



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

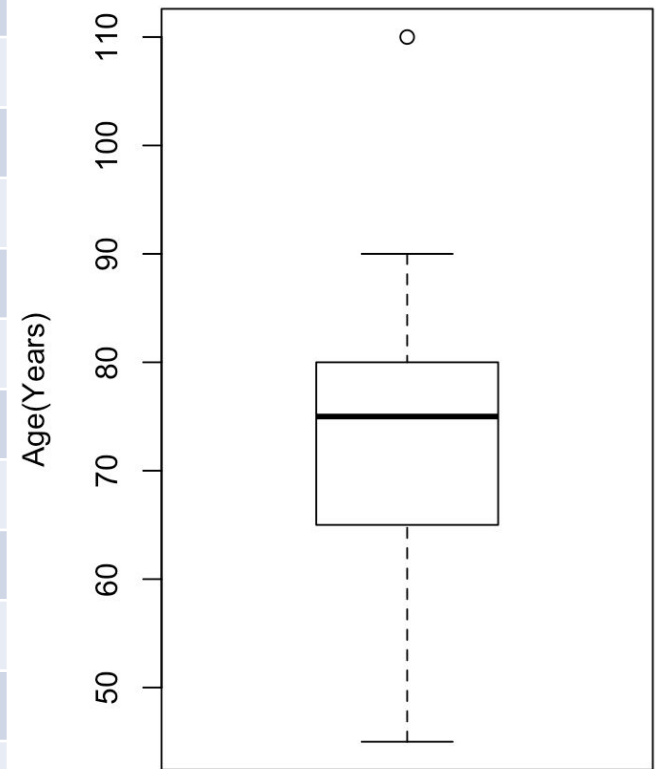
- Dados faltantes
- Dados duplicados
- Dados irrelevantes
- Dados incorretos



