

Package ‘NPBayesImpute’

September 30, 2014

Type Package

Title Non-parametric Bayesian Multiple Imputation for Categorical Data

Version 0.3

Date 2014-04-05

Author Quanli Wang, Daniel Manrique-Vallier, Jerome P. Reiter and Jingchen Hu

Maintainer Quanli Wang <quanli@stat.duke.edu>

Description These routines create multiple imputations of missing at random categorical data, with or without structural zeros. Imputations are based on Dirichlet process mixtures of multinomial distributions, which is a non-parametric Bayesian modeling approach that allows for flexible joint modeling.

License GPL (>= 3)

Depends methods, Rcpp (>= 0.10.2)

LinkingTo Rcpp

RcppModules clcm

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-09-30 16:46:17

R topics documented:

NPBayesImpute-package	2
CreateModel	3
GetDataFrame	4
GetMCZ	5
Lcm	6
MCZ	6
Rcpp_Lcm	6
Rcpp_Lcm-class	6
X	8

NPBayesImpute-package *Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros*

Description

This package implements a fully Bayesian, joint modeling approach to multiple imputation for categorical data based on latent class models with structural zeros. The idea is to model the implied contingency table of the categorical variables as a mixture of independent multinomial distributions, estimating the mixture distributions nonparametrically with Dirichlet process prior distributions. Mixtures of multinomials can describe arbitrarily complex dependencies and are computationally expedient, so that they are effective general purpose multiple imputation engines. In contrast to other approaches based on loglinear models or chained equations, the mixture models avoid the need to specify (potentially many) models, which can be a very time-consuming task with no guarantee of a theoretically coherent set of models. The package is designed to include for structural zeros, i.e., certain combinations of variables are not possible a priori.

Details

Package: NPBayesImpute
Type: Package
Version: 0.1
Date: 2014-04-05
License: GPL(>=3)

Author(s)

Quanli Wang, Daniel Manrique-Vallier, Jerome P. Reiter and Jingchen Hu
Maintainer: Quanli Wang<quanli@stat.duke.edu>

References

Manrique-Vallier, D. and Reiter, J.P. (2013), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros", JCGS.
Manrique-Vallier, D. and Reiter, J.P. (2014), "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros", Survey Methodology.

Examples

```
demo(example_short)  
demo(example)
```

CreateModel	<i>Create and initialize the Rcpp_Lcm model object</i>
-------------	--

Description

CreateModel creates and initializes an Rcpp_Lcm [Rcpp_Lcm-class](#) object for non-parametric multiple imputation of discrete multivariate categorical data with structural zeros.

Usage

```
CreateModel(X, MCZ, K, Nmax, aalpha, balpha)
```

Arguments

X	a data frame with the dataset with missing values. All variables must be unordered factors.
MCZ	a dataframe with the definition of the structural zeros. Placeholder components are represented with NAs. Variables in MCZ must be factors with the same levels as X. Rows do not need to define disjoint regions of the contingency table. See Manrique-Vallier and Reiter (2014) for details of the definition of structural zeros.
K	the maximum number of mixture components.
Nmax	An upper truncation limit for the augmented sample size.
aalpha	the hyper parameter 'a' for alpha in stick-breaking prior distribution.
balpha	the hyper parameter 'b' for alpha in stick-breaking prior distribution.

Details

This should be the first function one should call to use the library. The returned model object will be referenced in all subsequent calls.

Value

CreateModel returns an Rcpp_lcm object. The returned model object will be referenced in all subsequent calls.

References

Manrique-Vallier, D. and Reiter, J.P. (2013), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros", JCGS.
Manrique-Vallier, D. and Reiter, J.P. (2014), "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros", Survey Methodology.

