Pré-Processamento de Dados

Luciano Barbosa

Motivação

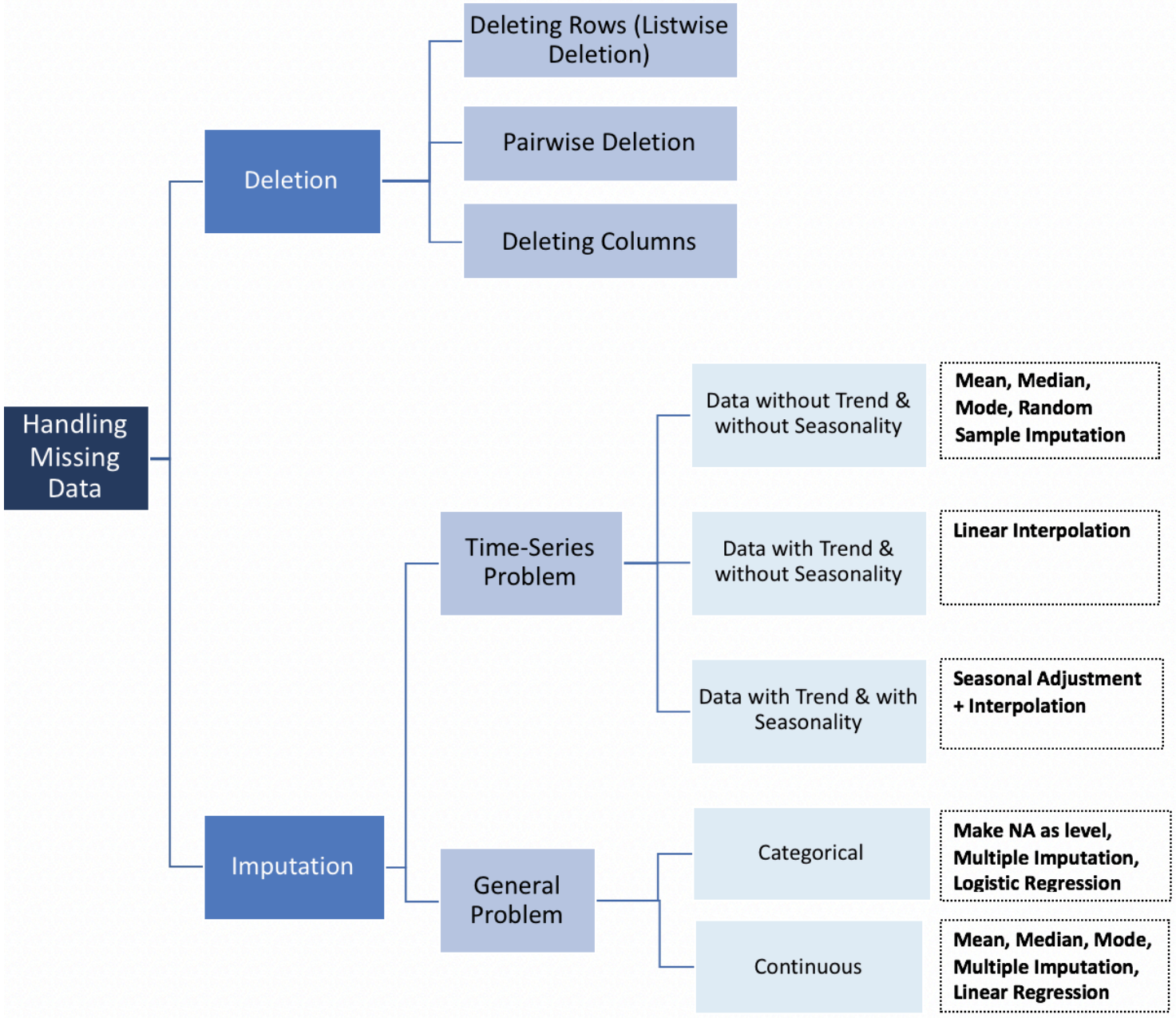
- “Todo mundo faz mas ninguém fala”
- Boa parte do tempo gasto nesta etapa
- Dados incorretos podem distorcer os resultados das análises e modelos
- “Better data beats fancier algorithms”

