MCAR 45% missing - Generative Adversarial Imputation Nets (GAIN)

```
# sample MCAR dataset from PUMS
source("../../utils/sampleMCAR.R")
n = 10000
missing_col = c(1,3,7,9,10,11)
missing_prob = 0.45
set.seed(1)

output_list <- sampleMCAR(n, missing_prob)
df <- output_list[['df']]
df_observed <- output_list[['df_observed']]
```

Generative Adversarial Imputation Nets (GAIN)

reference: https://arxiv.org/abs/1806.02920

```
# Load imputed dataset
d1 = read.csv('../../GAIN/imputed_dataset/MCAR_45percent_1.csv', header = FALSE, sep = ',')
d2 = read.csv('../../GAIN/imputed_dataset/MCAR_45percent_2.csv', header = FALSE, sep = ',')
d3 = read.csv('../../GAIN/imputed_dataset/MCAR_45percent_3.csv', header = FALSE, sep = ',')
d4 = read.csv('../../GAIN/imputed_dataset/MCAR_45percent_4.csv', header = FALSE, sep = ',')
d5 = read.csv('../../GAIN/imputed_dataset/MCAR_45percent_5.csv', header = FALSE, sep = ',')
colnames(d1) = colnames(df)
colnames(d2) = colnames(df)
colnames(d3) = colnames(df)
colnames(d4) = colnames(df)
colnames(d5) = colnames(df)
imputed_df = rbind(d1, d2, d3, d4, d5)
```

