Introdução à Inteligência Computacional

Métodos de Reamostragem

Cristiano Leite de Castro - crislcastro@ufmg.br
André Paim Lemos – andrepaim@ufmg.br

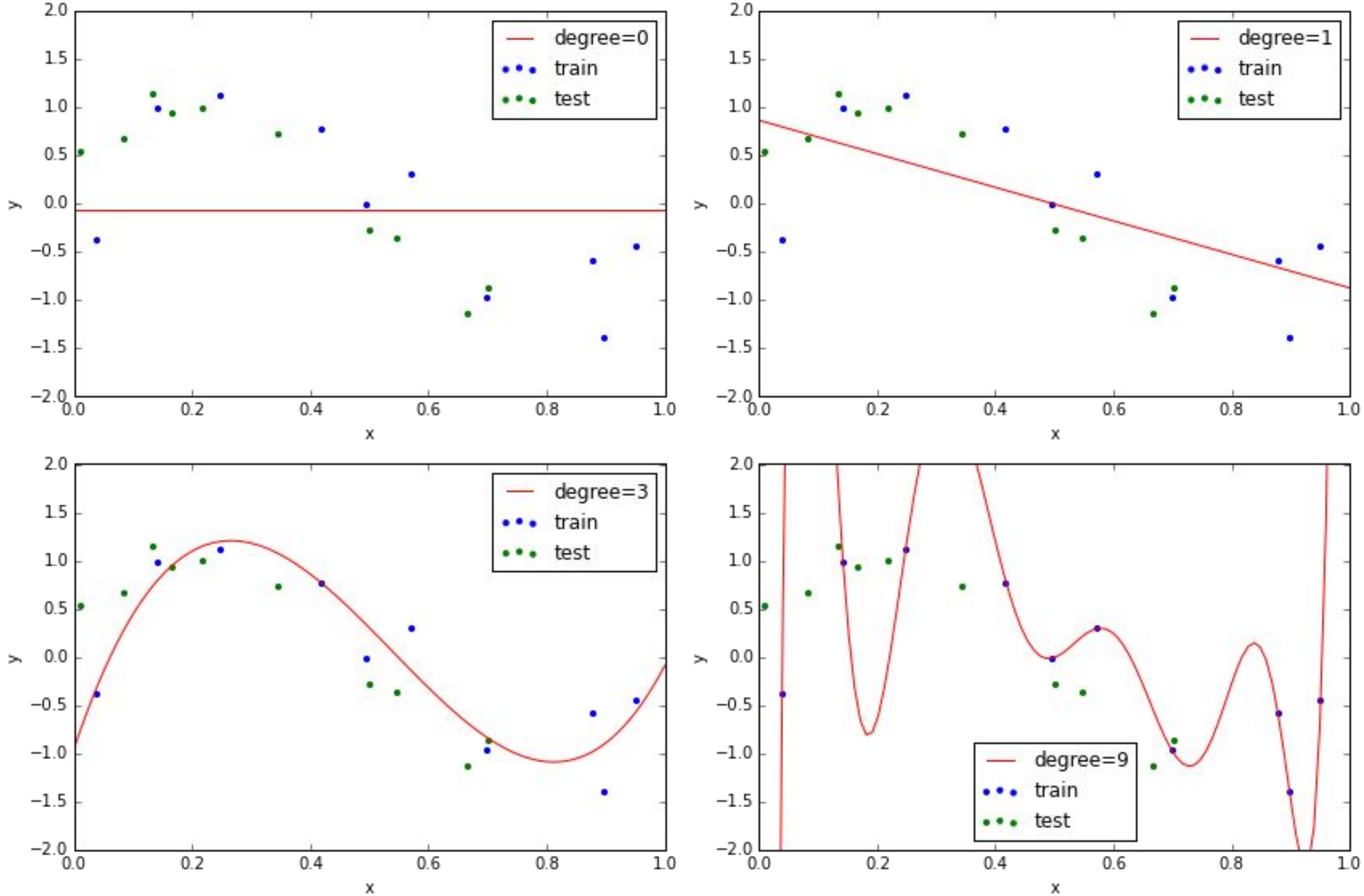
Sumário

- Validação Cruzada
  - Conjunto de validação
  - Leave-one-out cross validation
  - K-fold cross validation
  - Compromisso entre Viés-Variância
  - Aplicação em Classificação de Padrões
- Bootstrap

Métodos de Reamostragem

- Técnicas que envolvem várias reamostragens de observações do conjunto de treinamento seguida de ajustes de modelos
- Importantes para inferir informações do modelo ajustado
  - Estimar o erro de teste do modelo *(model assessment)*
  - Selecionar o modelo com a capacidade adequada para o problema *(model selection)*
- Alto Custo computacional

Erro de Teste

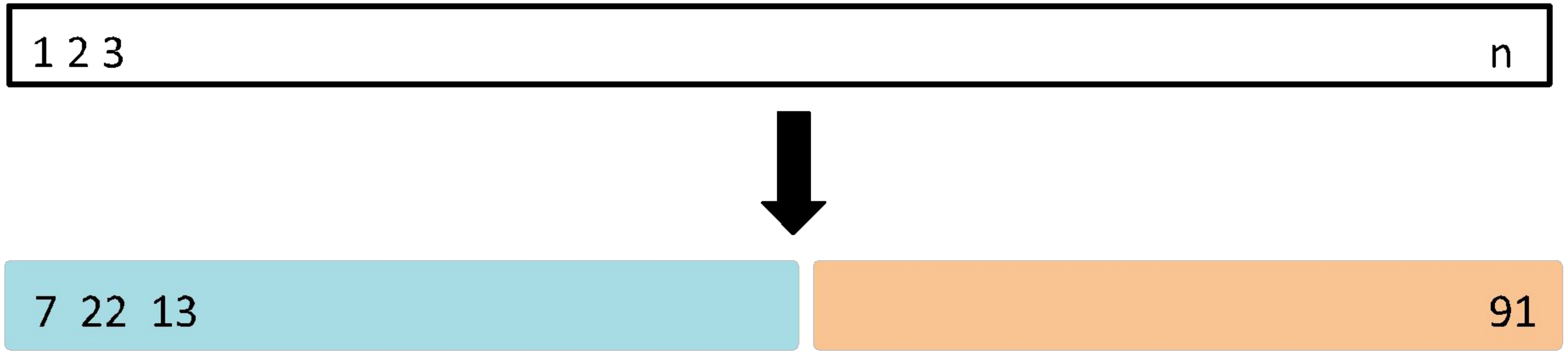


Conjunto de Validação

- Se o conjunto de dados for suficientemente grande (*Hold-out cross validation*)
  - Divide-se o conjunto em um conjunto de treinamento e outro de validação
  - Utiliza-se o conjunto de treinamento para estimar modelos com todas as combinações das variáveis
  - Escolhe-se o subconjunto das variáveis associada ao modelo com menor erro no conjunto de validação

