

# Testing different imputation methods on PUMS (MCAR) - Generative Adversarial Imputation Nets (GAIN)

```
# 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 scenario with 30% chance of missing: df_observed
missing_prob = 0.3
df_observed <- df
missing_col = c(1,3,7,9,10,11)
for (col in missing_col) {
  missing_ind <- rbernoulli(n,p = missing_prob)
  df_observed[missing_ind, col] <- NA
}
```

## Generative Adversarial Imputation Nets (GAIN)

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

```
# Load imputed dataset
d1 = read.csv('../..//GAIN/imputed_dataset/MCAR_PUMS01_30percent_1.csv', header = FALSE, sep = ',')
d2 = read.csv('../..//GAIN/imputed_dataset/MCAR_PUMS01_30percent_2.csv', header = FALSE, sep = ',')
d3 = read.csv('../..//GAIN/imputed_dataset/MCAR_PUMS01_30percent_3.csv', header = FALSE, sep = ',')
d4 = read.csv('../..//GAIN/imputed_dataset/MCAR_PUMS01_30percent_4.csv', header = FALSE, sep = ',')
d5 = read.csv('../..//GAIN/imputed_dataset/MCAR_PUMS01_30percent_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

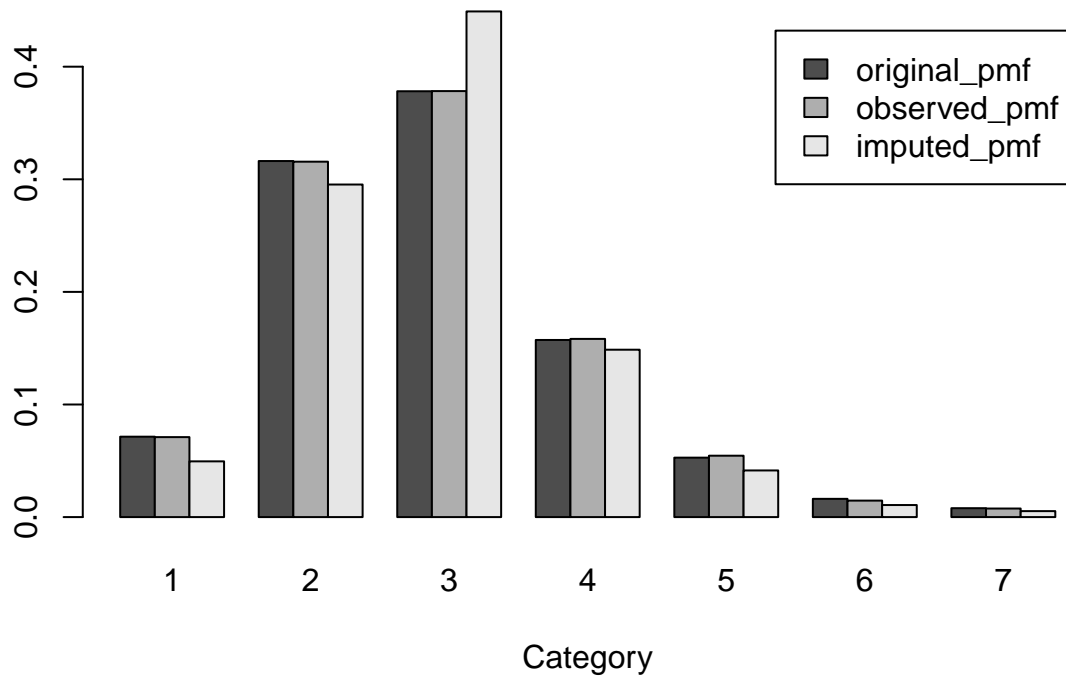
Assess bivariate joint distribution

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)
```