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 |





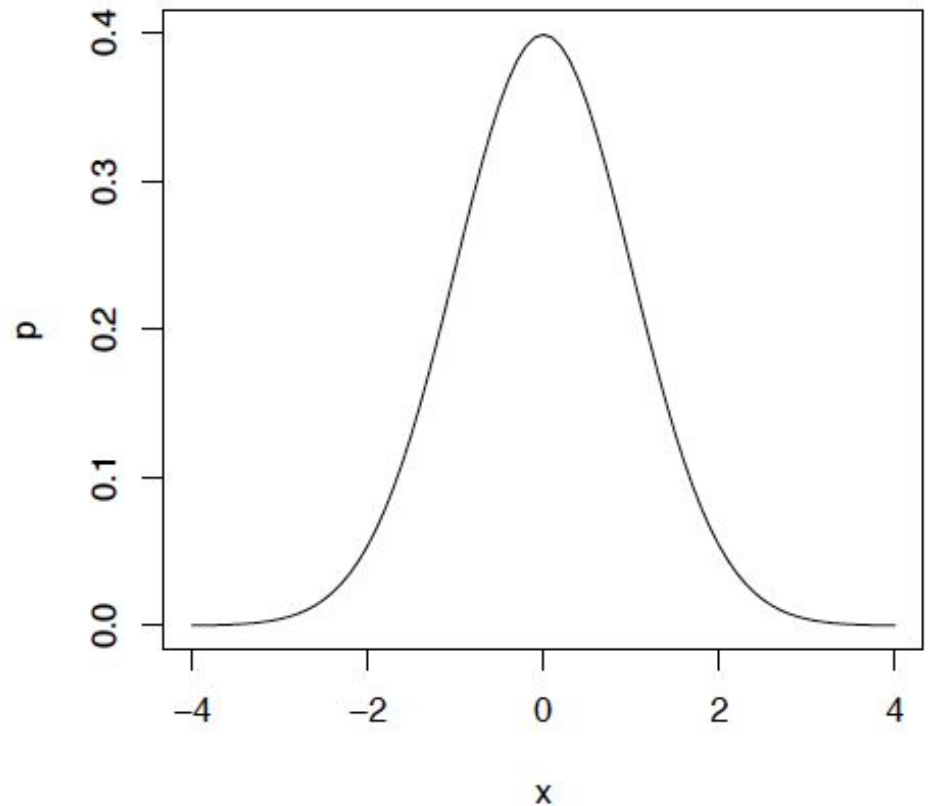
Métodos de Detecção Univariado

- 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

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



Z-Score Robusto

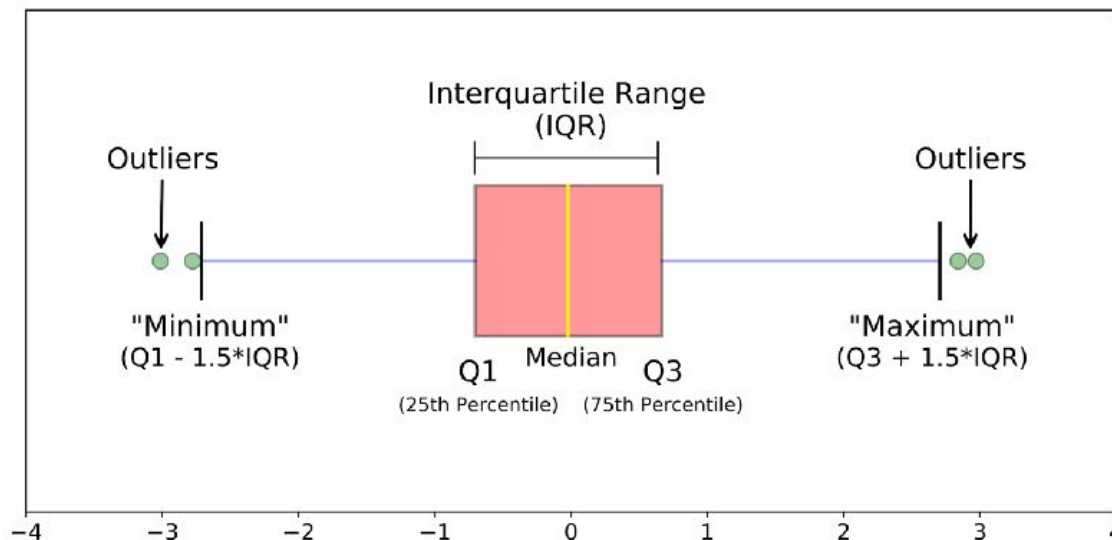
$$M_i = \frac{0.6745 (x_i - \bar{x})}{MAD}$$

