Testing different imputation methods on PUMS (MCAR)

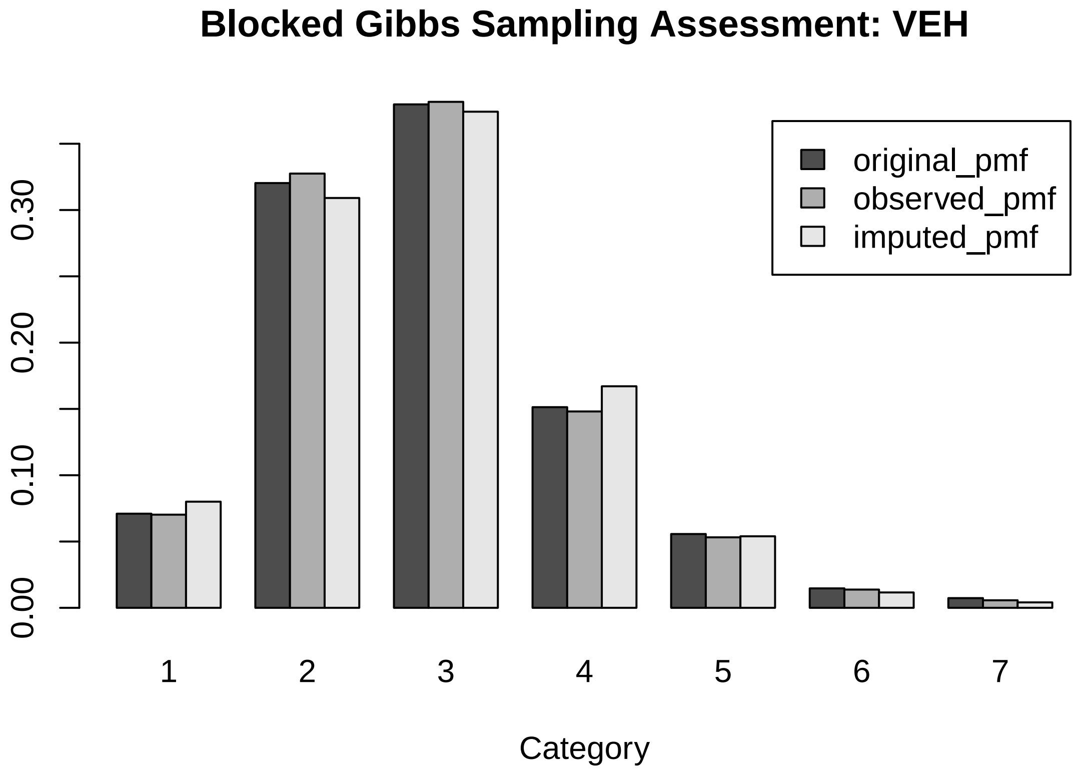
```r
# load dataset: df
load('../Datasets/ordinalPUMS.Rdata')

# take 2000 samples: df
set.seed(0)
n = 3000
sample <- sample(nrow(df), size = n)
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
}
```

