Pré-Processamento de Dados: Detecção de Outliers

Luciano Barbosa

Fontes de Erro

- Inserção dos dados
- Coleta dos dados

Tarefa Exploratória

- Ferramentas para limpeza
- Visualização dos dados
- Human in the loop

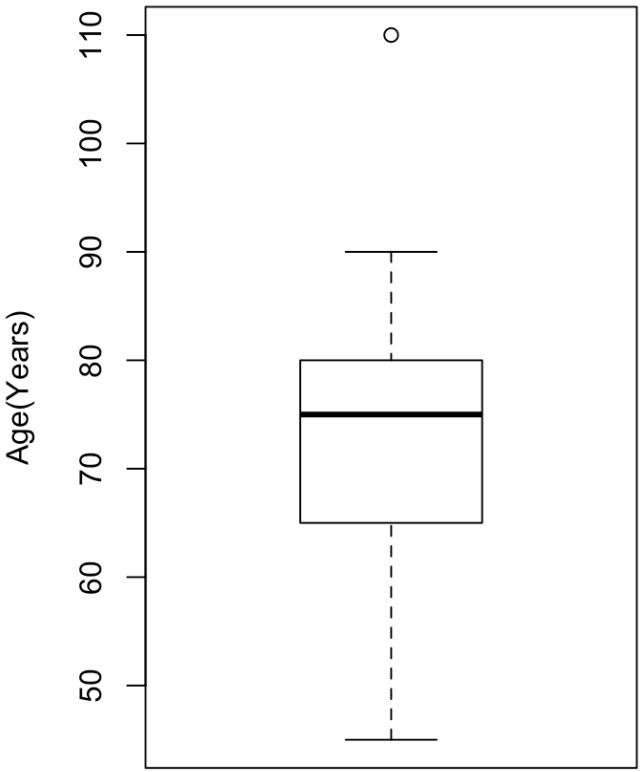
Tipos de Problemas nos Dados

- Dados faltantes
- Dados duplicados
- Dados irrelevantes
- Dados incorretos

Dados Incorretos (Outliers)

- Observação que não está próxima ao centro

Age(Years)
75
80
65
55
67
78
88
90
45
58
69
80
110

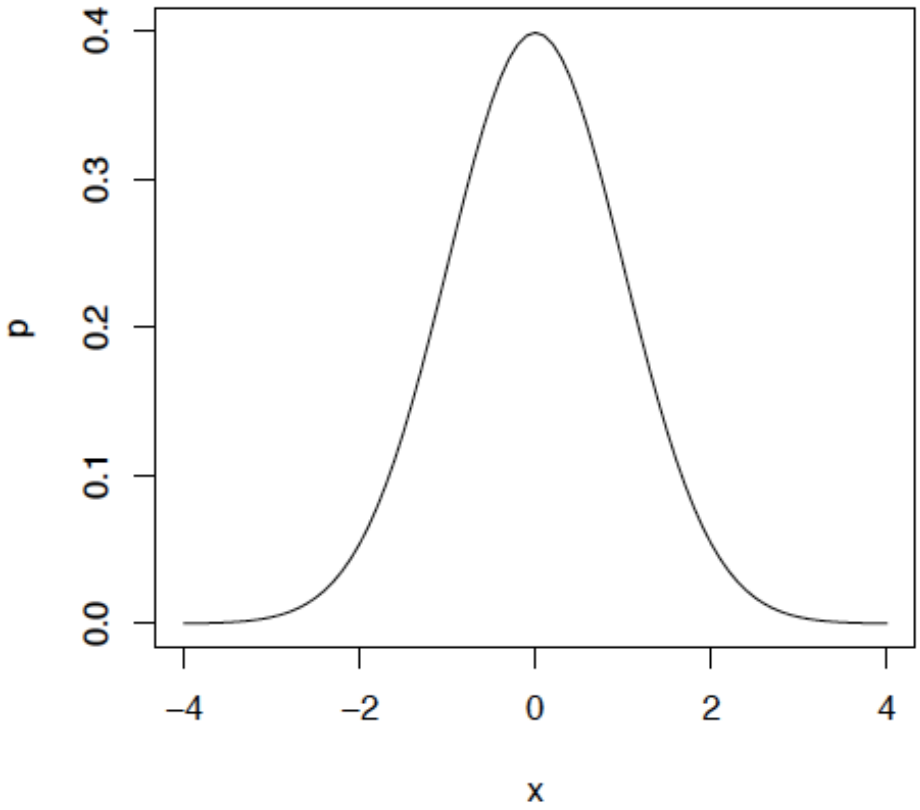


Métodos de Detecção Univariado

- Z-Score robusto
- Tukey

Centro and Dispersão

- Centro
  - Valor médio
  - Ex: média, mediana
- Dispersão
  - Desvio do centro
  - Ex: variância, desvio padrão



Z-Score Robusto

- Distribuição precisa ser simétrica
- Centro: Mediana
  - Metade dos valores são menores e metade são maiores
  - É influenciado pelas posições dos outliers mas não pelos seus valores

Z-Score Robusto

- Dispersão: Median absolute deviation
  - Mediana da distância da diferença de todos valores da mediana