- 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)



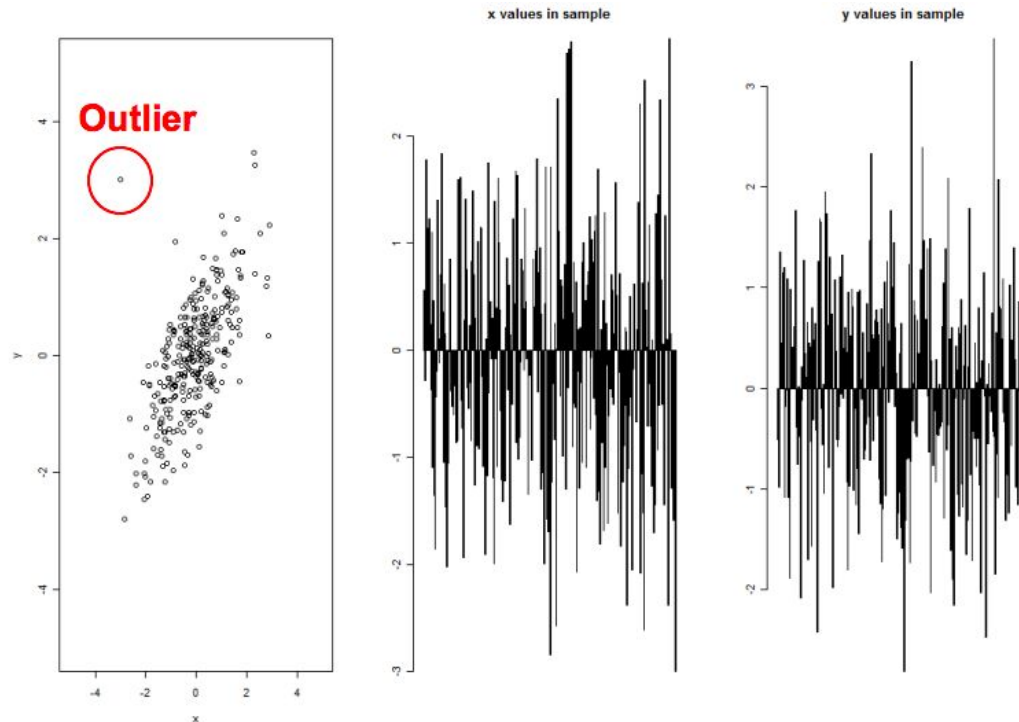
Método de Tukey

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





Bivariado



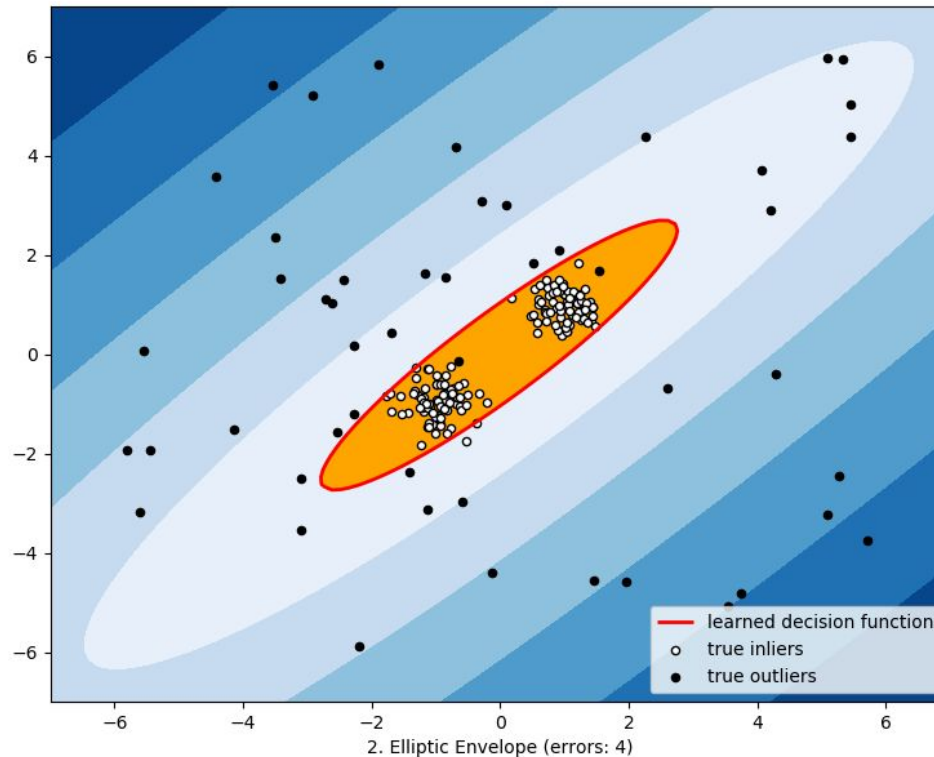
- Solução: transformar a relação em univariada (ex.: razão de uma variável pela outra)