Conjunto de Validação

- Seleciona-se observações aleatoriamente para cada partição



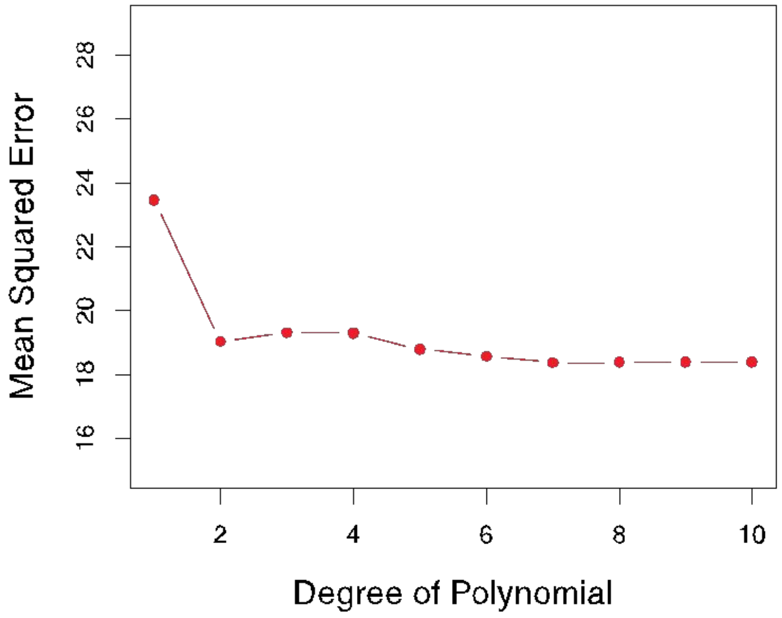
Conjunto de Validação

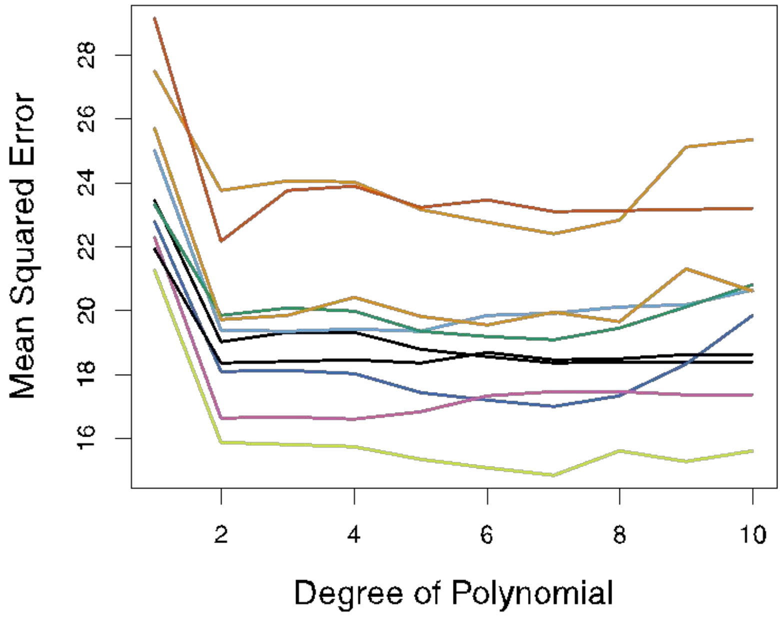
- Exemplo:
  - Base de dados *Auto*

- Dois modelos
  - mpg ~ horsepower
  - mpg ~ horsepower + horsepower^2

- Qual modelo gera o melhor erro de teste?
  - Divide-se a base de dados em dois conjuntos: treinamento (196 observações), validação (196)
  - Estima-se os dois modelos (conjunto de treinamento)
  - Avalia-se os modelos usando MSE (conjunto validação)
  - O modelo associado ao menor MSE é o escolhido

Conjunto de Validação





Conjunto de Validação

- Vantagens
  - Simples
  - Baixo custo computacional
- Desvantagens
  - Alta variabilidade do MSE de validação
  - Apenas um subconjunto dos dados é utilizado para estimar o modelo
    - O desempenho de métodos estatísticos tendem a degenerar com a redução do número de observações de treinamento
    - Tende a *sobre-estimar* o erro de teste

Leave-One-Out Cross Validation

- LOOCV
- Similar ao método anterior, porém tenta atacar as deficiências
- Dado um conjunto de dados contendo *n* observações
  - Conjunto de treinamento: n-1 observações
  - Conjunto de validação: 1 observação

Leave-One-Out Cross Validation

- Caso as observações de [2, n] sejam utilizadas no treinamento e a primeira observação no conjunto de validação

$$MSE_{(1)} = (y_1 - \hat{y}_1)^2$$

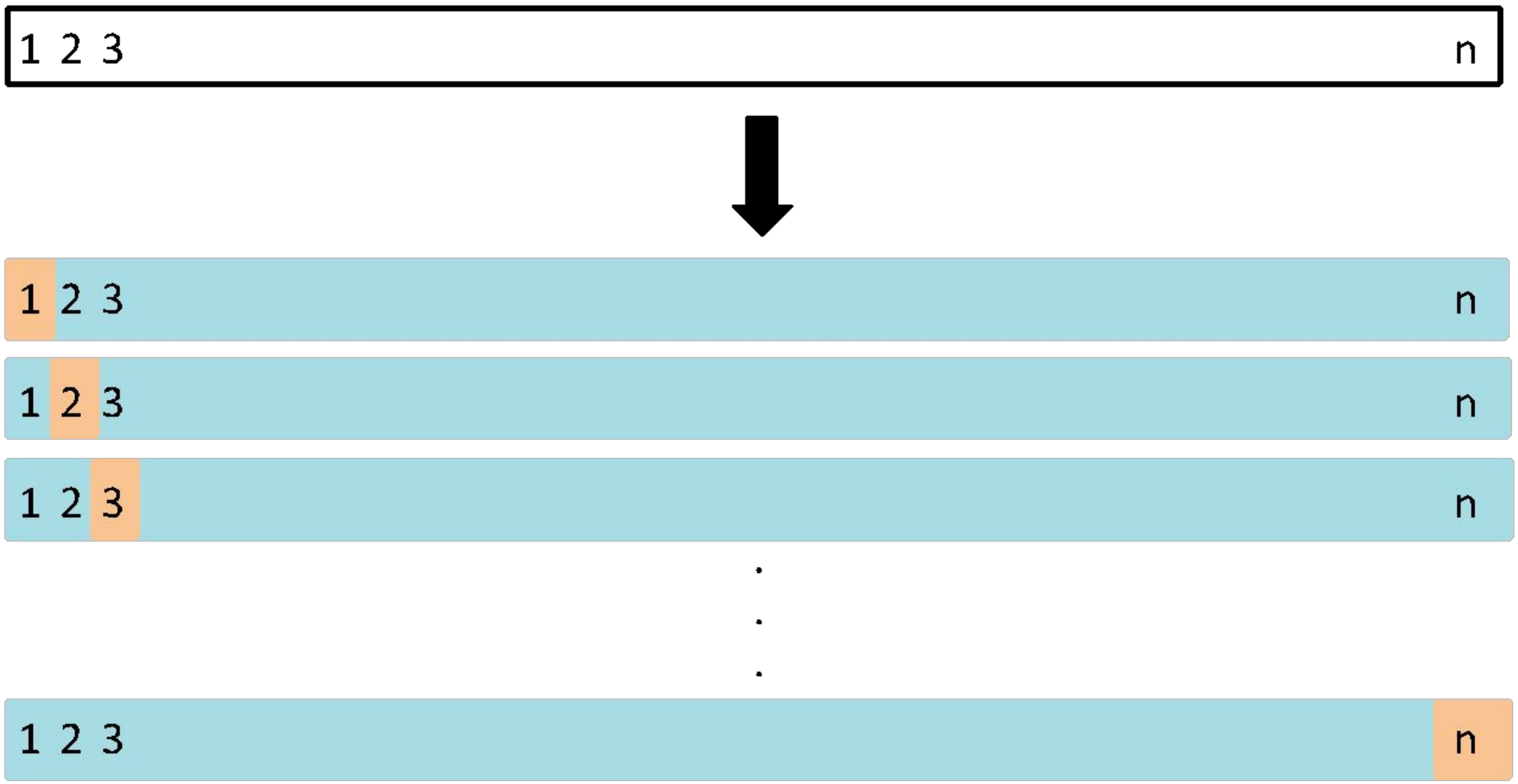
- É um estimador não viesado para o erro de teste
  - Alta variância, pois é baseado em apenas uma observação

