Package 'statMatchLCM'

October 31, 2026

Type Package

Title Statistical Matching Using Latent Class Models

Version 2.1

Description Provides a revised framework for statistical matching of categorical data based on latent class models, including multiple imputation, predictive modeling, nearest-neighbour hot-deck matching, and probabilistic assignment of categorical target variables. It allows researchers to perform data integration when no unique identifiers are available by modeling the joint distribution of variables through latent categorical structures. The package supports estimation of latent class models, probabilistic matching between donor and recipient data sets, and generation of synthetic linked data under uncertainty. It is particularly useful in survey research and data fusion applications where combining information from multiple sources is required while preserving statistical properties and accounting for measurement error and missing data mechanisms.
The latent class matching methods are described by Lewaa, Hafez and Ismail (2023) <doi:10.18576/jsap/120123>.
(2.0->2.1): in smc2_step2() Gower distance is replaced by Euclidean distance in hot deck method.

License GPL-3

Encoding UTF-8

RoxygenNote 7.3.3

Imports nnet, StatMatch

Suggests NPBayesImputeCat

Depends R (>= 4.1.0)

LazyData true

NeedsCompilation no

Author Alicja Wolny-Dominiak [aut, cre],
Israa Lewaaelhamd [aut],
Mohammed Ali Ismail [aut]

Maintainer Alicja Wolny-Dominiak <woali@ue.katowice.pl>

Repository CRAN

Date/Publication 2026-10-31 11:00:02 UTC

Contents

datA . . . 2
datABC_to_SM . . . 3
datAB_to_SM . . . 3
datB . . . 4
datC . . . 5
fact_to_num . . . 5
num_to_fact . . . 6
sma1_step2 . . . 7
sma2_step2 . . . 8
sma3_step2 . . . 9
sma_step1 . . . 11
smc1_step2 . . . 12
smc2_step2 . . . 14
smc3_step2 . . . 15
smc_step1 . . . 17
sm_quality . . . 18

Index **20**

`datA` *Dataset datA*

Description

A simple dataset with categorical variables.

Usage

```
datA
```

Format

A data frame with 20 observations and 2 variables:

gender Category "F", "M"

color Color category "blue", "green", "red"

Source

Simulated data

datABC_to_SM *Create stacked A/B/C dataset for MDA*

Description

Creates a joint data set with missing Y1/Z1, required for mass data combination.

Usage

```
datABC_to_SM(datA, datB, datC, Y, Z)
```

Arguments

datA	data.frame A
datB	data.frame B
datC	data.frame C
Y	the target variable in datA
Z	the target variable in datB

Value

data.frame of class() "datABC_SM" with harmonized structure

Examples

```
data(datA)
data(datB)
data(datC)
datABC_to_SM(datA, datB, datC, color, targetB)
```

datAB_to_SM *Create stacked A/B dataset for MDC*

Description

Creates a joint data set with missing Y1/Z1, required for mass data combination.

Usage

```
datAB_to_SM(datA, datB, Y, Z)
```

Arguments

datA	data.frame A
datB	data.frame B
Y	the target variable in datA
Z	the target variable in datB

Value

data.frame of class() "datAB_SM" with harmonized structure

Examples

```
data(datA)
data(datB)
datAB_to_SM(datA, datB, color, targetB)
```

---

datB *Dataset datB*

Description

A simple dataset with categorical variables.

Usage

```
datB
```

Format

A data frame with 25 observations and 2 variables:

gender Category "F", "M"

targetB Letter category "d", "e", "f"

Source

Simulated data

datC *Dataset datC*

Description

A simple dataset with categorical variables.

Usage

```
datC
```

Format

A data frame with 30 observations and 3 variables:

gender Category "F", "M"

color Color category "blue", "green", "red"

targetB Letter category "d", "e", "f"

Source

Simulated data

fact_to_num *Convert factor/character variables to numeric with mapping*

Description

Converts all factor or character columns in a data.frame to numeric codes and stores the mapping tables.

Usage

```
fact_to_num(df)
```

Arguments

df	data.frame with factor or character columns

Value

A list with:

data data.frame with numeric-coded variables

levels list of factor levels

tables list of mapping tables (factor → numeric)

Examples

```
data(datA)
fact_to_num(datA)
```

`num_to_fact` *Convert numeric codes back to factor*

Description

Restores a factor variable from numeric codes using a mapping table created by `fact_to_num()`.

