Impute evaluation vignette

## Package install and data import

library(magrittr)  
library(knitr)  
library(ggplot2)  
library(reshape2)  
source('Imputation evaluations.R')  
data\_test <- read.csv('OB\_data/Real\_data\_DM.csv', row.names = 1)  
group <- rownames(data\_test) %>% gsub('()-.\*', '\\1', .) %>% as.factor()

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | var\_1 | var\_2 | var\_3 | var\_4 | var\_5 | var\_6 | var\_7 | var\_8 |
| DM-1 | 70.791 | 514.177 | 206.219 | 411.484 | 4.721 | 21.626 | 7.300 | 0.648 |
| DM-10 | 50.623 | 128.610 | 41.453 | 227.880 | 0.540 | 47.970 | 4.781 | 2.123 |
| DM-11 | 42.720 | 125.704 | 46.343 | 160.441 | 2.990 | 9.155 | 9.994 | 0.524 |
| DM-12 | 15.682 | 67.553 | 21.916 | 71.591 | 0.728 | 23.039 | 5.737 | 2.506 |
| DM-13 | 65.920 | 72.615 | 18.242 | 123.881 | 1.381 | 18.932 | 3.326 | 0.708 |
| DM-14 | 161.750 | 379.706 | 26.088 | 274.699 | 0.490 | 43.720 | 15.053 | 4.504 |

## [1] DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM  
## [24] DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM  
## [47] DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM  
## [70] DM N N N N N N N N N N N N N N N N N N N N N N   
## [93] N N N N N N N N N N N N N N N N N N N N N N N   
## [116] N N N N N N N N N N N N N N N N N N N N N N N   
## [139] N N N N N N N N N N N N N N N N N N N N N N N   
## [162] N N N N N N N N N N N N N N N N N N N N N N N   
## [185] N N N N N N N N N N N N N N   
## Levels: DM N

# MCAR

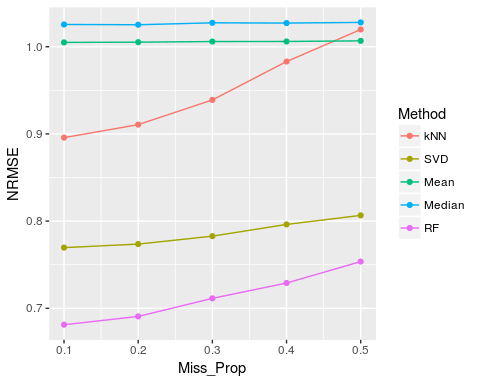
## MCAR generation and imputation

MCAR\_list <- MCAR\_gen\_imp(data\_c = data\_test, prop = seq(.1, .5, .1), impute\_list = c('kNN\_wrapper', 'SVD\_wrapper', 'Mean\_wrapper', 'Median\_wrapper', 'RF\_wrapper'), cores = 10)

## MCAR NRMSE evaluation and plot

MCAR\_NRMSE\_list <- NRMSE\_cal\_plot(MCAR\_list, plot = T, x = 'Miss\_Prop')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.8957382 | 0.7696032 | 1.004958 | 1.025544 | 0.6811728 | 0.1 | 130 |
| 0.9106655 | 0.7736548 | 1.005224 | 1.025213 | 0.6906557 | 0.2 | 130 |
| 0.9389917 | 0.7827661 | 1.005945 | 1.027410 | 0.7113980 | 0.3 | 130 |
| 0.9830233 | 0.7961935 | 1.006039 | 1.027144 | 0.7289833 | 0.4 | 130 |
| 1.0197809 | 0.8065801 | 1.006747 | 1.027885 | 0.7535703 | 0.5 | 130 |

## The above table shows the NRMSE of different imputaion methods

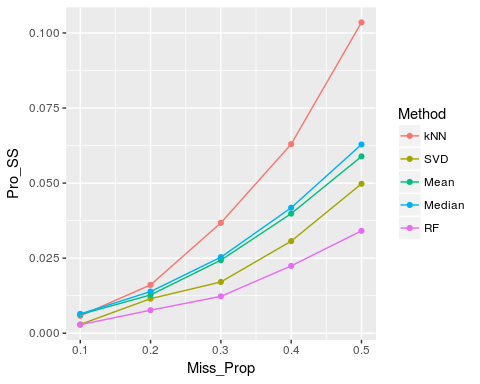
|  |  |  |
| --- | --- | --- |
| Miss\_Prop | Method | NRMSE |
| 0.1 | kNN | 0.8957382 |
| 0.2 | kNN | 0.9106655 |
| 0.3 | kNN | 0.9389917 |
| 0.4 | kNN | 0.9830233 |
| 0.5 | kNN | 1.0197809 |
| 0.1 | SVD | 0.7696032 |

