Další datové struktury a práce s nimi, matice, datové tabulky, seznamy

—

17VSADR – Skriptování a analýza dat v jazyce R

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova v Praze

ČVUT
FBMI

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

Obsah

Datové struktury

- vektor (`vector`)
- faktor (`factor`)
- matice (`matrix`)
- pole (`array`)
- tabulka dat (`data.frame`)
- seznam (`list`)

Tvorba matic a základní příkazy

- matice (`matrix`) je dvojrozměrným polem, které obsahuje prvky stejného datového typu
- všechny sloupce matice mají shodnou délku, podobně všechny řádky matice mají shodnou délku
- ať je

$$\boldsymbol{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

- v R matici $\boldsymbol{A}$ a $\boldsymbol{B}$ získáme příkazy

```
1       A <- matrix(c(1, 2, 3, 4), nrow = 2,
2                   ncol = 2)
3       B <- matrix(c(1, 3, 2, 4), nrow = 2,
4                   ncol = 2)
5       B <- matrix(c(1, 2, 3, 4), nrow = 2,
6                   ncol = 2, byrow = TRUE)
7       # vždy jeden z argumentů "nrow" či "ncol" je zbytný
```

Manipulace s maticemi

- ať je

$$\boldsymbol{C} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix}$$

- v R pak pomocí

```
C <- matrix(letters[1:12], nrow = 3,
            byrow = T)
```

- některé užitečné příkazy

```
is.matrix(C)    # TRUE
class(C)        # "matrix"
mode(C)         # "character"; datový typ pruků
str(C)          # chr [1:3, 1:4] "a" "e" "i" ...
dim(C)          # c(3, 4); rozměry matice C
```

Manipulace s maticemi

- ať je

$$\boldsymbol{C} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix}$$

- další užitečné příkazy

```
colnames(C) <- c("c1", "c2", "c3", "c4")
rownames(C) <- c("r1", "r2", "r3")
                # přidá jmenovky sloupcům i řádkům
C <- unname(C)
                # zbaví sloupce i řádky jmenovek
dimnames(C) <- list(
                    c("r1", "r2", "r3"),
                    c("c1", "c2", "c3", "c4")
                )
                # opět přidá jmenovky sloupcům i řádkům
```

Manipulace s maticemi

- stále mějme

$$\boldsymbol{C} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix}$$

- další užitečné příkazy

```
1  rbind(C, c("x", "x", "x", "x"))
2                  # přidání řádku c("x", "x", "x", "x")
3                  # k matici C
4  cbind(C, c("x", "x", "x"))
5                  # přidání sloupce c("x", "x", "x")
6                  # k matici C
7  C[-1, ]         # odebrání 1. řádku matici C
8  C[, -2]         # odebrání 2. sloupce matici C
```

Submatice, indexování, adresace

- ať je

$$\boldsymbol{C} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix}$$

- v R pomocí

```
C <- matrix(letters[1:12], nrow = 3,
            byrow = T, dimnames = list(
            c("r1", "r2", "r3"),
            c("c1", "c2", "c3", "c4")))
```

- adresace

```
C[2, 3]          # "g"; prvek 2. řádku, 3. sloupce
C["r2", "c3"]    # "g"; prvek 2. řádku, 3. sloupce
C[1, ]           # c("a", "b", "c", "d");
                 # tedy vektor 1. řádku matice C
                 # s popisky
```

Submatice, indexování, adresace

- stále mějme

$$\boldsymbol{C} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix}$$

- adresace

```
1  C[, 3]                  # c("c", "g", "k");
2                          # tedy vektor 3. sloupce matice C
3                          # s popisky
4  C[c(1, 3), c(2, 4)]
5                          # matrix(c("b", "j", "d", "l"), 2)
6                          # submatice 1. a 3. řádku,
7                          # 2. a 4. sloupce matice C
8                          # s popisky
9  C["r2", ]               # c("e", "f", "g", "h");
10                         # tedy vektor 2. řádku matice C
11                         # s popisky
```

