Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study

統計碩一 胡芷瑄

Contents

01

Background

Background

在代謝組學中最常用的技術--液相色譜聯用質譜（LC-MS）

問題:通常可能包含大量遺失值

解決： 插補(imputation)

02

Methods

Three missing mechanisms

MCAR
(Missing Com-pletely at Random)
完全隨機缺失

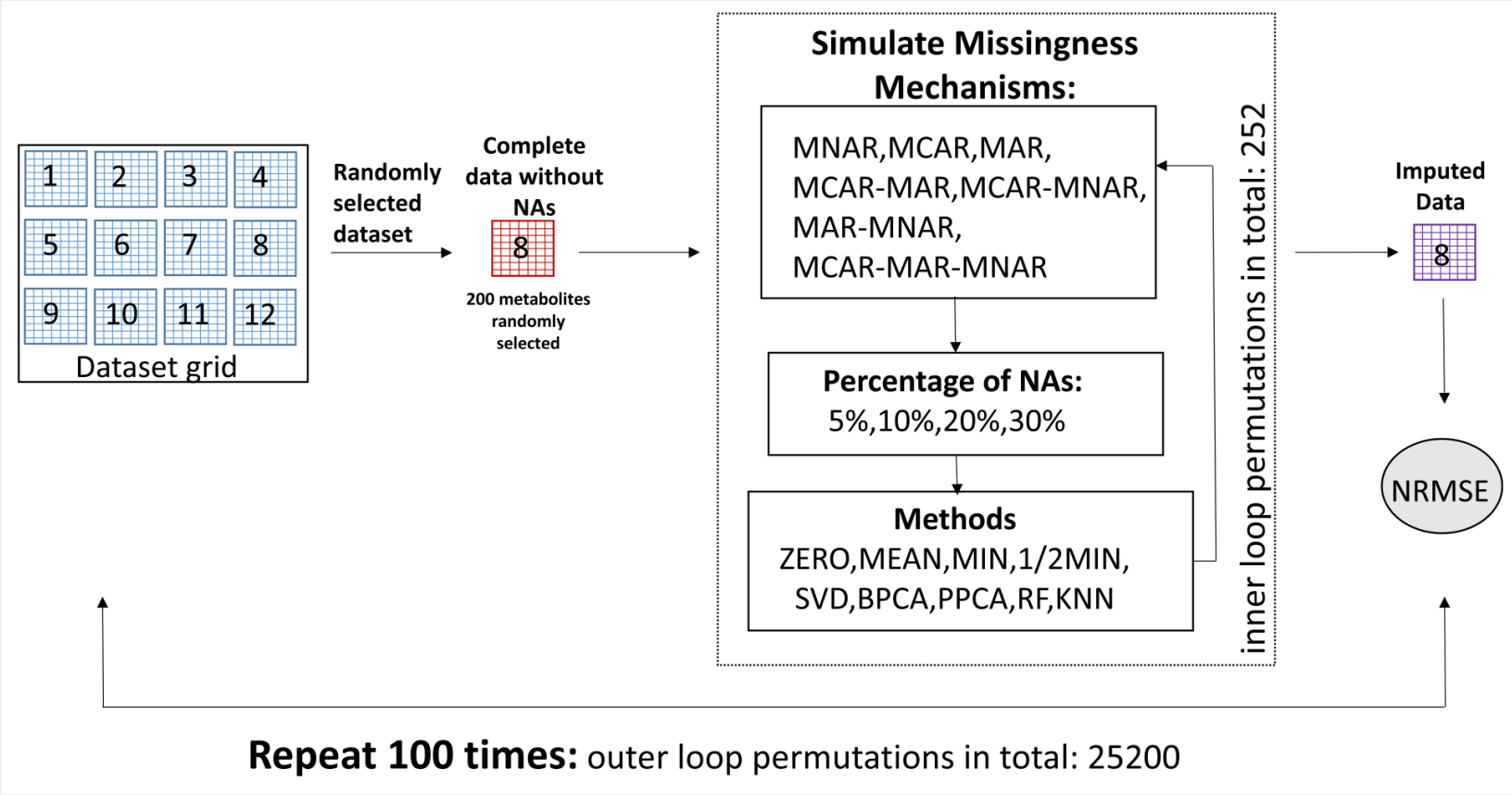
缺失數據是隨機發生

MAR
(Missing at Random)
隨機缺失

特徵缺失的機率可由其他觀察到的特徵決定

MNAR
(Missing Not at Random)
非隨機缺失

分子特徵的值是遺漏的原因，在許多代謝組學研究中被描述為左截切數據（分子特徵出現在檢測限制以下）



Repeat 100 times: outer loop permutations in total: 25200

Normalized Root Mean Square Error (NRMSE)