Multivariado: Elliptic Envelope

- Suposição: atributos seguem gaussiana

Outlier detection via Elliptic Envelope





Multivariado: Elliptic Envelope

- Suposição: atributos seguem gaussiana
- Utiliza distância Mahalanobis

$$d_{L2}(x, y) = \sqrt{(x - y)^T (x - y)}$$

Euclidiana

$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

Mahalanobis

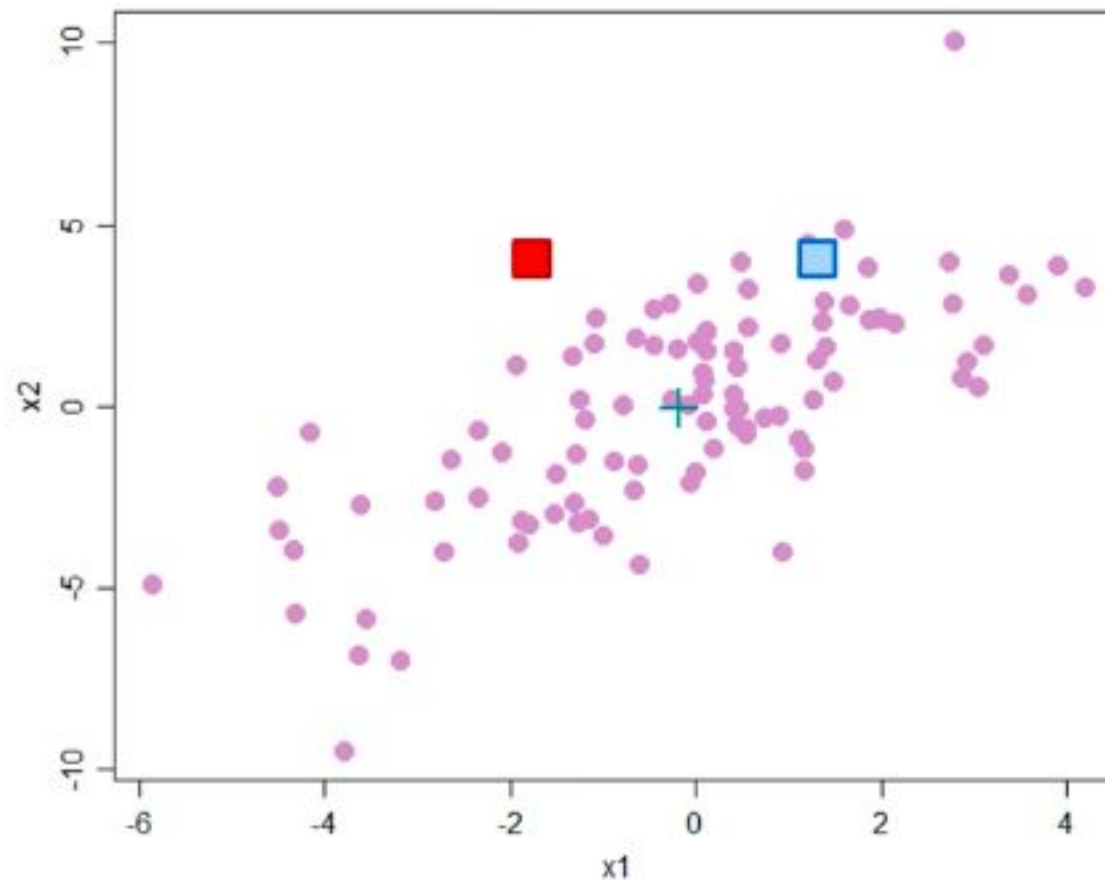


Matriz de covariância



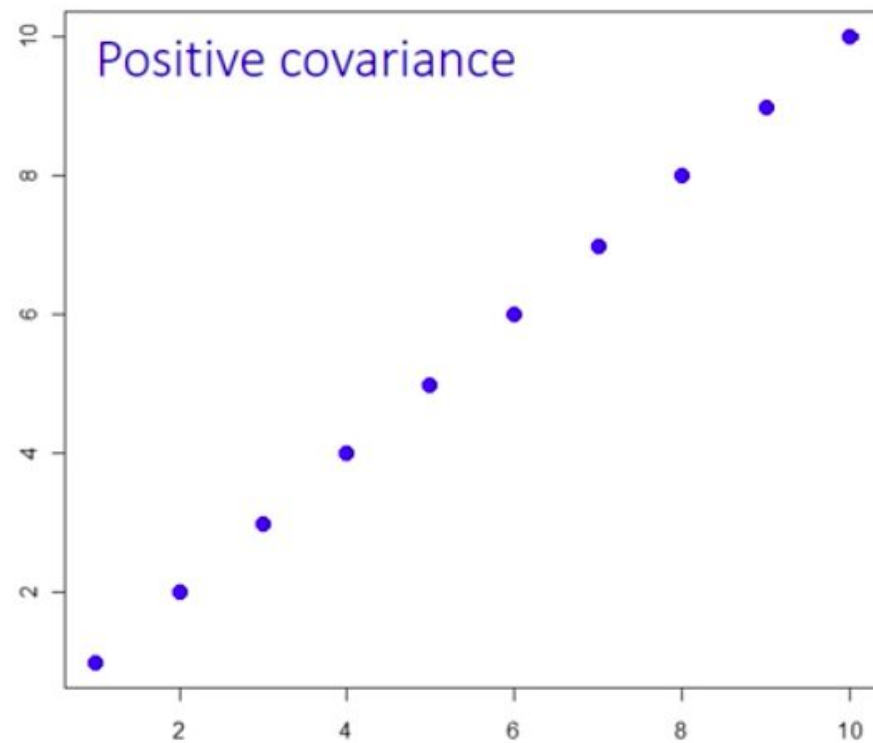
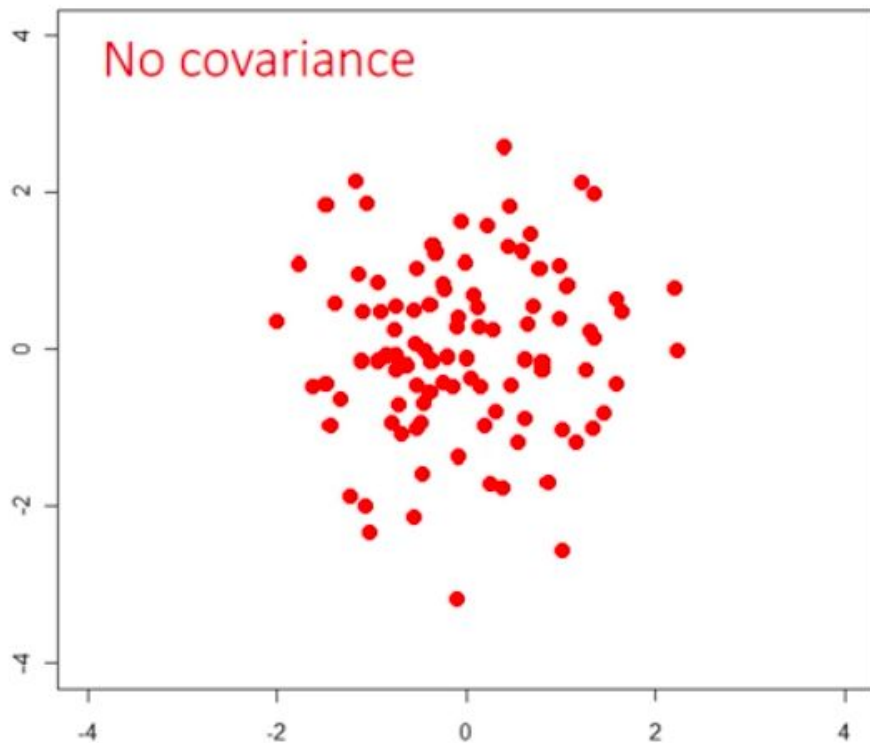
Limitação da Distância Euclidiana