Usage

```
num_to_fact(x, table)
```

Arguments

`x`	numeric vector
`table`	data.frame with columns level_fact and level_num

Value

A factor vector:

levels Defined by `table$level_num`.

labels Defined by `table$level_fact`.

Examples

```
# Convert factors to numeric codes and restore the original labels
data(datA)

converted <- fact_to_num(datA)

color_restored <- num_to_fact(
  x = converted$data$color,
  table = converted$tables$color
)

color_restored
levels(color_restored)
```

`sma1_step2` *SMA1 – Nearest-neighbour hot deck on shared variables*

Description

Second step of the SMA procedure using nearest-neighbour hot deck matching on shared variables (Y1, Z1) to fuse datasets while preserving their joint distribution.

Usage

```
sma1_step2(step1)
```

Arguments

`step1`	list returned by the SMA step 1 procedure

Value

A list with:

datA_fused_2 The final fused dataset A.

out_nndd NNDD matching results.

match.vars Matching variables in the nearest-neighbour distance matching procedure.

References

- Lewaa, I., Hafez, M. S., Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)

  data(datC)
```

```
  datABC <- datABC_to_SM(datA, datB, datC, "color", "targetB")

  # call DPMPM (reduced settings for speed)
  output_list_AII <- NPBayesImputeCat::DPMPM_nozeros_imp(
    X = datABC,
    nrun = 50,
    burn = 10,
    thin = 10,
    K = 20,
    aalpha = 0.25,
    balpha = 0.25,
    m = 2,
    seed = 1234,
    silent = TRUE
  )

  step_first_aii <- sma_step1(datABC, output_list_AII)
  set.seed(1234)
  step_second_aii1 <- sma1_step2(step_first_aii)

  result_aii1 <- step_second_aii1$datA_fused_2
  head(result_aii1)
}
```

`sma2_step2` *SMA2 – Hot deck on fitted multinomial probabilities*

Description

Second step of SMA using multinomial models and nearest-neighbour hot deck on fitted probabilities to improve data fusion accuracy.

Usage

```
sma2_step2(step1)
```

Arguments

`step1`	list returned by the SMA step 1 procedure

Value

A list with:

datA_fused_2 A data frame containing the final fused (imputed) version of dataset A.

out_nnd A list with results from the nearest-neighbour distance matching procedure.

match_vars Matching variables in the nearest-neighbour distance matching procedure.

modA A multinomial logistic model obtained for dataset datA_imp.

modC A multinomial logistic model obtained for dataset datC.

References

- Lewaa, I., Hafez, M. S., Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)
  data(datC)

  datABC <- datABC_to_SM(datA, datB, datC, "color", "targetB")

    output_list_AII <- NPBayesImputeCat::DPMPM_nozeros_imp(
    X = datABC,
    nrun = 50,
    burn = 10,
    thin = 10,
    K = 20,
    aalpha = 0.25,
    balpha = 0.25,
    m = 2,
    seed = 1234,
    silent = TRUE
  )

  step_first_aii <- sma_step1(datABC, output_list_AII)
  set.seed(1234)
  step_second_aii2 <- sma2_step2(step_first_aii)

  result_aii2 <- step_second_aii2$datA_fused_2
  head(result_aii2)
}
```

`sma3_step2` *SMA3 – Multinomial simulation approach*

Description

Second step of SMA is to generate the target variable using multinomial probability distributions derived from a donor-based model to achieve statistically coherent data fusion

Usage

```
sma3_step2(step1)
```

Arguments

step1	list returned by the SMA step 1 procedure

Value

A list with:

datA_fused_2 A data frame containing the final fused version of dataset A.

modB A multinomial logistic model obtained for dataset datB_imp.

zero_one A binary matrix (one-hot encoded) sampled from prob_Z_all, with categories in rows and observations in columns; each column contains a single 1 indicating the selected category of Z.