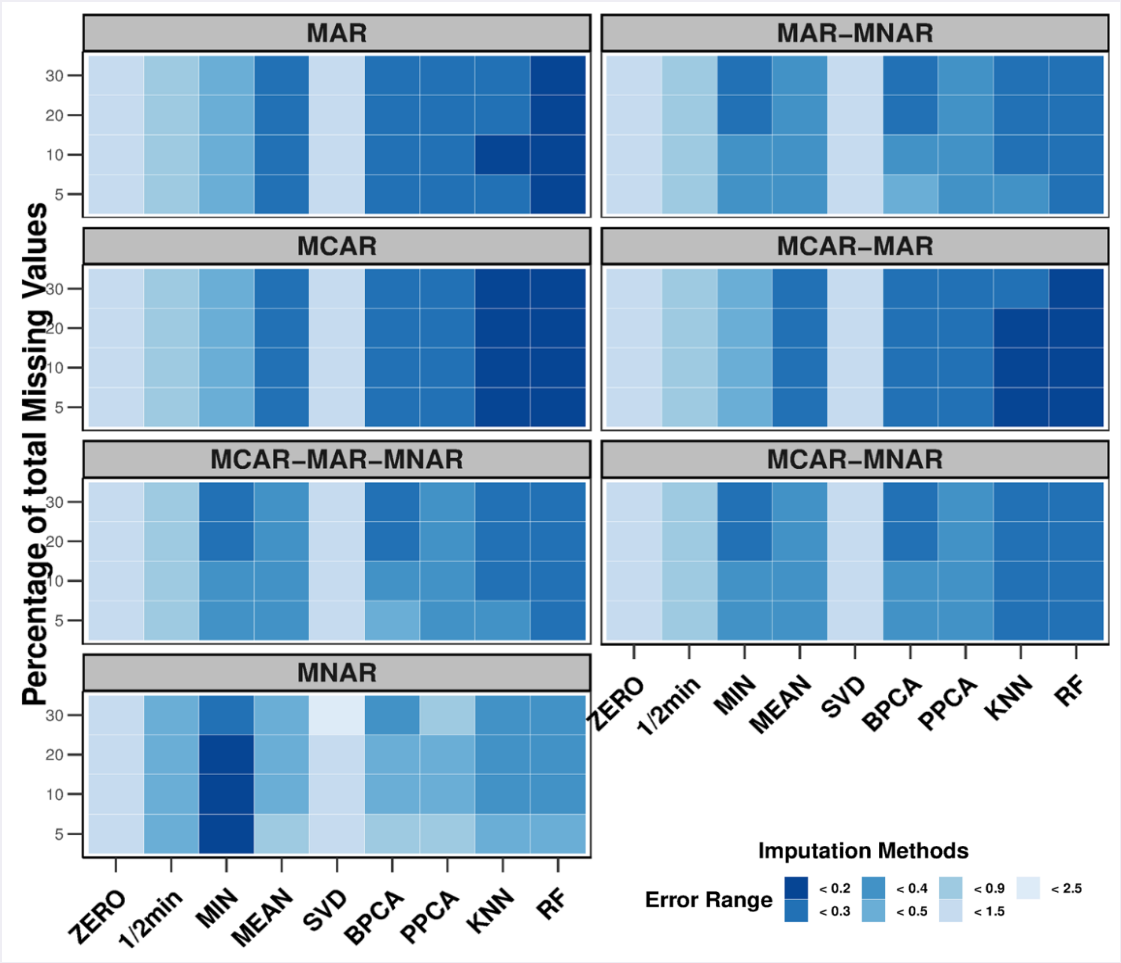
$$NRMSE = \sqrt{\frac{mean\left(\left(X^{comp} - X^{imp}\right)^2\right)}{var\left(X^{comp}\right)}}$$

