



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



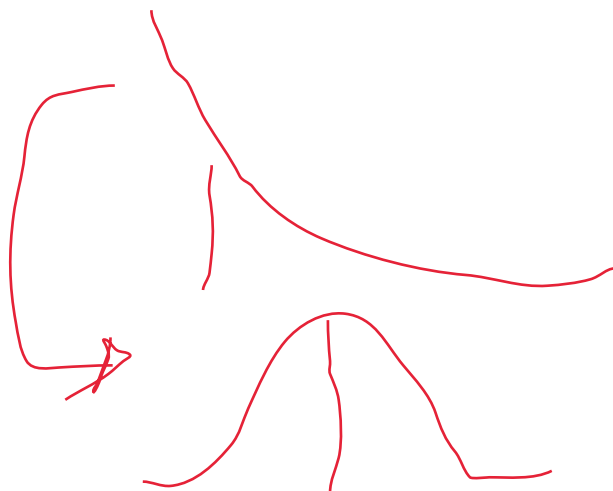
Tipos de Problemas nos Dados

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

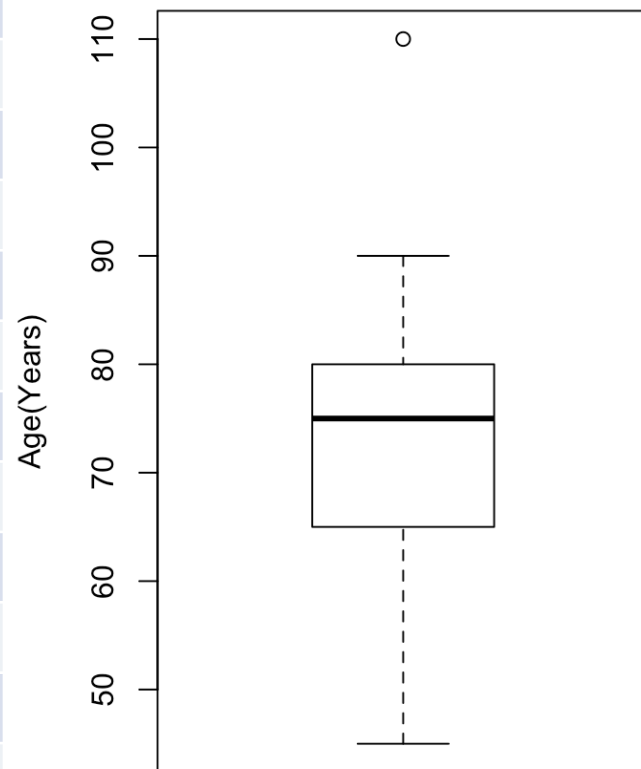


Dados Incorretos (Outliers)

- Observação que não está próxima ao centro



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





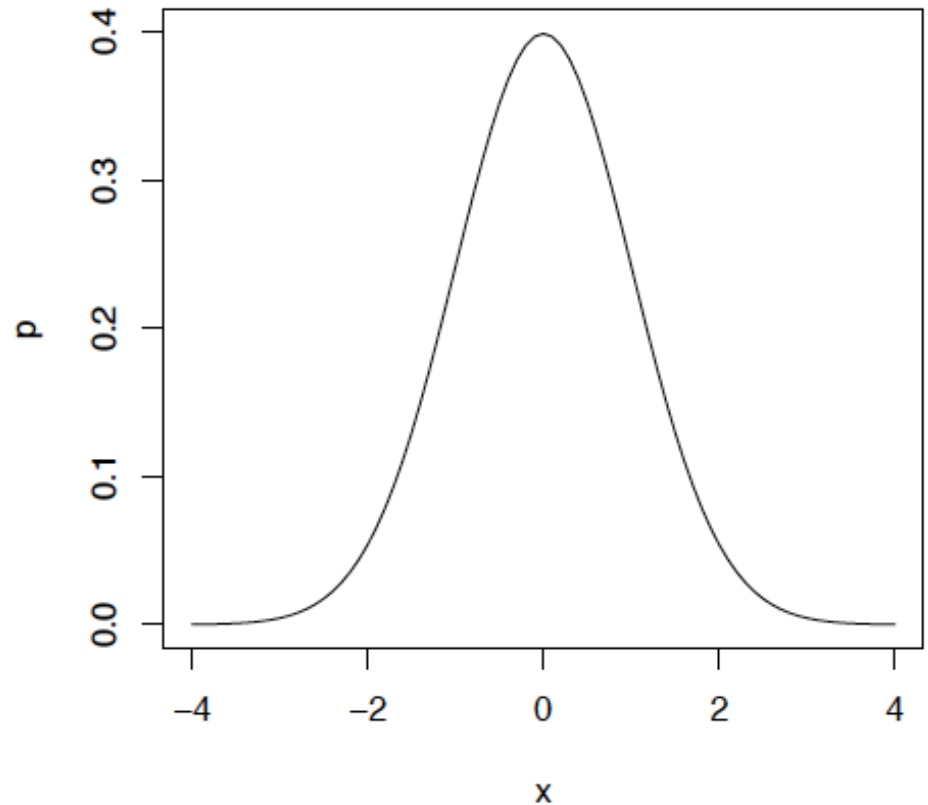
Métodos de Detecção Univariado

- 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

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



Z-Score Robusto

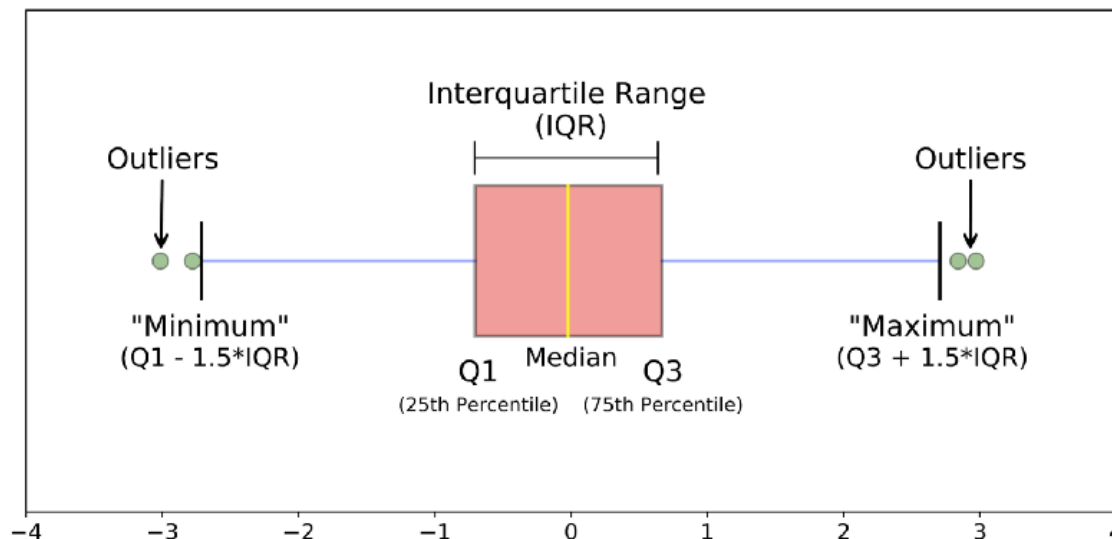
$$M_i = \frac{0.6745 (x_i - \bar{x})}{MAD}$$

