Testing different imputation methods on PUMS (MCAR) - Null Model

---

```r
# load dataset: df
load('../Datasets/ordinalPUMS.Rdata')

# take 10,000 samples: df
set.seed(0)
n = 10000
sample <- sample(nrow(df), size = 10000)
df <- df[sample,]

# create MCAR scneario with 30% chance of missing: df_observed
missing_prob = 0.3
df_observed <- df
missing_col = colnames(df)[c(1,3,5,7,9,11)]
for (col in missing_col) {
  missing_ind <- rbernoulli(n,p = missing_prob)
  df_observed[missing_ind, col] <- NA
}
```

Null Model

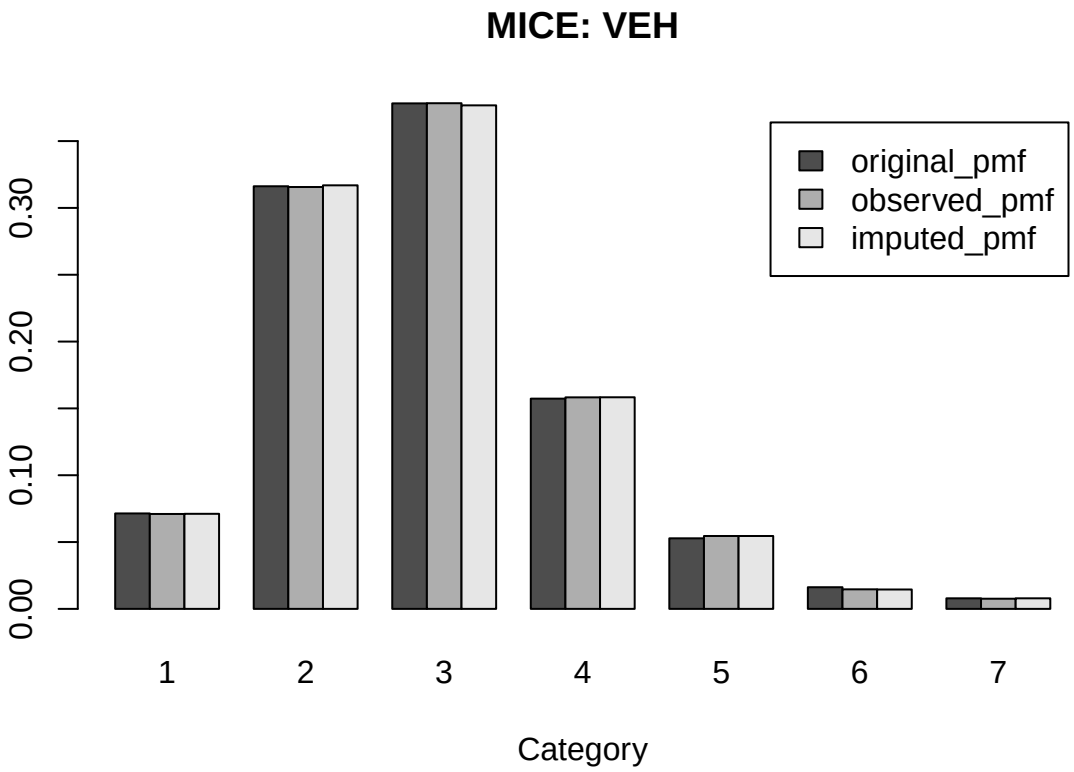
Using observed pmf for missing data imputation

```r
pmf_imputation <- function(data){
  # Missing data imputation with observed pmf
  # data: dataframe with missing values
  d = data
  missing_matrix = is.na(d)
  for (col in 1:dim(d)[2]) {
    # impute with observed pmf
    pmf = table(d[, col])
    pmf = pmf/sum(pmf)
    missing_indicator = missing_matrix[,col]

    n_missing = sum(missing_indicator)