## The above melted table is good for ggplot2

## MCAR PCA Procrustes analysis and plot

MCAR\_PCA\_ProSS\_list <- Procrustes\_cal\_plot(MCAR\_list, DR = 'PCA', nPCs = 2, x = 'Miss\_Prop', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.0058081 | 0.0029227 | 0.0062499 | 0.0064181 | 0.0028319 | 0.1 | 130 |
| 0.0160971 | 0.0115052 | 0.0127537 | 0.0138729 | 0.0076673 | 0.2 | 130 |
| 0.0367555 | 0.0170723 | 0.0243381 | 0.0253273 | 0.0122743 | 0.3 | 130 |
| 0.0629755 | 0.0306623 | 0.0398333 | 0.0417957 | 0.0224246 | 0.4 | 130 |
| 0.1035155 | 0.0497668 | 0.0588806 | 0.0628540 | 0.0340962 | 0.5 | 130 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

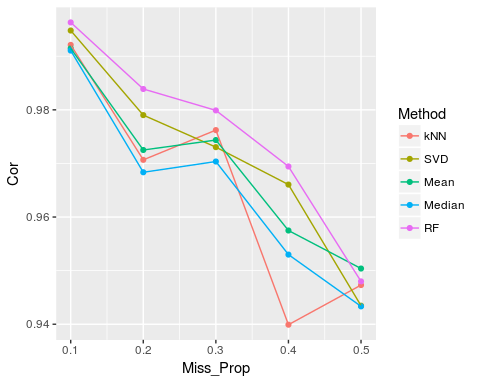
|  |  |  |
| --- | --- | --- |
| Miss\_Prop | Method | Pro\_SS |
| 0.1 | kNN | 0.0058081 |
| 0.2 | kNN | 0.0160971 |
| 0.3 | kNN | 0.0367555 |
| 0.4 | kNN | 0.0629755 |
| 0.5 | kNN | 0.1035155 |
| 0.1 | SVD | 0.0029227 |

## The above melted table is good for ggplot2

## MCAR T-test results correlation

MCAR\_Ttest\_cor\_list <- Ttest\_cor\_cal\_plot(MCAR\_list, group = group, plot = T, x = 'Miss\_Prop', cor = 'P')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.9921166 | 0.9947967 | 0.9914626 | 0.9910585 | 0.9963095 | 0.1 | 130 |
| 0.9706710 | 0.9790125 | 0.9725072 | 0.9683502 | 0.9838864 | 0.2 | 130 |
| 0.9762177 | 0.9730374 | 0.9743628 | 0.9703540 | 0.9798917 | 0.3 | 130 |
| 0.9399089 | 0.9660529 | 0.9574841 | 0.9529879 | 0.9694339 | 0.4 | 130 |
| 0.9473047 | 0.9434554 | 0.9503873 | 0.9433293 | 0.9479859 | 0.5 | 130 |

## The above table shows the Pearson Correlation of log T-test P-values between imputed data and complete data

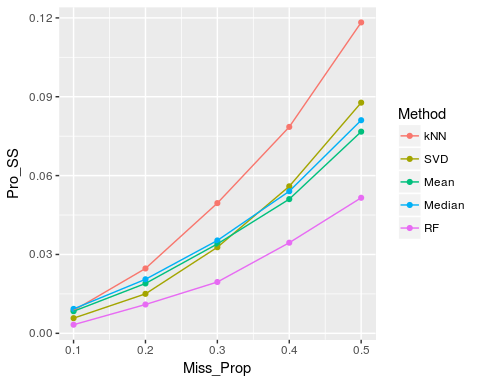
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.9906485 | 0.9914788 | 0.9890754 | 0.9883762 | 0.9939915 | 0.1 | 130 |
| 0.9803247 | 0.9781016 | 0.9776427 | 0.9761952 | 0.9877699 | 0.2 | 130 |
| 0.9700610 | 0.9571263 | 0.9714758 | 0.9686299 | 0.9732510 | 0.3 | 130 |
| 0.9372598 | 0.9338513 | 0.9505933 | 0.9454861 | 0.9560120 | 0.4 | 130 |
| 0.9098718 | 0.9009081 | 0.9193762 | 0.9063868 | 0.9426621 | 0.5 | 130 |