- Covariância entre as variáveis





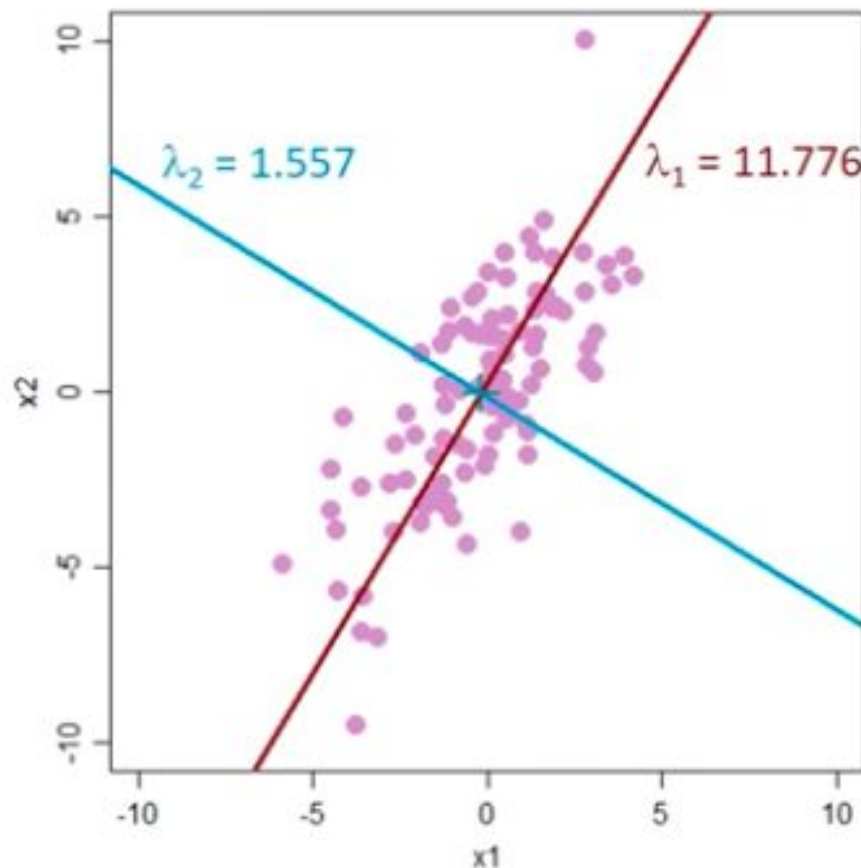
Covariância





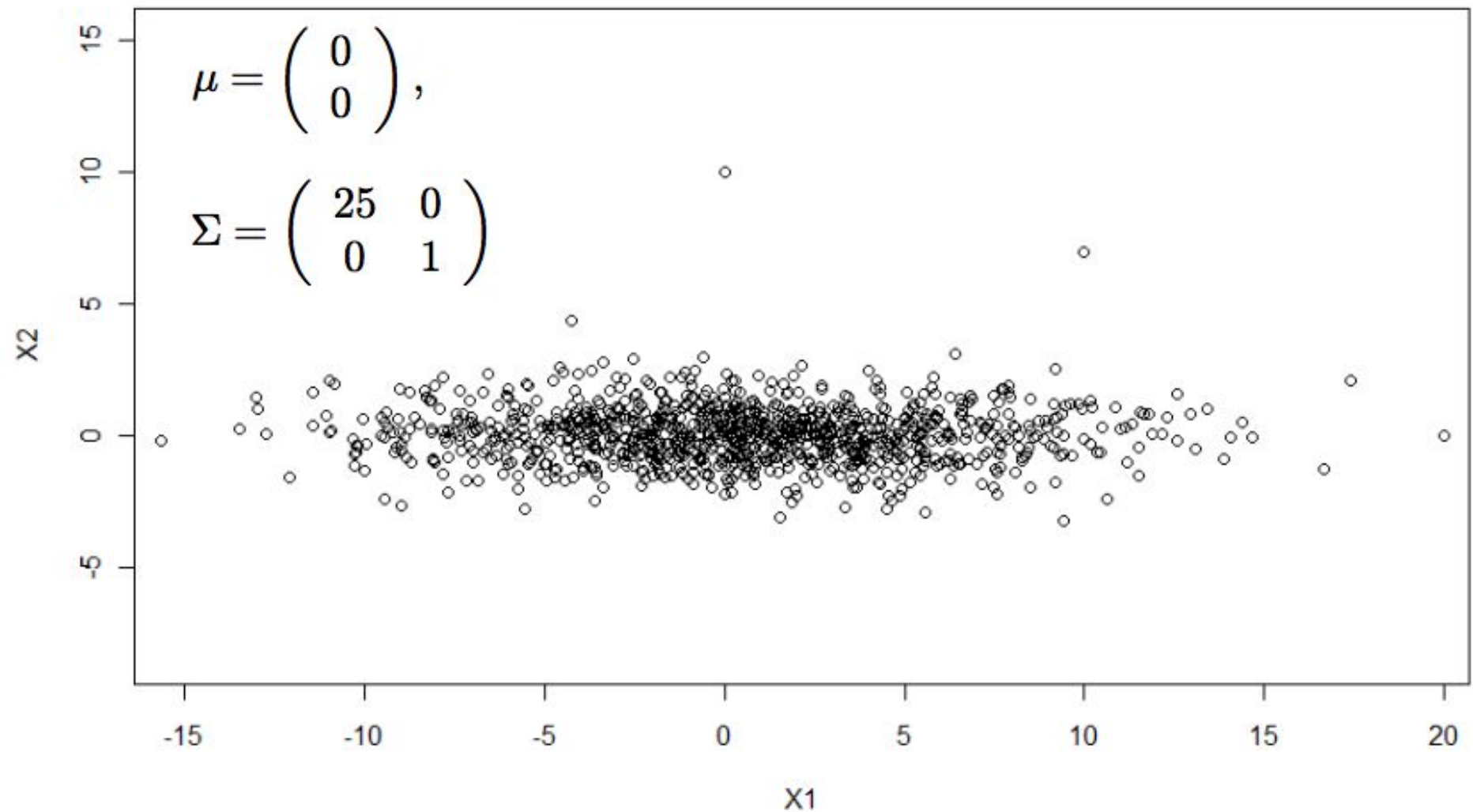
Remover a Covariância

- 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