    sample = rmultinom(n_missing,1,prob = pmf)
    d[missing_indicator,col] = apply(sample*(1:nrow(sample)),
                                     MARGIN = 2, FUN = sum)
  }
  # return imputed dataset
  return(d)
}
```

```r
# do random imputation 5 times
d1 = pmf_imputation(df_observed)
d2 = pmf_imputation(df_observed)
d3 = pmf_imputation(df_observed)
d4 = pmf_imputation(df_observed)
d5 = pmf_imputation(df_observed)
```

```
imputed_df = rbind(d1, d2, d3, d4, d5)
```

Diagnostics



Assess bivariate joint distribution

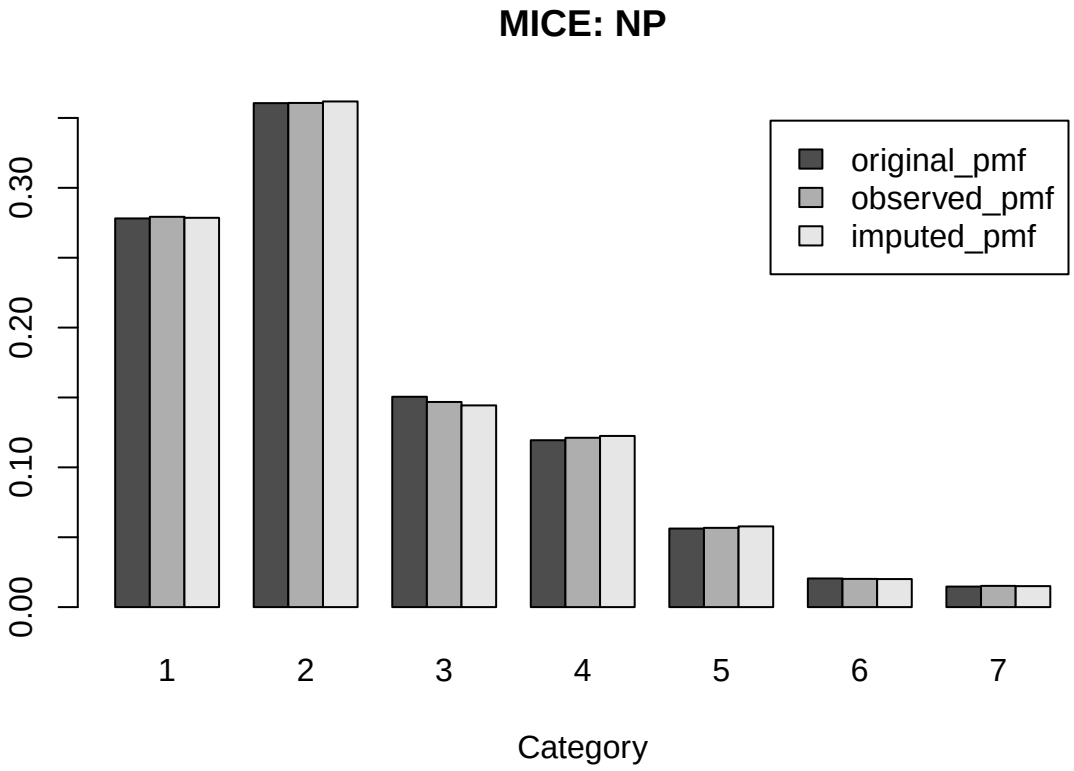
Assess bivariate joint distribution

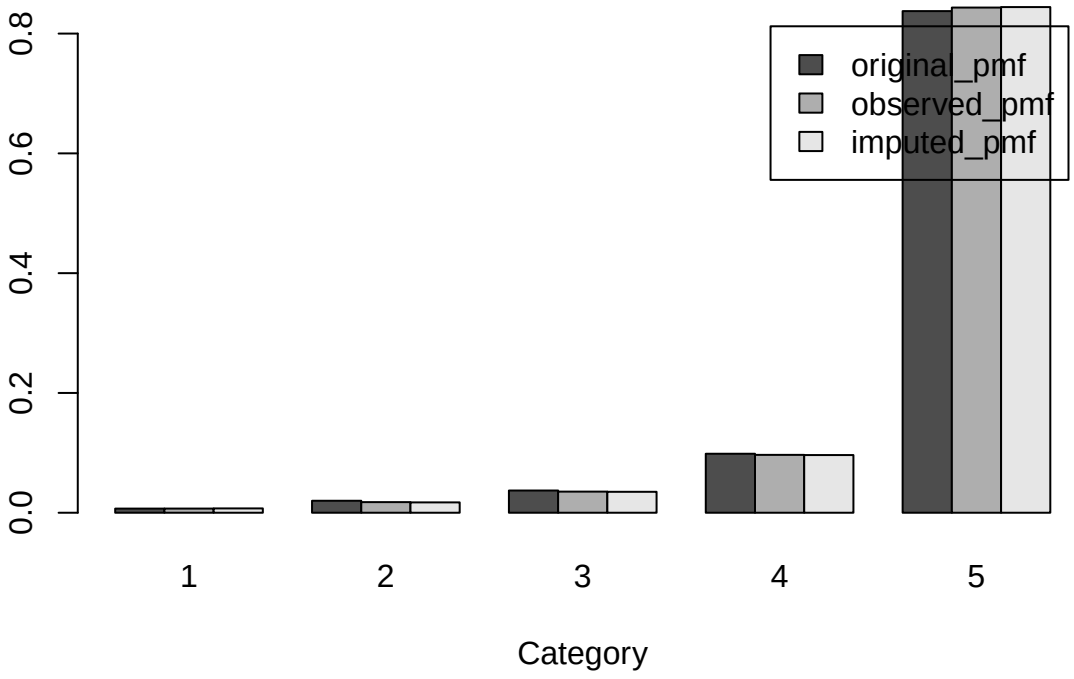
```
# calculate rmse
numeric_df = sapply(df, as.numeric)
normalized_df = t(t(numeric_df-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))
numeric_impute = sapply(d1, as.numeric)
normalized_impute = t(t(numeric_impute-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))

missing_matrix = is.na(df_observed)

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