Ordinal bayesian nonparametric model

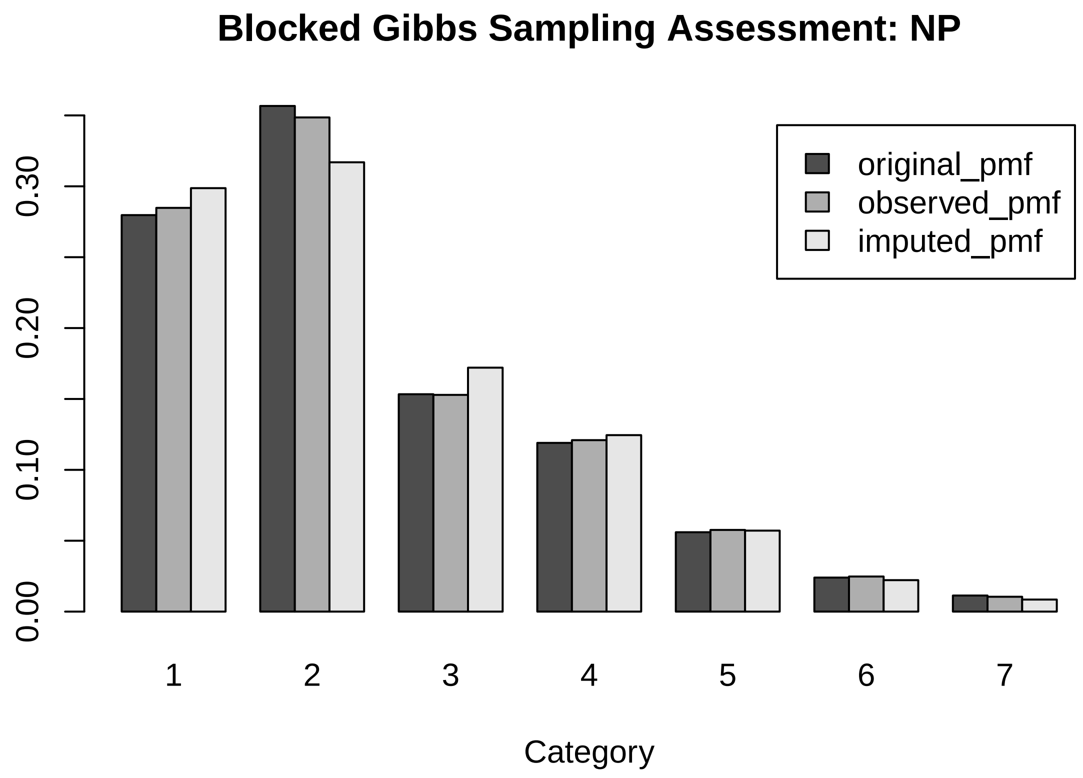
```r
source("../probitBayes.R")
N = 40
Mon = 300
B = 300
thin.int = 1
# function(y, N = 40, Mon = 2000, B = 300, thin.int = 5, seed = 0)
output_list <- probitBayesImputation(df_observed, N, Mon, B, thin.int)
```

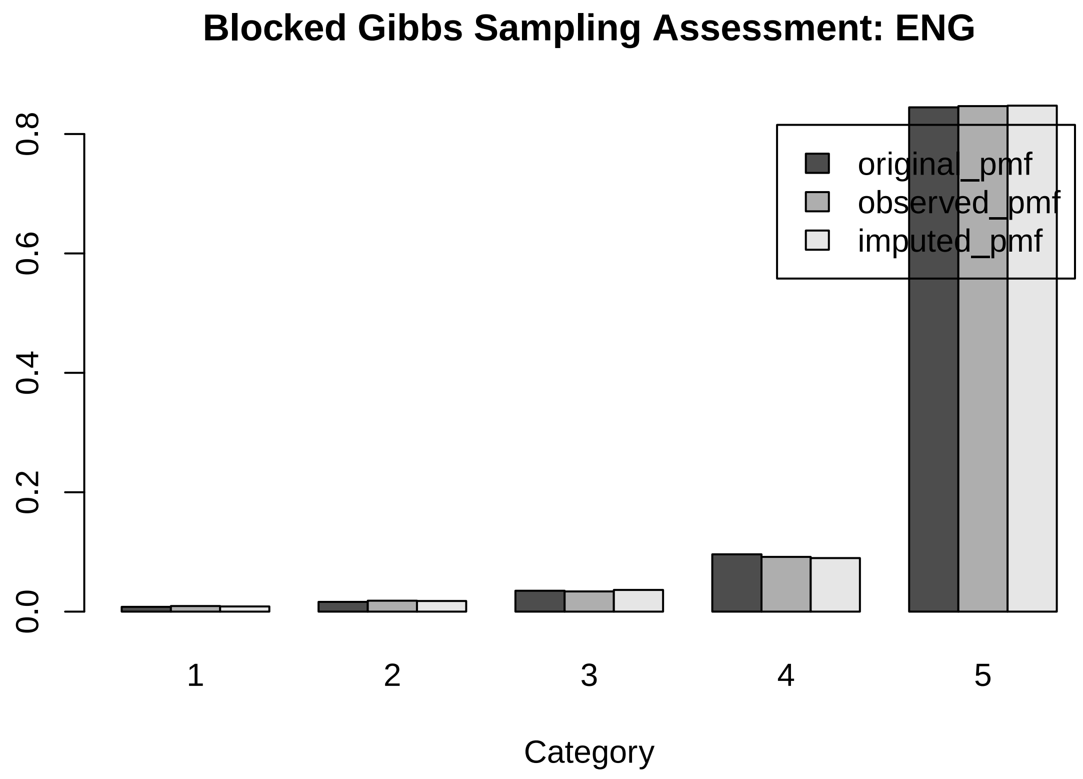
```r
sampled_y <- output_list[['sampled_y']]
sampled_z <- output_list[['sampled_z']]
```