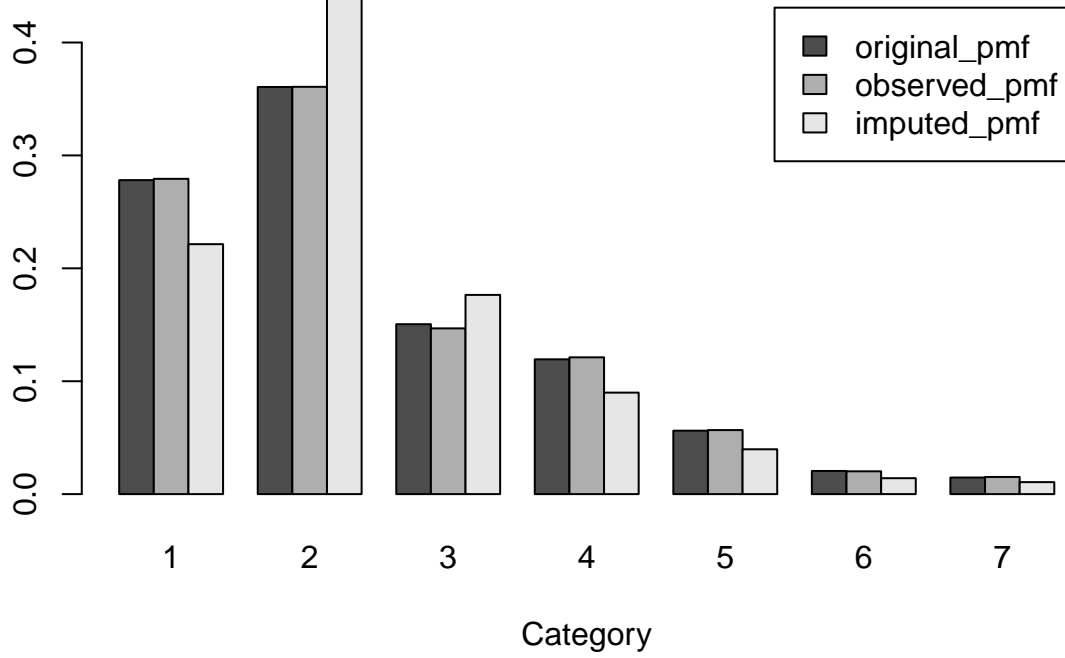
## MICE: VEH



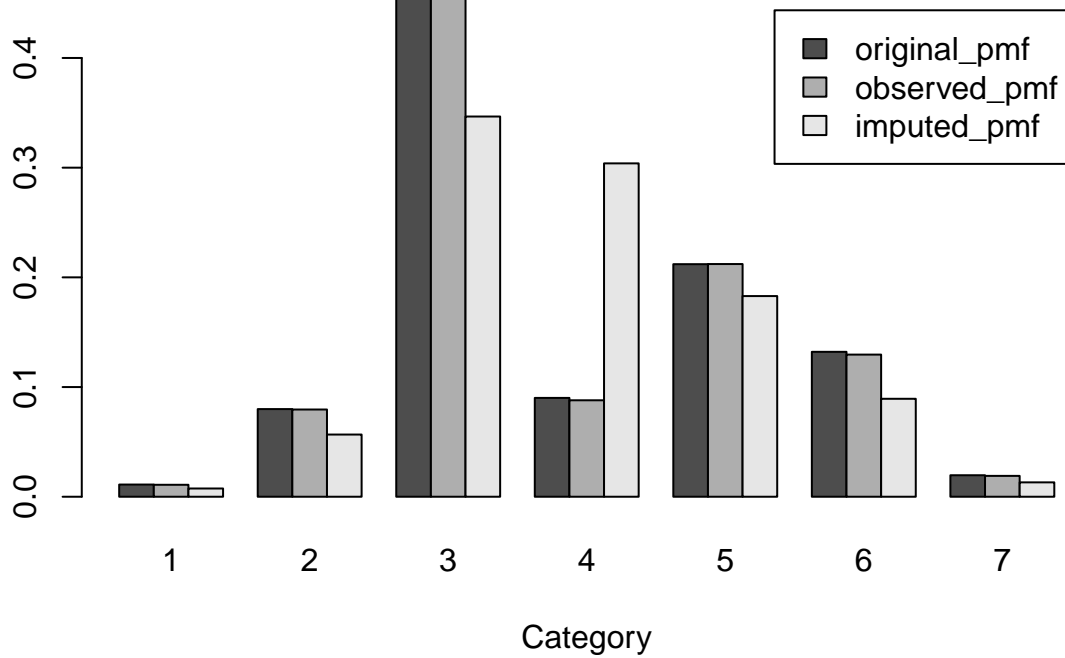
```
rmse = sqrt(sum((normalized_df[missing_matrix] - normalized_impute[missing_matrix])^2)/sum(missing_matrix == 1))
rmse
```

```
## [1] 0.2754284
```