Leave-One-Out Cross Validation

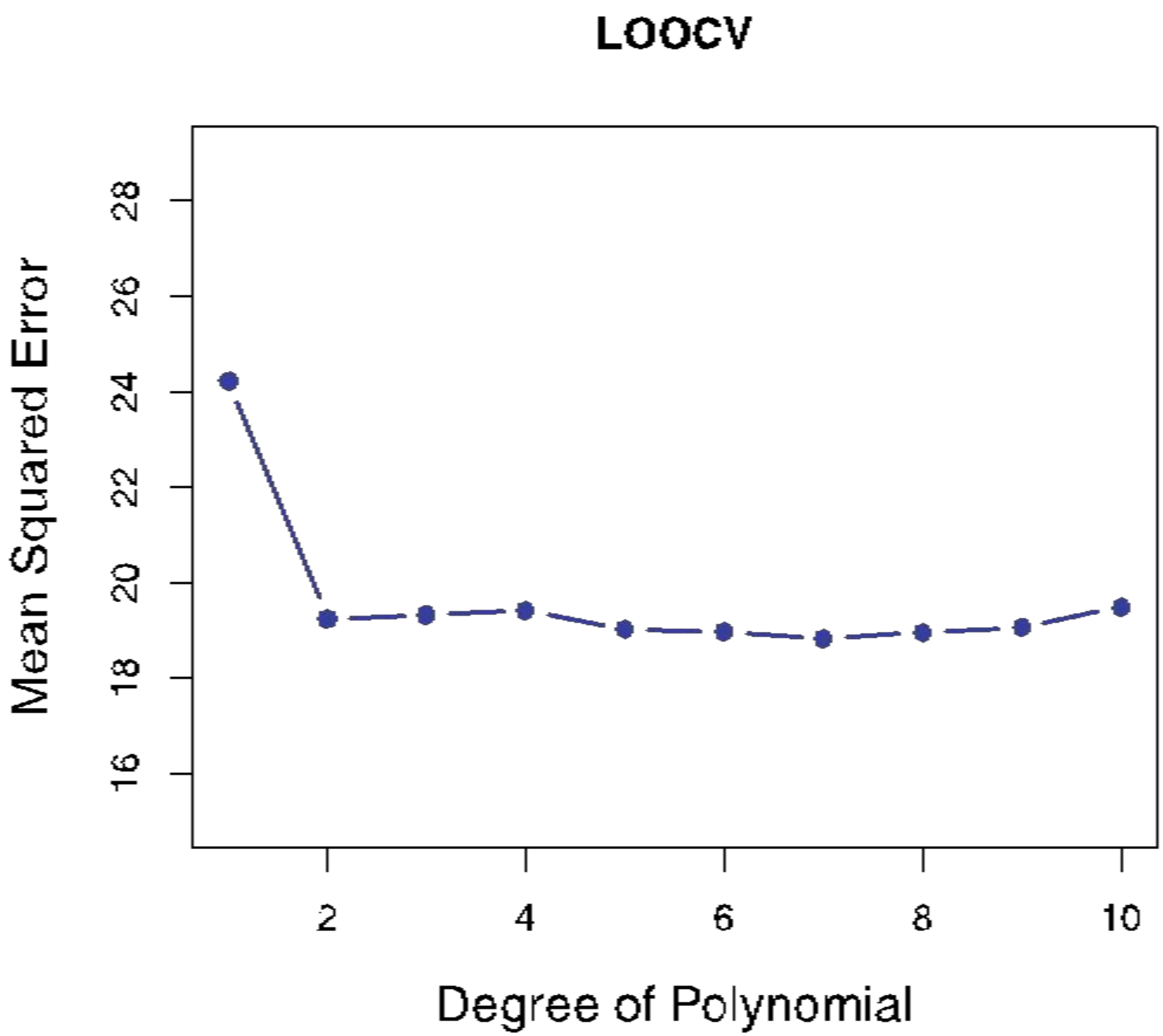
- Repete-se o processo *n* vezes
- O MSE do modelo é estimado como

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Leave-One-Out Cross Validation



Leave-One-Out Cross Validation



LOOCV vs Conjunto de Validação

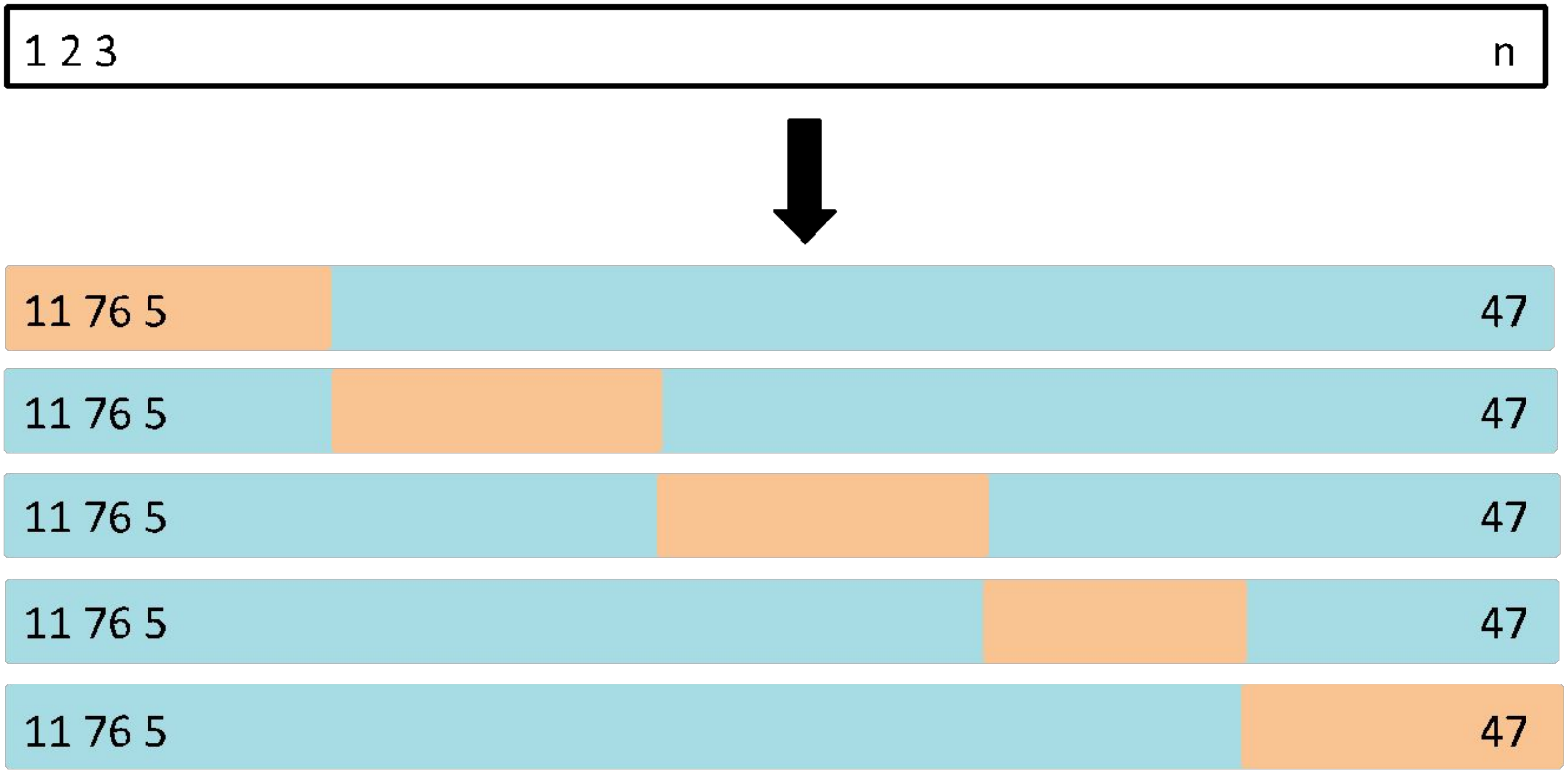
- LOOCV tem menor viés
  - *n* modelos são estimados utilizando um conjunto de treinamento que contém *n-1* observações
- LOOCV estima um MSE com menor variabilidade
  - A abordagem baseada no conjunto de validação estima um MSE diferente cada vez que é executado
  - LOOCV estima sempre o mesmo MSE, dado que o processo de reamostragem não é aleatório
- LOOCV tem um alto custo computacional!
  - Principalmente quando *n* é grande

K-fold Cross Validation

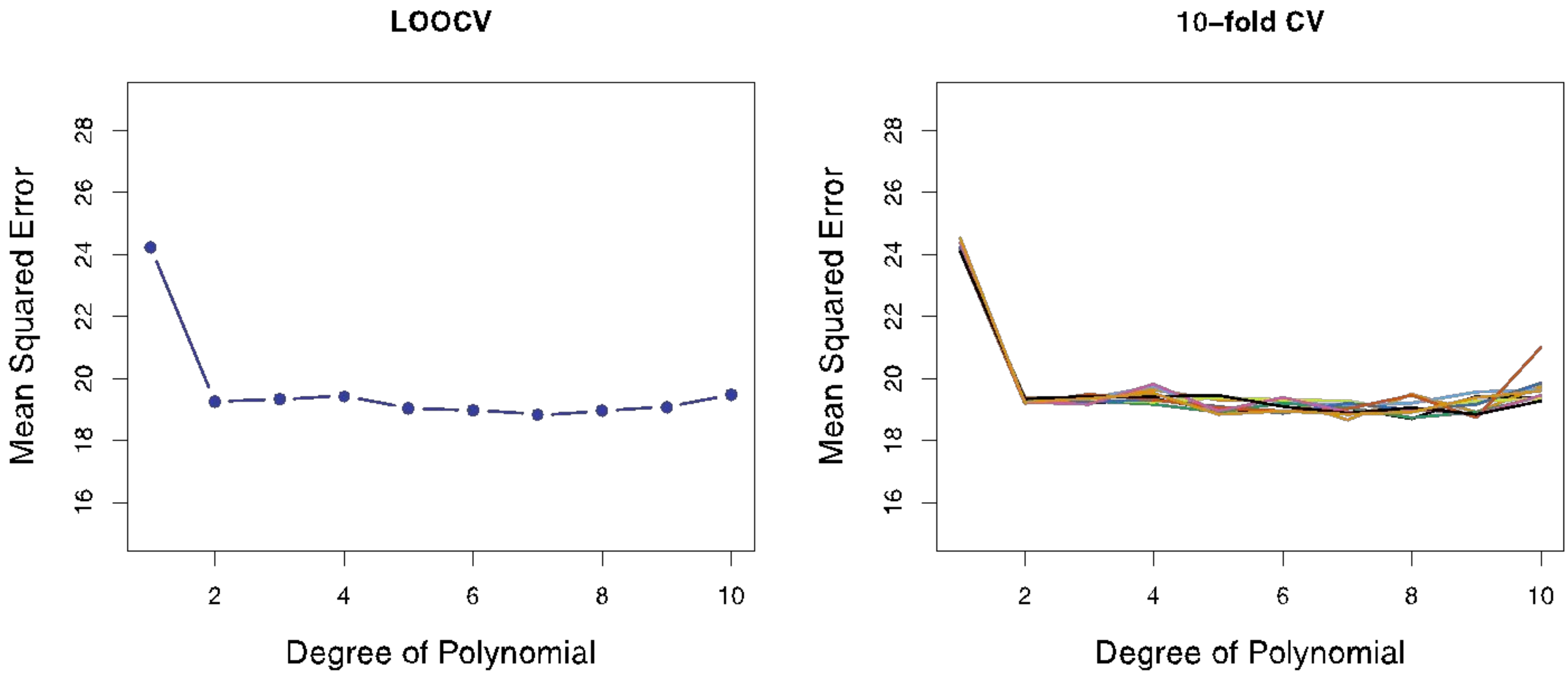
- Compromisso entre os dois métodos anteriores
- Divide-se o conjunto de dados em k partições (k=5 ou k=10, por exemplo)
- Estima-se o modelo com k-1 partições e calcula o MSE para a partição remanescente
- Repete-se esse processo k vezes

