Package 'FHDI'

January 20, 2025

Version 1.4.1

Date 2020-08-21

Title Fractional Hot Deck and Fully Efficient Fractional Imputation

Author Inho Cho [aut, cre],
Jaekwang Kim [aut],
Jongho Im [aut],
Yicheng Yang [aut]

Maintainer Inho Cho <icho@iastate.edu>

Depends R (>= 3.4.0)

Description Impute general multivariate missing data with the fractional hot deck imputation based on Jaekwang Kim (2011) <doi:10.1093/biomet/asq073>.

License GPL (>= 2)

URL https://www.r-project.org,
https://sites.google.com/view/jaekwangkim/software

BugReports https://sites.google.com/site/ichoddcse2017/home/type-of-trainings/r-package-fhdi

NeedsCompilation yes

Repository CRAN

Date/Publication 2020-09-22 06:30:30 UTC

Contents

FHDI-package *Fractional Hot Deck Imputation*

Description

Perform fractional hot deck imputation or perform fully efficient fractional imputation. This package is partially supported by the NSF grant CSSI 1931380.

Details

FHDI_Driver(daty, datr=NULL, datz=NULL, s_op_imputation="FEFI", i_op_SIS=0, s_op_SIS="global", s_op_cellmake="knn", top_corr_var=100, i_op_variance=1, M=5, k=5, w=NULL, id=NULL, s_op_merge="fixed", categorical=NULL)

Author(s)

Author: Inho Cho [aut, cre], Jaekwang Kim [aut], Jongho Im [aut], Yicheng Yang [aut] <icho@iastate.edu>

References

Im, J., Cho, I.H. and Kim, J.K. (2018). FHDI: An **R** Package for Fractional Hot-Deck Imputation. *The R Journal*. 10(1), pp. 140-154; Im, J., Kim, J.K. and Fuller, W.A. (2015). Two-phase sampling approach to fractional hot deck imputation, *Proceeding of the Survey Research Methods Section*, Americal Statistical Association, Seattle, WA.

See Also

FHDI_CellMake and FHDI_CellProb

Examples

```
### Toy Example ###
# y : multi-variate vector
# r : indicator corresponding to missingness in y

set.seed(1345)
n=100
rho=0.5
e1=rnorm(n,0,1)
e2=rnorm(n,0,1)
e3=rgamma(n,1,1)
e4=rnorm(n,0,sd=sqrt(3/2))

y1=1+e1
y2=2+rho*e1+sqrt(1-rho^2)*e2
y3=y1+e3
y4=-1+0.5*y3+e4

r1=rbinom(n,1,prob=0.6)
```

```
r2=rbinom(n,1,prob=0.7)
r3=rbinom(n,1,prob=0.8)
r4=rbinom(n,1,prob=0.9)

y1[r1==0]=NA
y2[r2==0]=NA
y3[r3==0]=NA
y4[r4==0]=NA

daty=cbind(y1,y2,y3,y4)

result_FEFI=FHDI_Driver(daty, s_op_imputation="FEFI", k=3)
result_FHDI=FHDI_Driver(daty, s_op_imputation="FHDI", M=5, k=3)
result_FHDI_merging=FHDI_Driver(daty, s_op_imputation="FHDI", s_op_cellmake="merging", M=5, k=3)
FEFI_SIS=FHDI_Driver(daty, i_op_SIS=2, s_op_SIS="intersection", k=3)

names(result_FEFI)
names(result_FHDI)
names(result_FHDI_merging)
names(FEFI_SIS)
```

`FHDI_CellMake` *Imputation Cell Creation*

Description

Perform a categorization procedure on the continuous raw data and then create imputation cells through a built-in merge algorithm. This package is partially supported by the NSF grant CSSI 1931380.