## The above table shows the Spearman Correlation of T-test P-values between imputed data and complete data

## MCAR PLS Procrustes analysis and plot

MCAR\_PLS\_ProSS\_list <- Procrustes\_cal\_plot(MCAR\_list, DR = 'PLS', nPCs = 2, outcome = group, x = 'Miss\_Prop', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.0086773 | 0.0057721 | 0.0084034 | 0.0092919 | 0.0032826 | 0.1 | 130 |
| 0.0246732 | 0.0150190 | 0.0189715 | 0.0205573 | 0.0109688 | 0.2 | 130 |
| 0.0495326 | 0.0327509 | 0.0340160 | 0.0353257 | 0.0195211 | 0.3 | 130 |
| 0.0785230 | 0.0559197 | 0.0510504 | 0.0540521 | 0.0344740 | 0.4 | 130 |
| 0.1182722 | 0.0877537 | 0.0767151 | 0.0810655 | 0.0515663 | 0.5 | 130 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

|  |  |  |
| --- | --- | --- |
| Miss\_Prop | Method | Pro\_SS |
| 0.1 | kNN | 0.0086773 |
| 0.2 | kNN | 0.0246732 |
| 0.3 | kNN | 0.0495326 |
| 0.4 | kNN | 0.0785230 |
| 0.5 | kNN | 0.1182722 |
| 0.1 | SVD | 0.0057721 |

## The above melted table is good for ggplot2

# MNAR

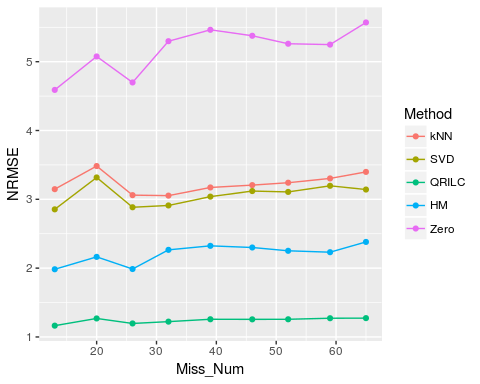
## MNAR generation and imputation

MNAR\_list <- MNAR\_gen\_imp(data\_c = data\_test, mis\_var\_prop = seq(.1, .5, .05), var\_mis\_prop = seq(.3, .6, .01), impute\_list = c('kNN\_wrapper', 'SVD\_wrapper', 'QRILC\_wrapper', 'HM\_wrapper', 'Zero\_wrapper'), cores = 1)

## MNAR NRMSE evaluation and plot

MNAR\_NRMSE\_list <- NRMSE\_cal\_plot(MNAR\_list, plot = T, x = 'Miss\_Num')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7  
## [1] 8  
## [1] 9



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 3.147346 | 2.854218 | 1.164114 | 1.983631 | 4.590021 | 0.0441725 | 13 |
| 3.483512 | 3.317546 | 1.268794 | 2.163684 | 5.076054 | 0.0681818 | 20 |
| 3.060679 | 2.884561 | 1.196039 | 1.986968 | 4.698292 | 0.0883450 | 26 |
| 3.053655 | 2.910843 | 1.222501 | 2.265631 | 5.297800 | 0.1084305 | 32 |
| 3.172240 | 3.039070 | 1.257001 | 2.323954 | 5.463234 | 0.1400544 | 39 |
| 3.206389 | 3.120068 | 1.255708 | 2.299862 | 5.376610 | 0.1648407 | 46 |

## The above table shows the NRMSE of different imputaion methods

|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | NRMSE |
| 13 | kNN | 3.147346 |
| 20 | kNN | 3.483512 |
| 26 | kNN | 3.060679 |
| 32 | kNN | 3.053655 |
| 39 | kNN | 3.172240 |
| 46 | kNN | 3.206389 |

## The above melted table is good for ggplot2

## MNAR NRMSE rank evaluation and plot

MNAR\_NRMSE\_rank\_list <- NRMSE\_rank\_cal\_plot(MNAR\_list, plot = T, x = 'Miss\_Num')