Submatice, indexování, adresace

- stále mějme

$$\boldsymbol{C} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix}$$

- adresace

```
1  C[dim(C)[1], dim(C)[2]]
2                  # "l"; obecná adresace prvku
3                  # vpravo dole (např. neznáme-li
4                  # číselné rozměry matice)
5  C[5]            # "f"; major-column ordering
6  C[c(8, 9)]      # c("g", "k")
7  C[13]           # NA
8  diag(C)         # c("a", "f", "k"); hlavní diagonála
9  diag(C[, dim(C)[2]:1])
10                 # c("d", "g", "j"); vedlejší diagonála
```

Intermezzo

- mějme matici

$$\boldsymbol{X} = \begin{pmatrix} 1 & 3 & 9 & 6 \\ -5 & 0 & 1 & 2 \\ -8 & 4 & 7 & 3 \end{pmatrix}$$

- pomocí R
  - (i) založme matici $\boldsymbol{X}$
  - (ii) najděme prvek třetího řádku druhého sloupce matice $\boldsymbol{X}$
  - (iii) jaký řádkový a sloupcový index mají prvky 0, 7 matice $\boldsymbol{X}$
  - (iv) najděme minimum matice $\boldsymbol{X}$ a řádkový a sloupcový index této hodnoty
  - (v) najděme stopu matice $\boldsymbol{X}$

Intermezzo

- mějme matici

$$\boldsymbol{X} = \begin{pmatrix} 1 & 3 & 9 & 6 \\ -5 & 0 & 1 & 2 \\ -8 & 4 & 7 & 3 \end{pmatrix}$$

- řešení

```
1  # (i)
2  X <- matrix(c(1, -5, -8, 3, 0, 4, 9, 1,
3                7, 6, 2, 3), nrow = 3)
4  # (ii)
5  X[3, 2]        # 4
6  # (iii)
7  c((which(X == 0) - 1) %% dim(X)[1] + 1,
8    ceiling(which(X == 0) / dim(X)[1]))
9  c((which(X == 7) - 1) %% dim(X)[1] + 1,
10   ceiling(which(X == 7) / dim(X)[1]))
```

Intermezzo

- mějme matici

$$\boldsymbol{X} = \begin{pmatrix} 1 & 3 & 9 & 6 \\ -5 & 0 & 1 & 2 \\ -8 & 4 & 7 & 3 \end{pmatrix}$$

- řešení

```
1  # (iii) lépe
2  which(X == 0, ind.arr = TRUE)
3  which(X == 7, ind.arr = TRUE)
4            # which(X == 1, ind.arr = TRUE)
5            # vrátí c(1, 2) i c(2, 3)
6  # (iv)
7  min(X); which(X == min(X), arr.ind = T)
8            # -8; c(3, 1)
9  # (v)
10 sum(diag(X))
11           # 8 := 1 + 0 + 7
```

Maticová algebra

- buďte

$$\boldsymbol{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} \qquad \boldsymbol{b} = (1, 0)^T$$

- v R

```r
A <- matrix(c(1, 2, 3, 4), nrow = 2)
B <- matrix(c(5, 6, 7, 8), nrow = 2)
b <- c(1, 0)
```

- Hadamardův součin (*element-wise*, *pairwise*) $\boldsymbol{A} \circ \boldsymbol{B} = \begin{pmatrix} 5 & 21 \\ 12 & 32 \end{pmatrix}$

```r
A * B       # matrix(c(5, 12, 21, 32), 2)
```

- maticový součin $\boldsymbol{A} \cdot \boldsymbol{B} = \begin{pmatrix} 23 & 31 \\ 34 & 46 \end{pmatrix}$

```r
A %*% B     # matrix(c(23, 34, 31, 46), 2)
```

Maticová algebra

- buďte

$$\boldsymbol{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} \qquad \boldsymbol{b} = (1, 0)^T$$

- transpozice $\boldsymbol{A}^T = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

```
t(A)          # matrix(c(1, 3, 2, 4), 2)
```

- jednotková matice $\boldsymbol{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

```
diag(2)       # matrix(c(1, 0, 0, 1), 2)
```