$$MAD = median\left|x_i - \hat{x}\right|$$

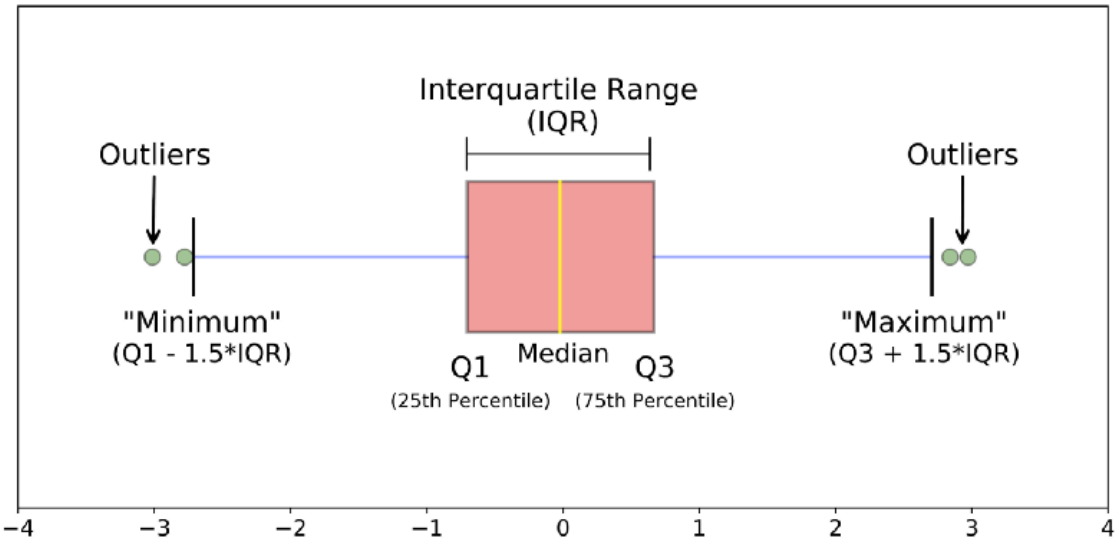
Z-Score Robusto

$$M_i = \frac{0.6745\,(x_i - \tilde{x})}{MAD}$$

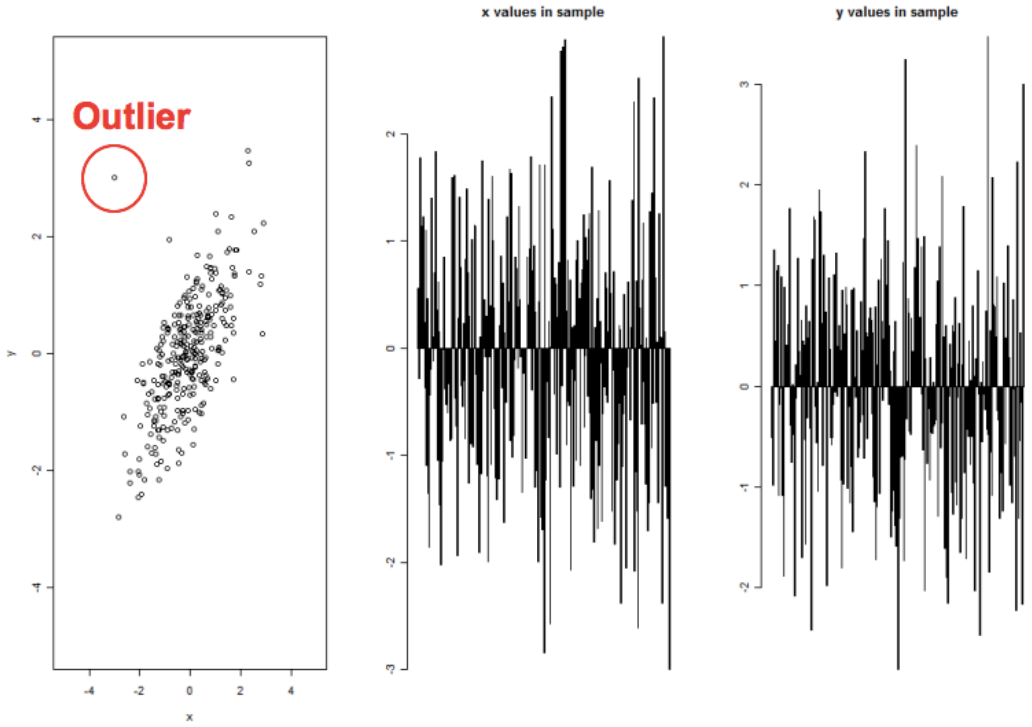
- Constante b = 0.6745: fator de escala que torna MAD um estimador não-enviesado do desvio padrão: $E(MAD) = 0.675\,\sigma$
- M_i> limiar: indica outlier (ex., 3 ou 3.5)

Método de Tukey

- Distribuição precisa ser simétrica
- Baseado em quartis
- Outliers:
  - Valores menores que Q1 − 1.5 * IQR
  - Valores maiores que Q3 + 1.5 * IQR



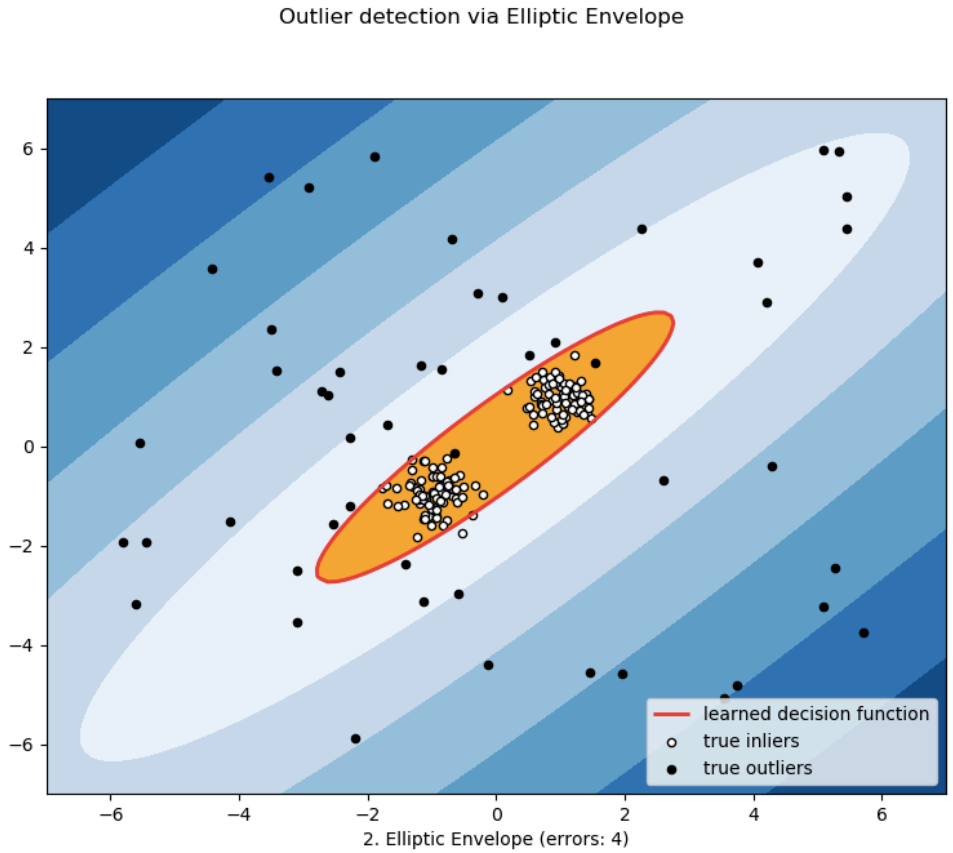
Bivariado



- Solução: transformar a relação em univariada (ex.: razão deu uma variável pela outra)

Multivariado: Elliptic Envelope

- Suposição: atributos seguem gaussiana



Multivariado: Elliptic Envelope

- Suposição: atributos seguem gaussiana
- Utiliza distância Mahalanobis

T indicates a transposed matrix

$$D^2 = (x - \overline{x})^T S^{-1} (x - \overline{x})$$

Matrix of distances from mean

Inverse of covariance matrix

Matrix of:

$(x_1, x_2, \ldots, x_n)$ - $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)$

or

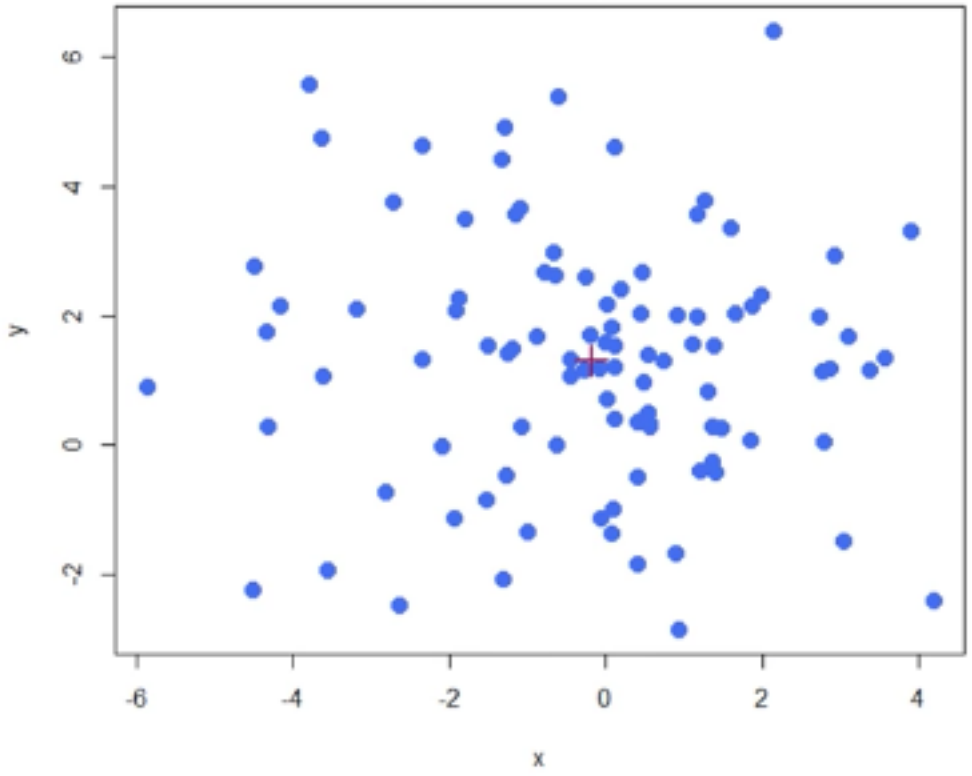
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_n \end{bmatrix}$$

$$\begin{bmatrix} s_1^2 & \cdots & \mathrm{Cov}(s_n, s_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(s_1, s_n) & \cdots & s_n^2 \end{bmatrix}$$

Matrix with diagonals = variance of samples 1 ... *n* and cells = covariance of samples (1,2) ... (1,*n*)