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



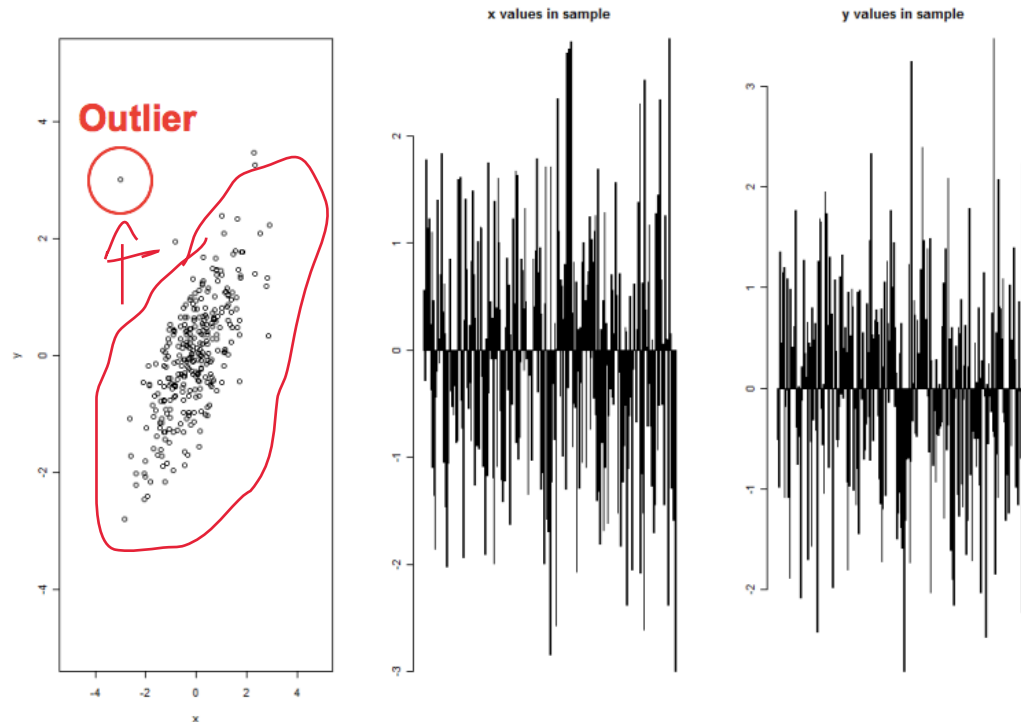
Método de Tukey

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





Bivariado



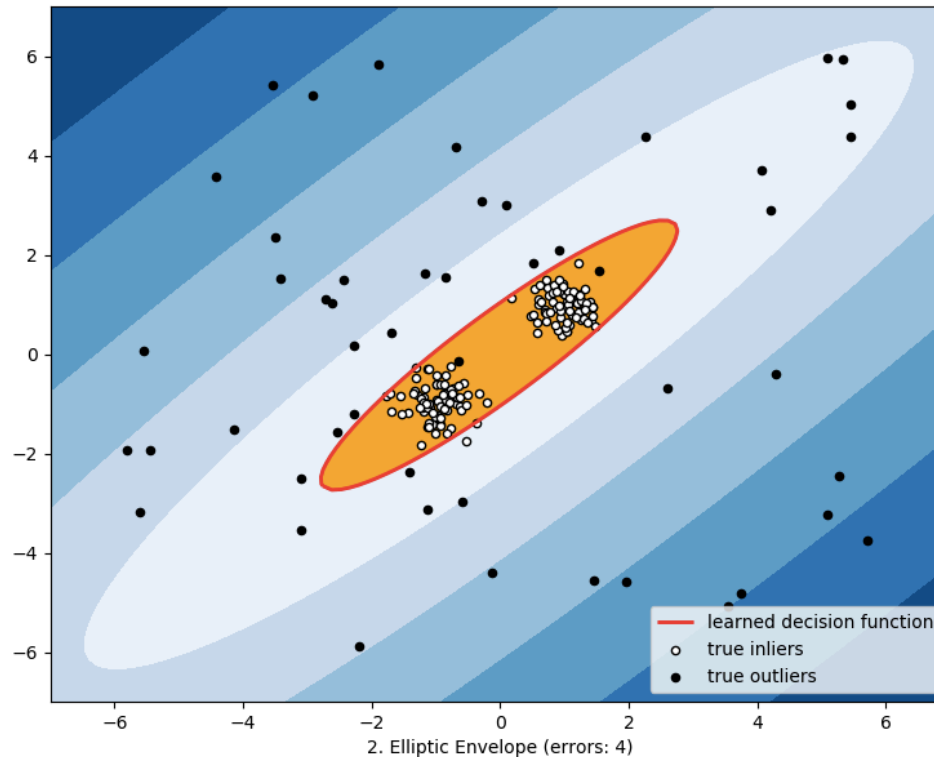
- Solução: transformar a relação em univariada (ex.: razão de uma variável pela outra)



Multivariado: Elliptic Envelope

- Suposição: atributos seguem gaussiana

Outlier detection via Elliptic Envelope





Multivariado: Elliptic Envelope

- Suposição: atributos seguem gaussiana
- Utiliza distância Mahalanobis

T indicates a transposed matrix

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

Matrix of distances from mean

Inverse of covariance matrix

Matrix of:
 $(x_1, x_2, \dots, x_n) - (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$