Usage

```
FHDI_CellMake(daty, datr=NULL, k=5, w=NULL, id=NULL, i_op_SIS=0,
              s_op_SIS="global", s_op_cellmake="knn", top_corr_var=100,
              s_op_merge="fixed", categorical=NULL)
```

Arguments

`daty`	raw data matrix (nrow_y, ncol_y) containing missing values. Each row must have at least one observed value, and no completely missing (blank) rows are allowed.
`datr`	response indicator matrix with the same dimensions as daty. Each response is recorded with 0 for missing value and 1 for observed value. If NULL, automatically filled with 1 or 0 according to daty.
`k`	the number of total categories per variable. Default = 5. The maximum is 35 since 9 integers (1-9) and 26 alphabet letters (a-z) are used. When a scalar value is given, all variables will have the same number of categories, while when a vector is given, i.e., k(ncol_y), each variable may have a different number of categories.

w	sampling weight for each row of daty. Default = 1.0 if NULL. When a scalar value is given, all rows will have the same weight, while when a vector is given, i.e., w(nrow_y), each row may have a different sampling weight.
id	index for each row. Default = 1:nrow_y if NULL.
i_op_SIS	(FHDI Version >= 1.4) the desired number of reduced variables after the sure independence screening per each missing pattern. Default = 0 means no variable reduction and uses all variables. Range must be <= ncol_y.
s_op_SIS	(FHDI Version >= 1.4) "intersection" for sure independence screening with an intersection of simple correlation, "union" for sure independence screening with a union of simple correlation, or default = "global" for sure independence screening with a global ranking of simple correlation.
s_op_cellmake	(FHDI Version >= 1.4) option for different methods of cell construction with deficient donors. "merging" for adopting cell collapsing and merging or default = "knn" for adopting the k-nearest neighbors in terms of the Euclidean distance.
top_corr_var	(FHDI Version >= 1.4) the number of top-ranking variables based on simple correlation with default 100.
s_op_merge	option for random cell make. Default = "fixed" using the same seed number; "rand" using a purely random seed number.
categorical	(FHDI Version >1.3) index vector indicating non-collapsible categorical variables. Default = zero vector of size ncol_y. For instance, when categorical=c(1,0), the first variable (i.e., 1st column) is considered strictly to be non-collapsible and categorical, thus no automatic cell-collapse will take place while the second variable (i.e., 2nd column) is considered as a continuous or collapsible categorical variable.

Details

This function creates imputation cells with the given number of categories k. If the input value k is given a scalar, the same number of categories is applied to all variables for initial discretization. Imputation cells are created to assign at least two donors on each missing unit. The donors have the same cell values with the observed parts of the missing unit. From version >= 1.4, the sure independence screening method (Fan and Lv 2008) has been incorporated to perform variable reduction for each missing pattern, which is useful for high dimensional (i.e., big-p) datasets. Besides, we provide an alternative method using k-nearest neighbors to speed up the convergence of cell construction with deficient donors, which is useful for big-p datasets.

Value

data	matrix of raw data (nrow_y, ncol_y) attached with id and weights, w.
cell	categorized matrix of y. A real value is categorized into 1~k categories with 0 meaning missing value.
cell.resp	unique patterns of respondents (donors) that are fully observed.
cell.non.resp	unique patterns of nonrespondents that have at least one missing item.
w	reprint of the sampling weights "w" initially defined by the user.
s_op_merge	reprint of the option "s_op_merge" initially defined by the user.

i_op_SIS	reprint of the option "i_op_SIS" initially defined by the user.
s_op_SIS	reprint of the option "s_op_SIS" initially defined by the user.
s_op_cellmake	reprint of the option "s_op_cellmake" initially defined by the user.
top_corr_var	reprint of the option "top_corr_var" initially defined by the user.
cell.selected	list of selected variables for each unique pattern of nonrespondents that have at least one missing item. Note that all the observed variables of a unique missing pattern will be selected if i_op_SIS is greater than the number of observed variables of the unique missing pattern; otherwise, the deficient selected variables are replaced by 0s.