- inverzní matice $\boldsymbol{A}^{-1} = \begin{pmatrix} -2 & 3/2 \\ 1 & -1/2 \end{pmatrix}$

```
solve(A)      # matrix(c(-2, 1, 1.5, -0.5), 2)
```

Maticová algebra

- buďte

$$\boldsymbol{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} \qquad \boldsymbol{b} = (1, 0)^T$$

- vlastní čísla matice $\boldsymbol{A}$ jsou $\boldsymbol{\lambda} = (\frac{1}{2}(5 + \sqrt{33}), \frac{1}{2}(5 - \sqrt{33}))^T$

```
1 ||      eigen(A)$values
```

- vlastní vektory matice $\boldsymbol{A}$ jsou například
  $\boldsymbol{v} = [(\frac{1}{4}(-3 + \sqrt{33}), 1)^T, (\frac{1}{4}(-3 - \sqrt{33}), 1)^T]$

```
1 ||      eigen(A)$vectors     # R vrací ortonormální
2 ||                           # vlastní vektory
```

- řešení $\boldsymbol{x}$ ze soustavy $\boldsymbol{Ax} = \boldsymbol{b}$ je $\boldsymbol{x} = (-2, 1)^T$

```
1 ||      solve(A, b)          # c(-2, 1)
```

Intermezzo

- uvažujme markovský proces s počátečním stavovým vektorem $\boldsymbol{p}^0 = (1, 0, 0)^T$ a přechodovou maticí

$$\boldsymbol{S} = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

  kde v i-tém řádku jsou vždy po řadě pravděpodobnosti přechodu z i-tého stavu do j-tého stavu pro $\forall i, j \in \{1, 2, 3\}$

- pomocí R
  - (i) ověřme, že matice $\boldsymbol{S}$ je stochastická
  - (ii) určeme pravděpodobnosti, s jakými proces dospěl po pěti krocích do prvního, druhého či třetího stavu
  - (iii) určeme, zda existuje dynamicky rovnovážný stav procesu

Intermezzo

- uvažujme markovský proces s počátečním stavovým vektorem $\boldsymbol{p}^0 = (1, 0, 0)^T$ a přechodovou maticí

$$\boldsymbol{S} = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

  kde v i-tém řádku jsou vždy po řadě pravděpodobnosti přechodu z i-tého stavu do j-tého stavu pro $\forall i, j \in \{1, 2, 3\}$

- řešení

```
1  S <- matrix(c(1/4, 1/2, 1/4,
2                2/3, 0, 1/3,
3                1/2, 1/2, 0), nrow = 3,
4                byrow = TRUE)
5  p <- c(1, 0, 0)
```

Intermezzo

- uvažujme markovský proces s počátečním stavovým vektorem $\boldsymbol{p}^0 = (1, 0, 0)^T$ a přechodovou maticí

$$\boldsymbol{S} = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

  kde v i-tém řádku jsou vždy po řadě pravděpodobnosti přechodu z i-tého stavu do j-tého stavu pro $\forall i, j \in \{1, 2, 3\}$

- řešení

```
1 # (i)
2 rowSums(S)          # c(1, 1, 1)
```

Intermezzo

- uvažujme markovský proces s počátečním stavovým vektorem $\boldsymbol{p}^0 = (1, 0, 0)^T$ a přechodovou maticí

$$\boldsymbol{S} = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

  kde v i-tém řádku jsou vždy po řadě pravděpodobnosti přechodu z i-tého stavu do j-tého stavu pro $\forall i, j \in \{1, 2, 3\}$

- řešení

```
1 # (ii)
2 install.packages("expm", dependency = T)
3 library("expm")
4 p %*% (S %^% 5)    # c(0.437, 0.344, 0.219)
```

Intermezzo

- uvažujme markovský proces s počátečním stavovým vektorem $\boldsymbol{p}^0 = (1, 0, 0)^T$ a přechodovou maticí

$$\boldsymbol{S} = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

  kde v i-tém řádku jsou vždy po řadě pravděpodobnosti přechodu z i-tého stavu do j-tého stavu pro $\forall i, j \in \{1, 2, 3\}$