or

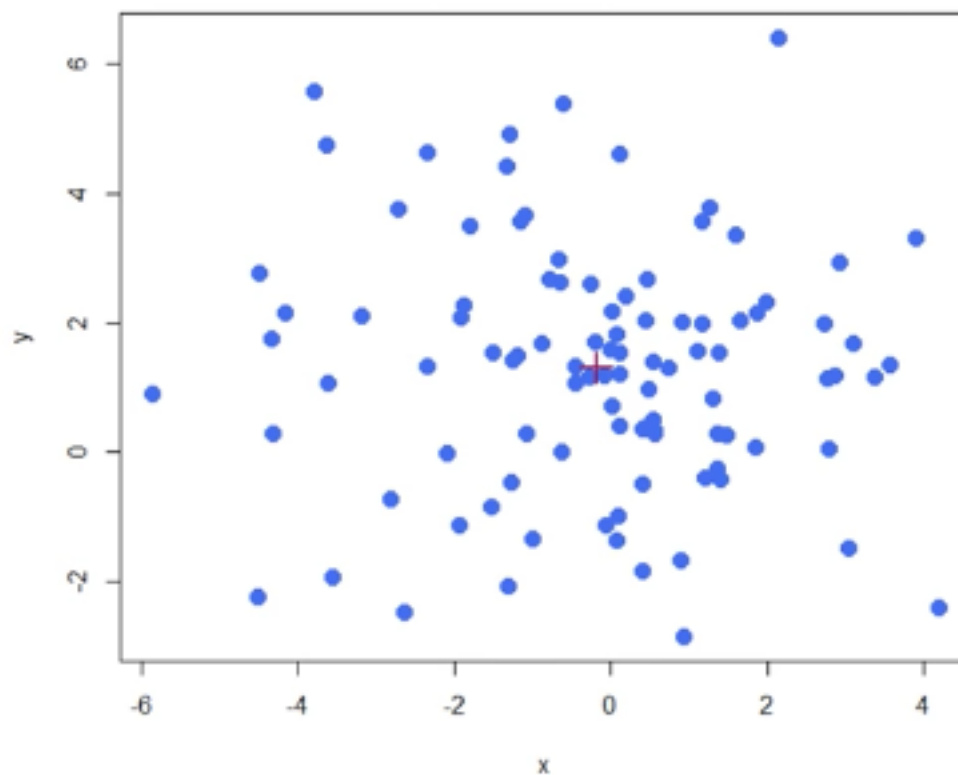
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{bmatrix}$$
$$\begin{bmatrix} s_1^2 & \dots & \text{Cov}(s_n, s_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}(s_1, s_n) & \dots & s_n^2 \end{bmatrix}$$

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



Por que não Distância Euclidiana?

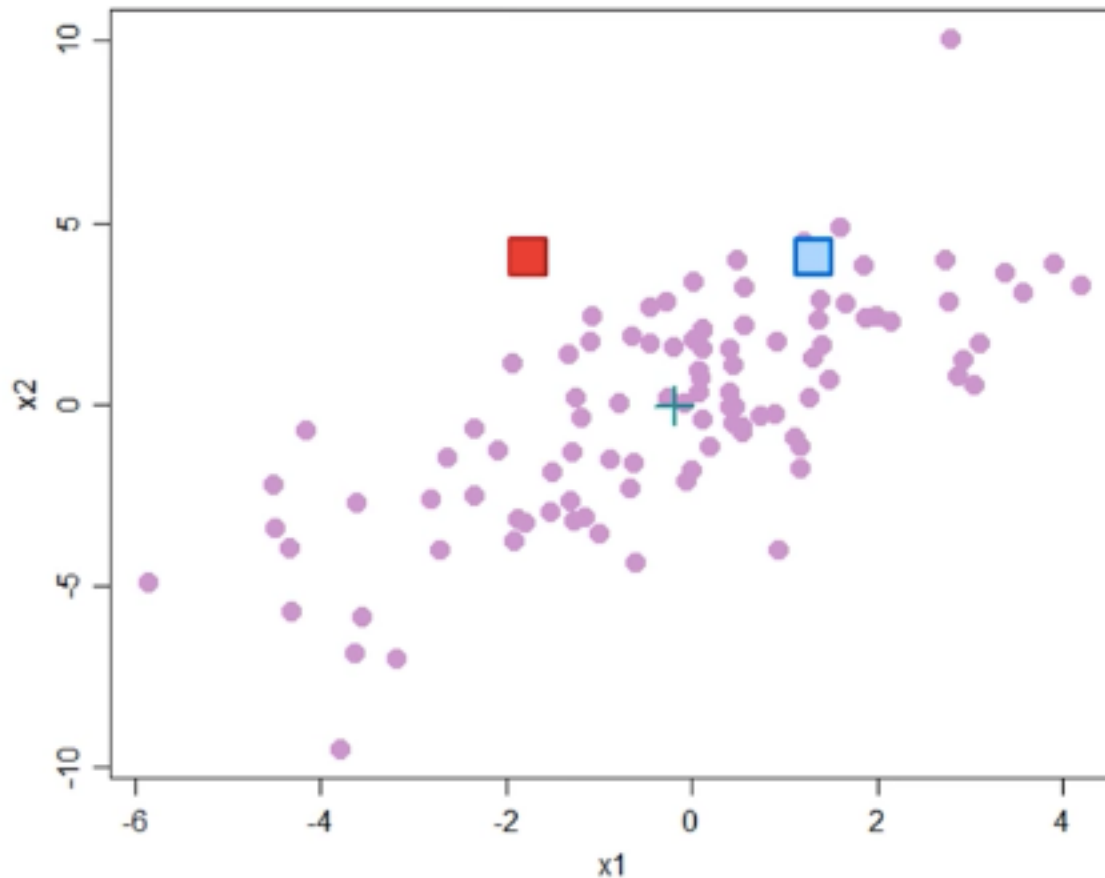
$$\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + \dots + (n_i - \bar{n})^2}$$