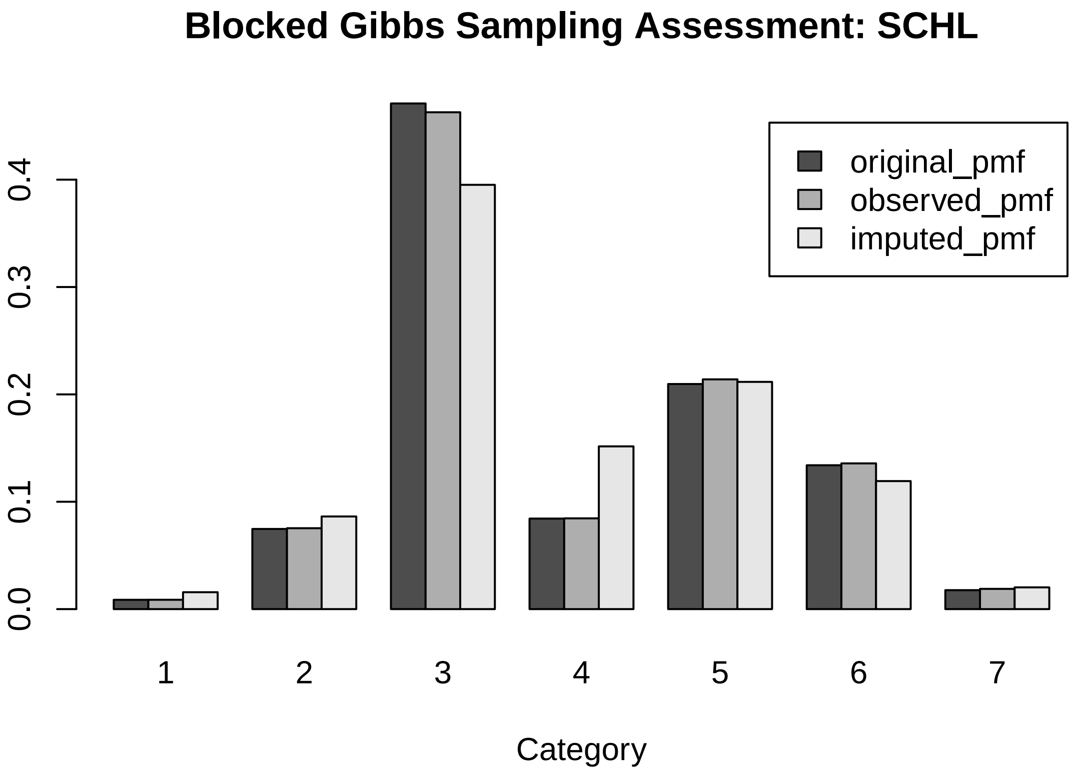
```r
for (var_index in c(1,3,5,7,9,11)) {
  y_original = df[,var_index]
  original_pmf = table(y_original)/length(y_original)

  # Observed distribution
  missing_indicator = is.na(df_observed)[,var_index]
  y_observed = y_original[!missing_indicator]
  observed_pmf = table(y_observed)/length(y_observed)

  # Extract variable from imputed data
  imputed_pmf = table(sampled_y[,,var_index])
  imputed_pmf = imputed_pmf/sum(imputed_pmf)

  results = rbind(original_pmf,observed_pmf,imputed_pmf)
  colnames(results)<- 1:dim(imputed_pmf)
  barplot(results, xlab = 'Category', beside = TRUE,
          legend = TRUE,
          main = paste('Blocked Gibbs Sampling Assessment:', colnames(df)[var_index]))
}
```