X^{comp} :沒有缺失值的數據集

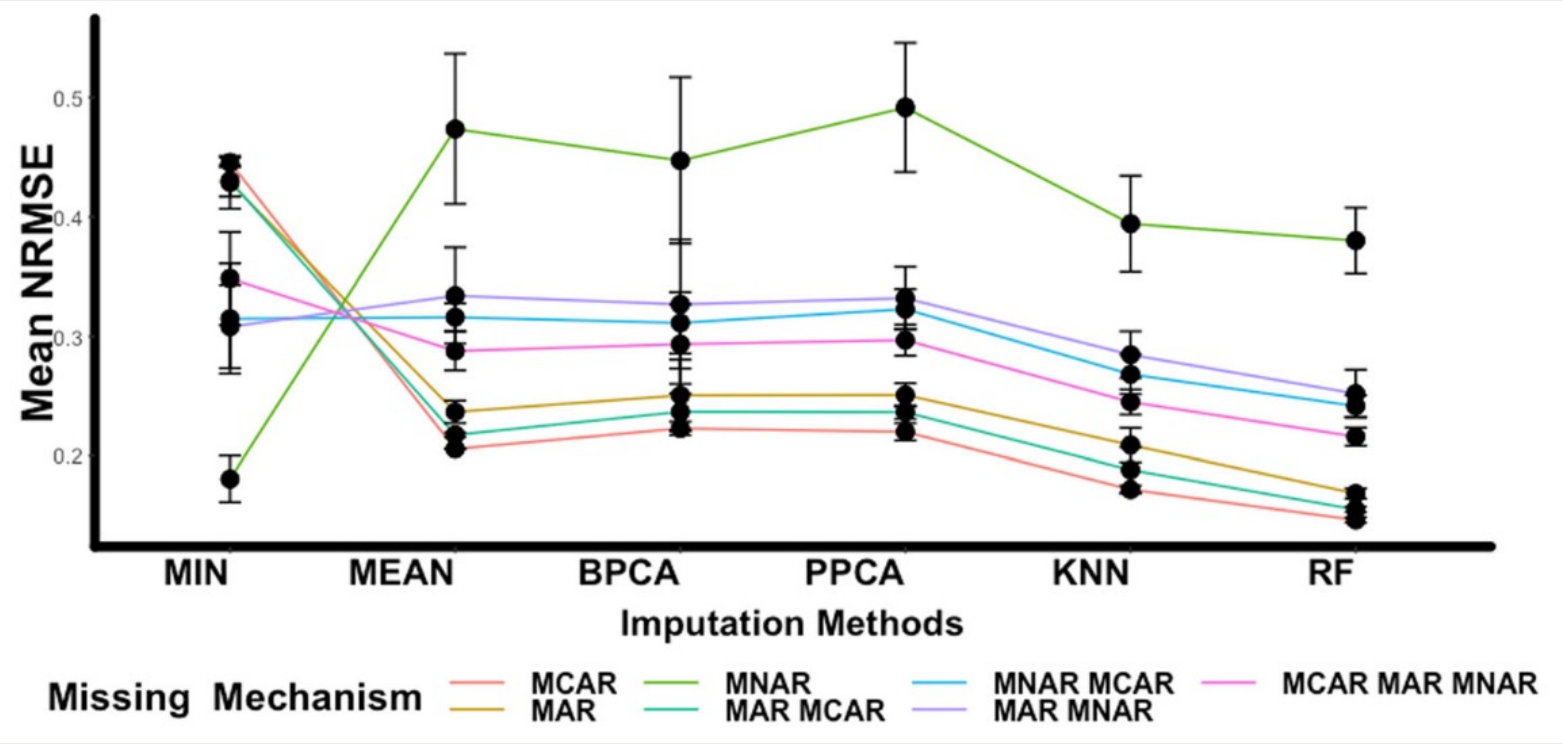
X^{imp} :進行了缺失值填補的同一數據集

03

Results



ZERO
½ MIN(½ - 最小値)
MIN(最小値)
MEAN(平均値)
SVD(奇異値分解)
BPCA(貝氏主成分分析)
PPCA(機率主成分分析)
KNN(K 最近鄰)
RF(隨機森林)


Mean NRMSE
0.5
0.4
0.3
0.2
MIN
MEAN
BPCA
PPCA
KNN
RF
Imputation Methods
Missing Mechanism
MCAR
MAR
MNAR
MAR MCAR
MNAR MCAR
MAR MNAR
MCAR MAR MNAR

04

Discussions

Discussion

RF(隨機森林)在填補代謝組學數據集方面是最有效方法。

多變量模型在跨數據集缺失情況下表現更好。

單一值替代（例如MIN）等缺失數據處理方法，
在數據來自非完全隨機缺失機制時表現較好。

其他基於局部結構的方法（例如RF）在涉及隨機性的情況下表現更好。

如果數據是左截尾的，那麼首選的選擇將是MIN或任何其他填補方法，
如KNN-TN和GSimp

05

Conclusion & Suggestion

Conclusion & Suggestion

缺失的類型和所佔比例會影響插補方法的性能和適用性。
在大多數測試的情境中,基於RF的填補方法表現最佳。
我們建議在填補缺失的代謝組學數據時使用RF-based填補方法,
因為通常事先並不知道缺失的原因。

在處理完整的代謝組學數據集時,需要考慮分組以避免偏差,
甚至要引入誤差,因此插補可能需要以分組方式進行

Resources

- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3110-0

Thanks!