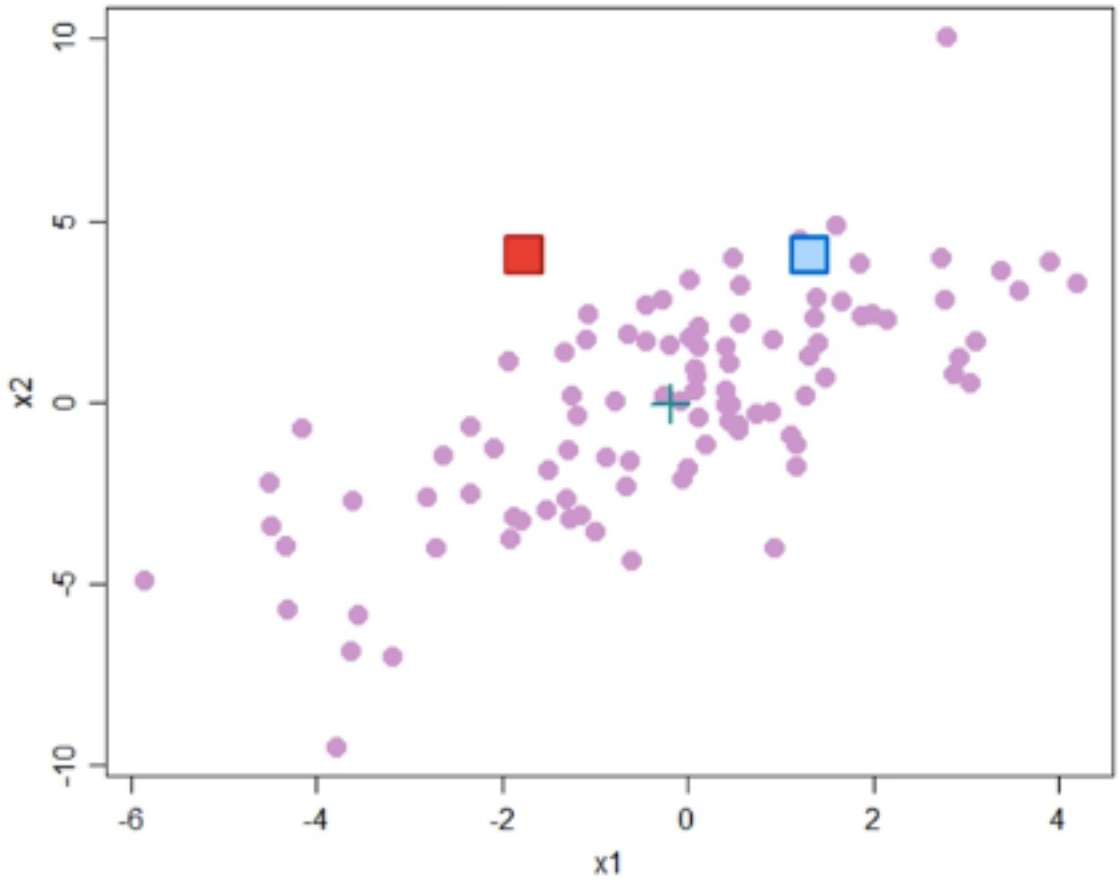
Por que não Distância Euclidiana?

$$\sqrt{(x_i - \overline{x})^2 + (y_i - \overline{y})^2 + \cdots + (n_i - \overline{n})^2}$$