- řešení

```
1  # (iii)
2  p %*% (S %^% 10)   # c(0.444, 0.333, 0.222)
3  p %*% (S %^% 20)   # c(0.444, 0.333, 0.222)
4  p %*% (S %^% 30)   # c(0.444, 0.333, 0.222)
5  p %*% (S %^% 40)   # c(0.444, 0.333, 0.222)
```

Tvorba polí a práce s nimi

- pole je objekt s vlastnostmi matice, ale libovolným počtem dimenzí
- např.

```
(X <- array(c(1:24), dim = c(4, 3, 2)))
# , , 1
# [,1] [,2] [,3]
# [1,] 1 5 9
# [2,] 2 6 10
# [3,] 3 7 11
# [4,] 4 8 12
# , , 2
# [,1] [,2] [,3]
# [1,] 13 17 21
# [2,] 14 18 22
# [3,] 15 19 23
# [4,] 16 20 24
```

Manipulace s poli

- navigace v poli

```
X[, , 1]      # první vrstva
X[, , 2]      # druhá vrstva
X[3, , 1]     # třetí řádek první vrstvy
X[, 2, ]      # druhé sloupce všech vrstev
```

- pole se příliš nepoužívají, protože vstupem pro analytické funkce jazyka R je obvykle matice či datová tabulka
- pole libovolných rozměrů lze převést na dvourozměrnou matici pomocí dummy proměnné „vrstva“

Tvorba datových tabulek a základní příkazy

- datová tabulka (`data.frame`) je dvojrozměrným polem, které obsahuje v každém sloupci prvky stejného datového typu, ale jednotlivé sloupce se mohou datovým typem lišit
- všechny sloupce datové tabulky mají shodnou délku, podobně všechny řádky datové tabulky mají shodnou délku
- v R je řada vestavěných datových tabulek

```
1  mtcars
2  str(mtcars)
3  class(mtcars)              # "data.frame"
4  mode(mtcars)               # "list"
5  is.data.frame(mtcars)      # TRUE
6  str(iris)
7       # 'data.frame': 150 obs. of 5 variables
8       # ...
9  dim(iris)                  # c(150, 5)
```

Manipulace s datovými tabulkami

- užitečné příkazy

```
1  data <- mtcars
2  colnames(data)
3  colnames(data) <- paste("c",
4                          1:dim(data)[2],
5                          sep = "_")
6  rownames(data) <- paste("r",
7                          1:dim(data)[1],
8                          sep = "_")
9                  # změní jmenovky sloupcům i řádkům
10 head(data)      # náhled na prvních 6 řádků
11 head(data, 10)
12                 # náhled na prvních 10 řádků
13 tail(data)      # náhled na posledních 6 řádků
14 tail(data, 10)
15                 # náhled na posledních 10 řádků
```

Manipulace s datovými tabulkami

- další užitečné příkazy

```
        rbind(data, rep(0, dim(data)[2]))
                    # přidání řádku c(0, 0, ..., 0)
                    # k data.frameu "data"
        cbind(data, rep(0, dim(data)[1]))
                    # přidání sloupce c(0, 0, ..., 0)
                    # k data.frameu "data"
        data.frame(data,
                   "ahoj" = rep(0, dim(data)[1]))
                    # přidání sloupce c(0, 0, ..., 0)
                    # se jménem "ahoj" k data.frameu
                    # "data"
        data[-1, ]  # odebrání 1. řádku data.frameu
                    # "data"
        data[, -1]  # odebrání 1. sloupce data.frameu
                    # "data"
```

Indexování, adresace

- adresace

```
1   data[2, 3]      # 160; prvek 2. řádku, 3. sloupce
2   data["r_2", "c_3"]
3                   # 160; prvek pod danými popisky
4   data[1, ]       # c(21, 6, 160, 110, ...);
5                   # tedy vektor 1. řádku data.frameu
6                   # "data" s popisky
7   data[, 2]       # c(6, 6, 4, 6, ...);
8                   # tedy vektor 2. sloupce
9                   # data.frameu "data" s popisky
10  data$c_5        # c(3.90, 3.90, 3.85, 3.08, ...);
11                  # tedy vektor 2. sloupce
12                  # data.frameu "data" s popisky
13  data$c_5[1]     # 3.9;
14                  # tedy první prvek 2. sloupce
15                  # data.frameu "data"
```