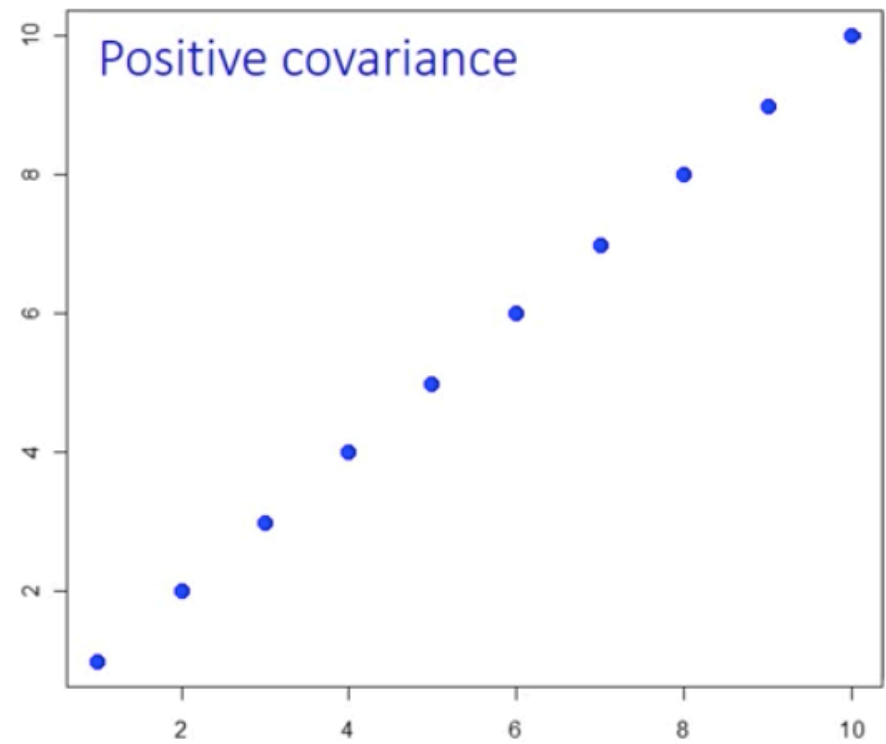
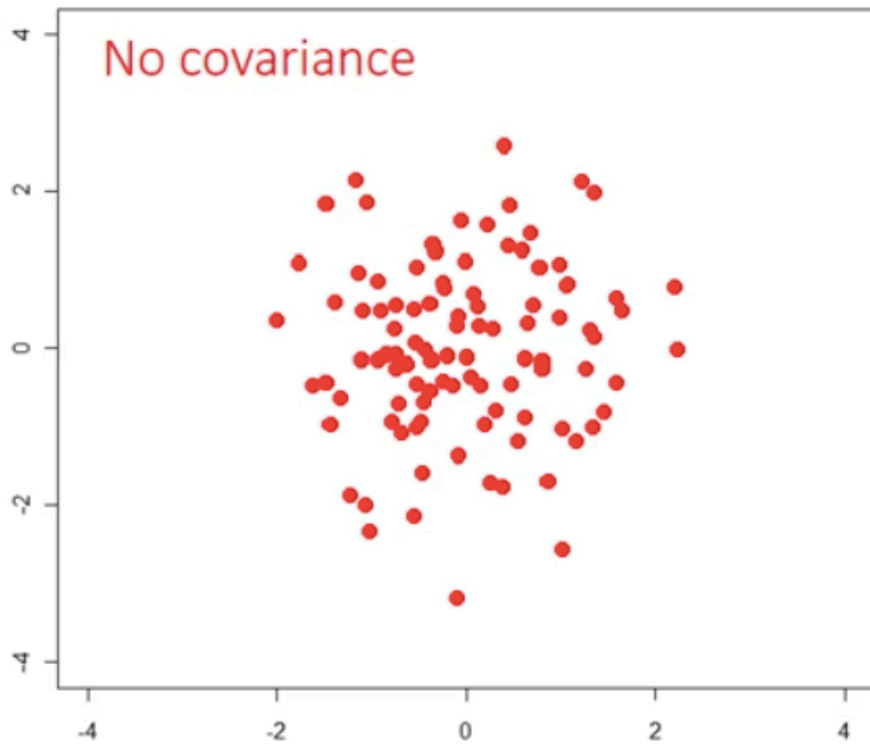
Limitação da Distância Euclidiana

