Package ‘BaBooN’

February 19, 2015

Version 0.1-6
Date 2011-02-24
Title Bayesian Bootstrap Predictive Mean Matching - Multiple and single imputation for discrete data
Author Florian Meinfelder <florian.meinfelder@uni-bamberg.de>
Maintainer Florian Meinfelder <florian.meinfelder@uni-bamberg.de>
Depends R (>= 2.12.0), MASS, nnet
Description The package contains two variants of Bayesian Bootstrap Predictive Mean Matching to multiply impute missing data. The first variant is a variable-by-variable imputation combining sequential regression and Predictive Mean Matching (PMM) that has been extended for unordered categorical data. The Bayesian Bootstrap allows for generating approximately proper multiple imputations. The second variant is also based on PMM, but the focus is on imputing several variables at the same time. The suggestion is to use this variant, if the missing-data pattern resembles a data fusion situation, or any other missing-by-design pattern, where several variables have identical missing-data patterns. Both variants can be run as 'single imputation’ versions, in case the analysis objective is of a purely descriptive nature.
License GPL (>= 2)
URL http://www.r-project.org
Repository CRAN
Date/Publication 2011-03-26 15:43:18
NeedsCompilation no

R topics documented:

BBPMM .............................................................. 2
BBPMM.row ............................................................ 4
MI.inference ............................................................ 6
Description

‘BBPMM’ performs single and multiple imputation (MI) of mixed-scale variables using a chained equations approach and (Bayesian Bootstrap) Predictive Mean Matching.

Usage

BBPMM(data, M=10, nIter=10, outfile=NULL, ignore=NULL, vartype=NULL, stepwise=TRUE, maxit=3, maxPerc = 0.98, verbose=TRUE, setSeed,...)

Arguments

data : A partially incomplete data frame or matrix.

M : Number of multiple imputations. If M=1, no Bayesian Bootstrap step is carried out. Default=10.

nIter : Number of iterations of the chained equations algorithm before the data set is stored as an 'imputed data set'. Default=10.

outfile : A character string that specifies the path and file name for the imputed data sets. If outfile=NULL (default), no data set is stored.

ignore : A character or numerical vector that specifies either column positions or variable names that are to be excluded from the imputation model and process, e.g. an ID variable. If ignore=NULL (default), all variables in data are used in the imputation model.

vartype : A character vector that flags the class of each variable in data (without the variables defined by the ignore argument), with either 'M' for metric-scale or 'C' for categorical. The default (NULL) takes over the classes of data. Overruling these classes can sometimes make sense: e.g. an ordinal-scale variable is originally classified as 'factor', but treating it as metric-scale variable within the imputation process might still be a better choice (considering the robust properties of PMM to model misspecification).

stepwise : Performs backwise selection for each imputation model based on the Schwarz (Bayes) Information criterion. Default=TRUE.

maxit : Imported argument from the nnet package that specifies the maximum number of iterations for the multinomial logit model estimation. Default=3.
maxPerc  The maximum percentage the mode category of a variable is allowed to have in order to try ‘regular’ imputation. If a variable is approximately Dirac distributed, i.e. if it has (almost) no variance, imputation is carried out by simple hot deck imputation. Default = 0.98.

verbose  The algorithm prints information on imputation and iteration numbers. Default=TRUE.

setSeed  Optional argument to fix the pseudo-random number generator in order to allow for reproducible results.

...  Further arguments passed to or from other functions.

Details

BBPMM is based on a chained equations approach that is using a Bayesian Bootstrap approach and Predictive Mean Matching variants for metric-scale, binary, and multi-categorical variables to generate multiple imputations. In order to emulate a monotone missing-data pattern as well as possible, variables are sorted by rate of missingness (in ascending order). If no complete variables exist, the least incomplete variable is imputed via hot-deck. The starting solution then builds the imputation model using the observed values of a particular y variable, and the corresponding observed or already imputed values of the x variables (i.e. all variables with fewer missing values than y). Due to the PMM element in the algorithm, auto-correlation of subsequent iterations is virtually zero. Therefore, a burn-in period is not required, and there is no need to administer ‘high’ values (> 20) to nIter either.