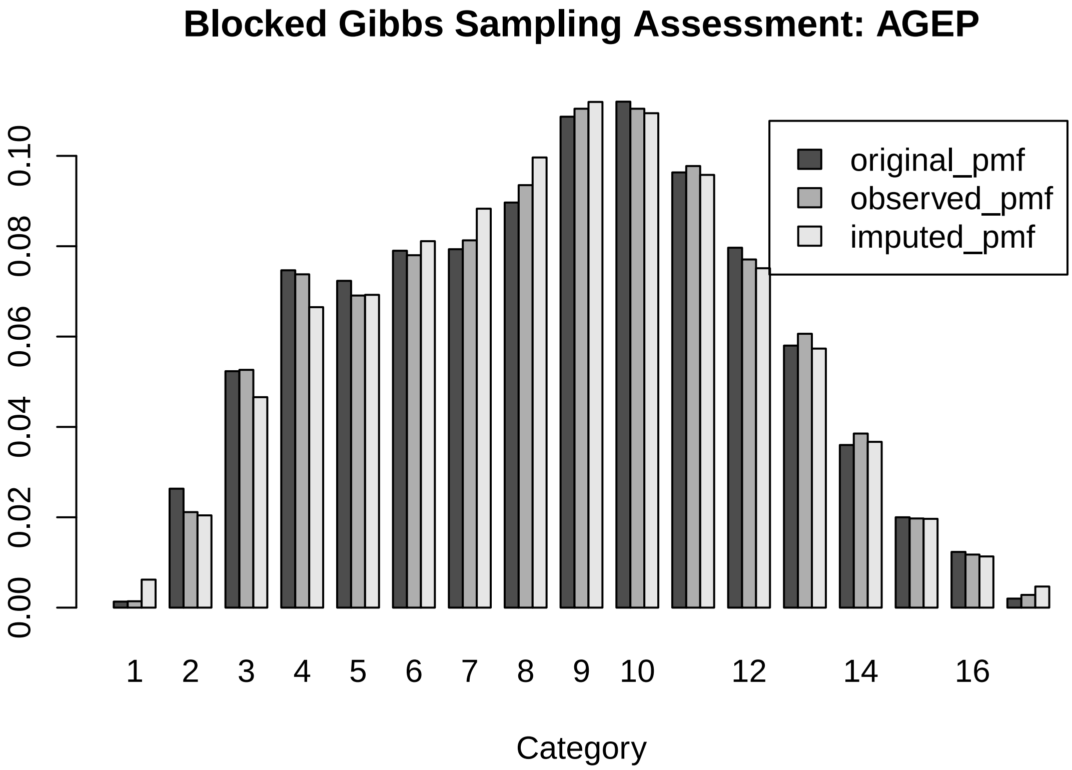
```
# trace plot
z.mcmc <- mcmc(sampled_z[,1,11], start=1)
plot(z.mcmc)
```


Blocked Gibbs Sampling Assessment: NP
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
Category
original_pmf
observed_pmf
imputed_pmf