# ## [1] 1 ## [1] 2 ## [1] 3 ## [1] 4 ## [1] 5 ## [1] 6

# MNAR

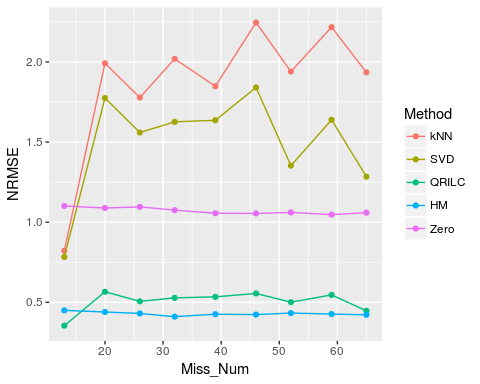
## MNAR generation and imputation

MNAR\_list <- MNAR\_gen\_imp(data\_c = data\_test, mis\_var\_prop = seq(.1, .5, .05), var\_mis\_prop = seq(.3, .6, .1), impute\_list = c('kNN\_wrapper', 'SVD\_wrapper', 'QRILC\_wrapper', 'HM\_wrapper', 'Zero\_wrapper'), cores = 1)

## MNAR NRMSE evaluation and plot

MNAR\_NRMSE\_list <- NRMSE\_cal\_plot(MNAR\_list, plot = T, x = 'Miss\_Num', sc=F)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7  
## [1] 8  
## [1] 9



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.8212974 | 0.7830408 | 0.3543918 | 0.4497508 | 1.101475 | 0.0454545 | 13 |
| 1.9923779 | 1.7752684 | 0.5659557 | 0.4396179 | 1.088549 | 0.0669775 | 20 |
| 1.7777514 | 1.5598421 | 0.5060922 | 0.4305481 | 1.095724 | 0.0870629 | 26 |
| 2.0196072 | 1.6265111 | 0.5279241 | 0.4104688 | 1.075491 | 0.1063326 | 32 |
| 1.8486088 | 1.6361416 | 0.5341109 | 0.4258740 | 1.056053 | 0.1394328 | 39 |
| 2.2466831 | 1.8408680 | 0.5554202 | 0.4232980 | 1.054718 | 0.1648019 | 46 |

## The above table shows the NRMSE of different imputaion methods

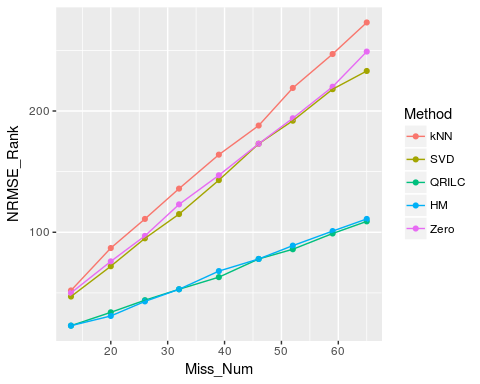
|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | NRMSE |
| 13 | kNN | 0.8212974 |
| 20 | kNN | 1.9923779 |
| 26 | kNN | 1.7777514 |
| 32 | kNN | 2.0196072 |
| 39 | kNN | 1.8486088 |
| 46 | kNN | 2.2466831 |

## The above melted table is good for ggplot2

## MNAR NRMSE rank evaluation and plot

MNAR\_NRMSE\_rank\_list <- NRMSE\_rank\_cal\_plot(MNAR\_list, plot = T, x = 'Miss\_Num')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7  
## [1] 8  
## [1] 9



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 52 | 47 | 23 | 23 | 50 | 0.0454545 | 13 |
| 87 | 72 | 34 | 31 | 76 | 0.0669775 | 20 |
| 111 | 95 | 44 | 43 | 97 | 0.0870629 | 26 |
| 136 | 115 | 53 | 53 | 123 | 0.1063326 | 32 |
| 164 | 143 | 63 | 68 | 147 | 0.1394328 | 39 |
| 188 | 173 | 78 | 78 | 173 | 0.1648019 | 46 |