References

- Lewaa, I., Hafez, M. S., Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)
  data(datC)

  datABC <- datABC_to_SM(datA, datB, datC, "color", "targetB")

  # call DPMPM (reduced settings for speed)
  output_list_AII <- NPBayesImputeCat::DPMPM_nozeros_imp(
```

```
    X = datABC,
    nrun = 50,
    burn = 10,
    thin = 10,
    K = 20,
    aalpha = 0.25,
    balpha = 0.25,
    m = 2,
    seed = 1234,
    silent = TRUE
  )

  step_first_aii <- sma_step1(datABC, output_list_AII)
  set.seed(1234)
  step_second_aii3 <- sma3_step2(step_first_aii)

  result_aii3 <- step_second_aii3$datA_fused_2
  head(result_aii3)
}
```

`sma_step1` *SMA step 1: selection of best imputed dataset*

Description

Selects the imputed dataset minimizing the Hellinger distance between the reference distribution of Z from dataset B and the synthetic distribution of Z from dataset A, in a three-sample statistical matching framework (A, B, C).

Usage

```
sma_step1(datABC, output_ll)
```

Arguments

`datABC`	data.frame of class datABC_SM returned by `datABC_to_SM`
`output_ll`	list with imputed datasets (impdata)

Value

A list with imputed datasets:

datABC_imp The full imputed dataset combining A, B, and C.

datA_imp Subset of the imputed data corresponding to dataset A.

datB_imp Subset corresponding to dataset B.

References

- Lewaa, I., Hafez, M. S., and Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications and Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)
  data(datC)

  datABC <- datABC_to_SM(datA, datB, datC, "color", "targetB")

  # call DPMPM
  output_list_AII <- NPBayesImputeCat::DPMPM_nozeros_imp(
    X = datABC, nrun = 500, burn = 50, thin = 50,
    K = 80, aalpha = 0.25, balpha = 0.25,
    m = 2, seed = 1234, silent = FALSE
  )

  step_first_aii <- sma_step1(datABC, output_list_AII)
  names(step_first_aii)
}
```

`smc1_step2` *SMC1 - Nearest-neighbour hot deck on original variables*

Description

Performs nearest-neighbour hot deck imputation on the selected imputed datasets by matching recipient units in dataset A with donor units in dataset B based on common variables. The procedure transfers the target variable from the nearest donor to construct a statistically coherent fused dataset.

Usage

```
smc1_step2(step1)
```

Arguments

step1	output from smc_step1()

Value

A list with:

datA_fused_2 The final fused dataset A after step 2 of the procedure.

out_nndd A list containing nearest-neighbour distance matching results.

match_vars Matching variables in the nearest-neighbour distance matching procedure.

References

- Lewaa, I., Hafez, M. S., and Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications and Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)

  datAB <- datAB_to_SM(datA, datB, "color", "targetB")

  # call DPMPM (reduced settings for speed)
  output_list <- NPBayesImputeCat::DPMPM_nozeros_imp(
    X = datAB,
    nrun = 50,
    burn = 10,
    thin = 10,
    K = 20,
    aalpha = 0.25,
    balpha = 0.25,
    m = 2,
    seed = 1234,
    silent = TRUE
  )

  step_first <- smc_step1(datAB, output_list)
```

```
  set.seed(1234)
  step_second <- smc1_step2(step_first)

  result <- step_second$datA_fused_2
  result
}
```

`smc2_step2` *SMC2 – Hot deck on fitted multinomial probabilities*

Description

Implements a model-based hot deck data fusion approach by estimating multinomial models on both datasets and matching observations based on their fitted probability distributions

Usage

```
smc2_step2(step1)
```

Arguments

`step1`	output from `smc_step1()`

Value

A list with:

datA_fused_2 A data frame containing the final fused (imputed) version of dataset A.

out_nndd A list with results from the nearest-neighbour distance matching procedure.

match_vars Matching variables in the nearest-neighbour distance matching procedure.

model_form A formula object used to fit the models.

fitA A data frame of fitted values obtained from the model for dataset A.

fitB A data frame of fitted values obtained from the model for dataset B.

References

- Lewaa, I., Hafez, M. S., and Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications and Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)

  datAB <- datAB_to_SM(datA, datB, "color", "targetB")

  # call DPMPM (reduced settings for speed)
  output_list <- NPBayesImputeCat::DPMPM_nozeros_imp(
    X = datAB,
    nrun = 50,
    burn = 10,
    thin = 10,
    K = 20,
    aalpha = 0.25,
    balpha = 0.25,
    m = 2,
    seed = 1234,
    silent = TRUE
  )

  step_first <- smc_step1(datAB, output_list)
  set.seed(1234)
  step_second <- smc2_step2(step_first)

  result <- step_second$datA_fused_2
  result
}
```

`smc3_step2` *SMC3 – Multinomial simulation approach*

Description

Generates the target variable using multinomial simulation based on fitted probability distributions to preserve uncertainty and variability

Usage

```
smc3_step2(step1)
```

Arguments

`step1`	output from `smc_step1()`

Value

A list with:

datA_fused_2 A data frame containing the final fused (imputed) version of dataset A.

fitA A data frame of fitted values obtained from the model for dataset A.

model_form A formula object used to fit the models.

modA A multinomial regression model on datA_imp.