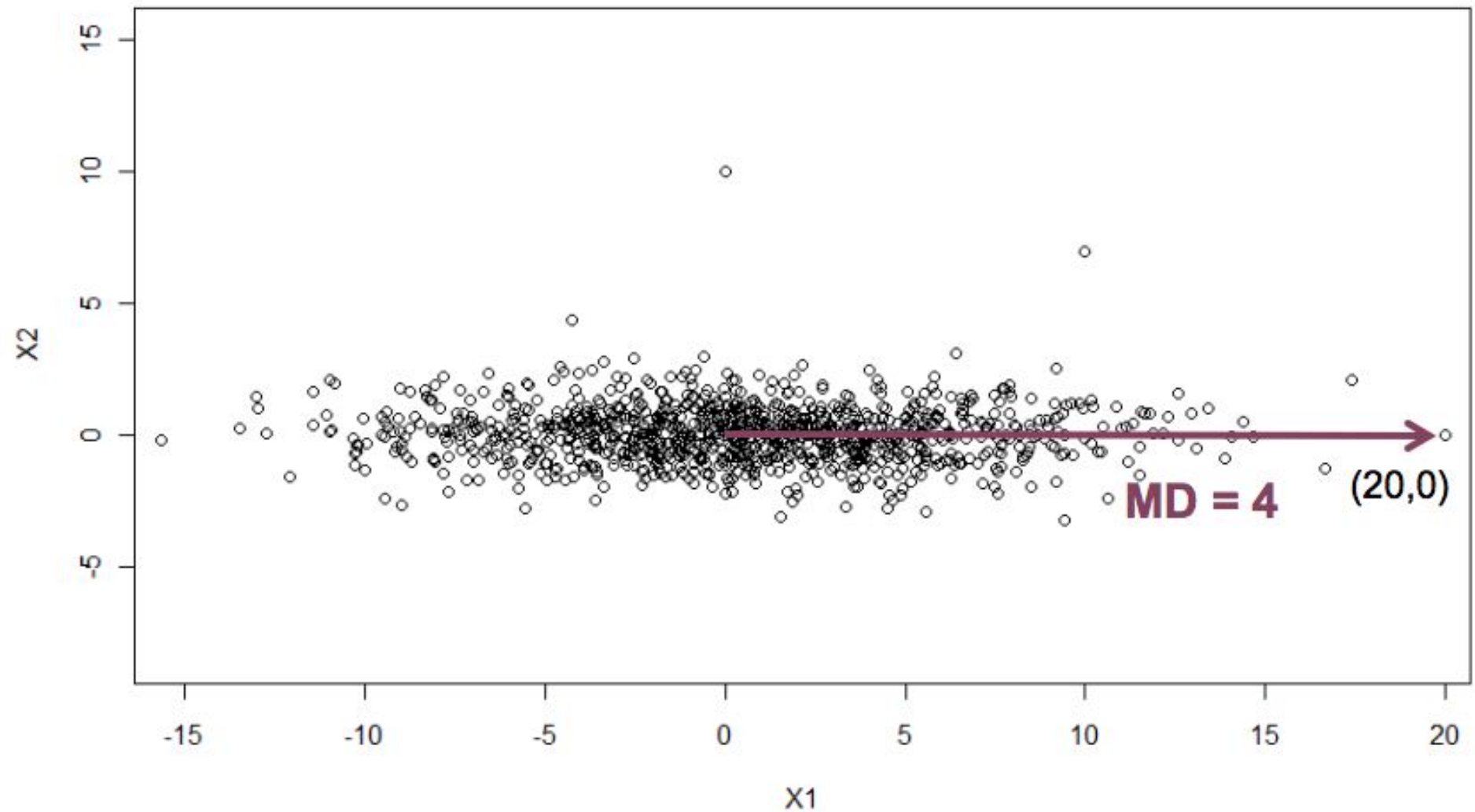


Exemplo: Distância Mahalanobis



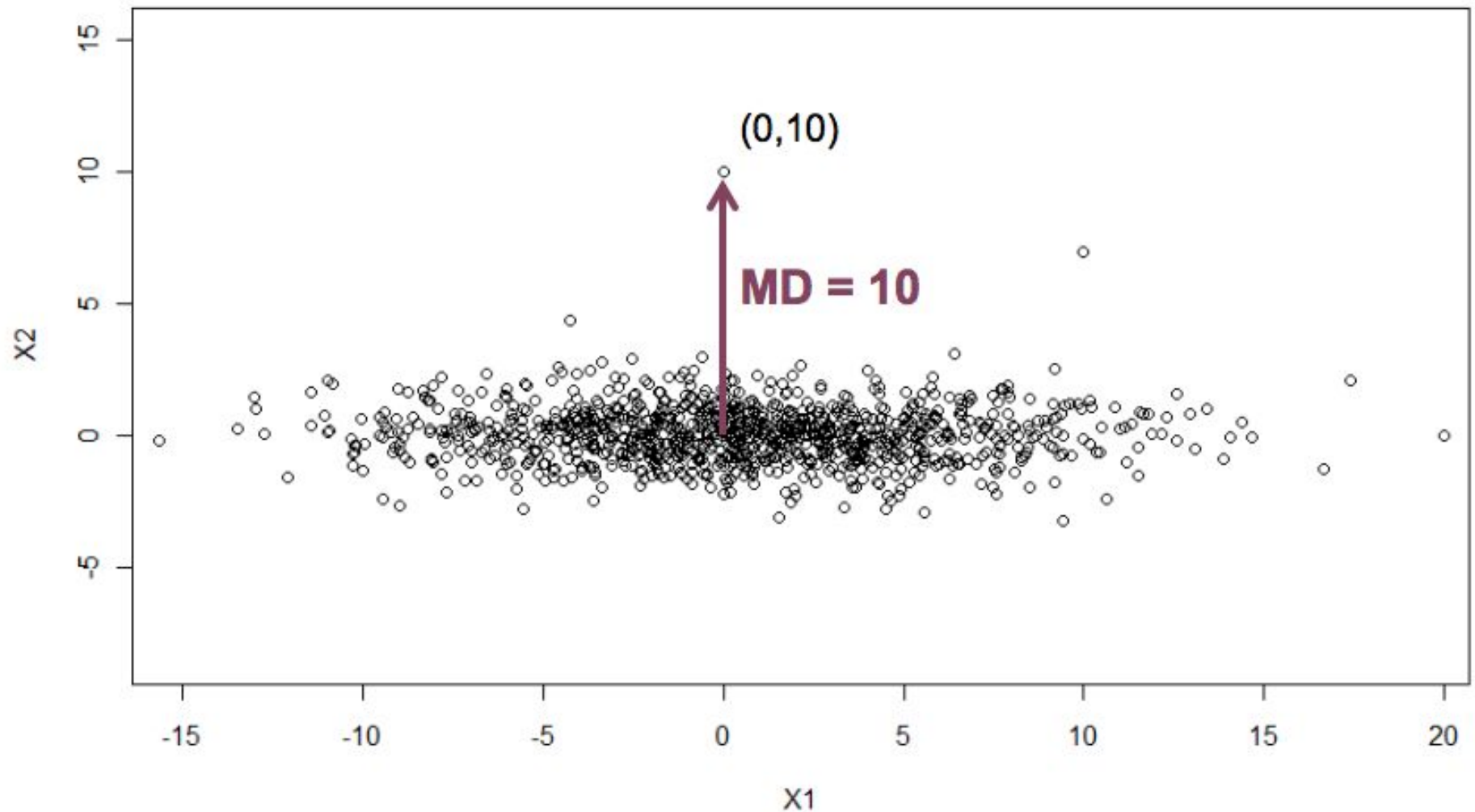


Exemplo: Distância Mahalanobis