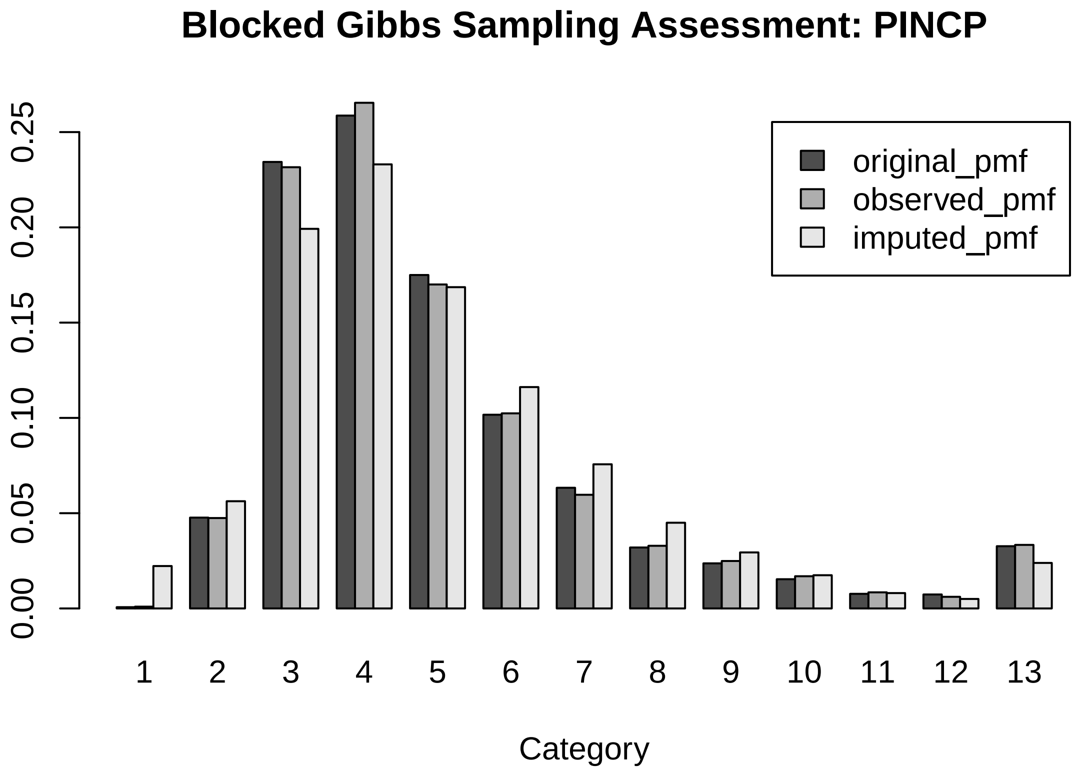

Blocked Gibbs Sampling Assessment: ENG
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Category
original_pmf
observed_pmf
imputed_pmf