Roteiro

- Ajuste de tipos
- Dados ausentes
- Discretização
- Normalização

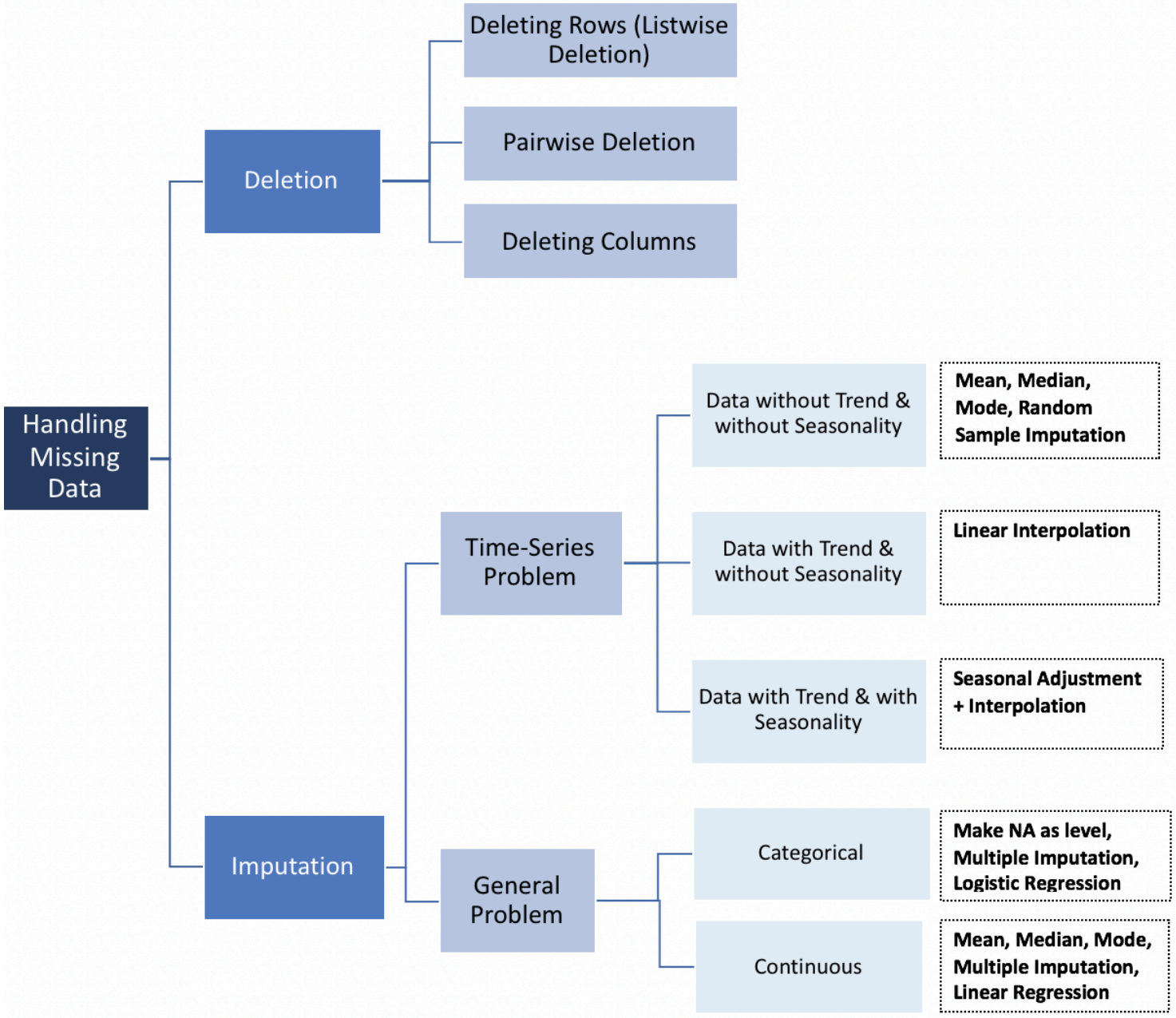
Dados Ausentes

- Viés na ausência de dados
  - Ex: pessoas mais ricas tendem a não dizer renda em pesquisas


Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

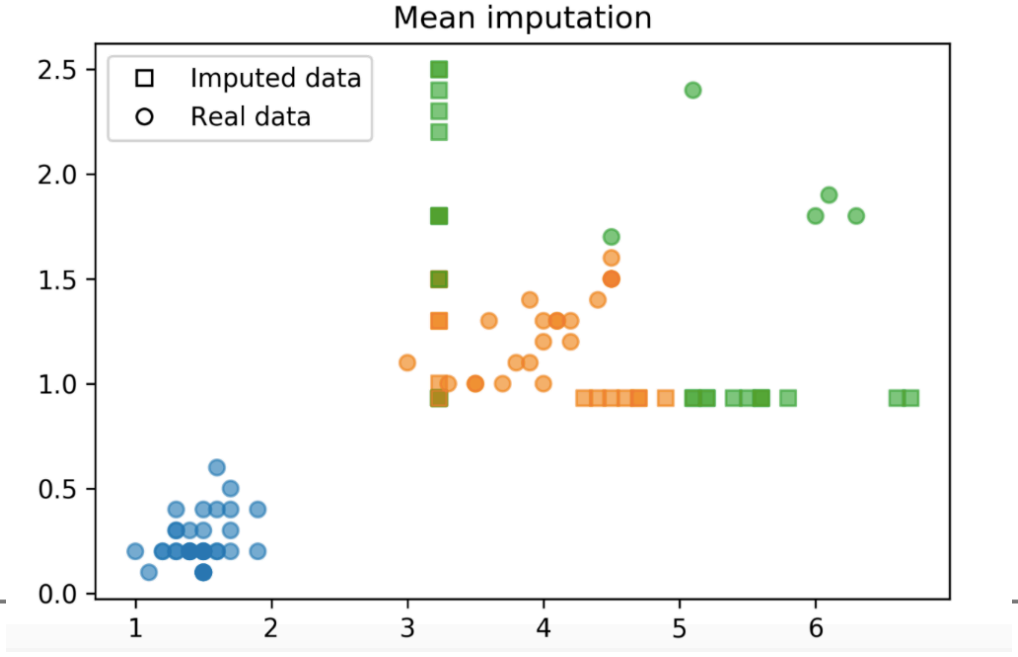
Remoção de Dados

- Remoção das instâncias: instâncias nas quais ao menos um dos atributos está faltando
  - Pro: simples
  - Con:
    - Possível viés se as instâncias removidas forem diferentes das não removidas
    - Pode remover boa parte dos dados
- Remoção de instâncias baseado na análise
- Remoção de variáveis com muitos valores ausentes


Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

Imputação de Dados

- Imputar dados ausentes
- Média, mediana ou moda da variável
  - Problema se existem muitos valores ausentes
  - Pode atrapalhar no cálculo de correlações entre variáveis



Formas de Imputação

- Imputar um valor e criar outra variável indicando que o dado foi imputado
- Para dados categóricos: adicionar uma categoria de dados ausentes
- Amostra aleatória dos valores da variável ou baseado em uma distribuição (ex.: normal)