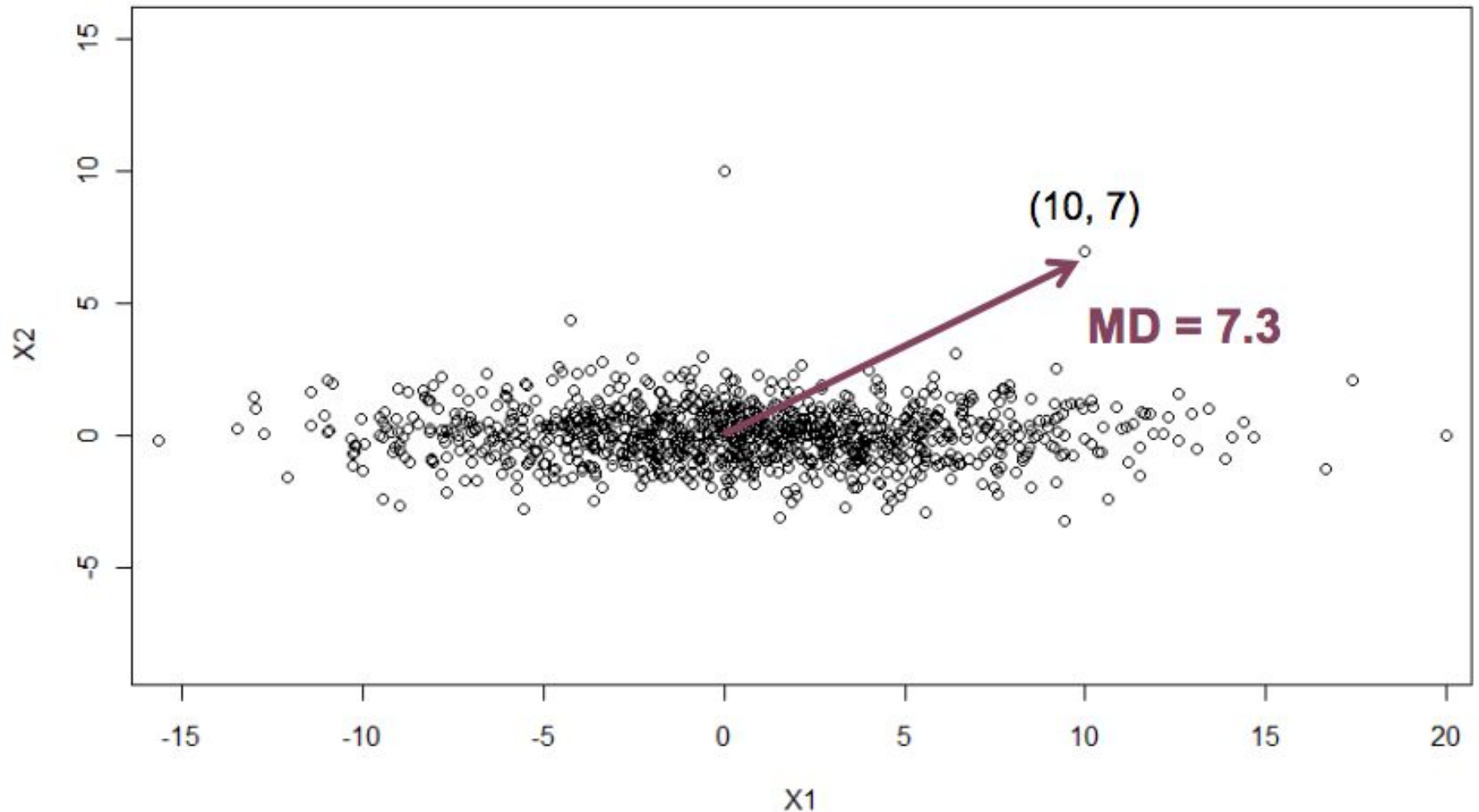


Exemplo: Distância Mahalanobis





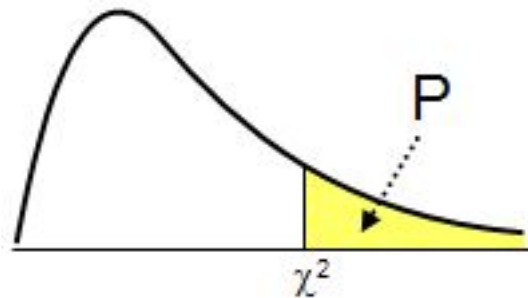
Exemplo: Distância Mahalanobis





Detectando Outliers