## The above table shows the NRMSE ranks of different imputaion methods

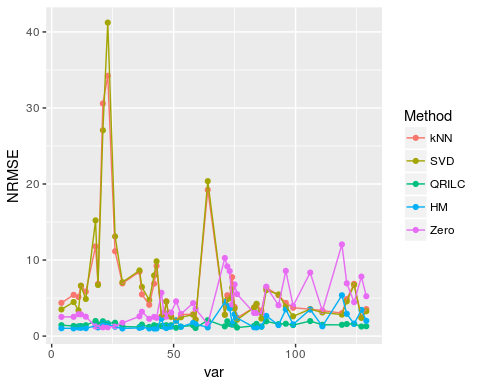
|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | NRMSE\_Rank |
| 13 | kNN | 52 |
| 20 | kNN | 87 |
| 26 | kNN | 111 |
| 32 | kNN | 136 |
| 39 | kNN | 164 |
| 46 | kNN | 188 |

## The above melted table is good for ggplot2

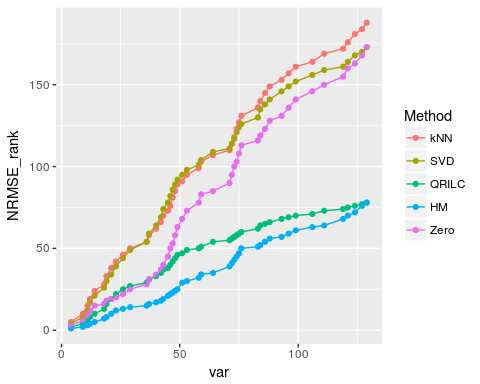
## NRMSE and SOR on MNAR

## [1] 6

## We randomly choose one cut point



## NRMSE for each missing variable, we observed that Zero was the worst at most variables, however, the overall NRMSE on all variables would ignore this problem

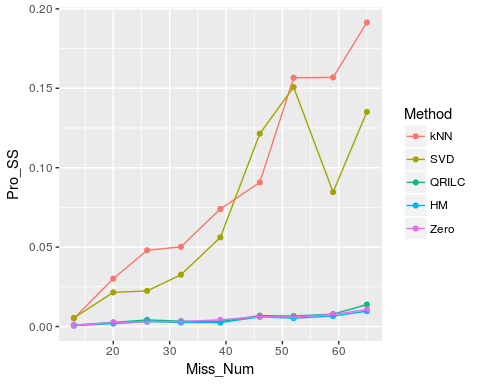


## Cumulative ranks for missing variables, we observed that Zero increased dramatically compared to other methods while the right end points represented the SOR

## MNAR PCA Procrustes analysis and plot

MNAR\_PCA\_ProSS\_list <- Procrustes\_cal\_plot(MNAR\_list, DR = 'PCA', nPCs = 2, x = 'Miss\_Num', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7  
## [1] 8  
## [1] 9



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.0050775 | 0.0054189 | 0.0008831 | 0.0005513 | 0.0007402 | 0.0454545 | 13 |
| 0.0301164 | 0.0215171 | 0.0025940 | 0.0018463 | 0.0024914 | 0.0669775 | 20 |
| 0.0480066 | 0.0224202 | 0.0041478 | 0.0031385 | 0.0031507 | 0.0870629 | 26 |
| 0.0501594 | 0.0326138 | 0.0033927 | 0.0025628 | 0.0031046 | 0.1063326 | 32 |
| 0.0740282 | 0.0561170 | 0.0031186 | 0.0024640 | 0.0042093 | 0.1394328 | 39 |
| 0.0907606 | 0.1214247 | 0.0068822 | 0.0061089 | 0.0062762 | 0.1648019 | 46 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

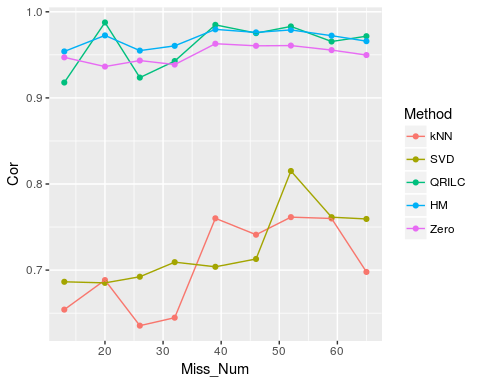
|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | Pro\_SS |
| 13 | kNN | 0.0050775 |
| 20 | kNN | 0.0301164 |
| 26 | kNN | 0.0480066 |
| 32 | kNN | 0.0501594 |
| 39 | kNN | 0.0740282 |
| 46 | kNN | 0.0907606 |

