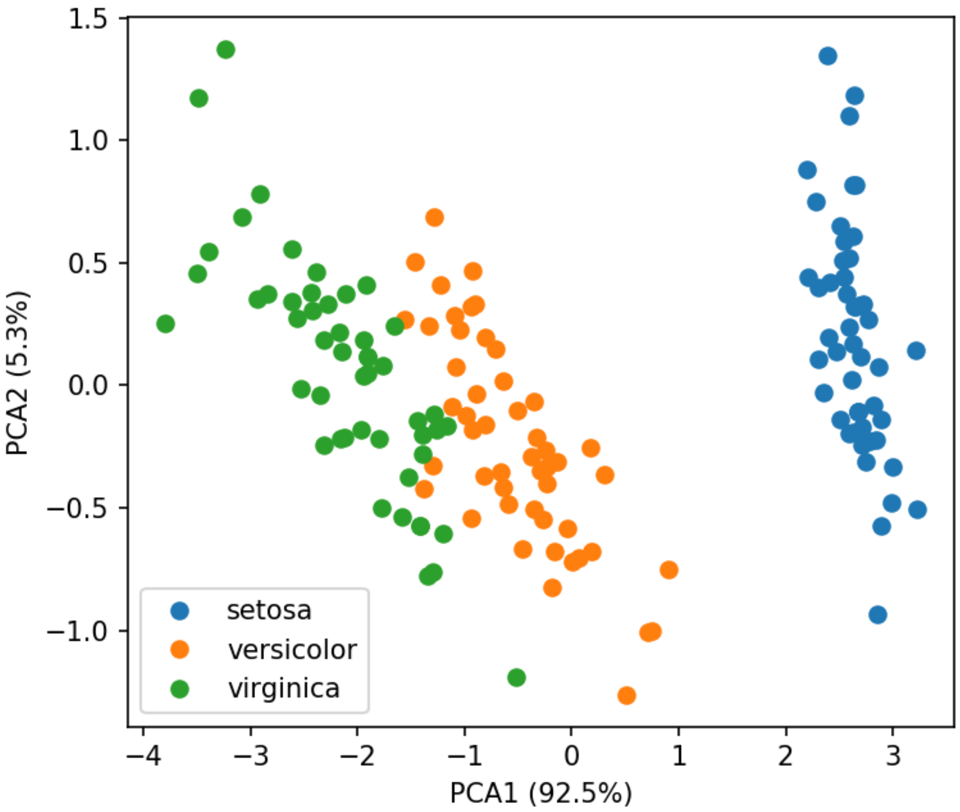
Exemplo de PCA em 4D – Dataset Iris

Essa base de dados possui 4 medidas para três tipos de flores (50 plantas para cada classe):

1. Comprimento da sépala em cm
2. Largura da sépala em cm
3. Comprimento da pétala, em cm
4. Largura da pétala, em cm



Exemplo de PCA em 4D – Dataset Iris



Análise de Componentes Principais - Python

- Em Python, podemos utilizar a função numpy corrcoef() para calcular a matriz de correlação a partir de uma matriz de dados:

```
numpy.corrcoef(X)
```

onde X é uma matriz contendo em cada coluna as propriedades medidas para um objeto

- Podemos calcular o PCA utilizando a função PCA() da biblioteca scikit-learn

```
pca = sklearn.decomposition.PCA(L)
pca.fit_transform(X)
```

onde L é o número de novos atributos a serem calculados (dimensão do espaço projetado) e X a mesma matriz descrita acima.