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



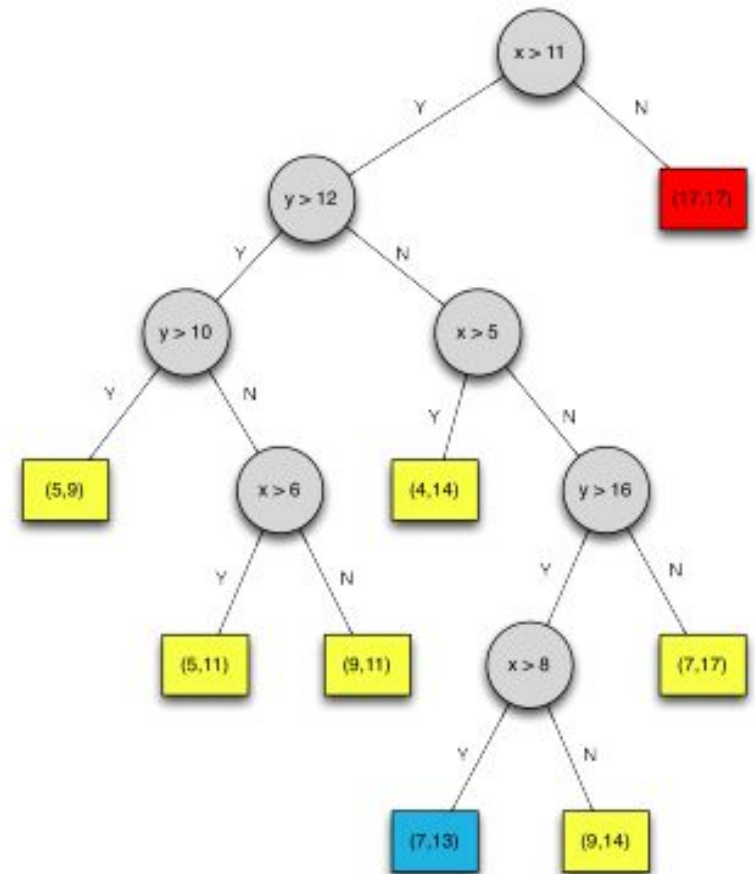
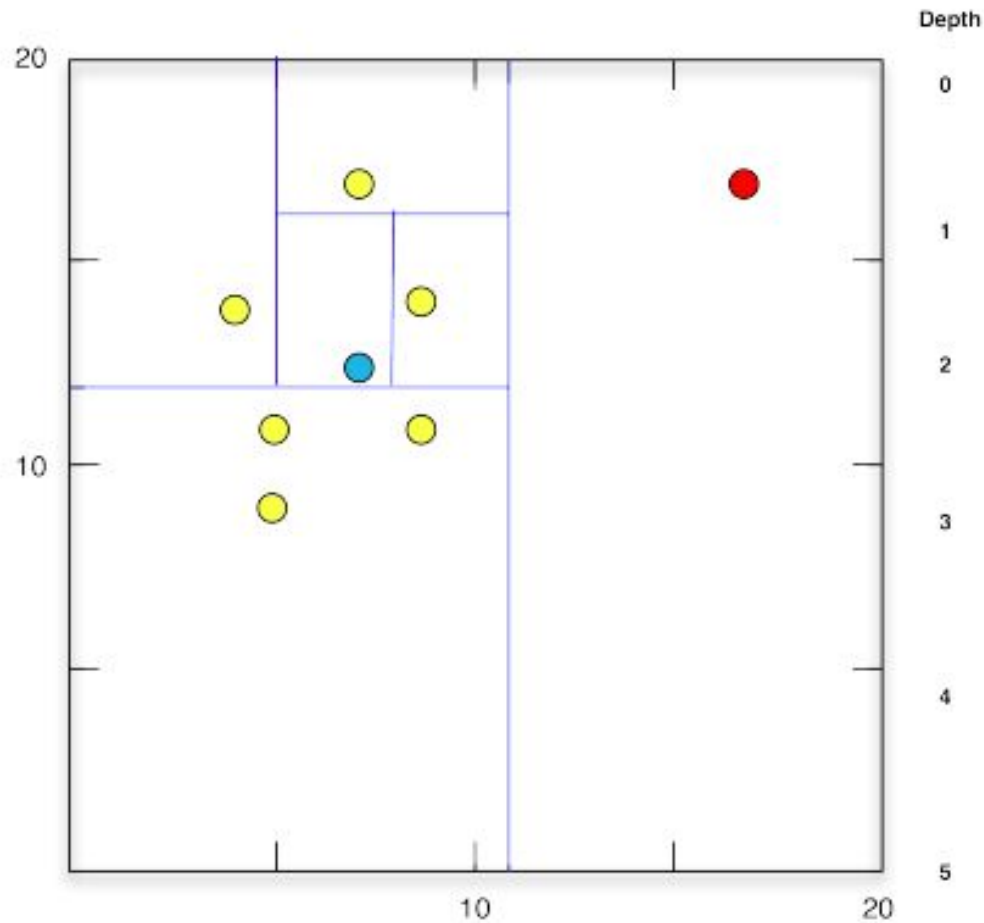


Isolation Forests

- 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