Author(s)

Dr. Cho, In Ho (maintainer) <icho@iastate.edu> Dr. Kim, Jae Kwang <jkim@iastate.edu> Dr. Im, Jong Ho <ijh38@yonsei.ac.kr> Yicheng Yang, Graduate Research Assistant

References

Im, J., Cho, I.H. and Kim, J.K. (2018). FHDI: An **R** Package for Fractional Hot-Deck Imputation. *The R Journal*. 10(1), pp. 140-154; Im, J., Kim, J.K. and Fuller, W.A. (2015). Two-phase sampling approach to fractional hot deck imputation, *Proceeding of the Survey Research Methods Section*, Americal Statistical Association, Seattle, WA.

Examples

```
### Toy Example ###
# y : multi-variate vector
# r : indicator corresponding to missingness in y

set.seed(1345)
n=100
rho=0.5
e1=rnorm(n,0,1)
e2=rnorm(n,0,1)
e3=rgamma(n,1,1)
e4=rnorm(n,0,sd=sqrt(3/2))

y1=1+e1
y2=2+rho*e1+sqrt(1-rho^2)*e2
y3=y1+e3
y4=-1+0.5*y3+e4

r1=rbinom(n,1,prob=0.6)
r2=rbinom(n,1,prob=0.7)
r3=rbinom(n,1,prob=0.8)
r4=rbinom(n,1,prob=0.9)

y1[r1==0]=NA
y2[r2==0]=NA
y3[r3==0]=NA
y4[r4==0]=NA
```

```
daty=cbind(y1,y2,y3,y4)

result_CM=FHDI_CellMake(daty, k=3, s_op_merge="fixed")
result_CM_merging=FHDI_CellMake(daty, k=3, s_op_cellmake ="merging", s_op_merge="fixed")
result_CM_reduced=FHDI_CellMake(daty, k=3, i_op_SIS=2, s_op_SIS="intersection", s_op_merge="fixed")
names(result_CM)
names(result_CM_merging)
names(result_CM_reduced)
```

`FHDI_CellProb` *Joint Cell Probabilities for Incomplete Categorical Data*

Description

Calculate the joint cell probabilities for multivariate missing data using the expectation-maximization (EM) algorithm. This package is partially supported by the NSF grant CSSI 1931380.

Usage

```
FHDI_CellProb(datz, w=NULL, id=NULL)
```

Arguments

`datz`	multivariate incomplete categorical data prepared by cell collapsing and merging algorithm.
`w`	samping weight. Default = 1.0 if NULL. a scalar or w(nrow_y).
`id`	index for each unit. Default = 1:nrow_y if NULL.

Details

The joint cell probabilities are estimated using EM by weighting method. The algorithm computes the maximum likelihood estimates of the joint cell probabilities under missing at random assumption. Note that the variable reduction (ver. >= 1.4) with sure independence screening method is not applicable to a separate CellProb task. The input incomplete categorical data should be generated by cell make with the cell collapsing and merging algorithm.

Value

`cellpr`	table of the joint cell probability. The name of a cell is linked to the user-defined categories in "k": e.g., name "325" denotes 3rd, 2nd, 5th categories for three variables, respectively, whereas "a1c" denotes 10th, 1st, 12th categories.
`w`	reprint of the sampling weights "w" initially defined by the user.

Author(s)

Dr. Cho, In Ho (maintainer) <icho@iastate.edu> Dr. Kim, Jae Kwang <jkim@iastate.edu> Dr. Im, Jong Ho <ijh38@yonsei.ac.kr> Yicheng Yang, Graduate Research Assistant

References

Im, J., Cho, I.H. and Kim, J.K. (2018). FHDI: An **R** Package for Fractional Hot-Deck Imputation. *The R Journal*. 10(1), pp. 140-154; Im, J., Kim, J.K. and Fuller, W.A. (2015). Two-phase sampling approach to fractional hot deck imputation, *Proceeding of the Survey Research Methods Section*, Americal Statistical Association, Seattle, WA.; Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Assocation* **85**, 765-769.