Indexování, adresace

- adresace

```
1       data[dim(data)[1], dim(data)[2]]
2                   # 2; obecná adresace pruku
3                   # vpravo dole (např. neznáme-li
4                   # číselné rozměry tabulky)
5       data[5]     # 5. sloupec, nikoliv major-column
6                   # ordering
```

Sloupcové přehledy

- může mít smysl znát agregovaný ukazatel nad všemi sloupci

```
colSums(data)
                # součty všech sloupců
apply(data, 2, sum)
                # totéž
colMeans(data)
                # průměry všech sloupců
apply(data, 2, mean)
                # totéž
data <- rbind(data, rep(NA, dim(data)[2]))
                # přidán řádek c(NA, NA, ...)
colMeans(data)
                # c(NA, NA, ...); pro získání průměrů
                # nutné přidat argument na.rm = TRUE
colMeans(data, na.rm = TRUE)
apply(data, 2, mean, na.rm = TRUE)
                # už funguje
```

Tvorba seznamů a základní příkazy

- seznam (`list`) je výčtem prvků, které mohou být různorodého datového typu, včetně listu
- délky jednotlivých prvků seznamu se mohou lišit

```
1  my_list <- list("a" = c(1:10),
2                  "b" = mtcars,
3                  "c" = matrix(1:8, 2),
4                  "z" = "ahoj")
5  str(my_list)
6  class(my_list)          # "list"
7  mode(my_list)           # "list"
8  is.list(my_list)        # TRUE
```

Manipulace se seznamy

- užitečné příkazy

```
1  names(my_list)
2  names(my_list) <- LETTERS[
3       1:length(my_list)
4  ]                  # přejmenovávám prvky listu
5
6  my_list[[length(my_list) + 1]] <- c(T, F)
7  names(length(my_list)) <- "XY"
8                     # přidání vektoru c(T, F)
9                     # k listu "my_list" pod jménem
10                    # "XY"
```

Indexování, adresace

- adresace

```
1  my_list[[2]]   # 2. prvek listu
2  my_list[["B"]]
3                 # prvek listu pod jmenovkou "B"
4                 # jde o původní datový typ
5                 # (data.frame)
6  my_list["B"]   # prvek listu pod jmenovkou "B"
7                 # jde (vždy) o list
8  my_list[c(2, 4)]
9                 # 2. a 4. prvek listu
10 my_list$C      # prvek listu pod jmenovkou "C"
11 my_list[[1]][2]
12                # 2; 2. prvek 1. prvku listu
13 my_list[[2]][3, 5]
14                # 3.85; z 2. prvku listu vybírám
15                # prvek o souřadnicích (3, 5)
```

Indexování, adresace pomocí funkce `lapply()`

- adresace

```
1   set.seed(1)
2   my_long_list <- lapply(
3       sample(c(80:120), 100, TRUE),
4       function(x) sample(
5           c(50:150), x, replace = TRUE
6       )
7   )    # list vektorů náhodné délky
8       # generovaných z náhodných čísel
9
10  lapply(my_long_list, "[[", 14)
11      # z každého prvku listu (vektoru)
12      # vybírám jen jeho 14. prvek
```

Prvkové přehledy

- může mít smysl znát agregovaný ukazatel nad všemi prvky seznamu

```
1  lapply(my_long_list, mean)
2      # pro každý prvek listu (vektor)
3      # vracím jeho průměr
4
5  lapply(my_long_list, length)
6      # pro každý prvek listu (vektor)
7      # vracím jeho délku
```

Literatura


Karel Zvára. *Základy statistiky v prostředí R*. Praha, Česká republika: Karolinum, 2013. ISBN: 978-80-246-2245-3.


Hadley Wickham. *Advanced R*. Boca Raton, FL: CRC Press, 2015. ISBN: 978-1466586963.

Děkuji za pozornost!

lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz

▸ GitHub

github.com/LStepanek/17VSADR_Skriptovani_a_analyza_dat_v_jazyce_R