Eigenfaces

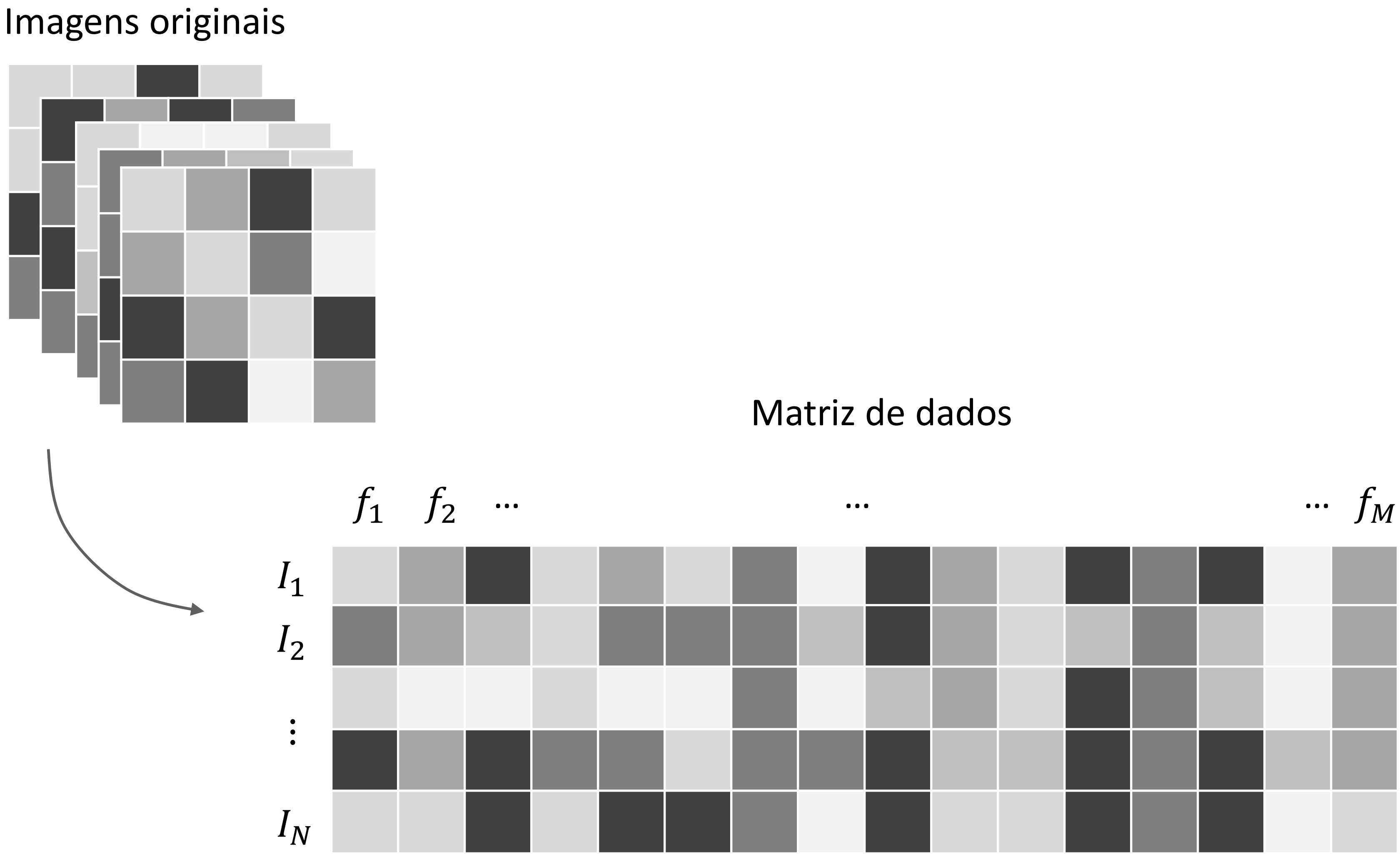
- O método eigenfaces busca definir novas imagens que representem bem um grande conjunto de imagens
- As imagens são dadas pelos autovetores do PCA