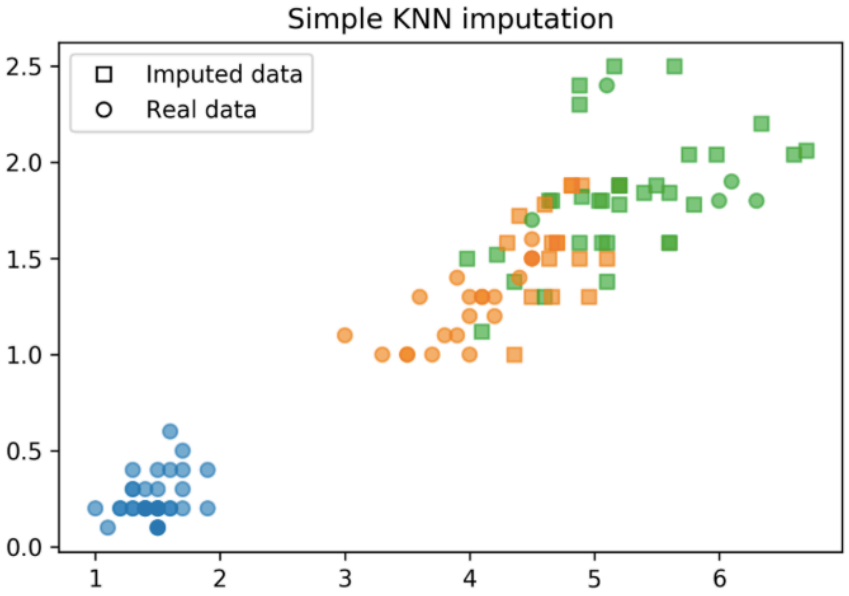
Formas de Imputação

- Para atributos categóricos:
  - Usar classificação para prever os valores ausentes de uma variável a partir das outras

- Para atributos numéricos:
  - Usar regressão para prever os valores ausentes de uma variável a partir das outras

Formas de Imputação

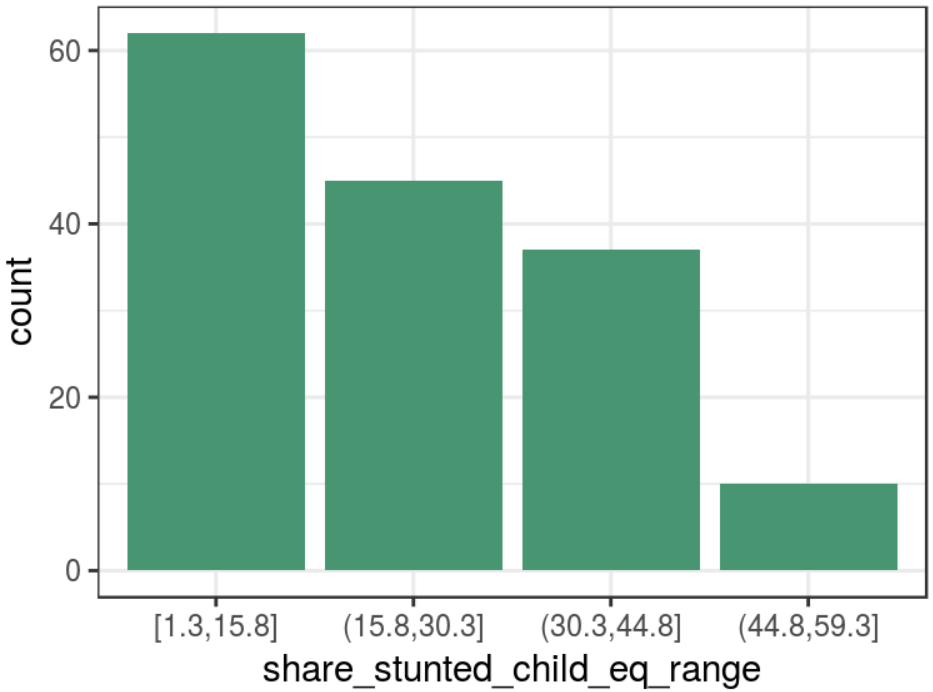
- KNN (https://github.com/iskandr/fancyimpute):
  - Valor ausente definido pelos valores das instâncias mais próximas
  - Valor: média dos valores dos vizinhos