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

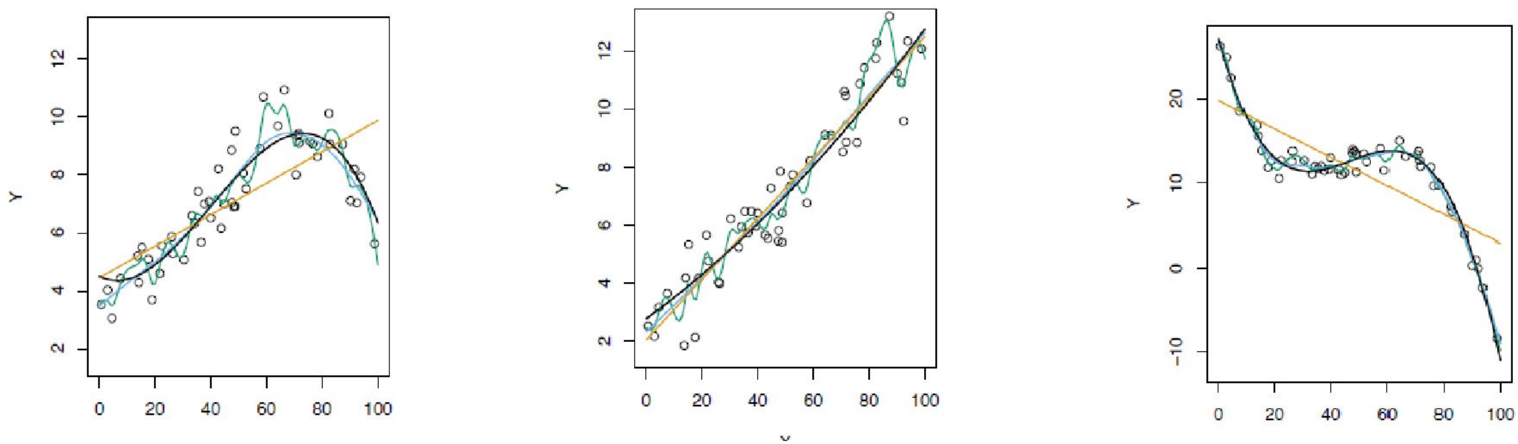
K-fold Cross Validation

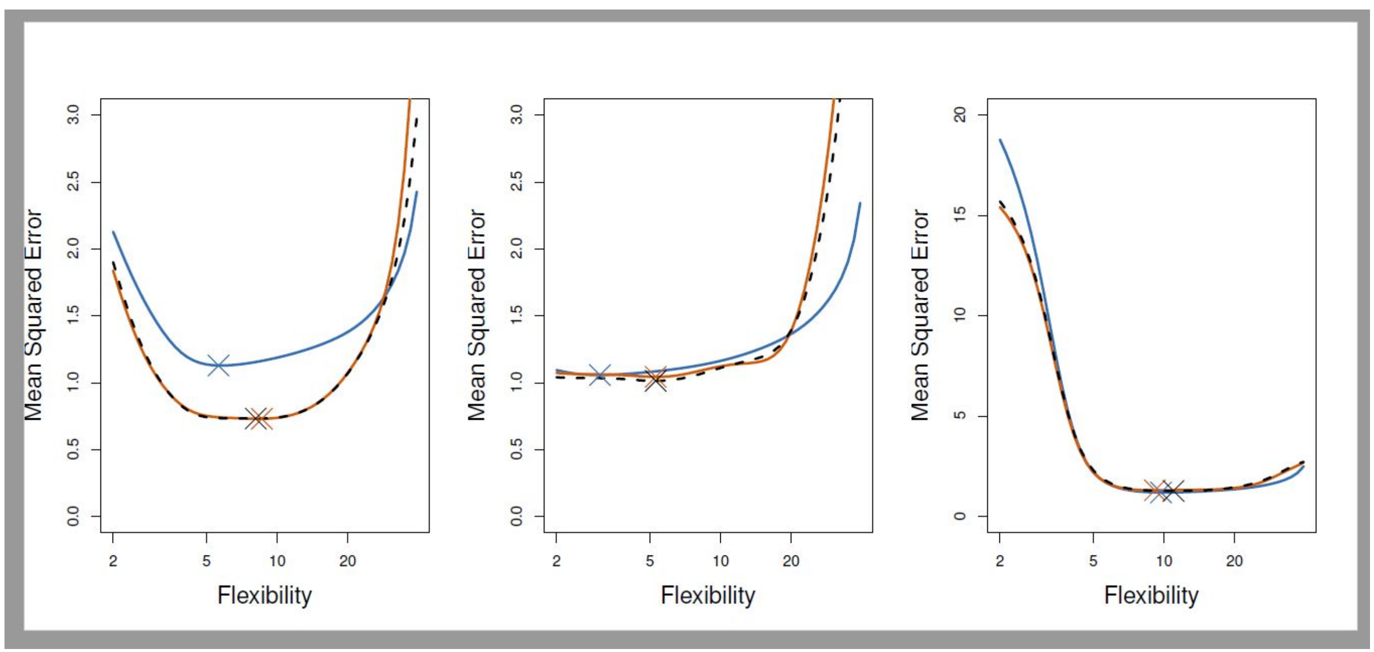


K-fold Cross Validation



LOOCV vs K-fold Cross





Compromisso entre Viés-Variância

- A abordagem do conjunto de validação tem um alto viés
  - Apenas 50% das observações são usadas no treinamento
- LOOCV possui um viés muito baixo
  - n-1 amostras utilizadas no treinamento
- K-fold CV tem um viés intermediário
  - K-1 partições no treinamento

Compromisso entre Viés-Variância

- O viés não é o único quesito a ser levado em consideração
- LOOCV tem uma variância maior que K-fold CV
- LOOCV calcula a média de *n* modelos muito similares
  - As saídas são muito correlacionadas
- K-fold CV calcula a média de *k* *(k<n)* modelos
  - Menor correlação entre as saídas
- A média de um conjunto de valores altamente correlacionados possui uma variância maior que a de valores com menor correlação
- Na prática, k-fold CV é muito utilizado (com k=5 ou k=10)