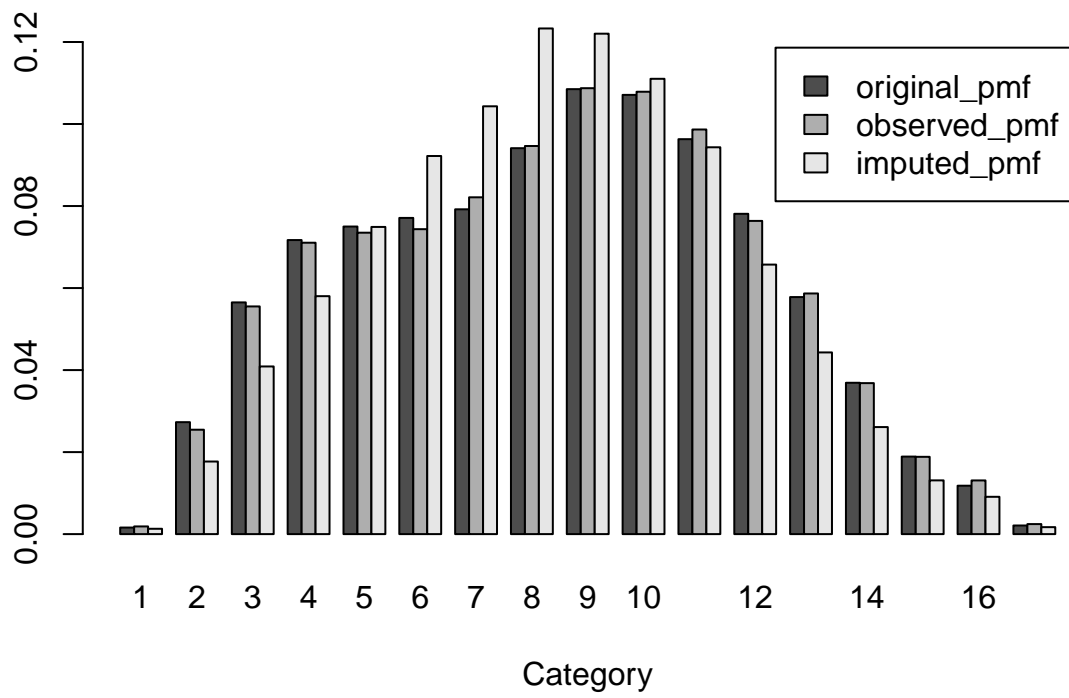
### MICE: NP



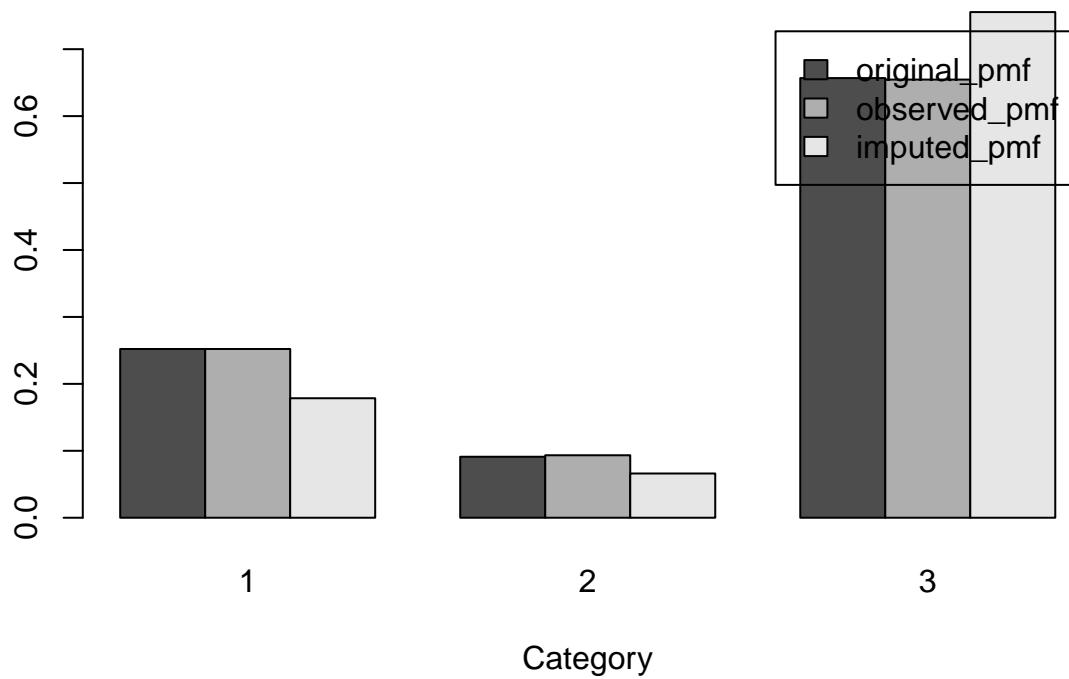
### MICE: SCHL



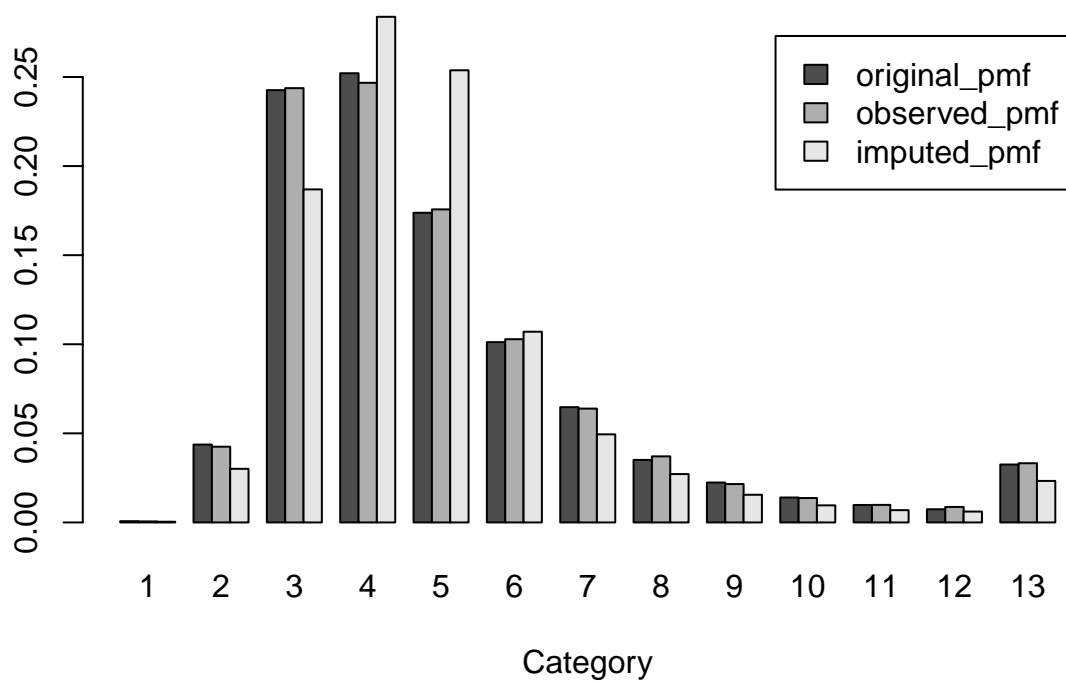