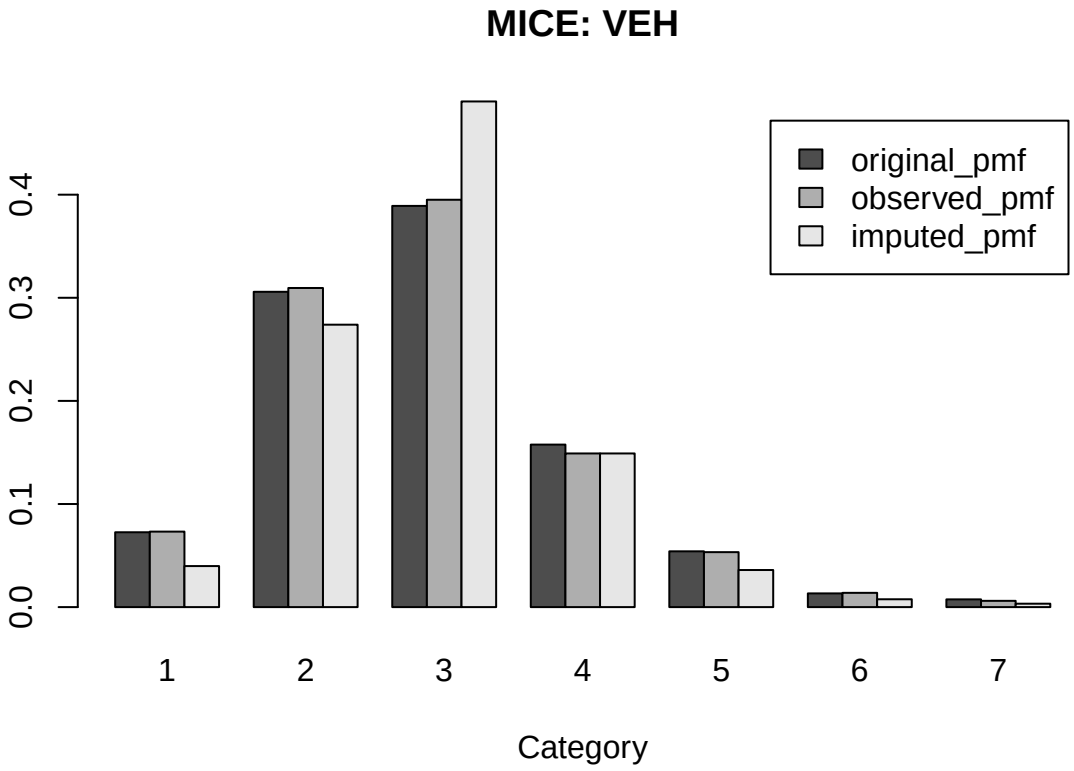
Diagnostics

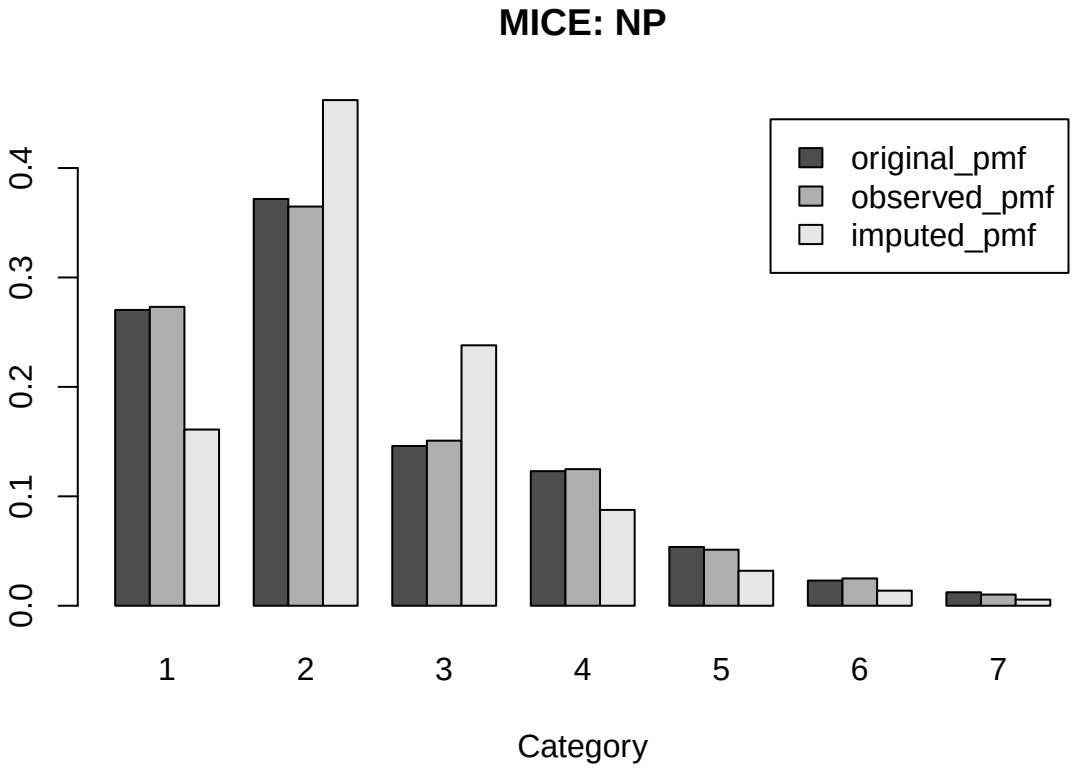
Assess bivariate joint distribution

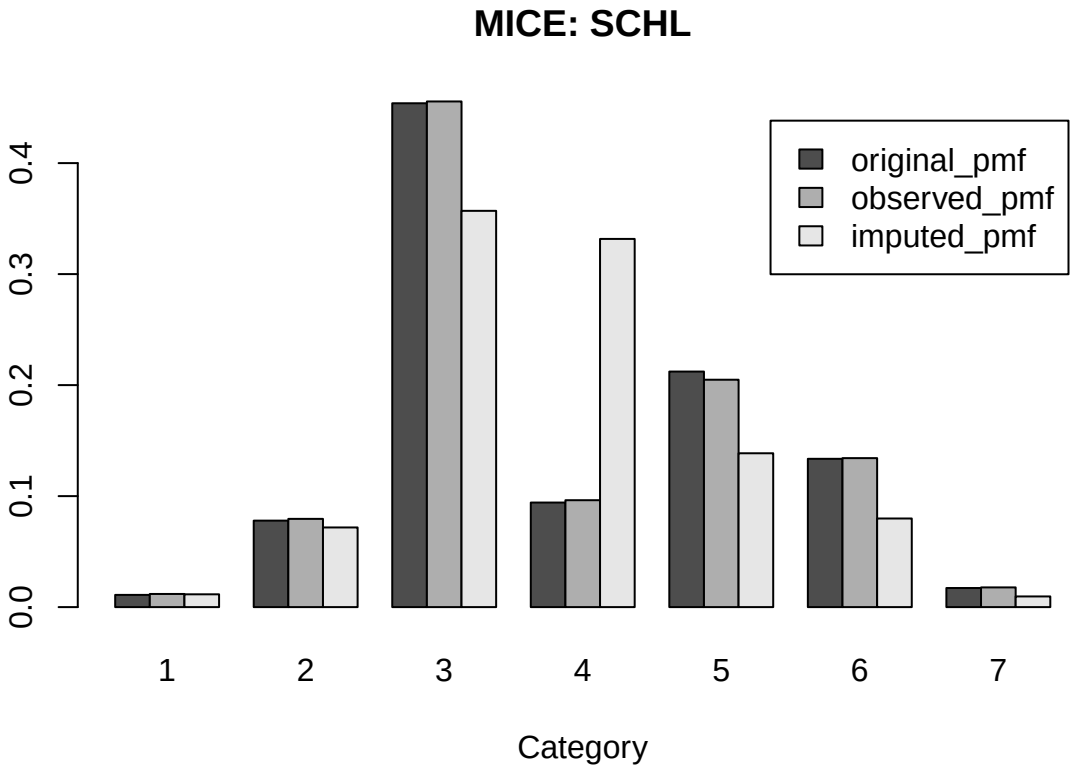
Assess bivariate joint distribution

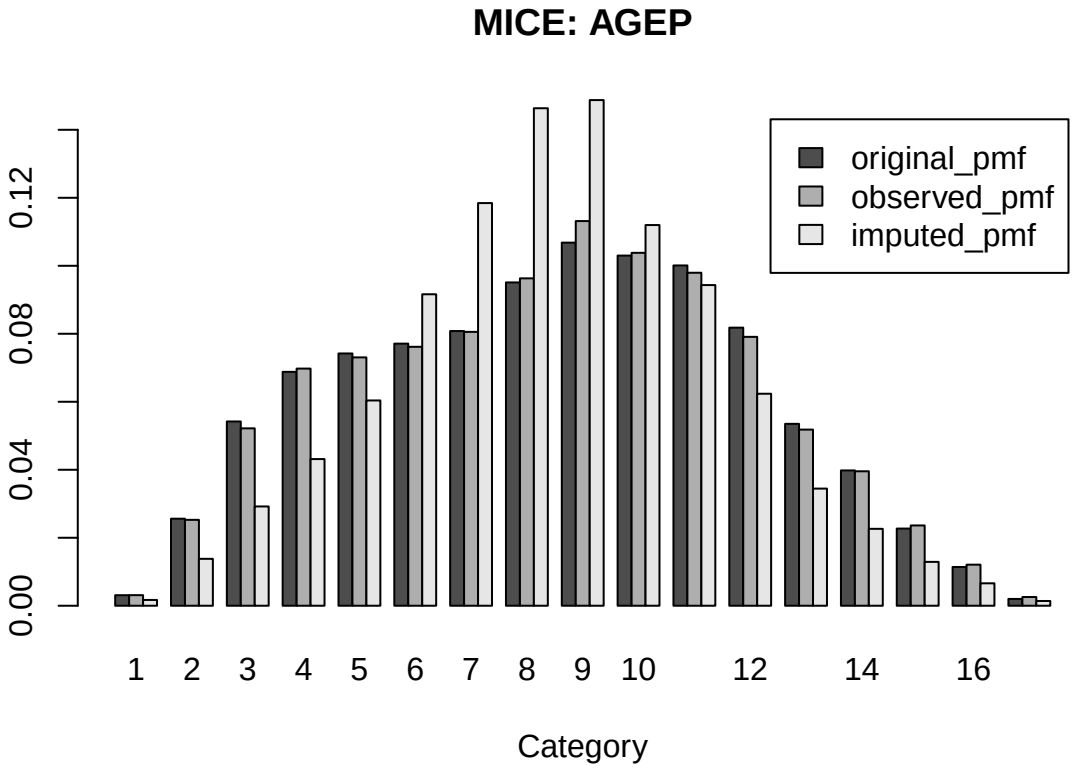
```
# calculate rmse
numeric_df = sapply(df, as.numeric)
normalized_df = t(t(numeric_df-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))
numeric_impute = sapply(d1, as.numeric)
normalized_impute = t(t(numeric_impute-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))

missing_matrix = is.na(df_observed)

rmse = sqrt(sum((normalized_df[missing_matrix] - normalized_impute[missing_matrix])^2)/sum(missing_matri
rmse
```