Eigenfaces

Imagem original

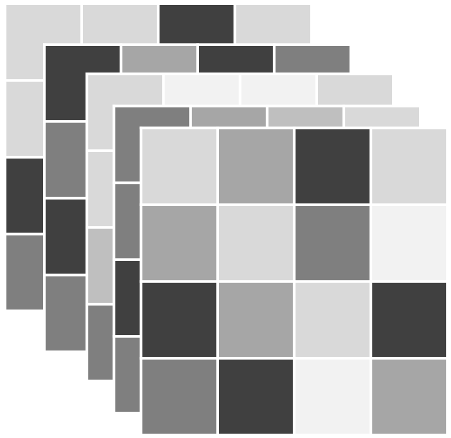
Imagem representada em 1D

Eigenfaces



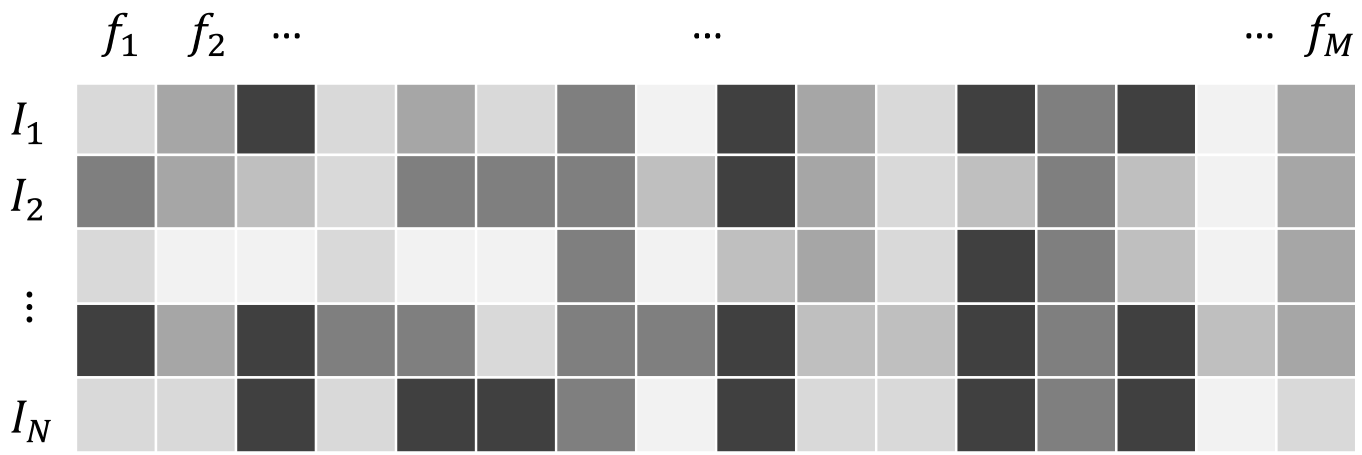
Eigenfaces

Imagens originais



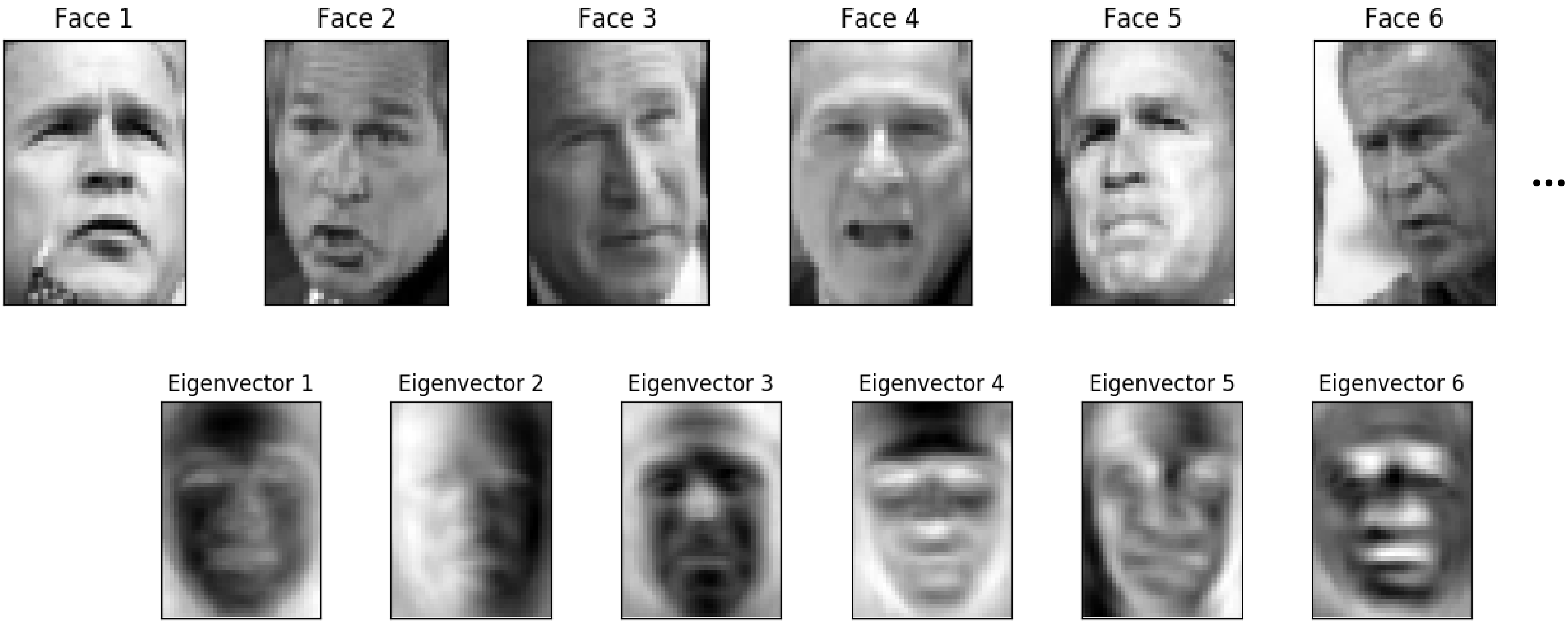
Aplicamos o PCA na matriz de dados e armazenamos os autovetores