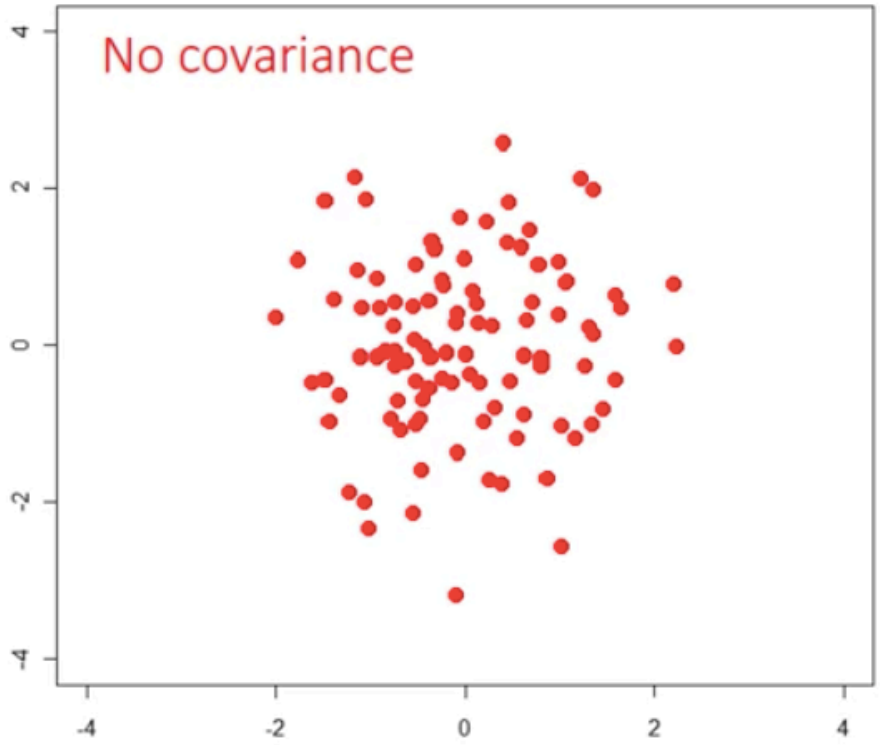


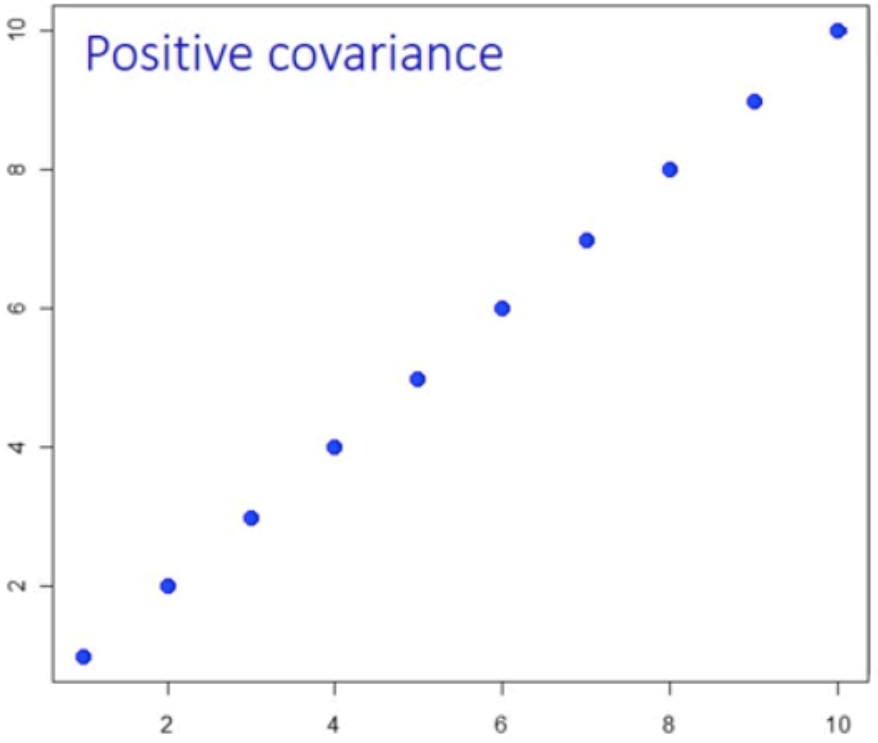
Limitação da Distância Euclidiana

- Covariância entre as variáveis



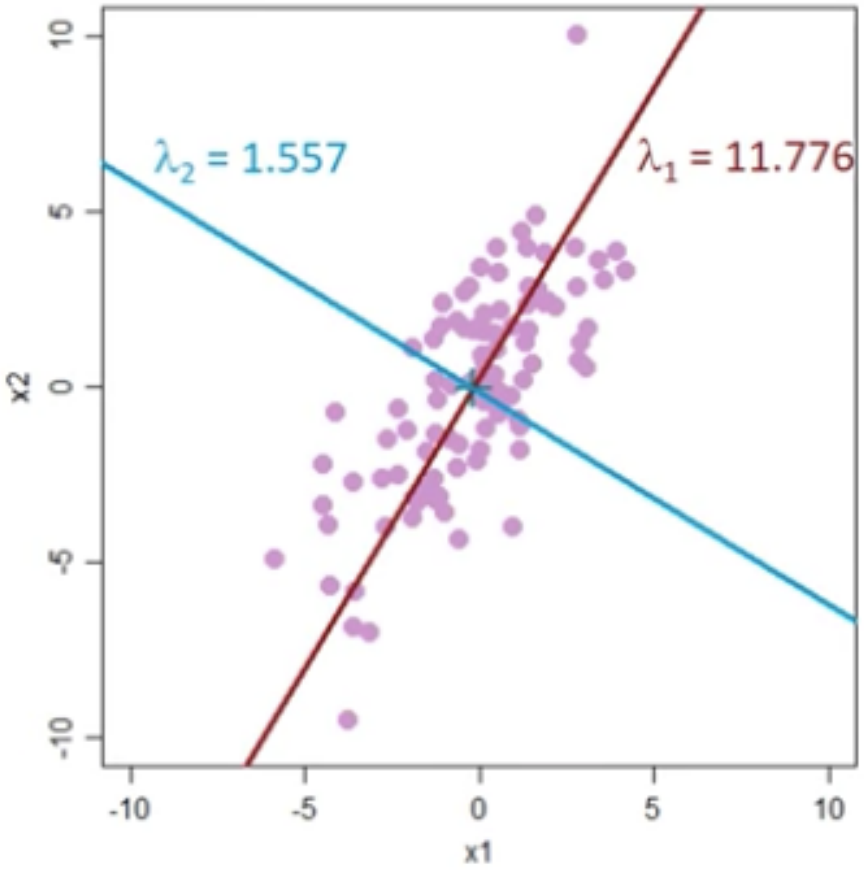
Covariância





Remover a Covariância

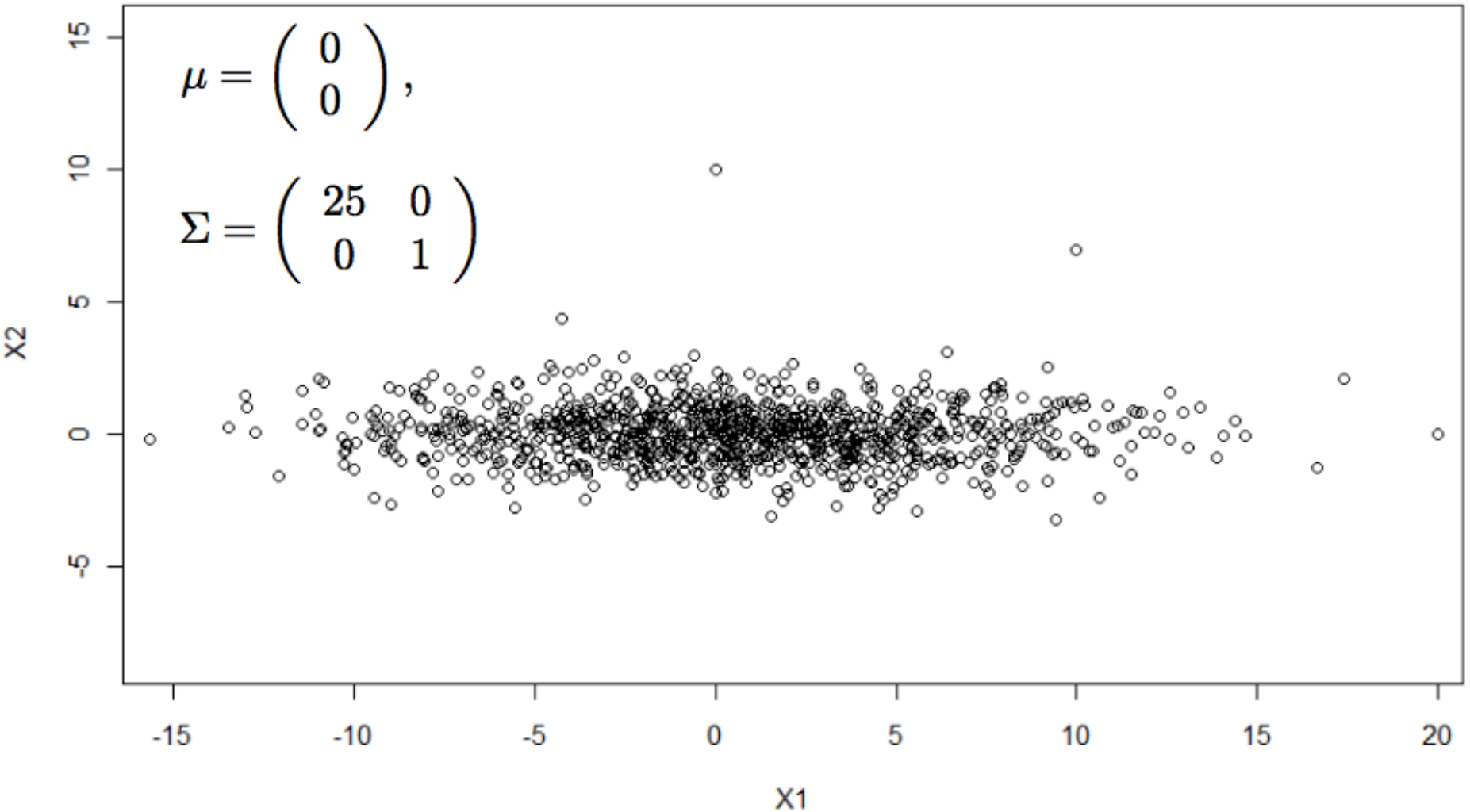
- Projetar os pontos nos Autovetores
- Rotacionar os pontos
- Novos eixos são os autovetores
- Reescalar os valores dos pontos em cada eixo pela raiz quadrada do auto-valor