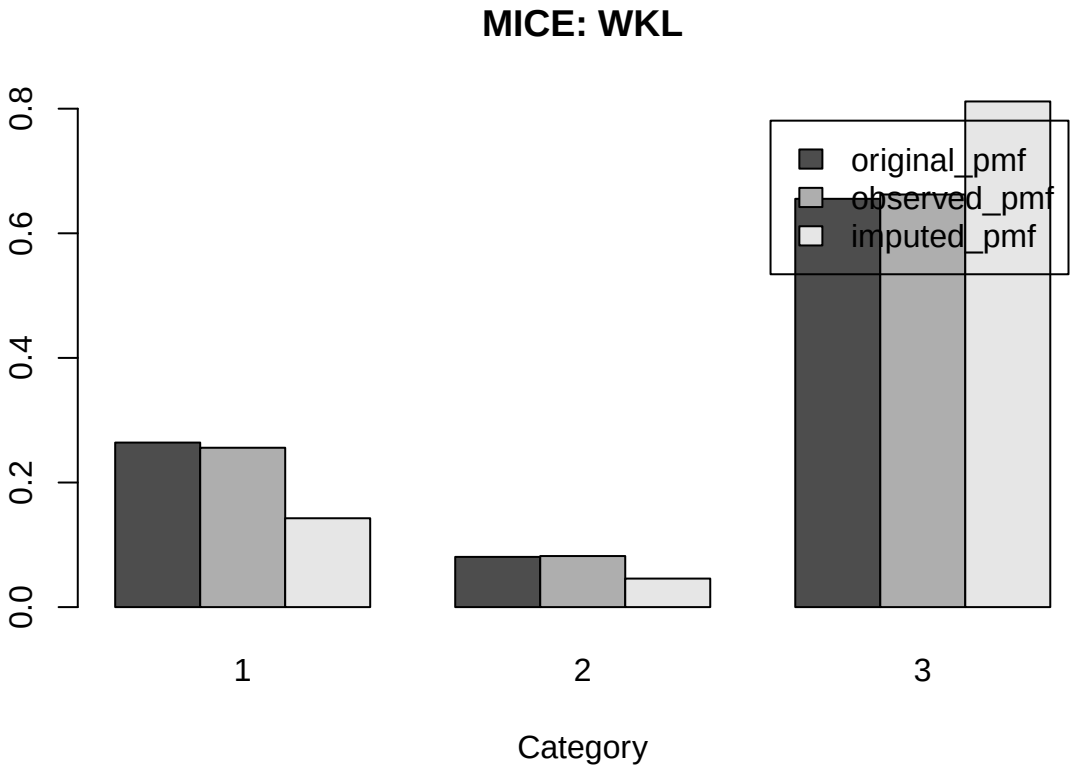
```
## [1] 0.2868933
```

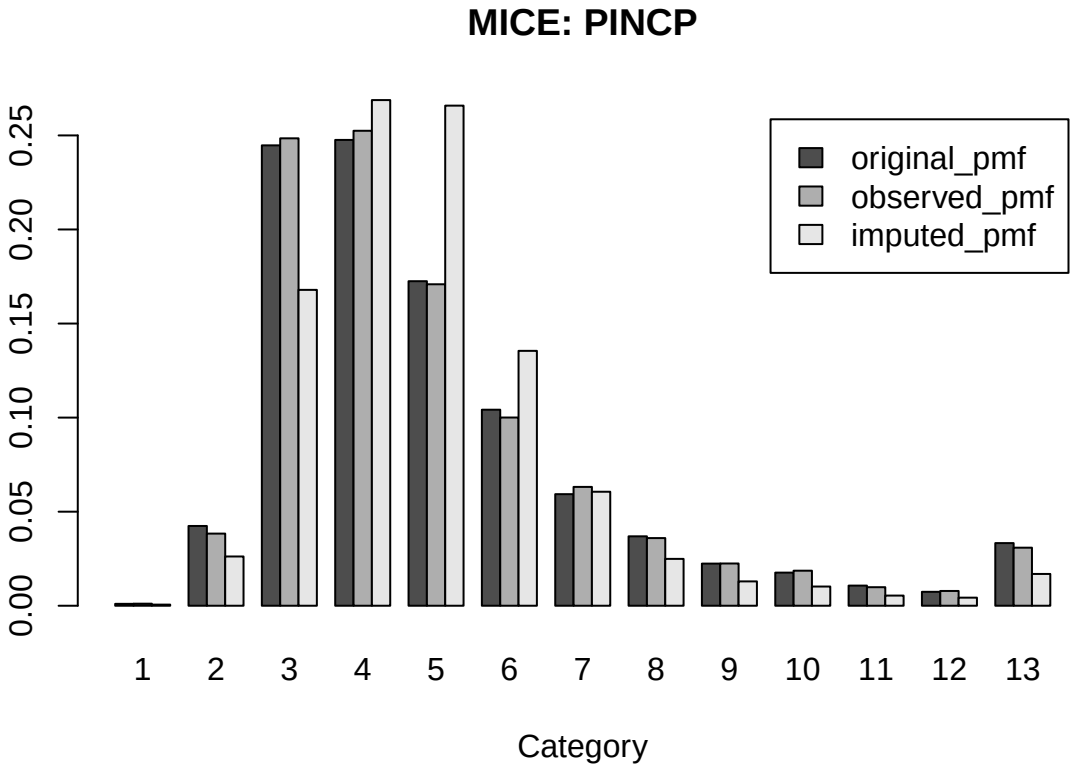

MICE: VEH
original_pmf
observed_pmf
imputed_pmf
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