If M=1, no Bayesian Bootstrap step is carried out for the chained equations. Note that in this case the algorithm is still unlikely to converge to a stable solution, because of the Predictive Mean Matching step.

Value

impdata  The imputed data set, if M=1, or a list containing M imputed data sets.

misOverview  The percentage of missing values per incomplete variable.

indMatrix  A matrix with the same dimensions as data minus ignore containing flags for missing values.

M  Number of (multiple) imputations.

nIter  Number of iterations between two imputations.

References


See Also

BBPMM.row
Examples

```r
### sample data set with non-normal variables
n <- 50
x1 <- round(runif(n,0,5,3.5))
x2 <- as.factor(c(rep(1,10),rep(2,25),rep(3,15)))
x3 <- round(rnorm(n,0,3))
y1 <- round(x1-0.25*(x2==2)+0.5*x3+rnorm(n,0,1))
y1 <- ifelse(y1<1,1,y1)
y1 <- as.factor(ifelse(y1>4,5,y1))
y2 <- x1+rnorm(n,0,0.5)
y3 <- round(x3+rnorm(n,0,2))
data1 <- as.data.frame(cbind(x1,x2,x3,y1,y2,y3))
misrow1 <- sample(n,20)
misrow2 <- sample(n,15)
misrow3 <- sample(n,10)
is.na(data1[misrow1, 4]) <- TRUE
is.na(data1[misrow2, 5]) <- TRUE
is.na(data1[misrow2, 6]) <- TRUE

### imputation
imputed.data <- BBPMM(data1, nIter=5, M=5)
```

**BBPMM.row** *(Multiple) Imputation of variable vectors*

Description

‘BBPMM.row’ performs single and multiple imputation (MI) of metric scale variable vectors. For MI, parameter draws from a posterior distribution are replaced by a Bayesian Bootstrap step. Imputations are generated using Predictive Mean Matching (PMM) as described in Little (1988).

Usage

```r
BBPMM.row(misDataPat, blockImp=length(misDataPat$blocks),
M=10, outfile=NULL, manWeights=NULL, stepwise=TRUE, verbose=TRUE,
tol=0.25, setSeed=NULL, ...)
```

Arguments

- **misDataPat**: An object created by rowimpPrep that contains information on all identified missing-data patterns.
- **blockImp**: A scalar or vector containing the number(s) of the block(s) considered for imputation. Per default only the last block is imputed.
- **M**: Number of multiple imputations. If M=1, no Bayesian Bootstrap step is carried out.
- **outfile**: A character string that specifies the path and file name for the imputed data sets. If outfile=NULL (default), no data set is stored.
manWeights Optional argument containing manual (non-negative) weights for the PMM step. manWeights can either be a list containing a vector for each missingness pattern, or just a vector, if only one missingness pattern/block exists. In either case, the number of elements in the vector(s) must match the number of variables in the corresponding block. Note that the higher the weight the higher the importance of a good match for the corresponding variable’s predictive means.

stepwise Performs backwise selection for each imputation model based on the Schwarz (Bayes) Information criterion. Default=TRUE.

verbose The algorithm prints information on weighting matrices and imputation numbers. Default=TRUE.

tol Imported argument from function qr that specifies the tolerance level for linear dependencies among the complete variables and defaults to 0.25.

setSeed Optional argument to fix the pseudo-random number generator in order to allow for reproducible results.

... Further arguments passed to or from other functions.

Details

The simultaneous imputation of several variables is useful for missing-by-design patterns, such as data fusion or split questionnaire designs. The predictive means of the imputation variables are weighted by the inverse of the covariance matrix of the residuals from the regression of these variables on the complete variables. The intuitive idea behind is that distances between predictive means should be punished more severely, if the particular variable can be explained well by the (completely observed) imputation model variables. Through partialization and subsequent usage of the residuals the weight matrix is transformed into a diagonal matrix. The calculated weights can be adjusted by manual weights. Since the weight matrix is a Mahalanobis type of distance matrix, the weights are in the denominator and therefore the lower the weight, the higher the influence. As this is somewhat counterintuitive, the reciprocal of the manual weights is taken. Therefore, the higher the manual weight the higher in the influence of the corresponding variable’s predictor on the overall distance. The donor/recipient ID pairlist for each imputation and identified pattern ('block') is stored. In general, weightMatrix, model and pairlist are list objects named ‘M1’ to ‘M<block>’, and each in return is a list object named ‘block1’ to ‘block<length(blockImp)>’. model contains another list object with lm-objects for all variables in a particular block. Unlike BBPMM this algorithm is not based on sequential regression. Therefore, imputed variables are conditionally independent given the completely observed variables (of which at least one must exist).