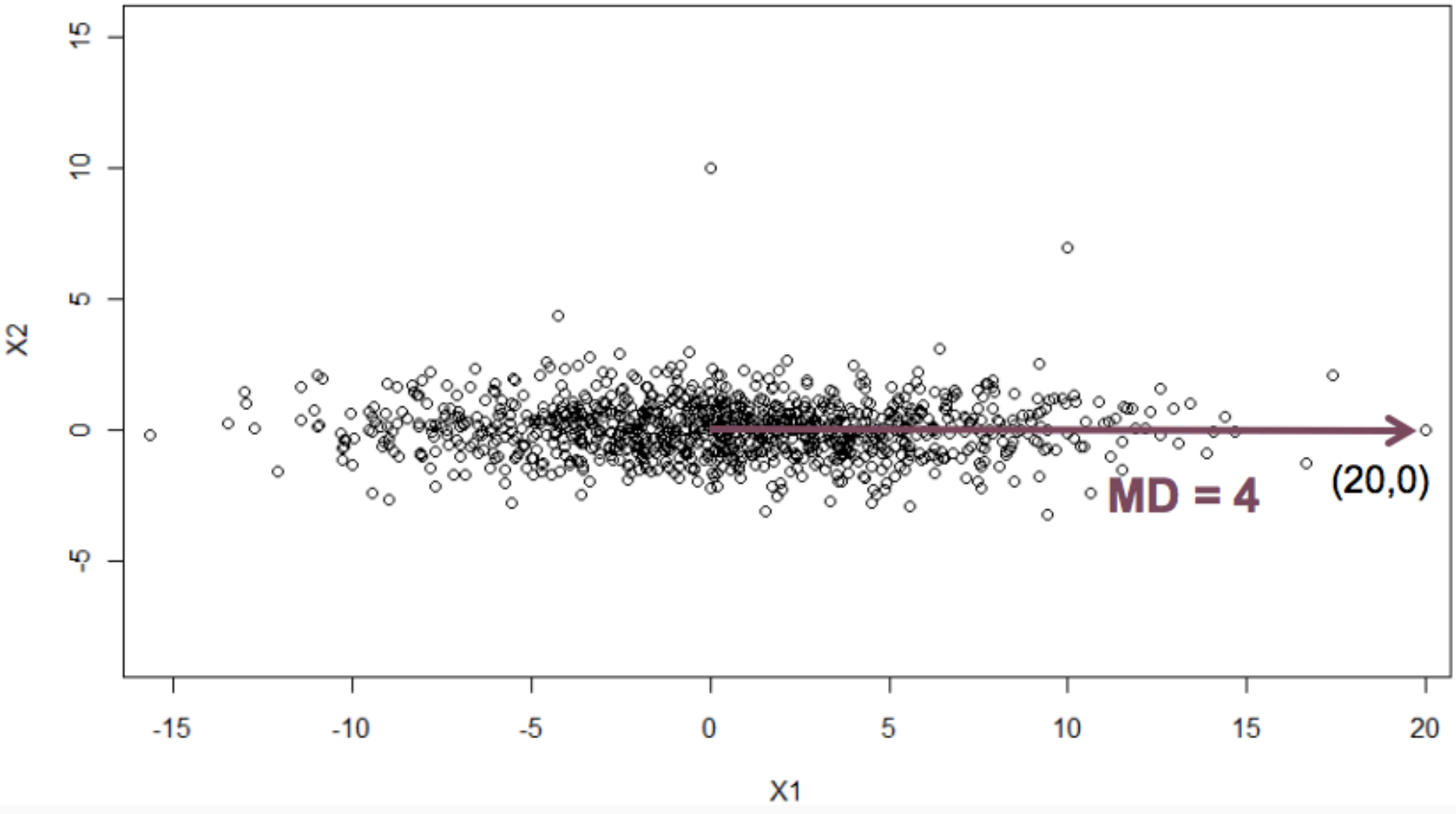
Exemplo: Distância Mahalanobis

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

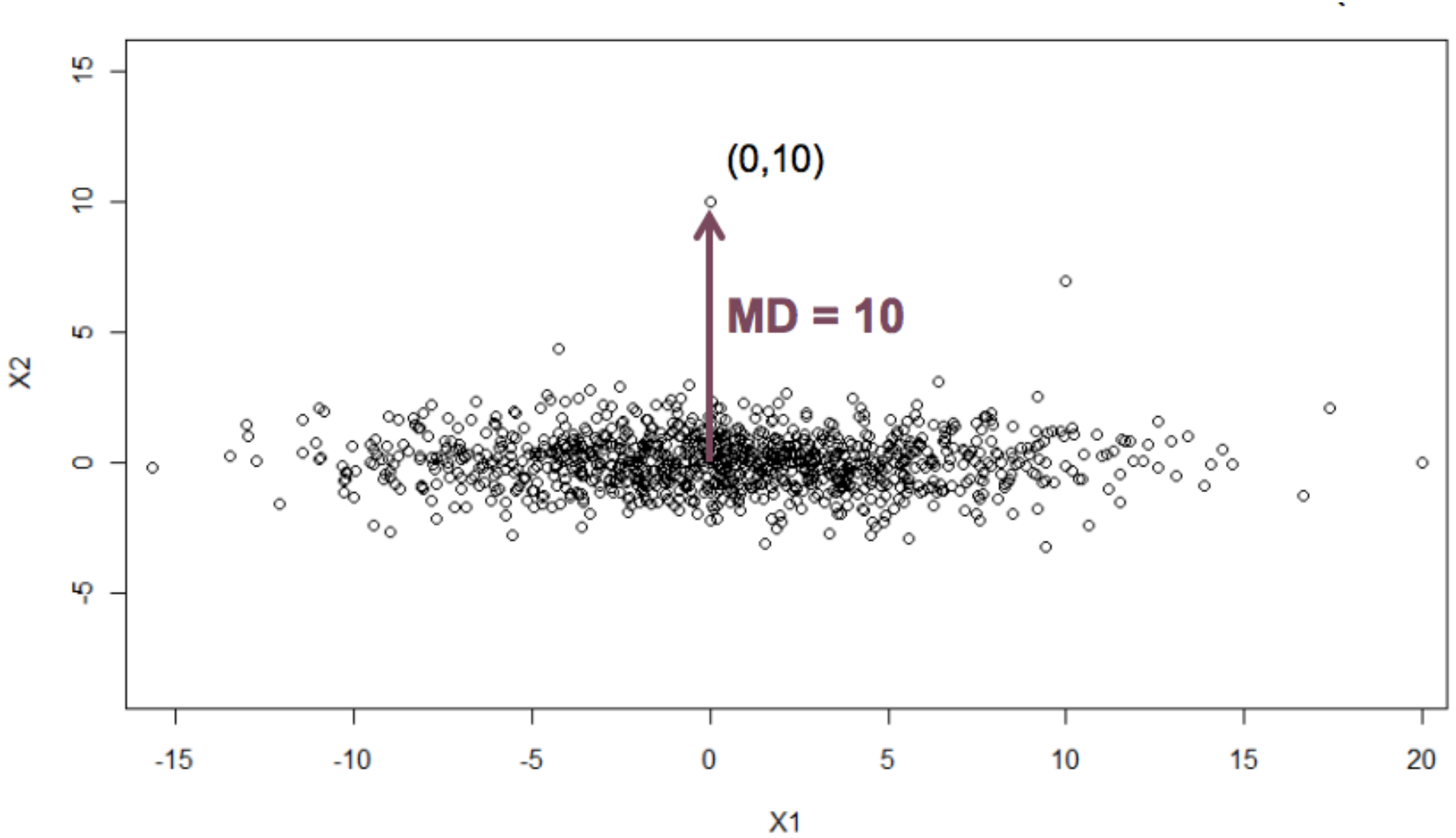
$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