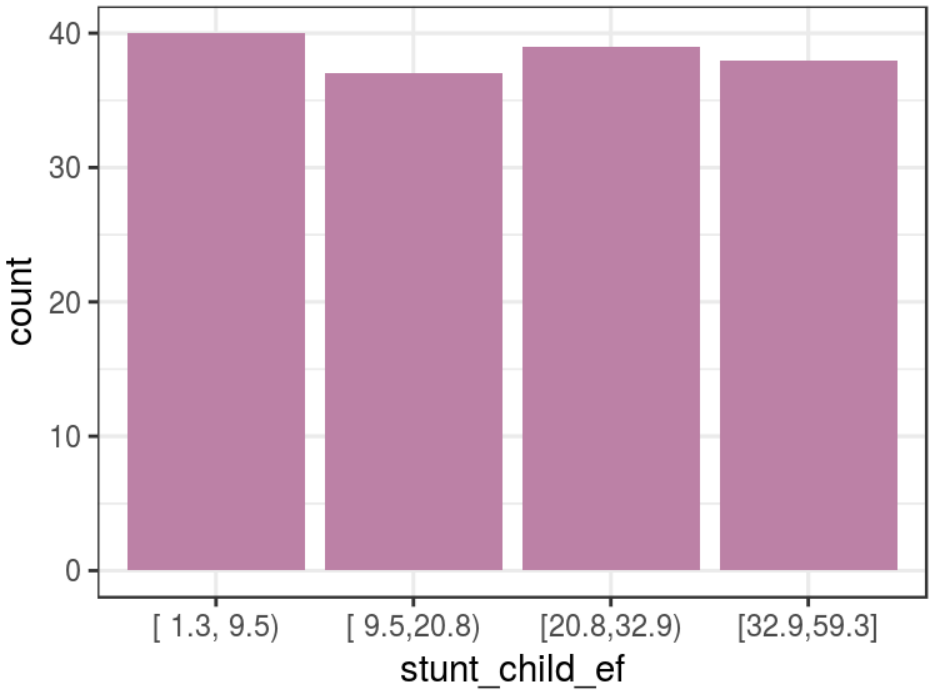
Discretização

- Transformar atributos contínuos em categóricos
- Decidir número de intervalos e seu tamanho
- Dois tipos
  - Supervisionada
    - Baseada em entropia: intervalos com mesma proporção de um rótulo
  - Não-supervisionada
    - Intervalos de mesma largura
    - Intervalos de mesma frequência

Não-Supervisionada



Mesmo Intervalo



Mesmo frequência

Normalização

- Útil para cálculo de distância entre instâncias com atributos com diferentes intervalos
- Intervalos: [0,1] ou [-1,1]
- Atributos com valores maiores podem dominar os com valores menores

Normalização por Min-Max

- Valores do atributo são ajustados para o intervalo [a,b] baseado no valor máximo e mínimo:

$$v[i] = \frac{v[i] - min(v)}{max(v) - min(v)}(b - a) + a$$

For:
$(X_1, Y_1) = (1, 40)$
$(X_2, Y_2) = (2, 100)$

Distance Equation Solution:

$$d = \sqrt{(2-1)^2 + (100-40)^2}$$

$$d = \sqrt{(1)^2 + (60)^2}$$

$$d = \sqrt{1 + 3600}$$

$$d = \sqrt{3601}$$

$$d = 60.008333$$

For:
$(X_1, Y_1) = (0, 0)$
$(X_2, Y_2) = (1, 1)$

Distance Equation Solution:

$$d = \sqrt{(1-0)^2 + (1-0)^2}$$

$$d = \sqrt{(1)^2 + (1)^2}$$

$$d = \sqrt{1 + 1}$$

$$d = \sqrt{2}$$

$$d = 1.414214$$

Normalização por Média e Desvio Padrão

- Média = 0
- Desvio padrão = 1

$$v[i] = \frac{v[i] - \bar{v}}{\sigma_v}$$

$$z = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix}.$$