Value

impdata A list containing M completed data sets.
weightMatrix A list containing weight matrices for all imputations and blocks.
model A list containing the lm-objects for all imputations and blocks.
pairlist A list containing the donor/recipient pairlist data frames for all imputations and blocks.
indMatrix A matrix with the same dimensions as data containing flags for missing values.
References


See Also

rowimpprep, BBPMM

Examples

```r
### sample data set with non-normal variables and a single
### missingness pattern
set.seed(1000)
n <- 50
x1 <- round(runif(n,0.5,3.5))
x2 <- as.factor(c(rep(1,10),rep(2,25),rep(3,15)))
x3 <- round(rnorm(n,0,3))
y1 <- round(x1-0.25*(x2==2)+0.5*x3+rnorm(n,0,1))
y1 <- ifelse(y1<1,1,y1)
y1 <- ifelse(y1>4,5,y1)
y2 <- y1+rnorm(n,0,0.5)
y3 <- round(x3+rnorm(n,0,2))
data <- as.data.frame(cbind(x1,x2,x3,y1,y2,y3))
misrow1 <- sample(n,20)
data[misrow1, c(4:6)] <- NA

### preparation step
impblock <- rowimpprep(data)

### imputation
imputed.data <- BBPMM.row(impblock, M=5)
```

---

**MI.inference**

### Multiple Imputation inference

**Description**

‘MI.inference’ applies Rubin’s combining rules to estimated quantities of interest that are based on multiply imputed data sets. The function requires as input two vectors of length M for the estimate and its variance.

**Usage**

```r
MI.inference(thetahat, varhat, thetahat, alpha=0.05)
```
Arguments

\texttt{thetahat} \hspace{1cm} A vector of length \( M \) containing estimates of the quantity of interest based on multiply imputed data sets.

\texttt{varhat.thetahat} \hspace{1cm} A vector of length \( M \) containing the corresponding variances of \texttt{thetahat}.

\texttt{alpha} \hspace{1cm} The significance level at which lower and upper bound are calculated. DEFAULT=0.05

Details

Multiple Imputation (Rubin, 1987) of missing data is a generally accepted way to get correct variance estimates for a particular quantity of interest in the presence of missing data. \texttt{MI.inference} estimates the \textit{within variance} \( W \) and \textit{between variance} \( B \), and combines them to the \textit{total variance} \( T \). Based on the output, further analysis figures, such as the \textit{fraction of missing information} can be calculated.

Value

\texttt{MI.Est} \hspace{1cm} A scalar containing the MI estimate of the quantity of interest (i.e. an estimator averaged over all \( M \) data sets).

\texttt{MI.Var} \hspace{1cm} The Multiple Imputation variance.

\texttt{CI.low} \hspace{1cm} The lower bound of the MI confidence interval.

\texttt{CI.up} \hspace{1cm} The upper bound of the MI confidence interval.

\texttt{BVar} \hspace{1cm} The estimated \textit{between variance}.

\texttt{WVar} \hspace{1cm} The estimated \textit{within variance}.

References


Examples

```r
### example 1
n <- 100
x1 <- round(runif(n,0.5,3.5))
x2 <- round(runif(n,0.5,4.5))
x3 <- runif(n,1,6)
y1 <- round(x10.5*2+0.5*x3+rnorm(n,0,1))
y1 <- ifelse(y1<2,2,y1)
y1 <- as.factor(ifelse(y1>4,5,y1))
y2 <- x3+rnorm(n,0,2)
y3 <- as.factor(ifelse(x2+rnorm(n,0,2)>2,1,0))

mis1 <- sample(100,20)
mis2 <- sample(100,30)
mis3 <- sample(100,25)
data1 <- data.frame("x1"=x1,"x2"=x2,"x3"=x3,"y1"=y1,"y2"=y2,"y3"=y3)
```
is.na(data1$y1[mis1]) <- TRUE
is.na(data1$y2[mis2]) <- TRUE
is.na(data1$y3[mis3]) <- TRUE
imputed.data <- BBPMM(data1, M=10, nIter=10)