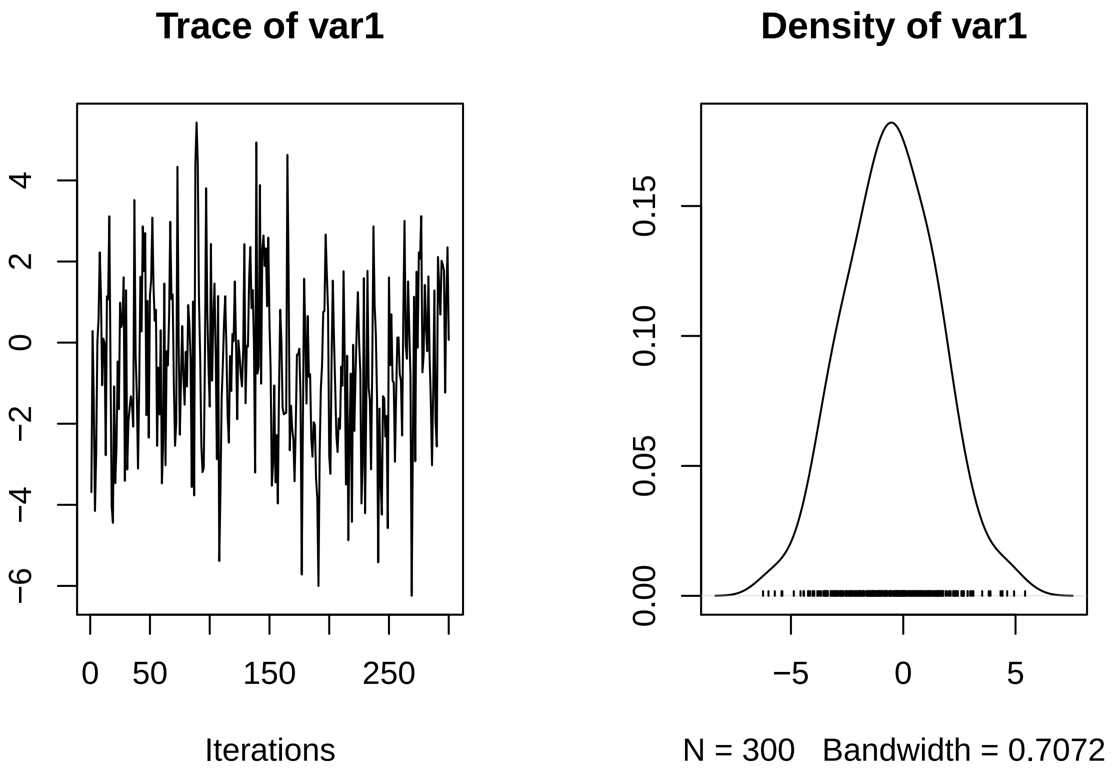


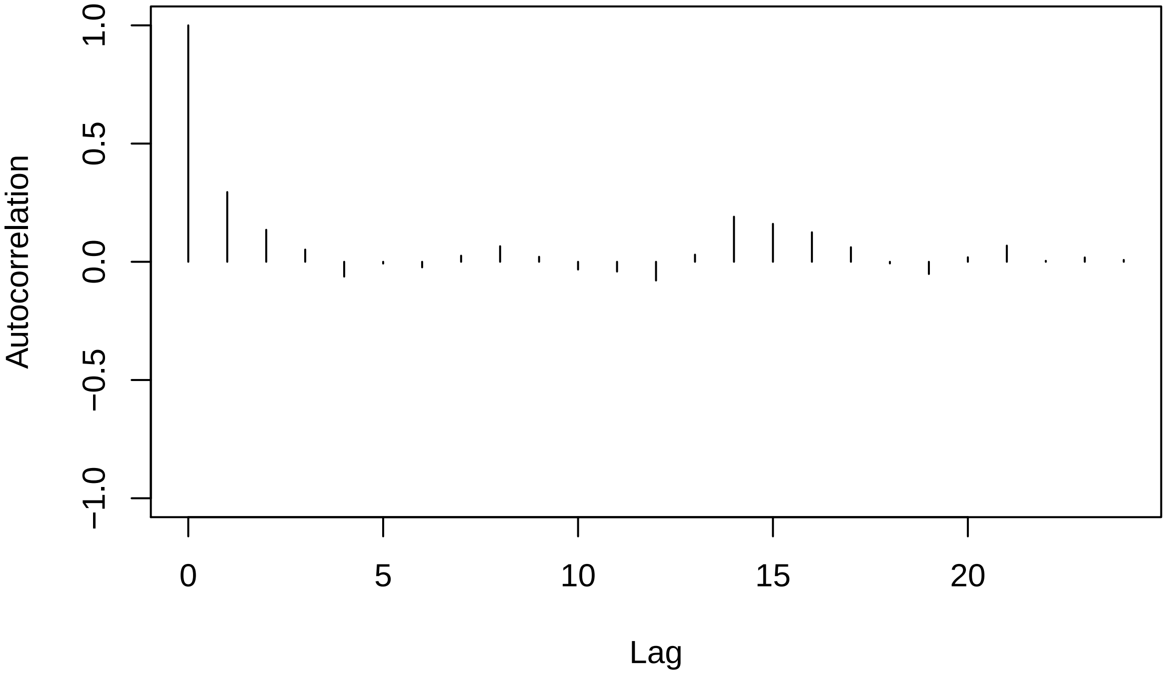
Blocked Gibbs Sampling Assessment: SCHL



Blocked Gibbs Sampling Assessment: AGEP


Blocked Gibbs Sampling Assessment: 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


Trace of var1
−6
−4
−2
0
2
4
0
50
150
250
Iterations
Density of var1
0.00
0.05
0.10
0.15
−5
0
5
N = 300   Bandwidth = 0.7072


Autocorrelation
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20
Lag