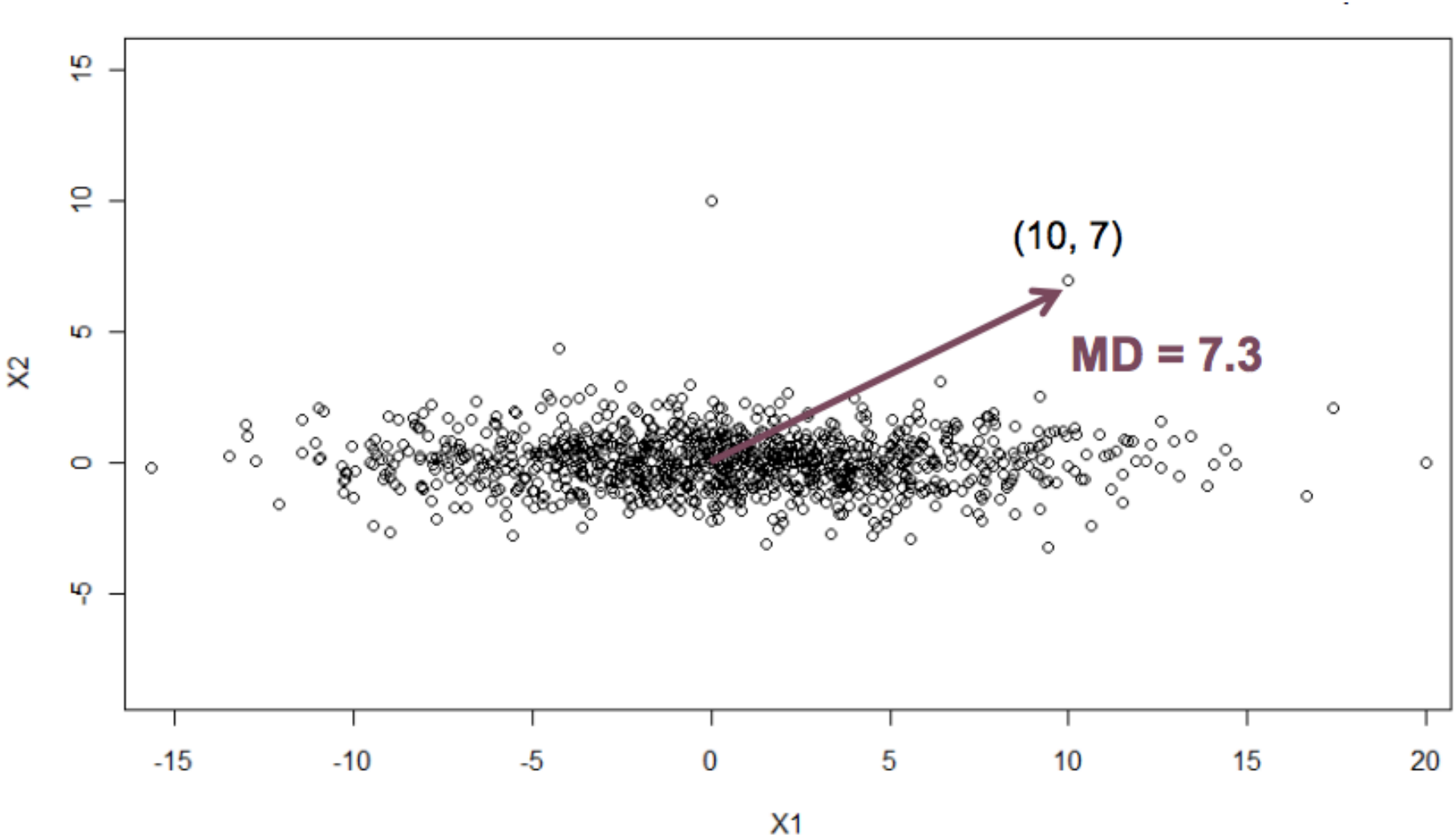
Exemplo: Distância Mahalanobis



Exemplo: Distância Mahalanobis



Exemplo: Distância Mahalanobis



Multivariado: Elliptic Envelope

- Suposição: atributos seguem gaussiana
- Utiliza distância Mahalanobis

T indicates a transposed matrix

$$D^2 = (x - \overline{x})^T S^{-1} (x - \overline{x})$$

Matrix of distances from mean

Inverse of covariance matrix

Matrix of:

$(x_1, x_2, \ldots, x_n)$ - $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)$

or

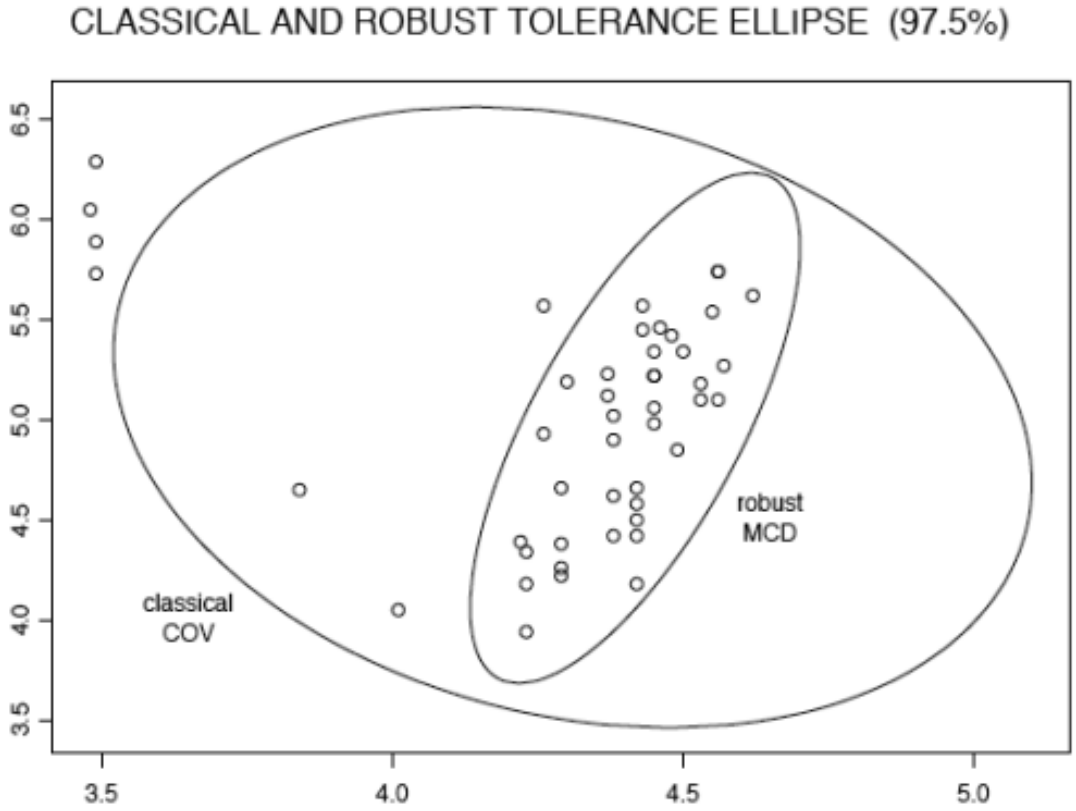
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_n \end{bmatrix}$$

$$\begin{bmatrix} s_1^2 & \cdots & \mathrm{Cov}(s_n, s_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(s_1, s_n) & \cdots & s_n^2 \end{bmatrix}$$

Matrix with diagonals = variance of samples 1 ... *n* and cells = covariance of samples (1,2) ... (1,*n*)

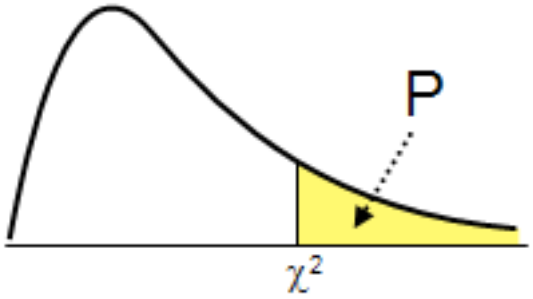
Minimum Covariance Determinant

- Encontrar a matriz de covariância com mínimo volume dentro de % dos dados



Detectando Outliers

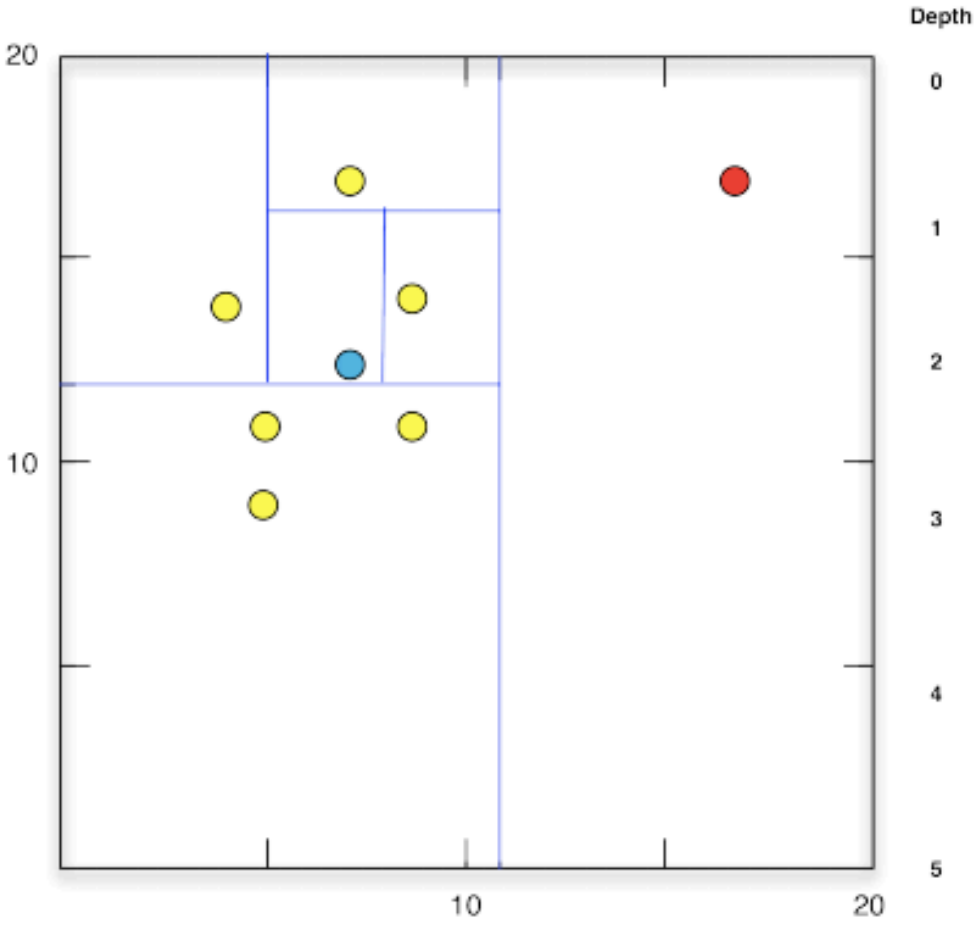
- Computar a distância Mahalanobis para cada amostra
- Outliers: amotras com distância maior que um determinado valor crítico da distribuição chi-square