Validação Cruzada em Problemas de Classificação

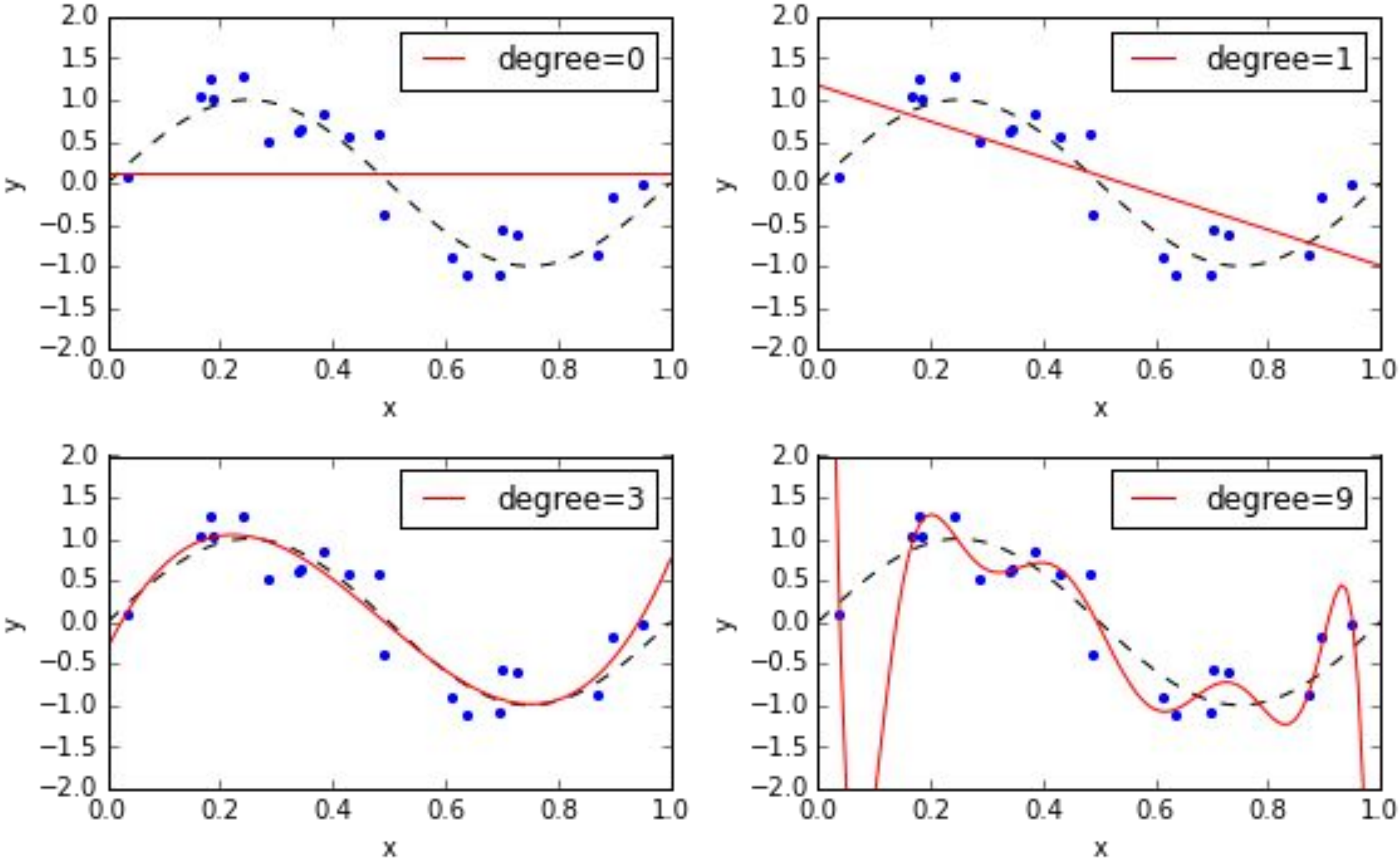
- Também pode ser utilizada para problemas em que a variável de resposta é qualitativa

$$CV_(n) = \frac{1}{n}\sum_{i=1}^{n} Err_i$$

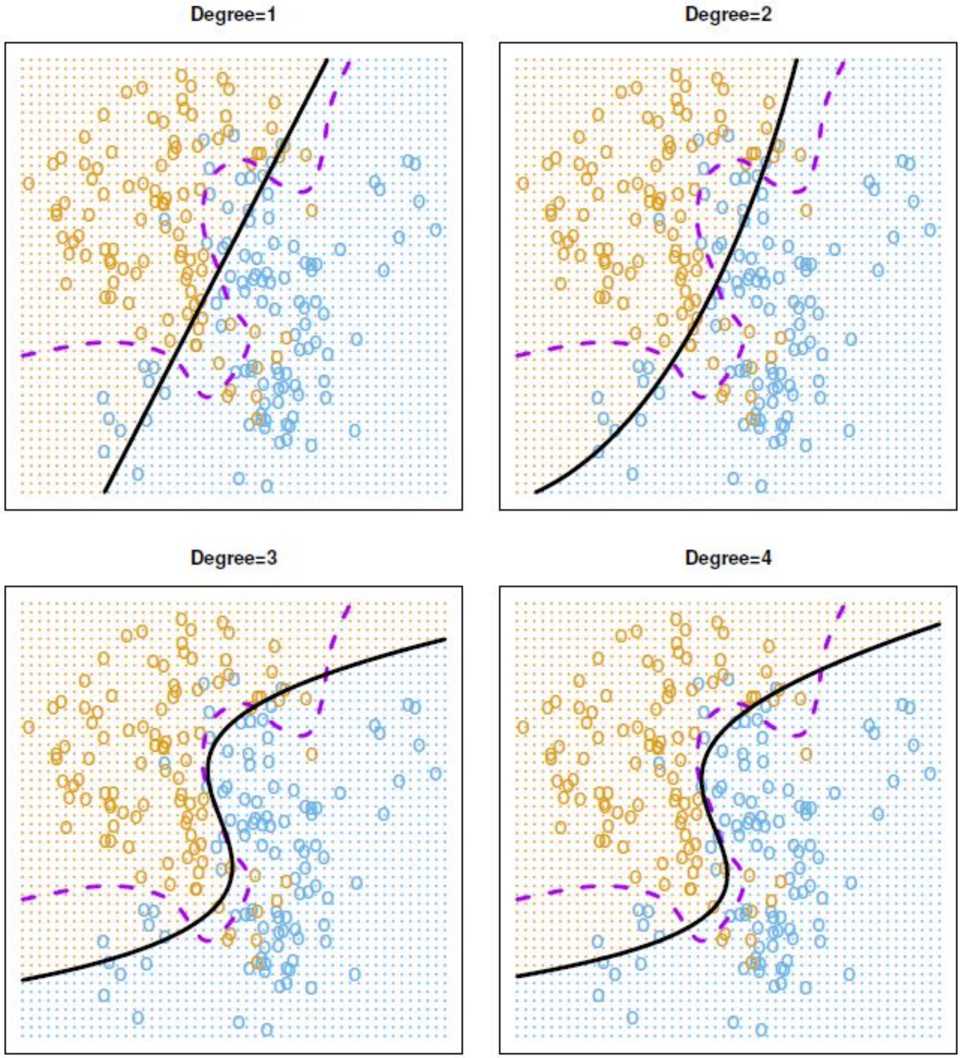
onde Err_i corresponde ao número de classificações erradas do modelo i

Model Selection

- Como escolher o modelo correto?



Model Selection

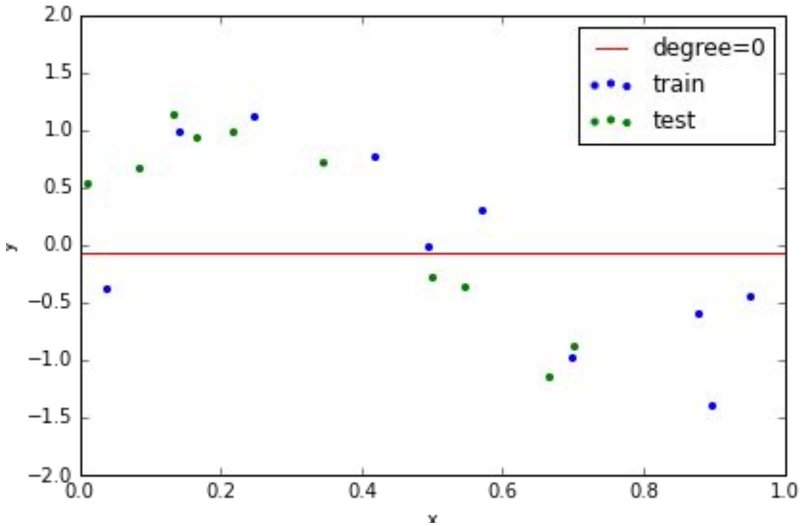


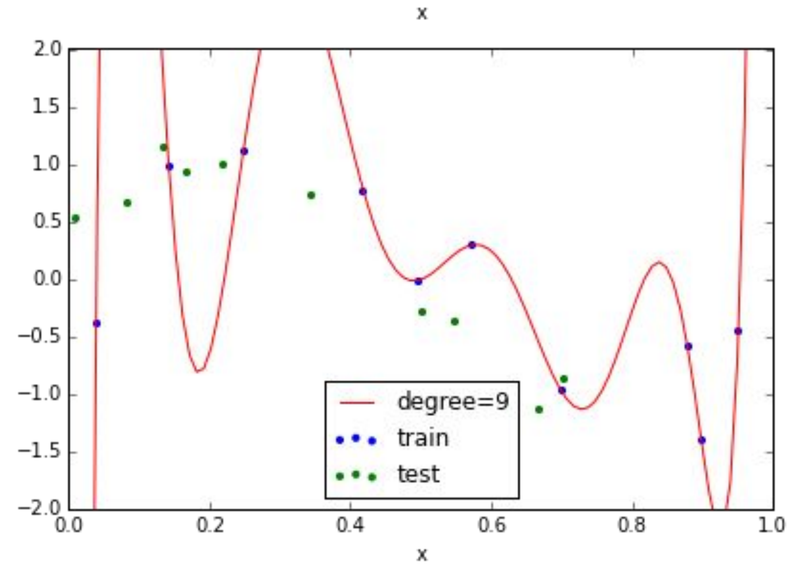
Capacidade do Modelo

- Cada classe de modelos possui uma **capacidade**
  - “Número de funções que o modelo é capaz de estimar”
  - Polinômio de 1º grau = retas
  - Polinônio de 2º grau = retas + parábolas
  - Assim por diante…
- Aprendizado = busca no “espaço de funções”

