Package 'MMDai'

May 7, 2026

Type Package

Title Multivariate Multinomial Distribution Approximation and Imputation for Incomplete Categorical Data

Version 2.0.0

Author Chaojie Wang

Maintainer Chaojie Wang <wang910930@163.com>

Description A method to impute the missingness in categorical data. Details see the paper <doi:10.4310/SII.2020.v13.n1.a2>.

License GPL (>= 2)

Encoding UTF-8

LazyData true

Depends stats

RoxygenNote 7.0.2

NeedsCompilation no

Repository CRAN

Date/Publication 2020-05-02 04:10:02 UTC

Contents

`GenerateData` *Generate random dataset*

Description

This function is used to generate random datasets following mixture of product multinomial distribution

Usage

```
GenerateData(
  n,
  p,
  d,
  k = 3,
  theta = rdirichlet(1, rep(10, k)),
  psi = InitialPsi(p, d, k)
)
```

Arguments

`n`	- number of samples
`p`	- number of variables
`d`	- a vector which denotes the number of categories for each variable. It could be distinct among variables.
`k`	- number of latent classes
`theta`	- probability for latent class
`psi`	- probability for specific category

Value

data - generated random dataset, a matrix with n rows and p columns.

Examples

```
# dimension parameters
n<-200; p<-5; d<-rep(2,p);
# generate complete data
Complete<-GenerateData(n, p, d, k = 3)
```

`Imputation` *Imputation*

Description

This function is used to perform multiple imputation for missing data given the joint distribution.

Usage

```
Imputation(data, theta, psi)
```

Arguments

`data`	- incomplete dataset
`theta`	- vector of probability for each component
`psi`	- specific probability for each variable in each component

Value

ImputedData - dataset has been imputated.

`InitialPsi` *initial psi*

Description

This function creates a psi list in that each component has equal weight

Usage

```
InitialPsi(p, d, k)
```

Arguments

`p`	- number of variables
`d`	- a vector which denotes the number of categories for each variable. It could be distinct among variables.
`k`	- number of components

Value

psi - a list in that each component has equal weight

kIdentifier *Identify the suitable number of components k*

Description

This function is used to find the suitable number of components k.

Usage

```
kIdentifier(data, d, TT = 1000, alpha = 0.25)
```

Arguments

data	- data in matrix formation with n rows and p columns
d	- number of categories for each variable
TT	- number of iterations in Gibbs sampler, default value is 1000. T should be an even number for 'burn-in'.
alpha	- hyperparameter that could be regarded as the pseudo-count of the number of samples in the new component

Value

k_est - posterior estimation of k

k_track - track of k in the iteration process

Examples

```
# dimension parameters
n<-200; p<-5; d<-rep(2,p);
# generate complete data
Complete<-GenerateData(n, p, d, k = 3)
# mask percentage of data at MCAR
Incomplete<-Complete
Incomplete[sample(1:n*p,0.2*n*p,replace = FALSE)]<-NA
# k identify
K<-kIdentifier(data = Incomplete, d, TT = 10)
```

MovieRate *Real application dataset*

Description

This is a real application dataset. The source of original data is the ratings dataset in (Harper and Konstan (2016) <DOI:10.1145/2827872>). This dataset is used to evaluate the performance of package in real applications.

Author(s)

Chaojie Wang

ParEst *Estimate theta and psi in multinomial mixture model*

Description

This function is used to estimate theta and psi in multinomial mixture model given the number of components k.

Usage

```
ParEst(data, d, k, TT = 1000)
```

Arguments

data	- data in matrix formation with n rows and p columns
d	- number of categories for each variable
k	- number of components
TT	- number of iterations in Gibbs sampler, default value is 1000. T should be an even number for 'burn-in'.

Value

theta - vector of probability for each component

psi - specific probability for each variable in each component

Examples

```
# dimension parameters
n<-200; p<-5; d<-rep(2,p);
# generate complete data
Complete<-GenerateData(n, p, d, k = 3)
# mask percentage of data at MCAR
Incomplete<-Complete
Incomplete[sample(1:n*p,0.2*n*p,replace = FALSE)]<-NA
# k identify
K<-kIdentifier(data = Incomplete, d, TT = 10)
Par<-ParEst(data = Incomplete, d, k = K$k_est, TT = 10)
```

`rdirichlet` *Estimate theta and psi in multinomial mixture model*

Description

This function is generate random sample from Dirichlet distribution

Usage

```
rdirichlet(n = 1, alpha = c(1, 1))
```

Arguments

`n`	- sample size
`alpha`	- parameters in Dirichlet distribution

Value

out - generated data

Examples

```
# dimension parameters
rdirichlet(n=10,alpha=c(1,1,1))
```

Index