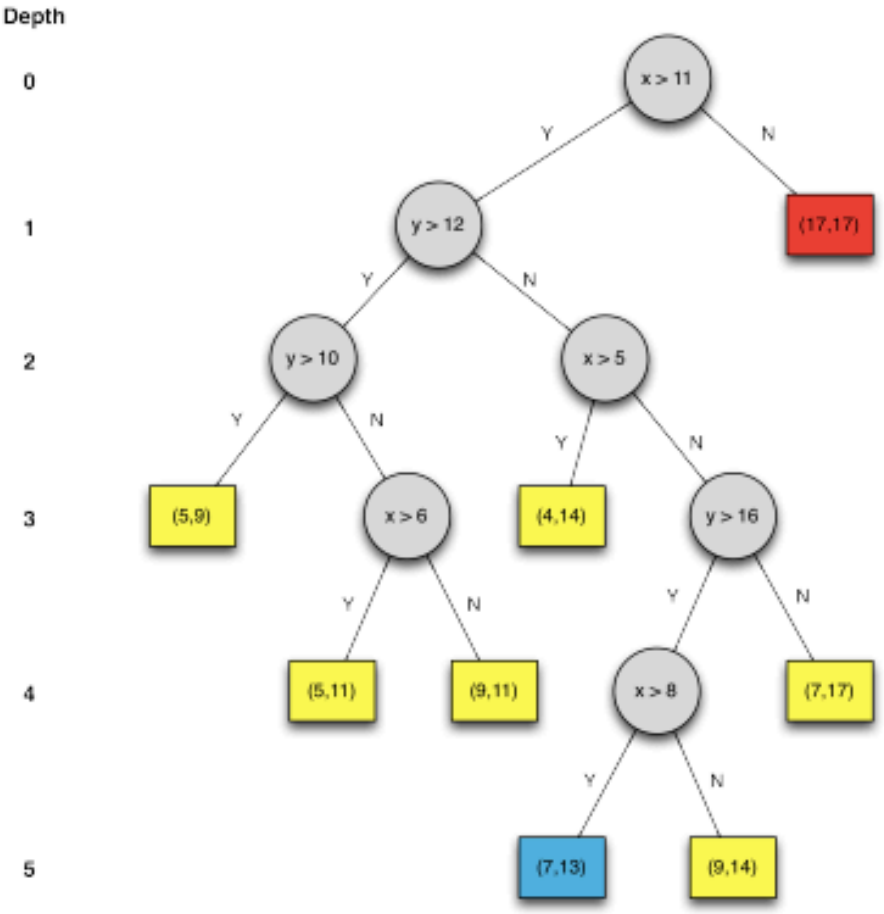


Isolation Forests

- Não-paramétrico
- Suposição: outliers são poucos e diferentes
- Passos:
  1. Seleciona aleatoriamente um feature
  2. Seleciona um valor aleatório dela entre o máximo e mínimo
  3. Repete passos 1 e 2 n vezes

Isolation Forests



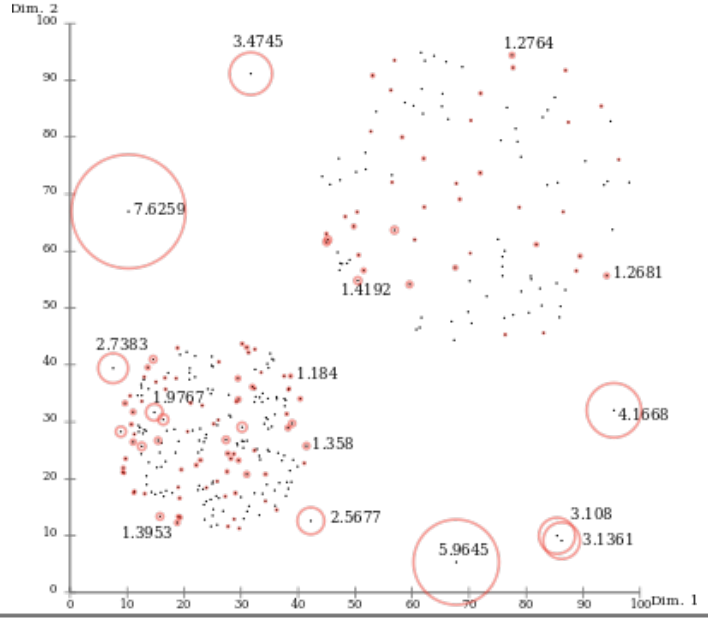


Isolation Forests

- Fácil de isolar outliers: poucas condições necessárias para separar dos demais
- Score: profundidade média do ponto na árvore necessária para isolar o ponto
  - Perto de 1 indica outlier

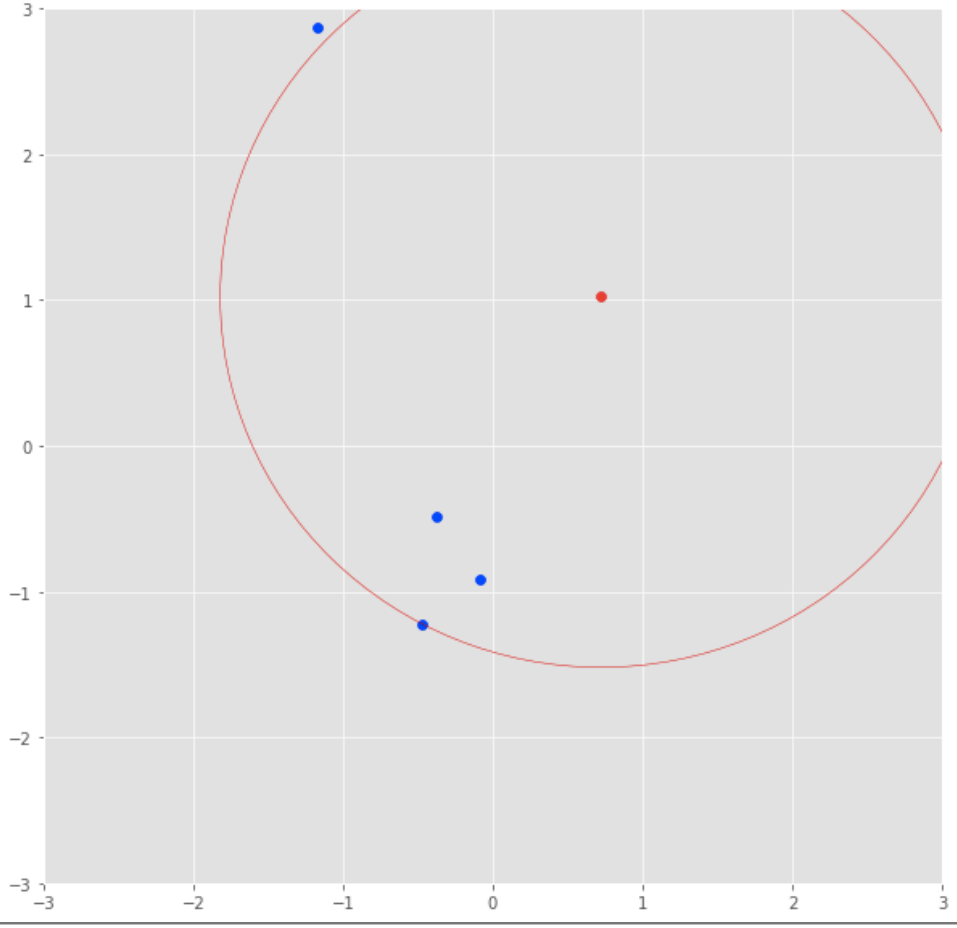
Local Outlier Factor

- Baseado na densidade dos dados
- Outliers: instâncias com menor densidade que os vizinhos
- Média da densidade dos k-vizinhos mais próximos sobre a densidade do ponto