Matriz de dados

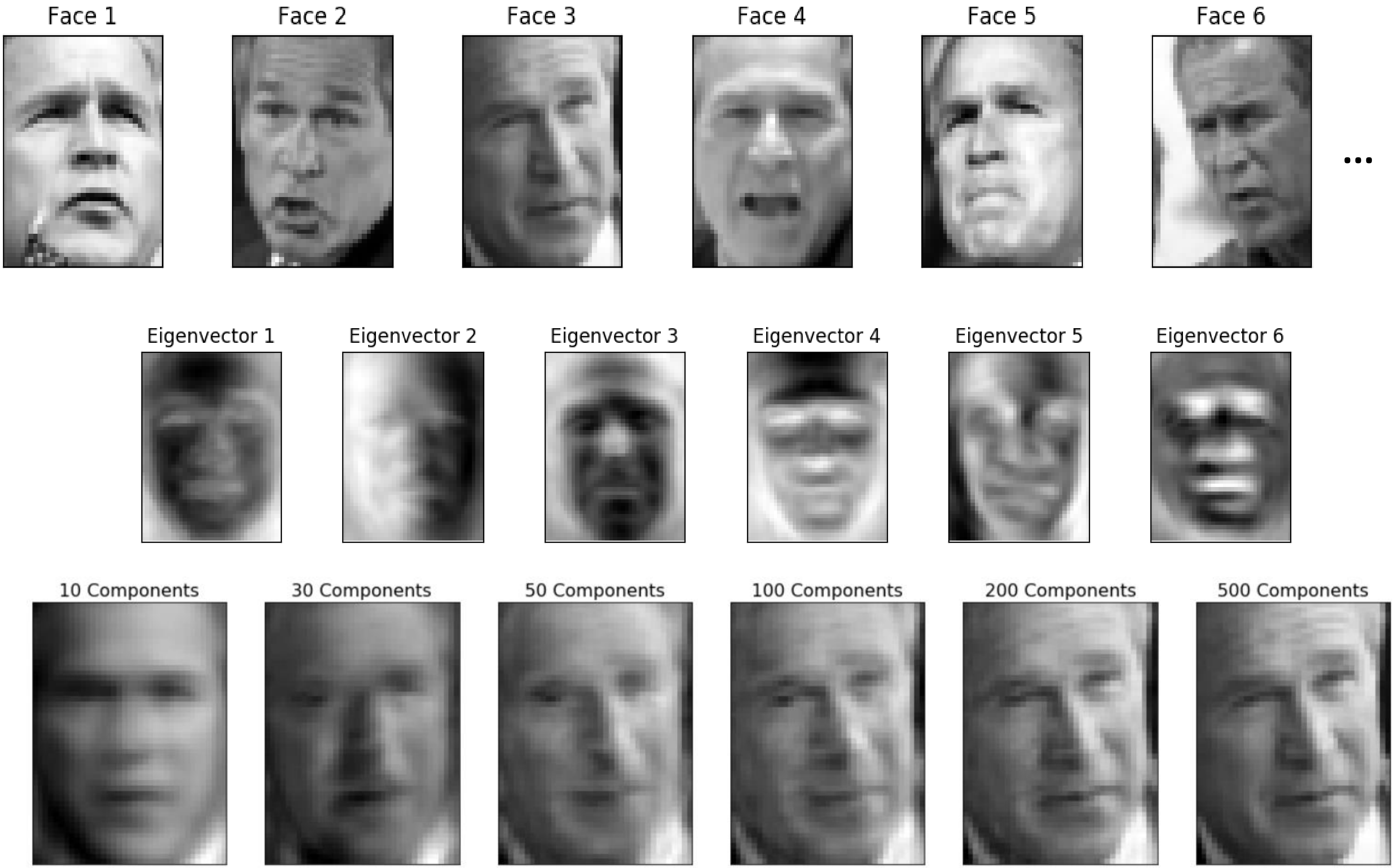


Eigenfaces

Eigenfaces


Face 1
Face 2
Face 3
Face 4
Face 5
Face 6
...
Eigenvector 1
Eigenvector 2
Eigenvector 3
Eigenvector 4
Eigenvector 5
Eigenvector 6

Eigenfaces


Face 1
Face 2
Face 3
Face 4
Face 5
Face 6
...
Eigenvector 1
Eigenvector 2
Eigenvector 3
Eigenvector 4
Eigenvector 5
Eigenvector 6
10 Components
30 Components
50 Components
100 Components
200 Components
500 Components

Uso do PCA em análise de dados

Notebook "**Análise de Componentes Principais (PCA)**"