Capacidade do Modelo

- Modelos com baixa capacidade podem gerar subajuste (*underfitting*)
- Modelos com alta capacidade podem gerar sobreajuste (*overfiting*)



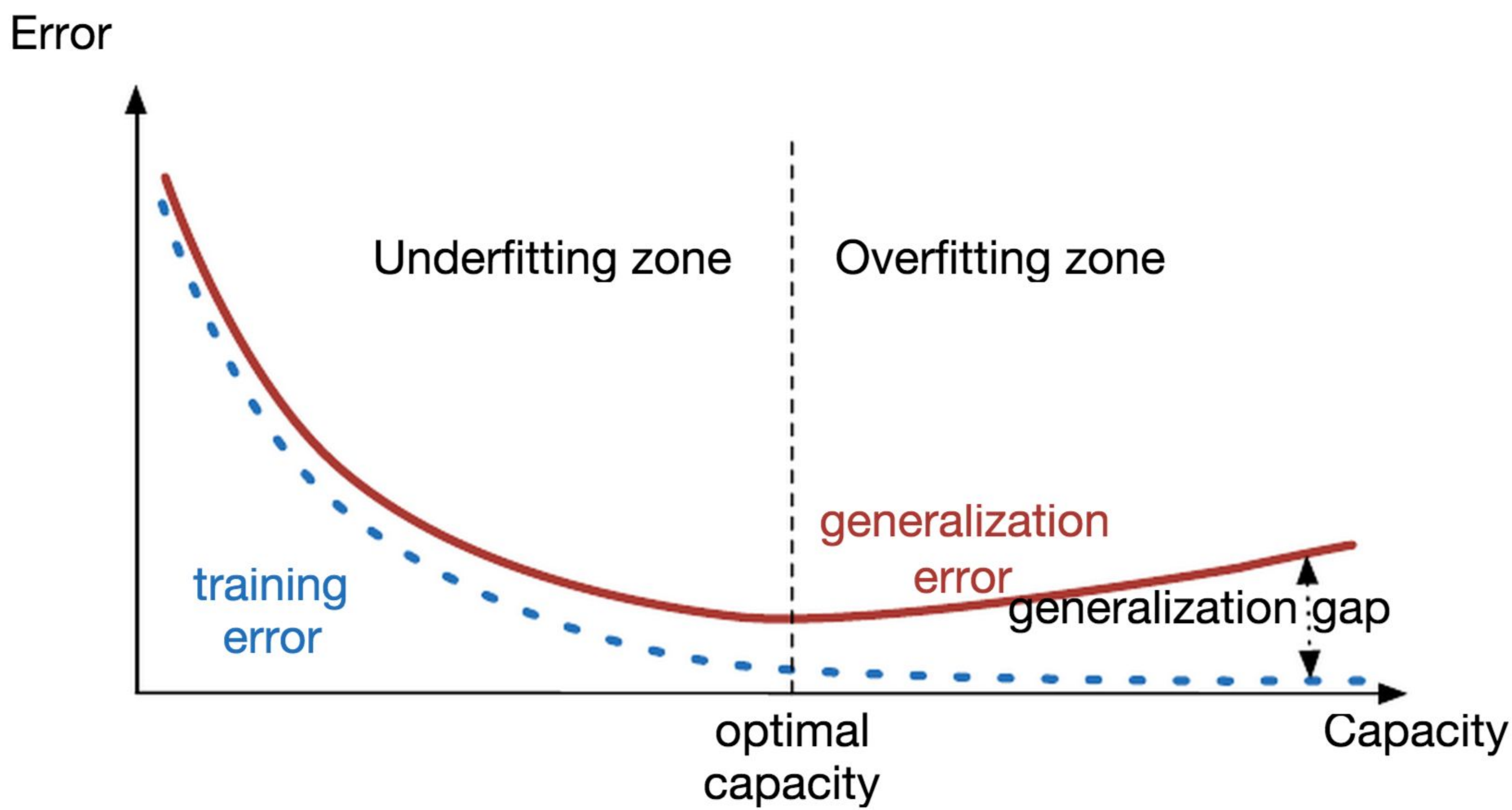


Capacidade do Modelo

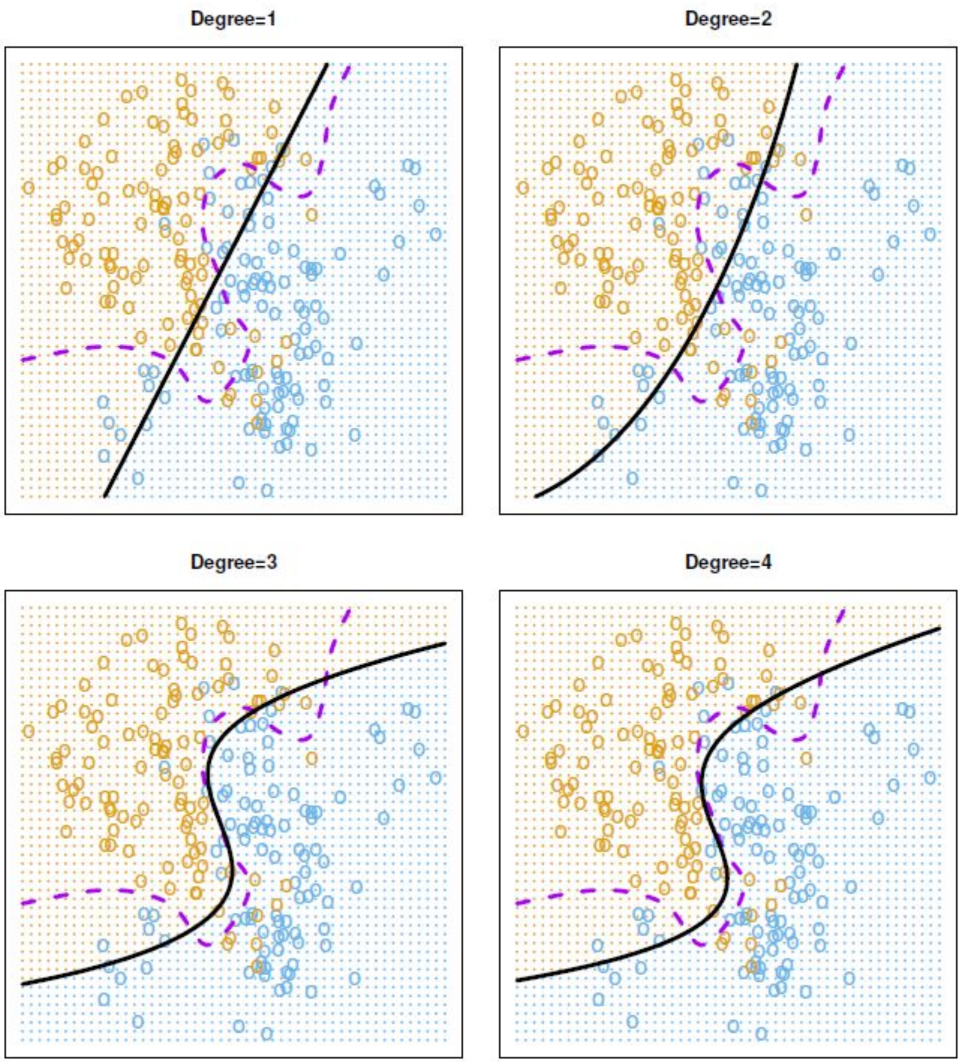
- O modelo terá um melhor desempenho se sua capacidade for compatível com a **capacidade do problema** a ser resolvido