Category

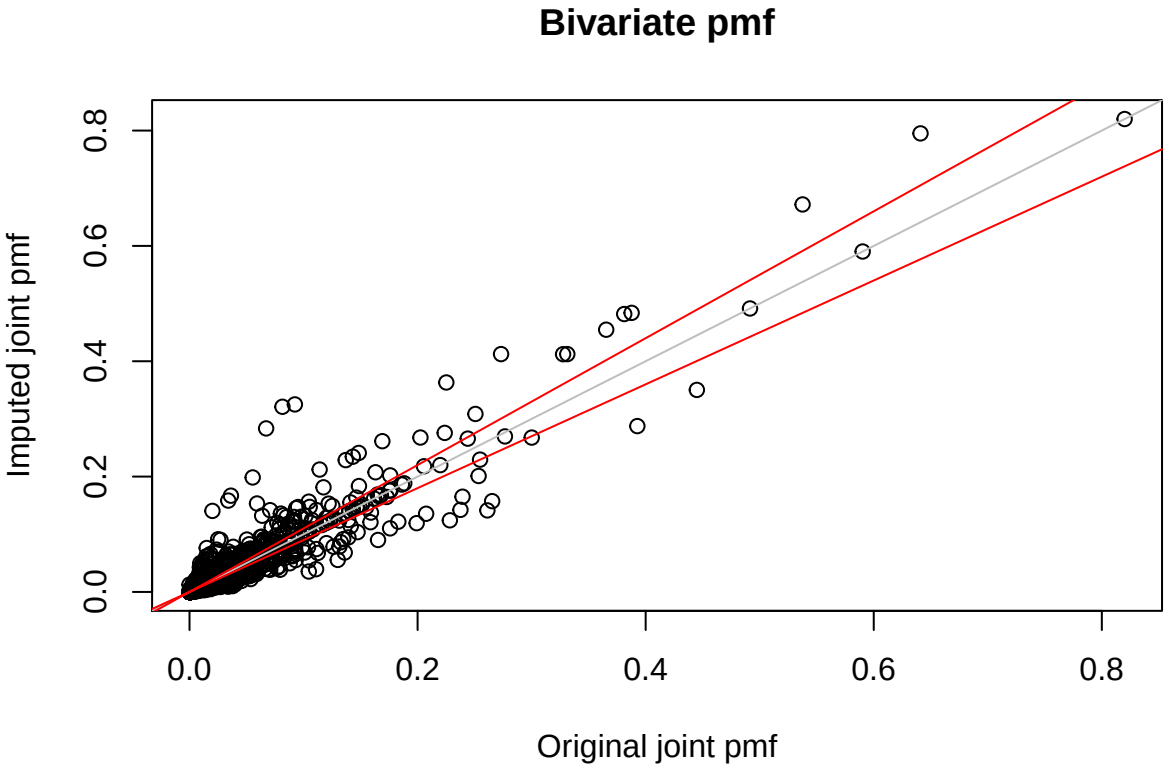

MICE: NP
original_pmf
observed_pmf
imputed_pmf
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
Category