MI.m.meany2.hat <- sapply(imputed.data$impdata,FUN=function(x) mean(x$y2))
MI.v.meany2.hat <- sapply(imputed.data$impdata,FUN=function(x) var(x$y2)/length(x$y2))

### MI inference
MI.y2 <- MI.inference(MI.m.meany2.hat, MI.v.meany2.hat, alpha=0.05)

MI.y2$MI.Est
MI.y2$MI.Var

### Not run:

---

### simple additional function to calculate coverages:    

coverage <- function(value, bounds) {
  ifelse(min(bounds) <= value && max(bounds) >= value, 1, 0)
}

### value                : true value            
### bounds               : vector with two elements (upper and lower bound of the CI)    

### sample size
n <- 100
### true value for the mean of y2
m.y2 <- 3.5
y2.cover <- vector(length=n)
set.seed(10000)

### 100 data generations
time1 <- Sys.time()
for (i in 1:100) {
  x1 <- round(runif(n,0.5,3.5))
  x2 <- round(runif(n,0.5,4.5))
  x3 <- runif(n,1,6)
  y1 <- round(x1-0.25*x2+0.5*x3+rnorm(n,0,1))
  y1 <- ifelse(y1<2,y1)
  y1 <- as.factor(ifelse(y1>4,5,y1))
  y2 <- x3+rnorm(n,0,2)
  y3 <- as.factor(ifelse(x2+rnorm(n,0,2)>2,1,0))
  mis1 <- sample(n,20)
  mis2 <- sample(n,30)
  mis3 <- sample(n,25)
  data1 <- data.frame(x1=x1,x2=x2,x3=x3,y1=y1,y2=y2,y3=y3)
  is.na(data1$y1[mis1]) <- TRUE
  is.na(data1$y2[mis2]) <- TRUE
  is.na(data1$y3[mis3]) <- TRUE
rowimpPrep

Missing-data pattern identifier

Description

‘rowimpPrep’ identifies all missingness patterns within an incomplete data set. Running rowimpPrep is a prerequisite for BBPM. row.

Usage

rowimpPrep(data, ID=NULL, verbose=TRUE)

Arguments

data

Either a data frame or matrix with missing values.

ID

A numeric or character string vector indicating the column positions or names of the ID variable (if two data sets were stacked that have a joint subset of variables). The first element refers to the ‘donor ID’, the second element refers to the ‘recipient ID’. This distinction is only of relevance, if the data set is 'L-shaped', i.e. if the data contains only one missing-data pattern (where incomplete cases are 'recipients'). If ID has only one element, the function assumes that the identifier variables of the two data sets are packed into a single variable. Default=NULL is used, if no ID variable is specified.

verbose

Prints information on identified missing-data patterns. Default=TRUE.

Details

rowimpPrep identifies all patterns, and allows to decide, whether to impute all missing-data patterns with BBPM. row or just some of them. This comes in handy if variables that were assumed to be completely observed have missing values. These variables are then likely to define an unexpected ‘block’ of their own. Of course, BBPM. row can be used to impute missing data that are not missing-by-design as well, but BBPM would probably be the better option. Note that all variables listed in
compNames are used for the imputation model in `BBPmm`. row, i.e. completely observed variables (ID variables aside) which are not to be used in the imputation model, have to be removed from the data set beforehand.

### Value

- **data**: The original data set minus the ID variable(s).
- **key**: The ID variable(s) from the original data set.
- **blocks**: A list containing the column positions of all identified missing-data patterns.
- **blockNames**: A list containing the variable names corresponding to object `blocks`.
- **compNames**: A character vector containing the variable names of the (completely observed) imputation model variables.