  – Modelos com baixa capacidade não são capazes de resolver o problema
  – Modelos com alta capacidade são capazes de resolver o problema, porém a busca pela função geradora dos dados se dá em um espaço maior
    - Aumenta a complexidade do aprendizado

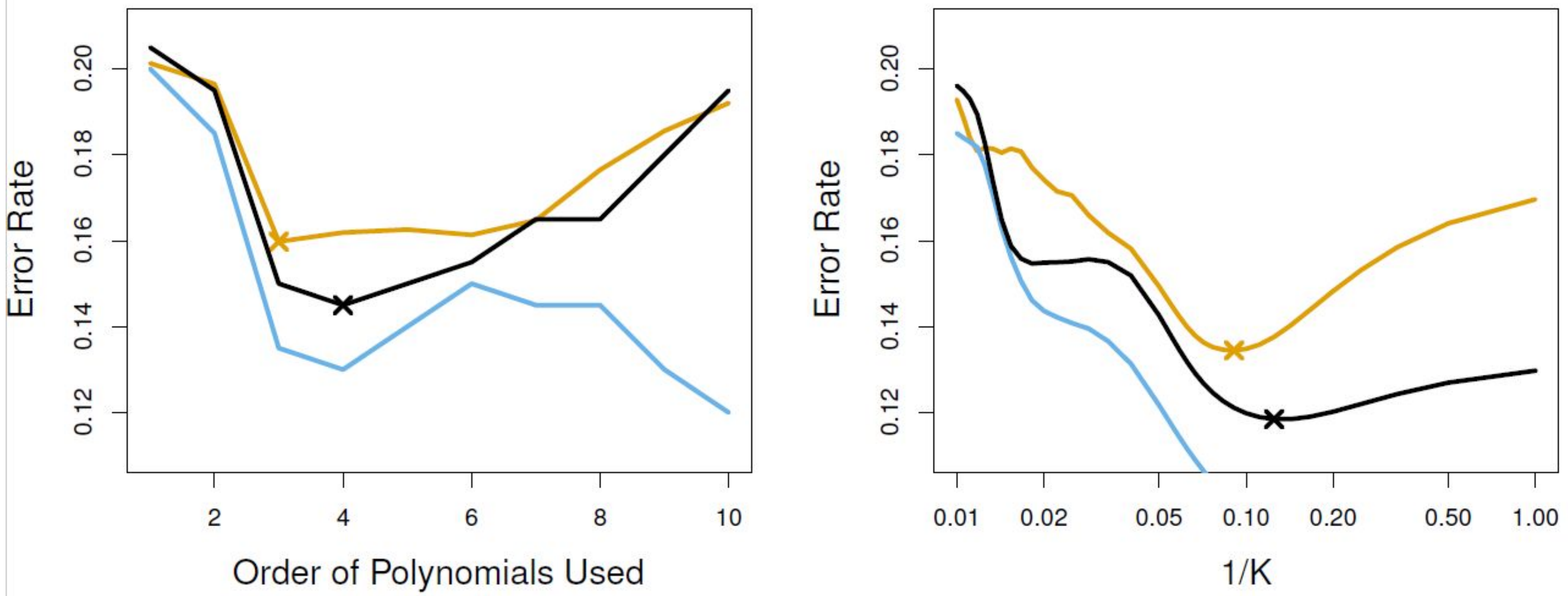
Model Selection



Model Selection



Model Selection



Bootstrap

- Técnica de reamostragem muito utilizada para quantificar incerteza associada a um estimador ou a um método de aprendizado estatístico

- Anteriormente, vimos que podemos calcular o erro padrão associado aos parâmetros de uma regressão linear analiticamente

- Bootstrap permite estimar incertezas para modelos estatísticos em que a solução analítica é difícil de ser obtida

Bootstrap

- Exemplo ilustrativo

- Suponha que tenhamos uma quantia de dinheiro a ser aplicada em dois investimentos financeiros que geram retornos X e Y, respectivamente (X e Y são valores aleatórios)

- Iremos investir uma fração da quantia em X (α) e o restante em Y ($1-\alpha$)

- Dado que existe uma variabilidade associada aos retornos de X e Y, desejamos encontrar o valor de α, que minimize o risco (variância) do investimento

Bootstrap

- Desejamos minimizar

$$Var(\alpha X + (1-\alpha)Y)$$

- O valor que minimiza o risco é dado por

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Bootstrap

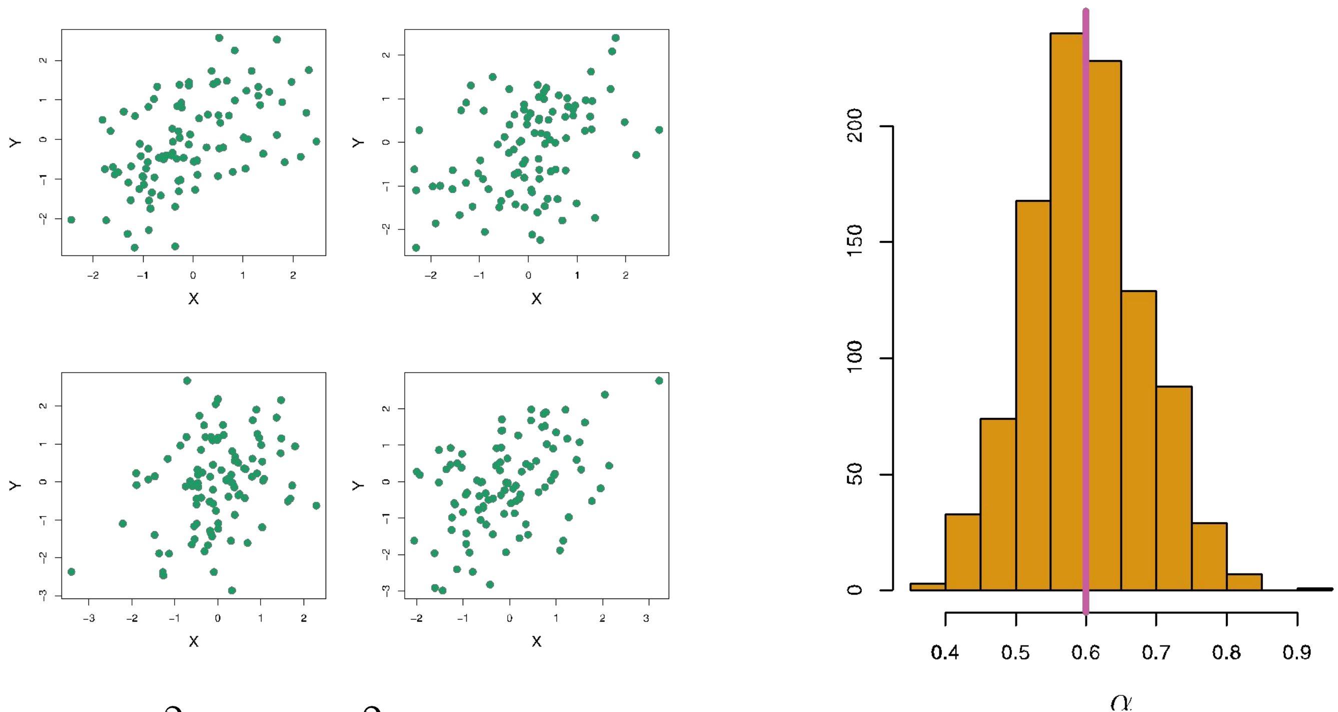
- As variâncias de X, Y e a covariância entre X e Y são desconhecidas
- Podemos calcular estimativas desses valores a partir de um conjunto de dados e em seguida realizar uma estimativa pontual do valor de $\hat{\alpha}$

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

Bootstrap

- Para calcular o erro associado a estimativa de $\hat{\alpha}$

seria necessário estimar o valor a partir de várias bases de dados

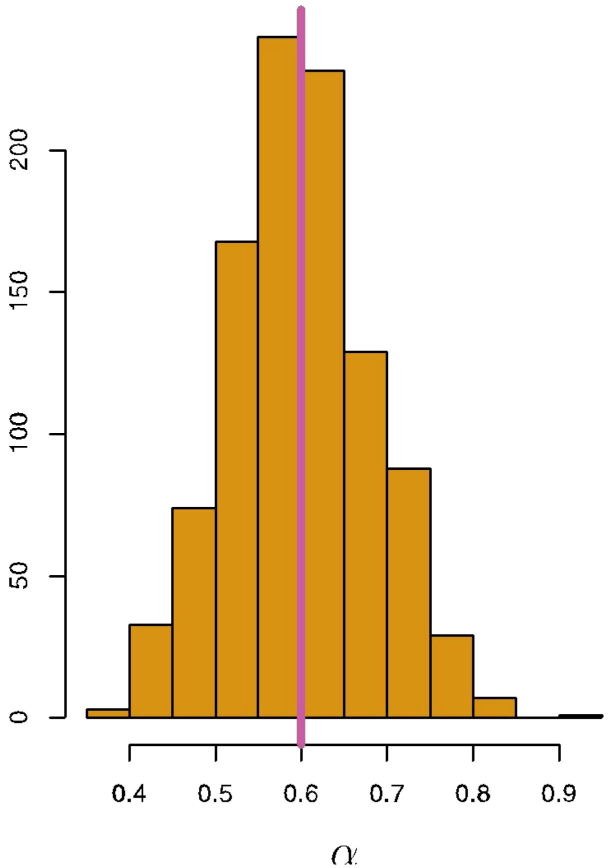


$$\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \text{ and } \sigma_{XY} = 0.5$$