Examples

```
### Toy Example ###
# y : trivariate variables
# r : indicator corresponding to missingness in y

set.seed(1345)
n=100
rho=0.5
e1=rnorm(n,0,1)
e2=rnorm(n,0,1)
e3=rgamma(n,1,1)
e4=rnorm(n,0,sd=sqrt(3/2))

y1=1+e1
y2=2+rho*e1+sqrt(1-rho^2)*e2
y3=y1+e3
y4=-1+0.5*y3+e4

r1=rbinom(n,1,prob=0.6)
r2=rbinom(n,1,prob=0.7)
r3=rbinom(n,1,prob=0.8)
r4=rbinom(n,1,prob=0.9)

y1[r1==0]=NA
y2[r2==0]=NA
y3[r3==0]=NA
y4[r4==0]=NA

daty=cbind(y1,y2,y3,y4)

result_CM=FHDI_CellMake(daty, k=5, s_op_cellmake="merging", s_op_merge="fixed")
datz=result_CM$cell
result_CP=FHDI_CellProb(datz)
names(result_CP)
```

FHDI_Driver *Main Driver of Fractional Hot Deck Imputation*

Description

Fully efficient fractional imputation (FEFI) or fractional hot deck imputation (FHDI) is implemented to fill in missing values in incomplete data. This package is partially supported by the NSF grant CSSI 1931380.

Usage

```
FHDI_Driver(daty, datr=NULL, datz=NULL, s_op_imputation="FEFI",
 i_op_SIS=0, s_op_SIS="global", s_op_cellmake="knn", top_corr_var=100,
 i_op_variance=1, M=5, k=5, w=NULL, id=NULL,
 s_op_merge="fixed", categorical=NULL)
```

Arguments

daty	raw data matrix (nrow_y, ncol_y) containing missing values. Each row must have at least one observed value, and no completely missing (blank) rows are allowed.
datr	response indicator matrix with the same dimensions as daty. Each response is recorded with 0 for missing value and 1 for observed value. If NULL, automatically filled with 1 or 0 according to daty.
datz	imputation cell matrix. If daty is a set of continuous data, datz can be obtained using FHDI_CellMake.
s_op_imputation	"FEFI" for fully efficient fractional imputation or "FHDI" for fractional hot deck imputation.
i_op_SIS	(FHDI Version >= 1.4) the desired number of reduced variables after the sure independence screening per each missing pattern. Default = 0 means no variable reduction and uses all variables. Range must be <= ncol_y.
s_op_SIS	(FHDI Version >= 1.4) "intersection" for sure independence screening with an intersection of simple correlation, "union" for sure independence screening with a union of simple correlation, or default = "global" for sure independence screening with a global ranking of simple correlation.
s_op_cellmake	(FHDI Version >= 1.4) option for different methods of cell construction with deficient donors. "merging" for adopting cell collapsing and merging or default = "knn" for adopting the k-nearest neighbors in terms of the Euclidean distance.
top_corr_var	(FHDI Version >= 1.4) the number of top-ranking variables based on simple correlation with default 100.
i_op_variance	1: perform Jackknife variance estimation; 0: no variance estimation.
M	the number of donors for FHDI with default 5.
k	the number of total categories per variable. Default = 5. The maximum is 35 since 9 integers (1-9) and 26 alphabet letters (a-z) are used. When a scalar value is given, all variables will have the same number of categories, while when a vector is given, i.e., k(ncol_y), each variable may have a different number of categories.

w	sampling weight for each row of daty. Default = 1.0 if NULL. When a scalar value is given, all rows will have the same weight, while when a vector is given, i.e., w(nrow_y), each row may have a different sampling weight.
id	index for each row. Default = 1:nrow_y if NULL.
s_op_merge	option for random cell make. Default = "fixed" using the same seed number; "rand" using a purely random seed number.
categorical	(FHDI Version >1.3) index vector indicating non-collapsible categorical variables. Default = zero vector of size ncol_y. For instance, when categorical=c(1,0), the first variable (i.e., 1st column) is considered strictly to be non-collapsible and categorical, thus no automatic cell-collapse will take place while the second variable (i.e., 2nd column) is considered as a continuous or collapsible categorical variable.