- Covariância entre as variáveis





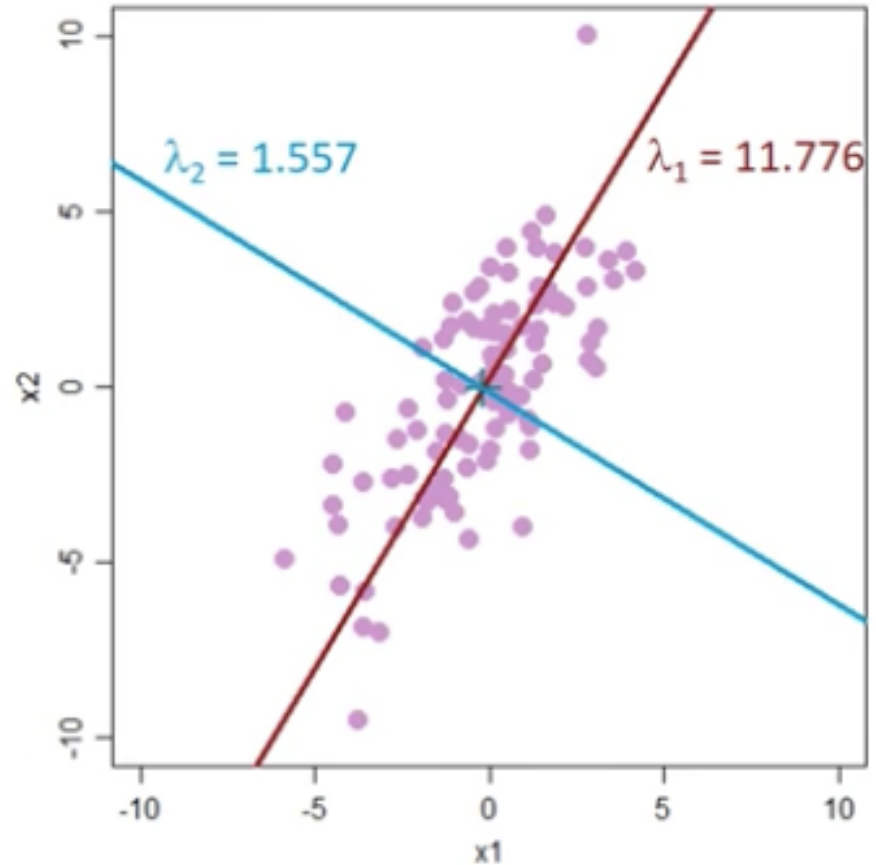
Covariância





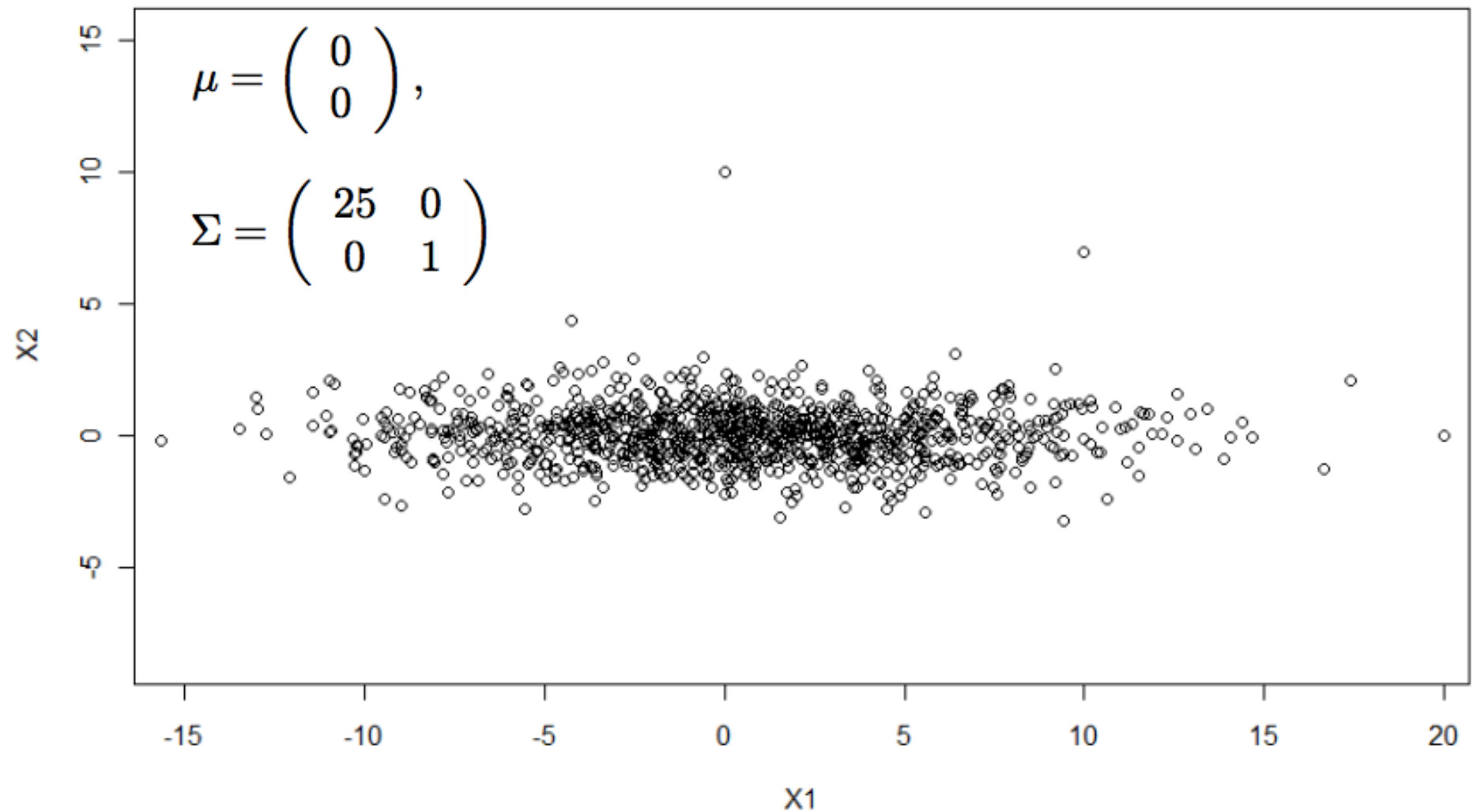
Remover a Covariância

- 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 autovalor