### MICE: AGEP



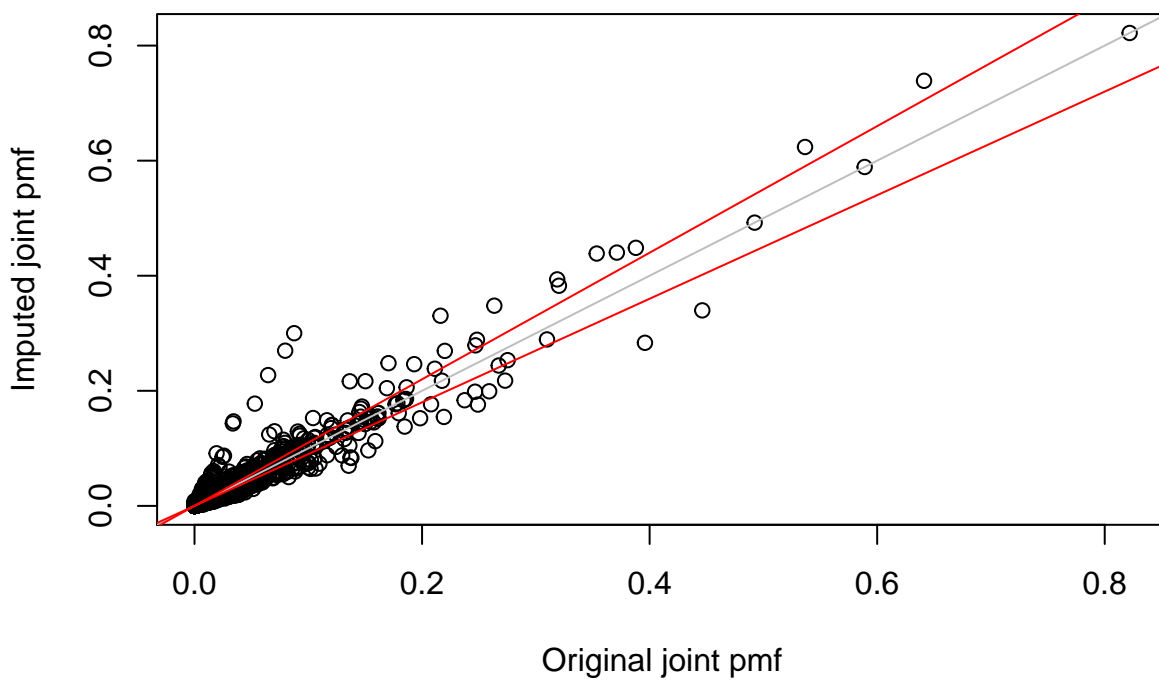
### MICE: WKL



### MICE: PINCP



### Bivariate pmf



Trivariate pmf