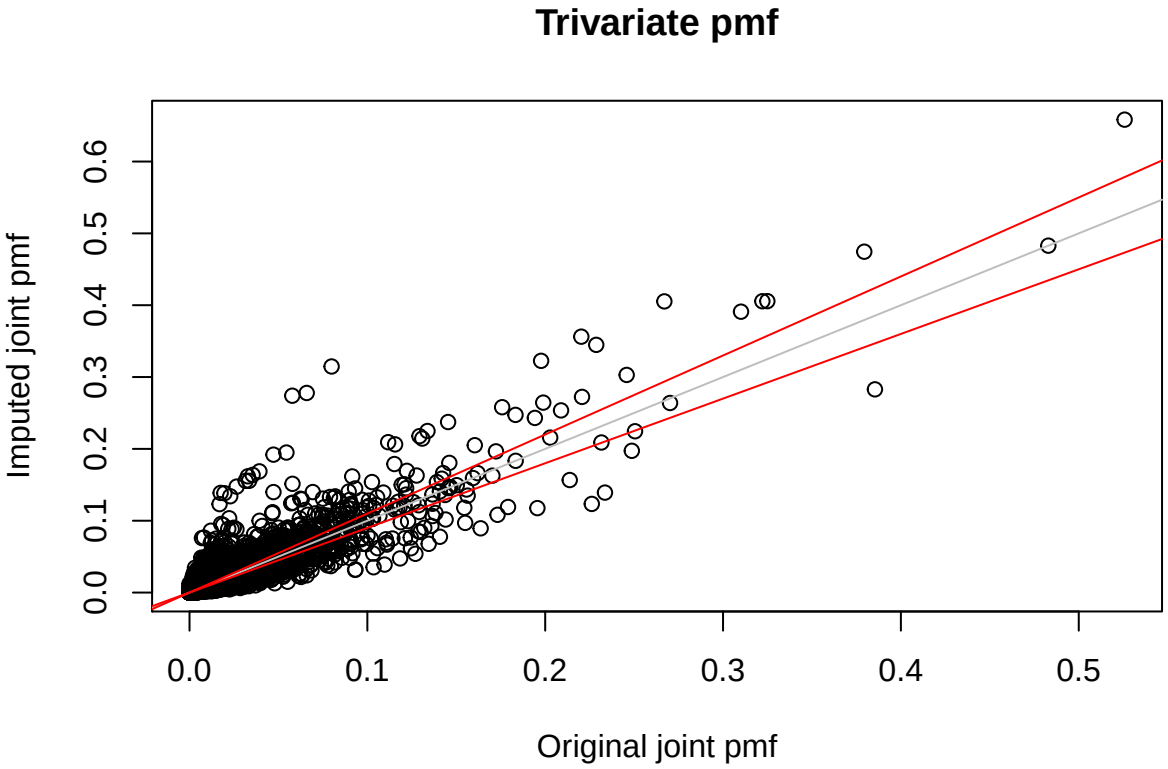

MICE: SCHL
original_pmf
observed_pmf
imputed_pmf
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
Category


MICE: AGEP
original_pmf
observed_pmf
imputed_pmf
0.00
0.04
0.08
0.12
1
2
3
4
5
6
7
8
9
10
12
14
16
Category


MICE: WKL
0.0
0.2
0.4
0.6
0.8
original_pmf
observed_pmf
imputed_pmf
1
2
3
Category


MICE: PINCP
0.00
0.05
0.10
0.15
0.20
0.25
original_pmf
observed_pmf
imputed_pmf
1
2
3
4
5
6
7
8
9
10
11
12
13
Category


Bivariate pmf
Imputed joint pmf
0.0
0.2
0.4
0.6
0.8
Original joint pmf


Trivariate pmf
Imputed joint pmf
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Original joint pmf