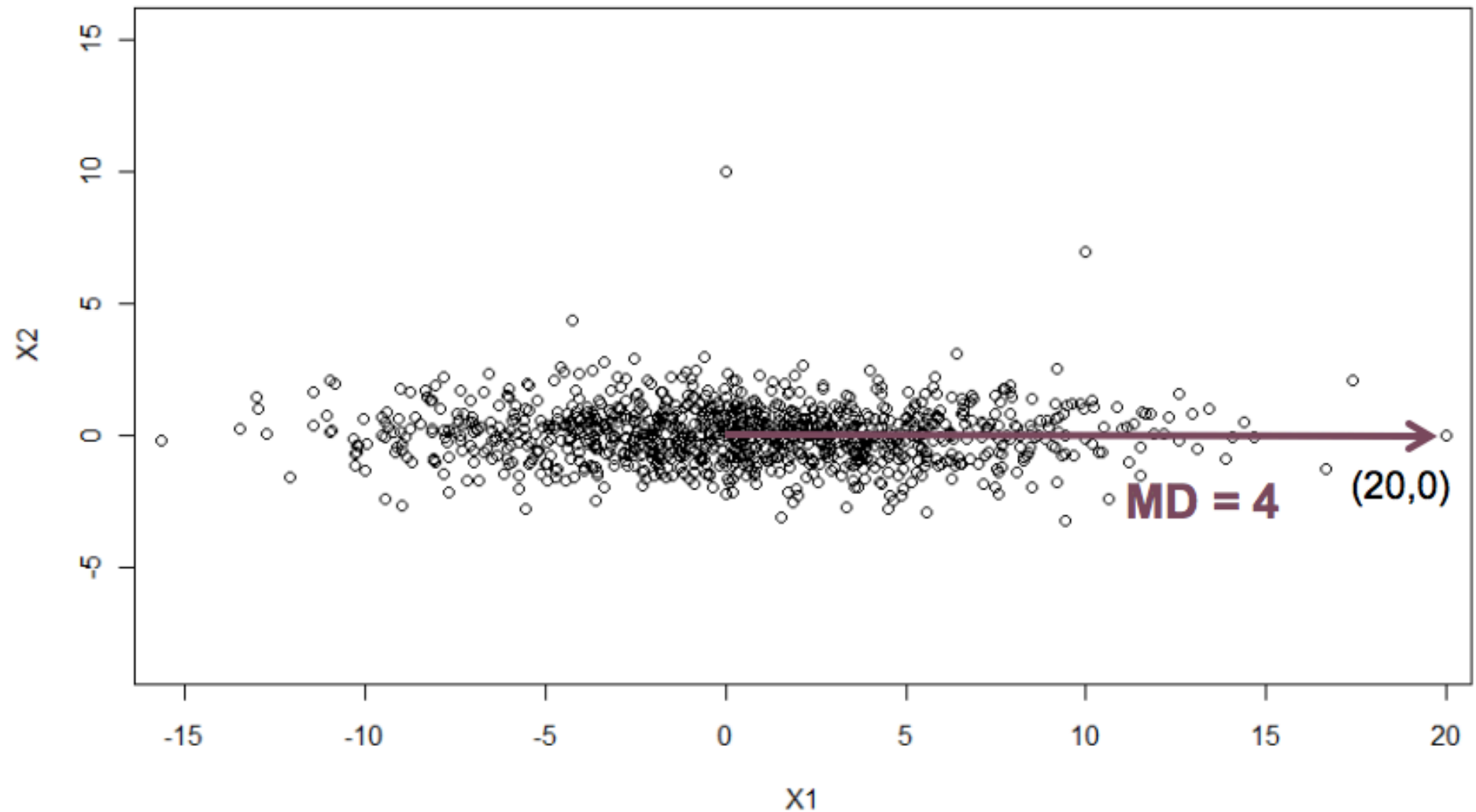


Exemplo: Distância Mahalanobis



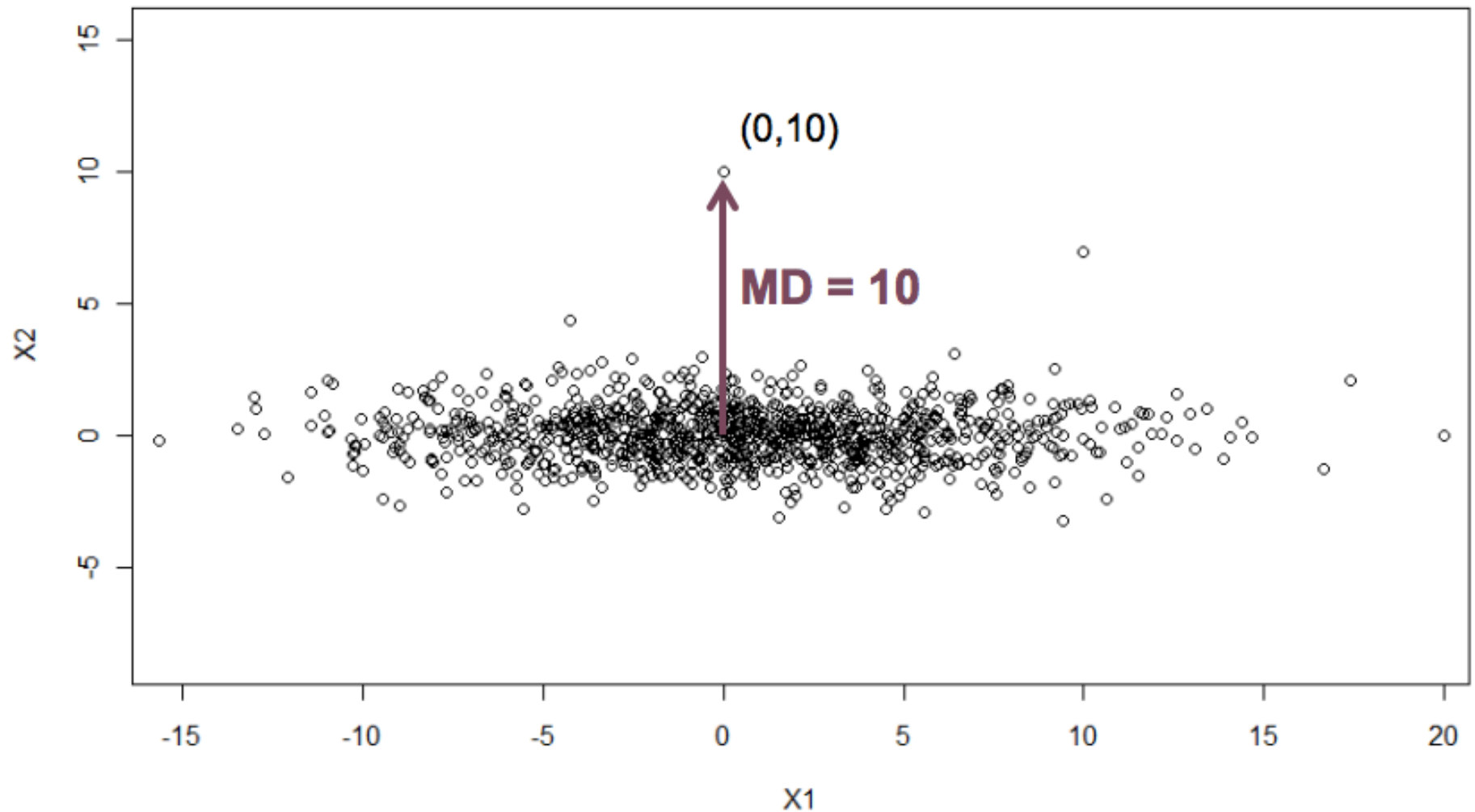


Exemplo: Distância Mahalanobis



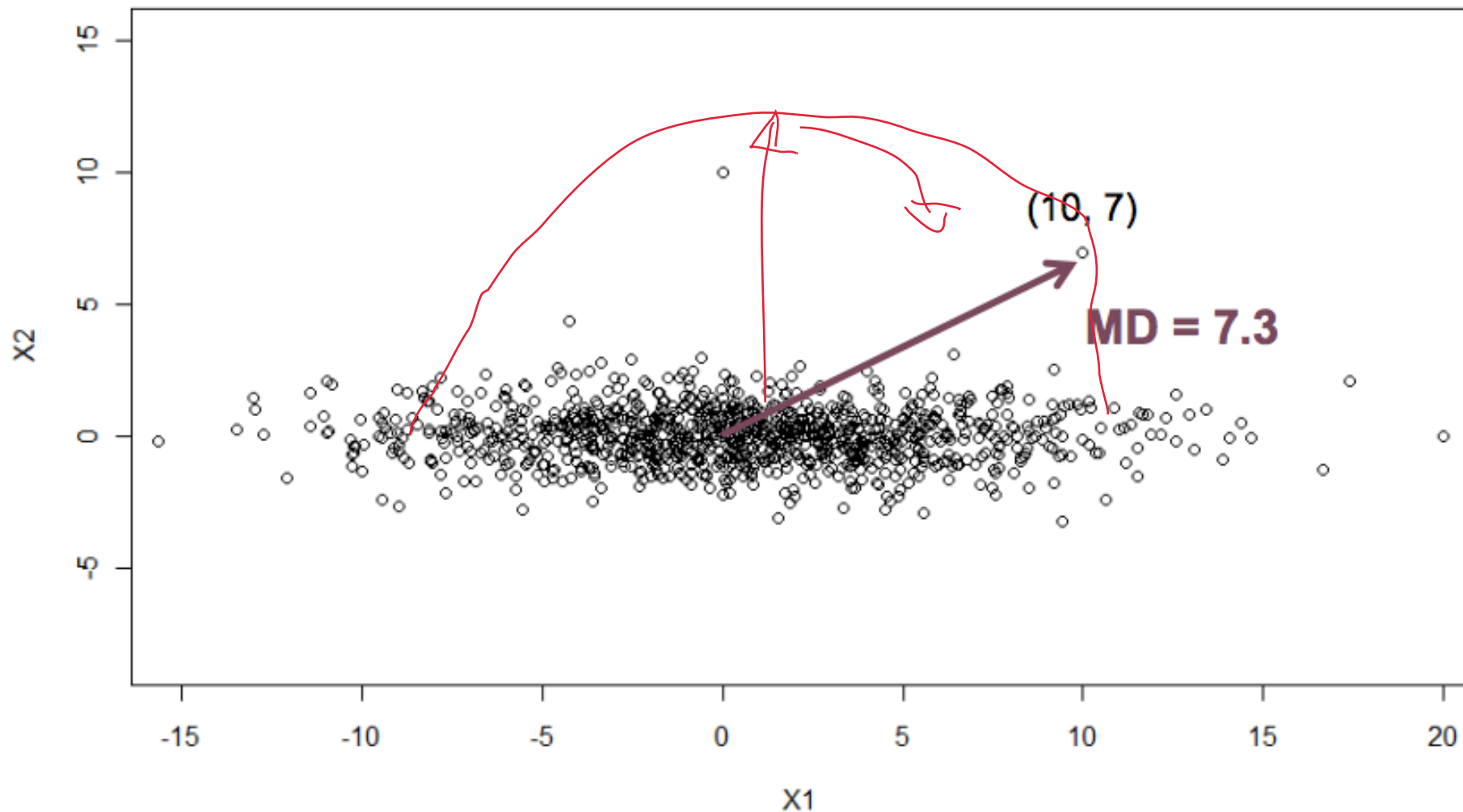


Exemplo: Distância Mahalanobis