Bootstrap

$\sigma_X^2 = 1, \sigma_Y^2 = 1.25$, and $\sigma_{XY} = 0.5$

$$\sqrt{\frac{1}{1,000-1}\sum_{r=1}^{1,000}\left(\hat{\alpha}_r - \bar{\alpha}\right)^2} = 0.083.$$

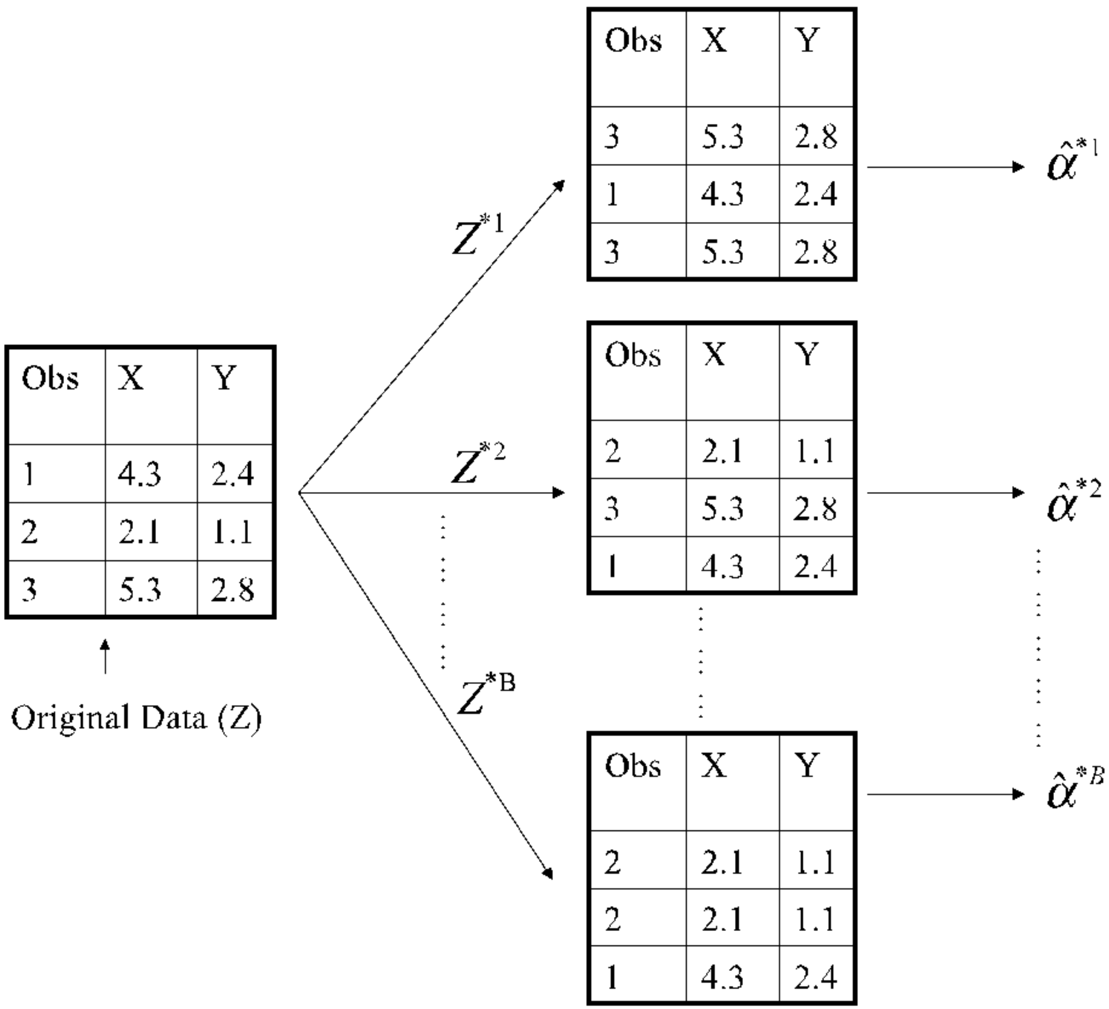


Bootstrap

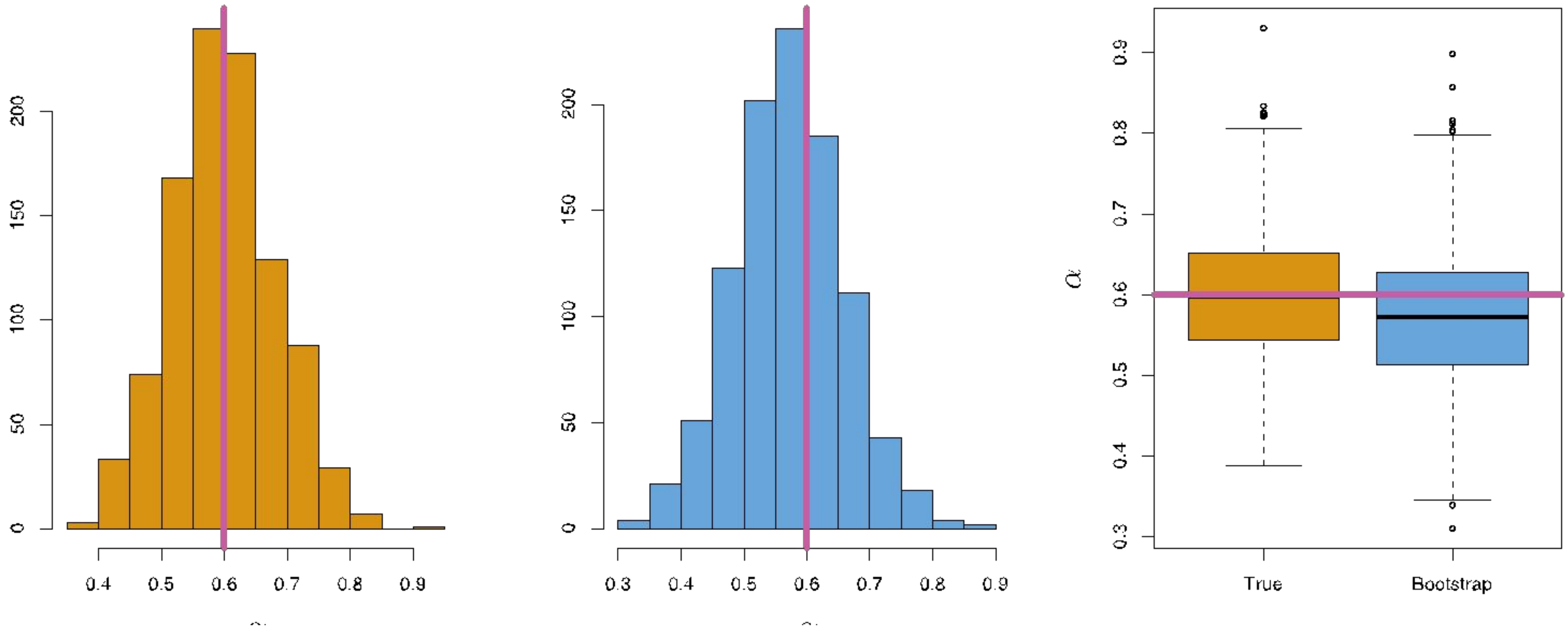
- Na prática, não podemos fazer isso, pois os valores das variâncias e covariância são desconhecidos

- Utilizamos a técnica de boostrap para emular o processo de gerar novas amostras de forma a simular a variabilidade do estimador

Bootstrap


Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$

Bootstrap



200
150
100
50
0
0.4
0.5
0.6
0.7
0.8
0.9
200
150
100
50
0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.9
0.8
0.7
0.6
0.5
0.4
0.3
α
True
Bootstrap

Referências

- Capítulo 5 do livro James, Gareth, et al. *An Introduction to Statistical Learning*. Vol. 112. New York: Springer, 2013 – Section 10.3