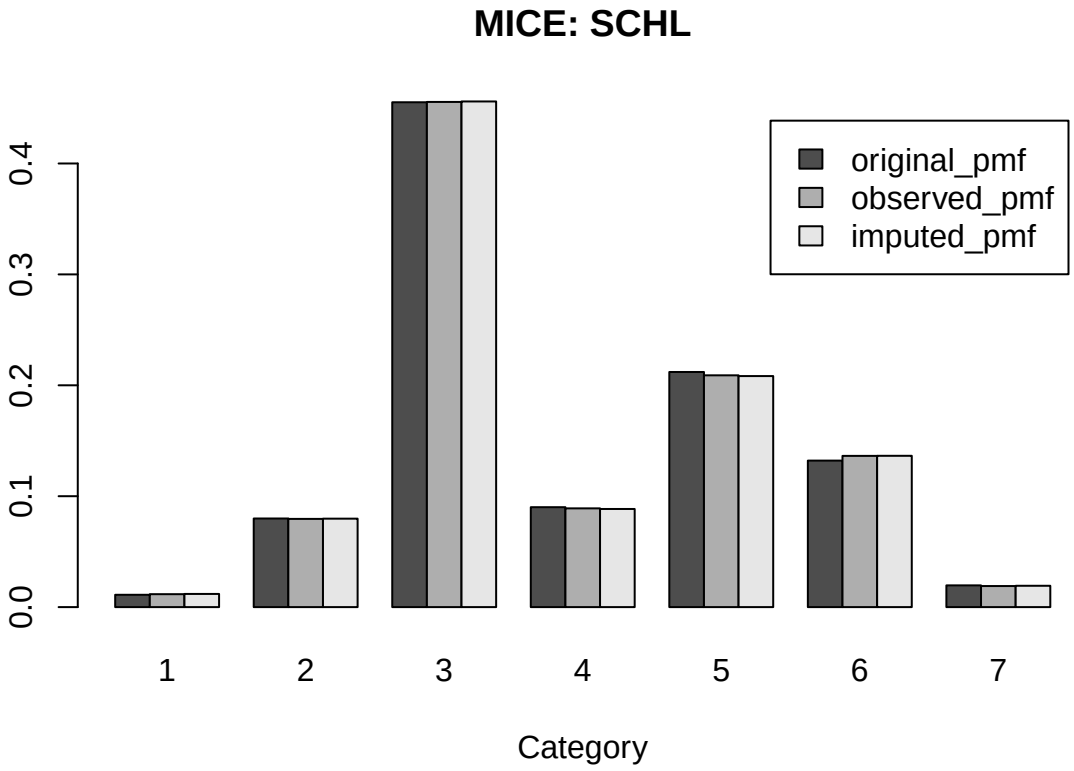
```
## [1] 0.2897228
```

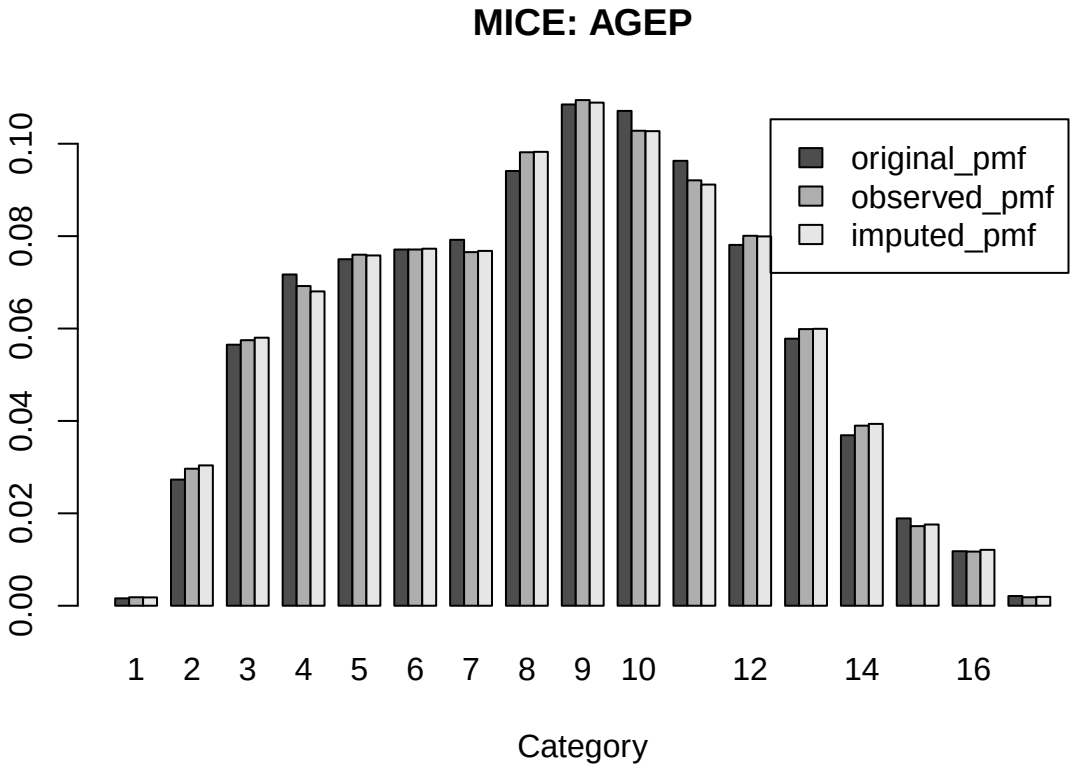
```
# accuracy
acc = sum(numeric_df[missing_matrix] == numeric_impute[missing_matrix])/sum(missing_matrix)
acc
```

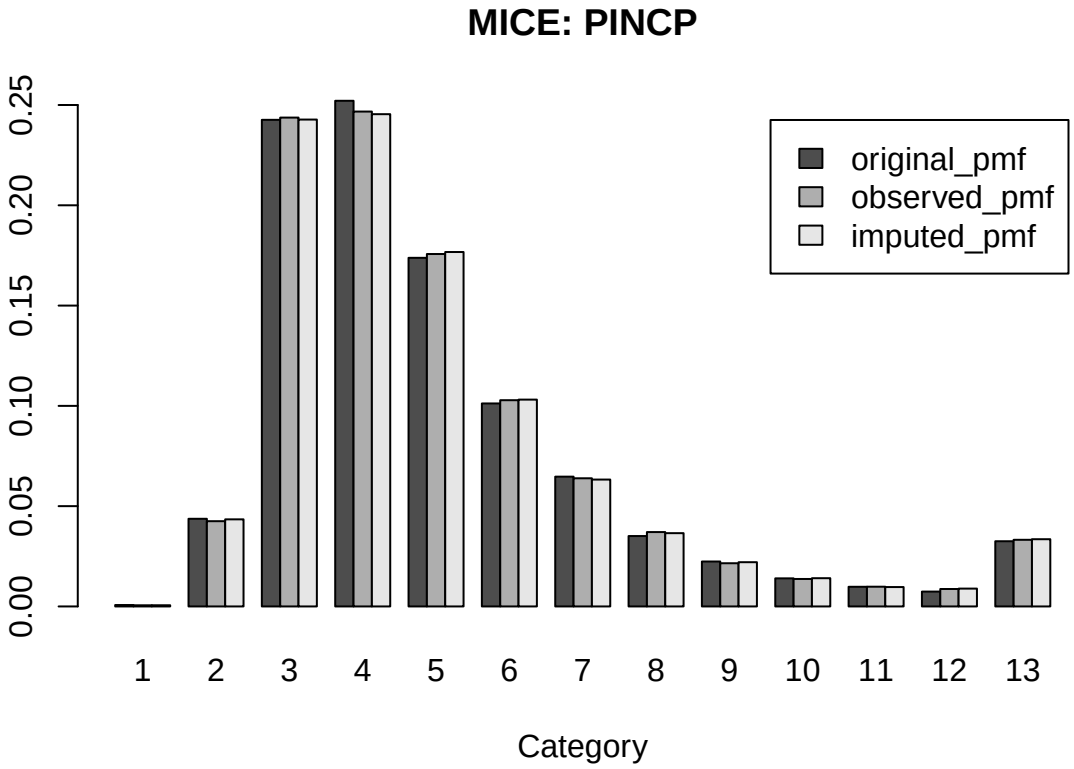
```
## [1] 0.2978077
```

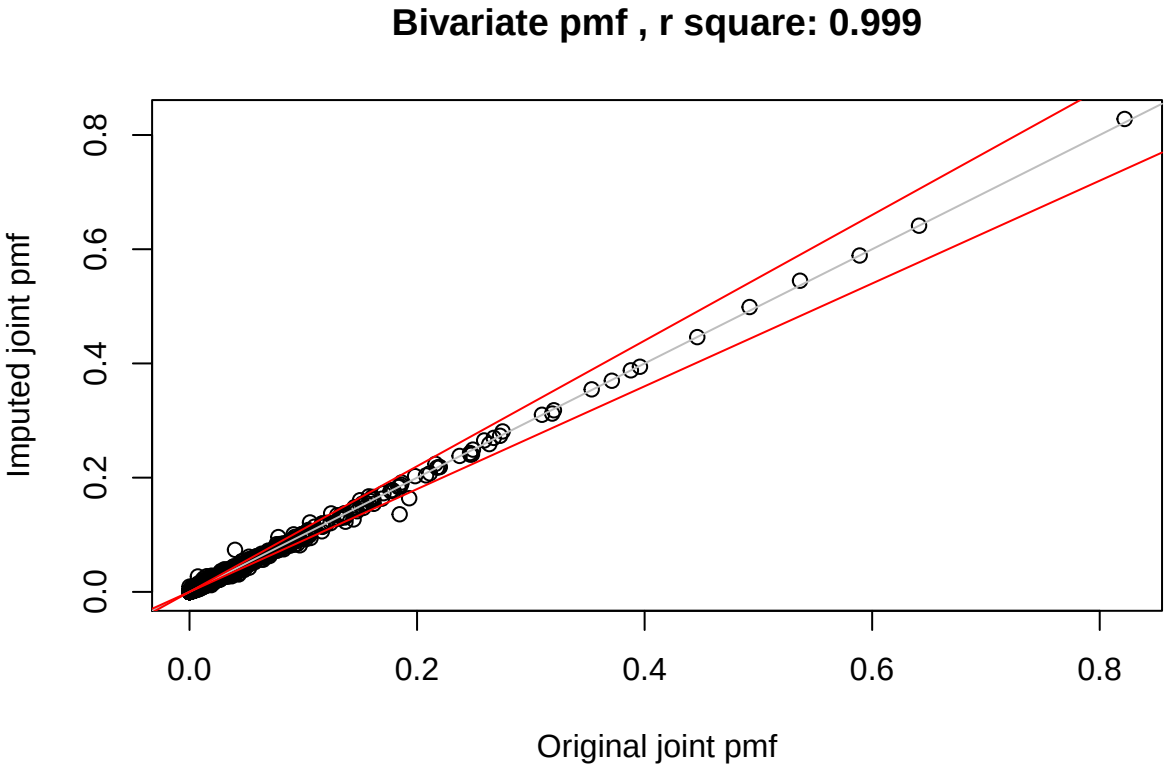

MICE: NP
original_pmf
observed_pmf
imputed_pmf
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
Category

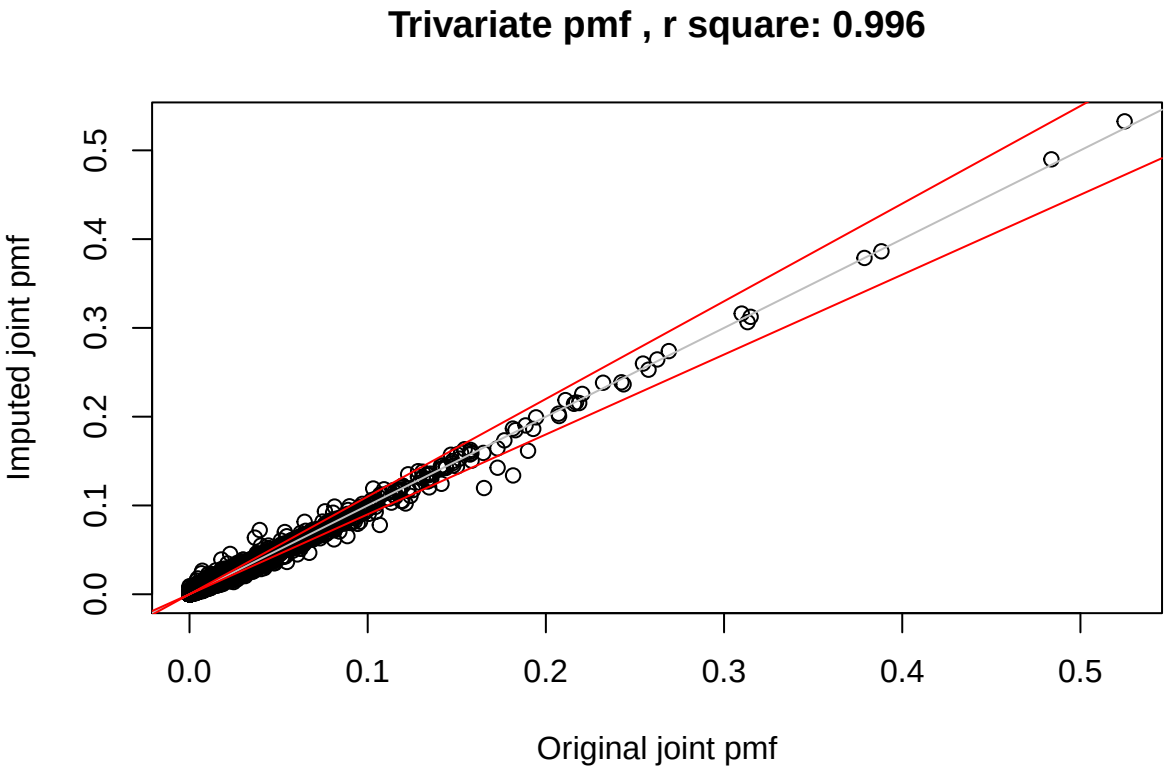

MICE: ENG
original_pmf
observed_pmf
imputed_pmf
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Category


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
0.02
0.04
0.06
0.08
0.10
1
2
3
4
5
6
7
8
9
10
12
14
16
Category


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


Bivariate pmf , r square: 0.999
Imputed joint pmf
0.0
0.2
0.4
0.6
0.8
Original joint pmf


Trivariate pmf , r square: 0.996
Imputed joint pmf
Original joint pmf
0.0
0.1
0.2
0.3
0.4
0.5