### Examples

```r
### sample data set with non-normal variables and a single missingness pattern
set.seed(1000)
n <- 50
x1 <- round(runif(n, 0.5, 3.5))
x2 <- as.factor(c(rep(1, 10), rep(2, 25), rep(3, 15)))
x3 <- round(rnorm(n, 0, 3))
y1 <- round(x1 - 0.25 * (x2 == 2) + 0.5 * x3 + rnorm(n, 0, 1))
y1 <- ifelse(y1 < 1, 1, y1)
y1 <- ifelse(y1 > 4.5, 4.5, y1)
y2 <- y1 + rnorm(n, 0, 0.5)
y3 <- round(x3 + rnorm(n, 0, 2))
data1 <- as.data.frame(cbind(x1, x2, x3, y1, y2, y3))
misrow1 <- sample(n, 20)
is.na(data1[misrow1, c(4:6)]) <- TRUE

### preparation step
impblock <- rowImpPrep(data1)
impblock$blockNames
```

### Description

Returns some information about the incomplete data set and the imputation process.
Usage

## S3 method for class 'imp'
summary(object,...)

Arguments

object Either with BBPMM or BBPMM.row generated object.
... Arguments to be passed to or from other functions.

Details

Returns information about the percentage of missing data as well as about the imputation variant, the number of (multiple) imputations and the number of iterations between two imputations.

Author(s)

Florian Meinfelder

See Also

BBPMM, BBPMM.row

Examples

### sample data set with non-normal variables and two different
### missingness patterns
n <- 50
x1 <- round(runif(n,0.5,3.5))
x2 <- as.factor(c(rep(1,10),rep(2,25),rep(3,15)))
x3 <- round(rnorm(n,0,3))
y1 <- round(x1-0.25*(x2==2)+0.5*x3+rnorm(n,0,1))
y1 <- ifelse(y1<1,1,y1)
y1 <- as.factor(ifelse(y1>4.5,y1))
y2 <- x1+rnorm(n,0,0.5)
y3 <- round(x3+rnorm(n,0,2))
data1 <- as.data.frame(cbind(x1,x2,x3,y1,y2,y3))
misrow1 <- sample(n,20)
misrow2 <- sample(n,15)
misrow3 <- sample(n,10)
is.na(data1[misrow1, 4]) <- TRUE
is.na(data1[misrow2, 5]) <- TRUE
is.na(data1[misrow2, 6]) <- TRUE

### imputation
imputed.data <- BBPMM(data1, nIter=5, M=5)
summary(imputed.data)
summary.impprep

Summary method for objects of class ‘impprep’

Description

Returns an overview of missing-data patterns, in particular of missing-by-design patterns.

Usage

```r
## S3 method for class 'impprep'
summary(object, nNames=10L,...)
```

Arguments

- `object`: An object generated by `rowimpprep`.
- `nNames`: Number of variable names per block to be printed (Default = 10).
- `...`: Arguments to be passed to or from other functions.

Details

Returns the number of identified missing-data patterns, the first `nNames` variable names per block and the names of the completely observed variables.

Author(s)

Florian Meinfelder

See Also

`rowimpprep`

Examples

```r
### sample data set with non-normal variables and a single
### missingness pattern
set.seed(1000)
n <- 50
x1 <- round(runif(n,0.5,3.5))
x2 <- as.factor(c(rep(1,10),rep(2,25),rep(3,15)))
x3 <- round(rnorm(n,0,3))
y1 <- round(x1-0.25*(x2==2)+0.5*x3+rnorm(n,0,1))
y1 <- ifelse(y1<1,1,y1)
y1 <- ifelse(y1>4.5,y1)
y2 <- y1+rnorm(n,0,0.5)
y3 <- round(x3+rnorm(n,0,2))
data1 <- as.data.frame(cbind(x1,x2,x3,y1,y2,y3))
misrow1 <- sample(n,20)
```
is.na(data1[misrow1, c(4:6)]) <- TRUE

### preparation step
impblock <- rowimpPrep(data1)

summary(impblock)
Index

*Topic datagen
    BBPMM, 2
    BBPMM.row, 4

*Topic robust
    BBPMM, 2
    BBPMM.row, 4

BBPMM, 2, 6, 9, 11
BBPMM.row, 3, 4, 9, 11

MI.inference, 6

rowimpPrep, 6, 9, 12

summary.imp, 10
summary.impprep, 12