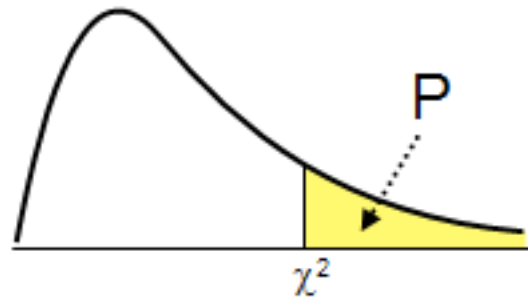
Exemplo: Distância Mahalanobis





Detectando Outliers

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



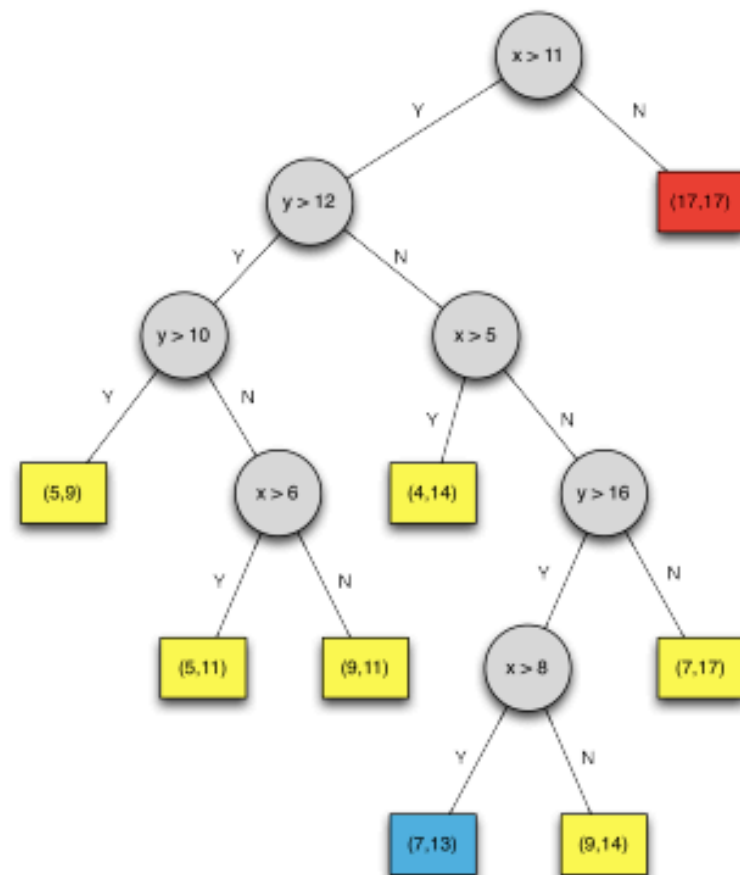
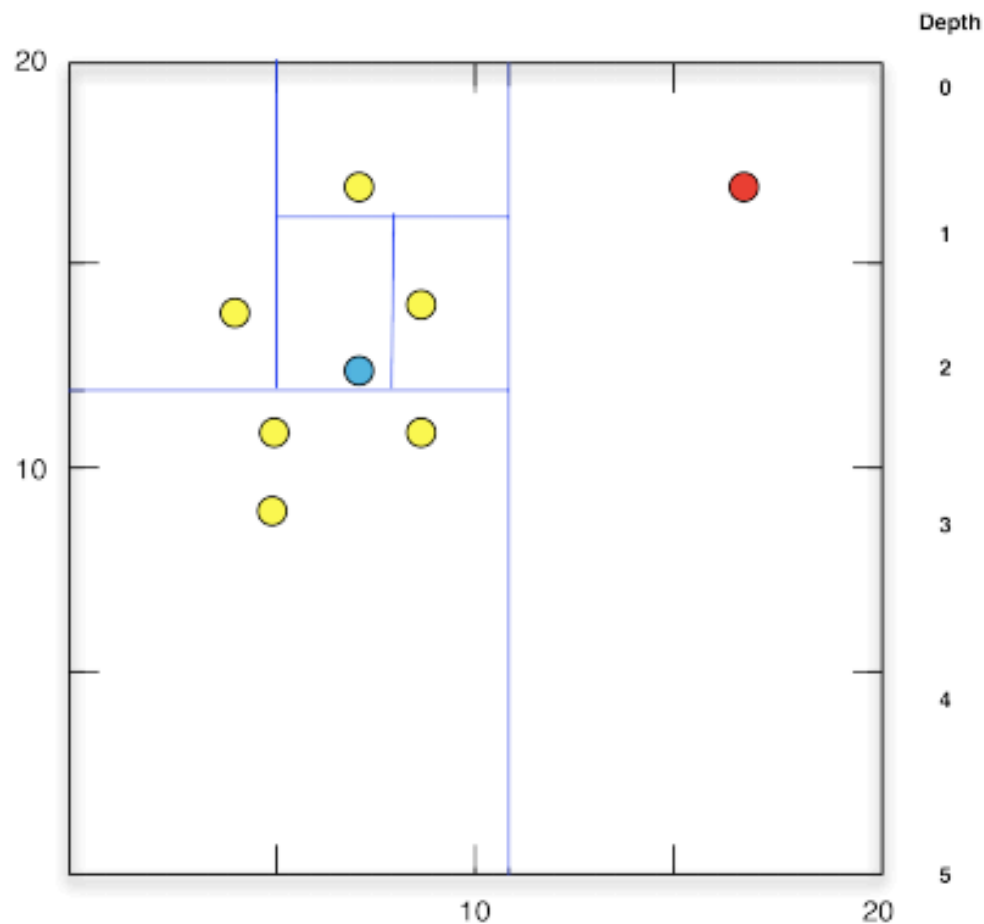