References

- Lewaa, I., Hafez, M. S., and Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications and Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)

  datAB <- datAB_to_SM(datA, datB, "color", "targetB")

  # call DPMPM (reduced settings for speed)
  output_list <- NPBayesImputeCat::DPMPM_nozeros_imp(
    X = datAB,
    nrun = 50,
    burn = 10,
    thin = 10,
    K = 20,
    aalpha = 0.25,
    balpha = 0.25,
    m = 2,
    seed = 1234,
    silent = TRUE
  )

  step_first <- smc_step1(datAB, output_list)
  set.seed(1234)
  step_second <- smc3_step2(step_first)
```

```
  result <- step_second$datA_fused_2
  result
}
```

`smc_step1` *SMC step 1: selection of best imputed dataset*

Description

Identifies and selects the imputed dataset that minimizes the Hellinger distance between the reference and synthetic distributions, thereby achieving the highest level of distributional similarity and improving the statistical consistency of the imputation.

Usage

```
smc_step1(datAB, output_ll)
```

Arguments

`datAB`	data.frame of the class datAB_SM return by the function `datAB_to_SM`
`output_ll`	list with imputed datasets (impdata)

Value

A list with imputed datasets:

datAB_imp The full imputed dataset obtained after combining A and B.

datA_imp Subset of `datAB_imp` corresponding to dataset A.

datB_imp Subset of `datAB_imp` corresponding to dataset B.

References

- Lewaa, I., Hafez, M. S., Ismail, M. A. (2023). *Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies*. Journal of Statistics Applications Probability, 12(1), 247–265.
- D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley and Sons.
- D'Orazio, M., Di~Zio, M., and Scanu, M. (2019). Auxiliary variable selection in a statistical matching problem.
- Zhang, L.-C., and Chambers, R. Analysis of integrated data. CRC/Chapman and Hall, pp. 101–120.
- Conti, P. L., Marella, D., and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111(516), 1715–1725.

Examples

```
if (requireNamespace("NPBayesImputeCat", quietly = TRUE)) {
  data(datA)
  data(datB)

  datAB <- datAB_to_SM(datA, datB, "color", "targetB")

  # call DPMPM (reduced settings for speed)
  output_list <- NPBayesImputeCat::DPMPM_nozeros_imp(
    X = datAB,
    nrun = 50,
    burn = 10,
    thin = 10,
    K = 20,
    aalpha = 0.25,
    balpha = 0.25,
    m = 2,
    seed = 1234,
    silent = TRUE
  )

  step_first <- smc_step1(datAB, output_list)
  str(step_first$datA_imp)
  str(step_first$datB_imp)
}
```

`sm_quality` *Quality assessment of synthetic Z1*

Description

Evaluates the quality of the synthetic target variable by computing the Hellinger distance between the reference and synthetic distributions. This measure quantifies the degree of similarity between the two distributions, providing an assessment of the accuracy and coherence of the data fusion process

Usage

```
sm_quality(Z_obs, Z_fused)
```

Arguments

`Z_obs`	the variable Z_obs from datB
`Z_fused`	the variable Z_fused from step2$datA_fused_2

Value

A list with:

heli_latent The complete list returned by `StatMatch::comp.prop()`, containing `meas`, `chi.sq`, and `p.exp`. The Hellinger distance is available as `heli_latent$meas["Hell"]`.

ref_distr A numeric vector or table representing the reference (original) distribution.

synt_distr A numeric vector or table representing the synthetic (generated) distribution.

Index

∗ **datasets**
datA, 2
datB, 4
datC, 5

datA, 2
datAB_to_SM, 3
datABC_to_SM, 3
datB, 4
datC, 5

fact_to_num, 5

num_to_fact, 6

sm_quality, 18
sma1_step2, 7
sma2_step2, 8
sma3_step2, 9
sma_step1, 11
smc1_step2, 12
smc2_step2, 14
smc3_step2, 15
smc_step1, 17