

Optimal matching

05.05.21 07:32

Greedy matching not optimal:

- greedy matching does not lead to the smallest global total distance

↳ Example:

greedy

T	C
45	44
38	36
41	35

$$\sum d = 1 + 2 + 6 \\ = 9$$

vs.

optimal

T	C
45	47
38	36
41	44

$$\sum d = 2 + 2 + 3 \\ = 7$$

} smaller total distance
with optimal matching

Optimal matching:

- minimizes the **global distance** measure (e.g. total distance)
- comp. very demanding on big datasets
- existing R packages:
 - o optmatch
 - o rebalance

Comp. feasibility of optimal matching:

- depends on the **"size"** of the problem: → **size: # of possible treatment-control pairings**

↳ examples:

$$(1) \left. \begin{array}{l} \# \text{ treated subjects: } 100 \\ \# \text{ control subjects: } 100 \end{array} \right\} 100^2 = \underbrace{100.000}_{\text{size}} \text{ possible pairings}$$

↳ comp. feasible

$$(2) \left. \begin{array}{l} \# \text{ treated: } 10000 = 10^4 \\ \# \text{ controls: } 100000 = 10^5 \end{array} \right\} 10^9 = \underbrace{1.000.000.000}_{\text{size: billion}} \text{ possible pairings}$$

↳ comp. infeasible w/ optimal matching

Sparse optimal matching: (R package: rebalance)

- impose constraints to make optimal matching comp. feasible for large data

↳ e.g.: (i) match within hospitals in a multi-site clinical study
↳ only consider matched pairs within same hospital

(ii) match within primary disease category (heart failure, COPD, etc.)

⇒ these are blocks

- mismatches can be tolerated if **fine balance** btw. the treated and control subjects can still be achieved