## The above melted table is good for ggplot2

## MNAR T-test results correlation

MNAR\_Ttest\_cor\_list <- Ttest\_cor\_cal\_plot(MNAR\_list, group = group, plot = T, x = 'Miss\_Num', cor = 'P')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7  
## [1] 8  
## [1] 9



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.6541117 | 0.6864177 | 0.9178544 | 0.9539387 | 0.9470735 | 0.0454545 | 13 |
| 0.6883988 | 0.6851539 | 0.9876039 | 0.9725649 | 0.9363864 | 0.0669775 | 20 |
| 0.6354761 | 0.6922887 | 0.9235852 | 0.9549122 | 0.9433824 | 0.0870629 | 26 |
| 0.6447259 | 0.7092143 | 0.9427681 | 0.9604551 | 0.9387029 | 0.1063326 | 32 |
| 0.7601306 | 0.7038311 | 0.9848421 | 0.9795935 | 0.9629277 | 0.1394328 | 39 |
| 0.7410988 | 0.7129162 | 0.9751929 | 0.9760754 | 0.9604850 | 0.1648019 | 46 |

## The above table shows the Pearson Correlation of log T-test P-values between imputed data and complete data

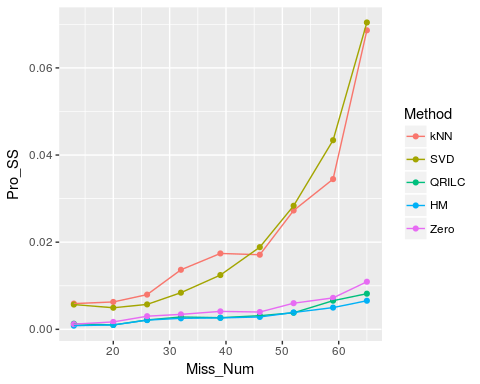
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.5274725 | 0.4670330 | 0.8956044 | 0.9505495 | 0.9340659 | 0.0454545 | 13 |
| 0.5488722 | 0.4030075 | 0.9804511 | 0.9609023 | 0.9353383 | 0.0669775 | 20 |
| 0.6348718 | 0.5829060 | 0.9589744 | 0.9254701 | 0.9295726 | 0.0870629 | 26 |
| 0.5513196 | 0.5898094 | 0.9600440 | 0.9450147 | 0.9325513 | 0.1063326 | 32 |
| 0.7064777 | 0.6530364 | 0.9777328 | 0.9672065 | 0.9506073 | 0.1394328 | 39 |
| 0.6980574 | 0.6752390 | 0.9555967 | 0.9574468 | 0.9469627 | 0.1648019 | 46 |

## The above table shows the Spearman Correlation of T-test P-values between imputed data and complete data

## MNAR PLS Procrustes analysis and plot

MNAR\_PLS\_ProSS\_list <- Procrustes\_cal\_plot(MNAR\_list, DR = 'PLS', nPCs = 2, outcome = group, x = 'Miss\_Num', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7  
## [1] 8  
## [1] 9



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.0059093 | 0.0056859 | 0.0012909 | 0.0008439 | 0.0011976 | 0.0454545 | 13 |
| 0.0062860 | 0.0049532 | 0.0010018 | 0.0010039 | 0.0017024 | 0.0669775 | 20 |
| 0.0079679 | 0.0057215 | 0.0021548 | 0.0021177 | 0.0030167 | 0.0870629 | 26 |
| 0.0136659 | 0.0084136 | 0.0028307 | 0.0025309 | 0.0034367 | 0.1063326 | 32 |
| 0.0174182 | 0.0124557 | 0.0026718 | 0.0026206 | 0.0041173 | 0.1394328 | 39 |
| 0.0171202 | 0.0188630 | 0.0031326 | 0.0028430 | 0.0039739 | 0.1648019 | 46 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | Pro\_SS |
| 13 | kNN | 0.0059093 |
| 20 | kNN | 0.0062860 |
| 26 | kNN | 0.0079679 |
| 32 | kNN | 0.0136659 |
| 39 | kNN | 0.0174182 |
| 46 | kNN | 0.0171202 |

## The above melted table is good for ggplot2