Isolation Forests

- 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





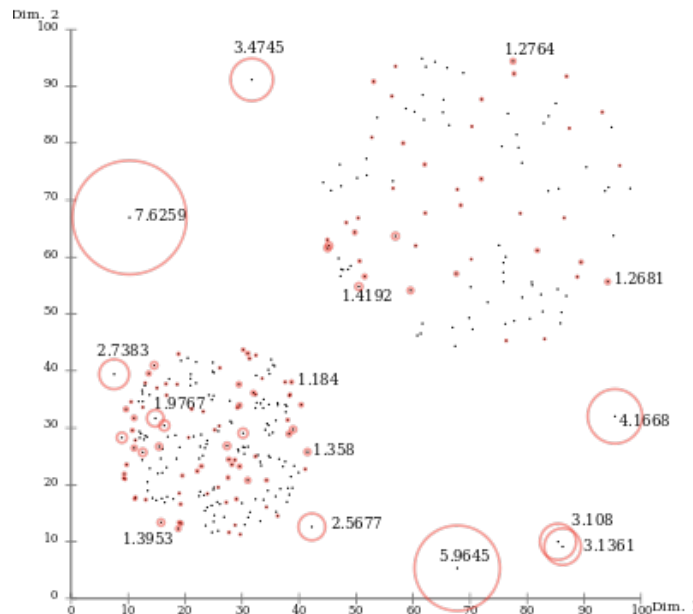
Isolation Forests

- 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

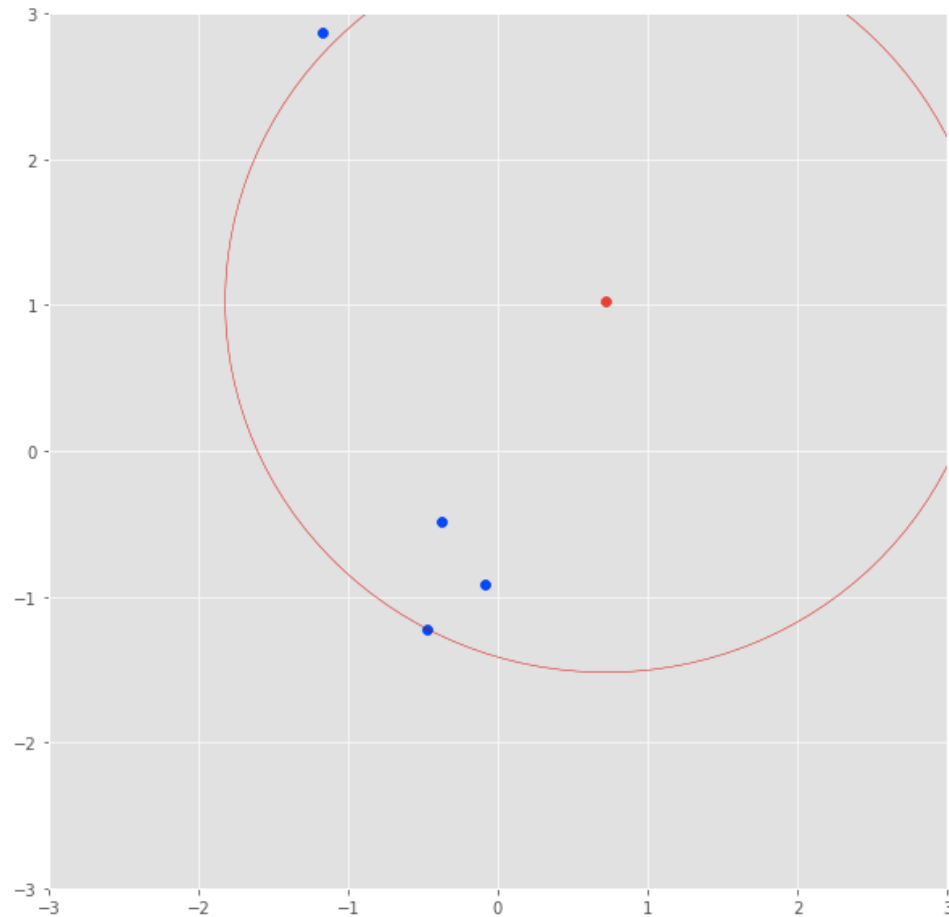
- Baseado na densidade dos dados
- Outliers: instâncias com menor densidade que os vizinhos
- Média da densidade dos k-vizinhos mais próximos sobre a densidade do ponto





K-Distance

- Distância para o k elemento mais próximo

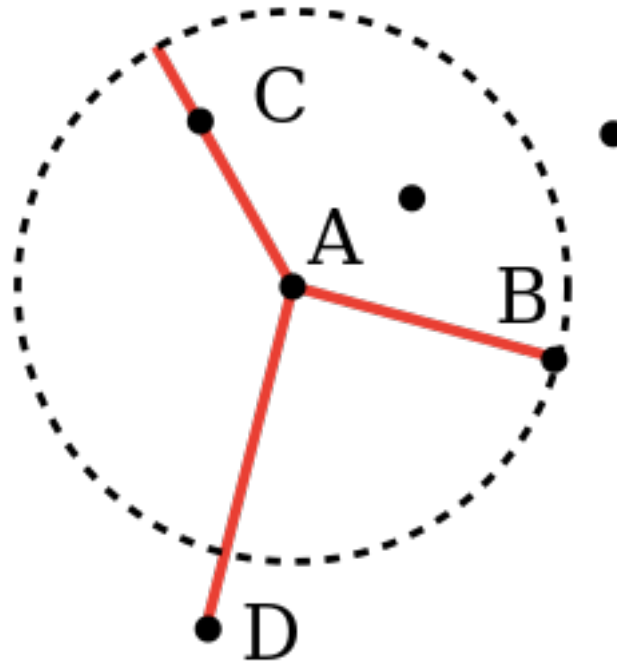




Reachability Distance

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

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

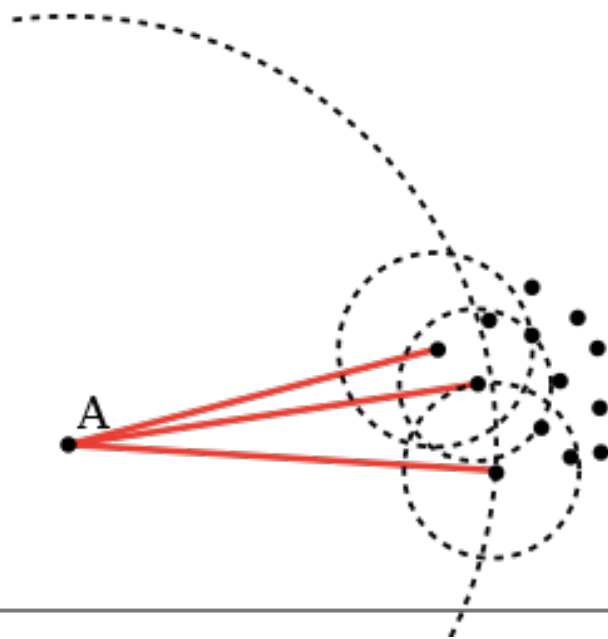




Local Reachability Density

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

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





Local Outlier Factor

- Média de lrd dos vizinhos sobre a do ponto calculado

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

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