Details

In the FEFI method, all possible donors are assigned to each missing unit with the FEFI fractional weights. In the FHDI method, M (>1) donors are selected with the probability proportional to the FEFI fractional weights. Thus, the imputed values have equal fractional weights in general.

The jackknife replicated weights are produced as the default output. The replicated weights are presented by the product of replicated sampling weights and replicated fractional weights. Thus, the replicated weights can be directly used to compute the variance estimate of the estimators. From version >= 1.4, the sure independence screening method (Fan and Lv 2008) has been incorporated to perform variable reduction for each missing pattern, which is useful for high dimensional (i.e., big-p) datasets. Besides, we provide an alternative method using k-nearest neighbors to speed up the convergence of cell construction with deficient donors, which is useful for big-p datasets.

Value

fimp.data	imputation results with fractional weights in the form of a matrix consisting of ID, donor id (FID), weight (WGT), fractional weight (FWGT), and fractionally imputed data.
simp.data	imputed data in the format of single imputation. The same shape as daty.
imp.mean	the mean estimates of each variable (first row) and the estimated standard error of each variable (second row). If input argument "i_op_variance=0" then this output is not produced.
rep.weight	replication fractional weights for variance estimation. If input argument "i_op_variance=0" then this output is not produced.
M	reprint of the number of donors M for FHDI defined by the user.
s_op_imputation	reprint of the option "s_op_imputation" initially defined by the user.
i_op_merge	reprint of the option "i_op_merge" initially defined by the user.
i_op_SIS	reprint of the option "i_op_SIS" initially defined by the user.
s_op_SIS	reprint of the option "s_op_SIS" initially defined by the user.
s_op_cellmake	reprint of the option "s_op_cellmake" initially defined by the user.
top_corr_var	reprint of the option "top_corr_var" initially defined by the user.

Author(s)

Dr. Cho, In Ho (maintainer) <icho@iastate.edu> Dr. Kim, Jae Kwang <jkim@iastate.edu> Dr. Im, Jong Ho <ijh38@yonsei.ac.kr> Yicheng Yang, Graduate Research Assistant

References

Im, J., Cho, I.H. and Kim, J.K. (2018). FHDI: An **R** Package for Fractional Hot-Deck Imputation. *The R Journal*. 10(1), pp. 140-154; Im, J., Kim, J.K. and Fuller, W.A. (2015). Two-phase sampling approach to fractional hot deck imputation, *Proceeding of the Survey Research Methods Section*, Americal Statistical Association, Seattle, WA.

Examples

```
### Toy Example ###
# y : multi-variate vector
# r : indicator corresponding to missingness in y

set.seed(1345)
n=100
rho=0.5
e1=rnorm(n,0,1)
e2=rnorm(n,0,1)
e3=rgamma(n,1,1)
e4=rnorm(n,0,sd=sqrt(3/2))

y1=1+e1
y2=2+rho*e1+sqrt(1-rho^2)*e2
y3=y1+e3
y4=-1+0.5*y3+e4

r1=rbinom(n,1,prob=0.6)
r2=rbinom(n,1,prob=0.7)
r3=rbinom(n,1,prob=0.8)
r4=rbinom(n,1,prob=0.9)

y1[r1==0]=NA
y2[r2==0]=NA
y3[r3==0]=NA
y4[r4==0]=NA

daty=cbind(y1,y2,y3,y4)

result_FEFI=FHDI_Driver(daty, s_op_imputation="FEFI", k=3)
result_FHDI=FHDI_Driver(daty, s_op_imputation="FHDI", M=5, k=3)
result_FHDI_merging=FHDI_Driver(daty, s_op_imputation="FHDI", s_op_cellmake="merging", M=5, k=3)
FEFI_SIS=FHDI_Driver(daty, i_op_SIS=2, s_op_SIS="intersection", k=3)

names(result_FEFI)
names(result_FHDI)
names(result_FHDI_merging)
names(FEFI_SIS)
```

Index