K-Distance

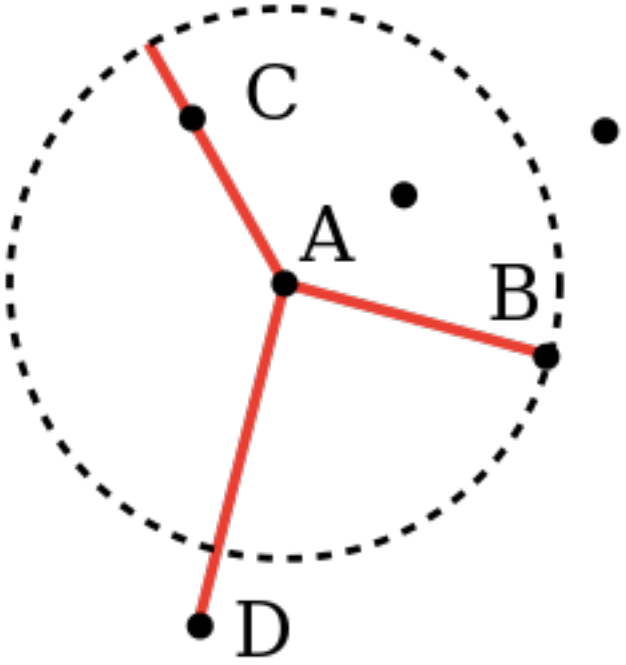
- Distância para o k elemento mais próximo



Reachability Distance

- Máximo entre a distância de dois pontos e o k-distance do segundo ponto:

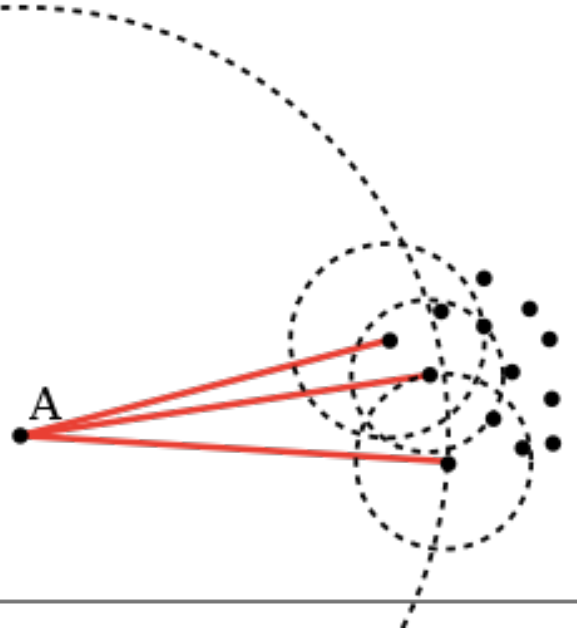
  Ex: RD(D,A) = max{k-distance(A),dist(A,D)}



Local Reachability Density

- Inverso da média da RD para todos os k-vizinhos

$$\mathrm{lrd}_k(A) := 1/\left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A,B)}{|N_k(A)|}\right)$$

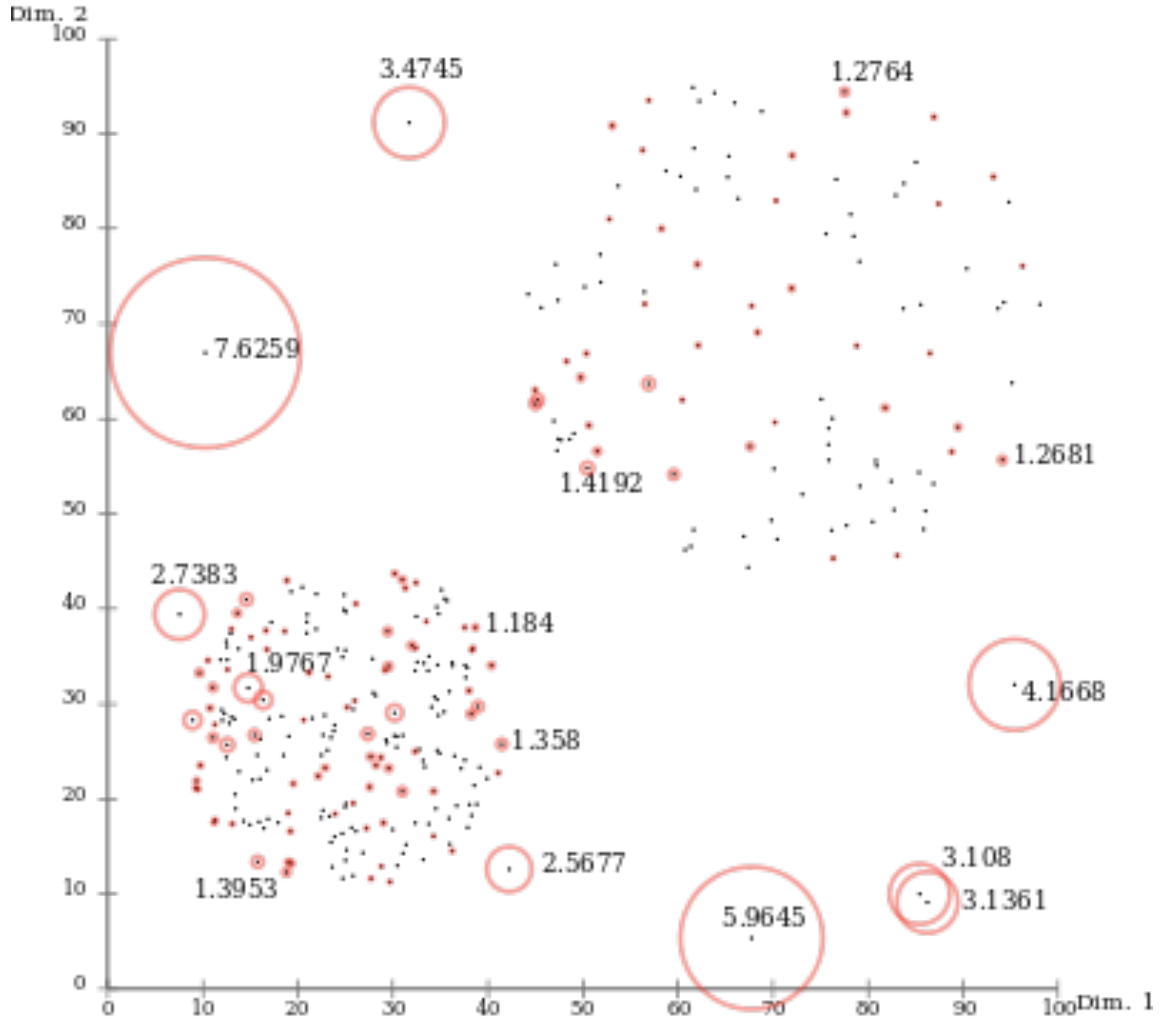


Local Outlier Factor

- Média de lrd dos vizinhos sobre a do ponto calculado

$$\mathrm{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\mathrm{lrd}(B)}{\mathrm{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \mathrm{lrd}(B)}{|N_k(A)|} / \mathrm{lrd}(A)$$

- LOF = 1: densidade similar a dos vizinhos
- LOF < 1: mais denso
- LOF > 1: menos denso


Dim. 2
Dim. 1
3.4745
1.2764
7.6259
1.2681
1.4192
2.7383
1.184
1.9767
4.1668
1.358
2.5677
3.108
3.1361
1.3953
5.9645