Examples

```

require(NPBayesImpute)
data('NYexample')

#create the model
model <- CreateModel(X,MCZ,50,200000,0.25,0.25)

#run 1 burnins, 20 mcmc iterations and thin every 2 iterations
model$Run(1,20,2)

#retrieve parameters from the final iteration
result <- model$snapshot

#convert ImputedX matrix to dataframe, using proper factors/names etc.
ImputedX <- GetDataFrame(result$ImputedX,X)
#View(ImputedX)

```

GetDataFrame	<i>Convert imputed data to a dataframe, using the same setting from original input data.</i>
--------------	--

Description

This is a utility function to convert the imputed data matrix to a dataframe. This function will be implemented as a RCPP internal function later on.

Usage

```
GetDataFrame(dest, from, cols = 1:NCOL(from))
```

Arguments

dest	the imputed output data matrix.
from	the original input dataframe.
cols	optinal. Always use default for now.

Value

The returned dataframe object for imputed data.

Examples

```

require(NPBayesImpute)
data('NYexample')

#create the model
model <- CreateModel(X,MCZ,50,200000,0.25,0.25)
#run 1 burnins, 20 mcmc iterations and thin every 2 iterations
model$Run(1,20,2)

#retrieve parameters from the final iteration
result <- model$snapshot
#convert ImputedX matrix to dataframe, using proper factors/names etc.
ImputedX <- GetDataFrame(result$ImputedX,X)
#View(ImputedX)

```

GetMCZ	<i>Convert disjointed structural zeros to a dataframe, using the same setting from original structural zero data.</i>
--------	---

Description

This is a utility function to convert the disjointed structural zero matrix to a dataframe. This function will be implemented as a RCPP internal function later on.

Usage

```
GetMCZ(dest, from, mcz, cols = 1:NCOL(from))
```

Arguments

dest	the output data matrix for disjointed structural zeros.
from	the original input dataframe.
mcz	the original input dataframe for structural zeros.
cols	optinal. Always use default for now.

Value

The returned dataframe object for disjointed structural zeros.

References

Manrique-Vallier, D. and Reiter, J.P. (2013), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros", JCGS.
 Manrique-Vallier, D. and Reiter, J.P. (2014), "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros", Survey Methodology.

Lcm	<i>RCPP implemenation of the library</i>
-----	--

Description

[Rcpp_Lcm-class](#)

MCZ	<i>Example dataframe for structural zeros.</i>
-----	--

Description

Example dataframe for structural zeros.

Rcpp_Lcm	<i>RCPP implemenation of the library</i>
----------	--

Description

[Rcpp_Lcm-class](#)

Rcpp_Lcm-class	<i>Class "Rcpp_Lcm"</i>
----------------	-------------------------

Description

This class implements the MCMC sampler for non-parametric imputation of discrete multivariate data described in Manrique-Vallier and Reiter (2014). It provides methods for updating and monitoring the sampler.

Details

Rcpp_lcm objects should be created with [CreateModel](#). Please see the examples in the demo folder for more detailed explanation on model fitting and parameter tracing.

Extends

Class "[C++Object](#)", directly.

All reference classes extend and inherit methods from "[envRefClass](#)".

Fields

- CurrentIteration:** the total number of iterations that have been run so far.
- EnableTracer:** to check tracer status or to enable/disable the tracer.
- MCZ:** the disjointed structural zero matrix.
- snapshot:** retrieve a list with the current state of all the parameters in the sampler, including the imputed sample. A call the the "snapshot" method returns a list with the following components:
- alpha:** the concentration parameter of the stick breaking prior.
 - k_star:** the effective number number of latent classes (mixture components)
 - Nmis:** the size of the augmented sample.
 - nu:** a vector with the mixture weights
 - z:** a matrix with the current latent class assignment of each member of the sample
 - ImputedX:** the current raw imputed dataset. Use [GetDataFrame](#) to convert the raw data to a data frame of factors as defined in the input data set.
 - psi:** The conditional multinomial probabilities. A $L_{\max} * K * J$ array, where L_{\max} is the maximum number of levels of all discrete factors in the dataset, J is the number of factors in the dataset, and K is the number of latent classes. Since variables might have different numbers of levels, unused entries in the first dimension are filled with NAs to complete L_{\max} .
- traceable:** list of model parameters that can be traced by the tracer.
- traced:** list of model parameters that are traced.

Methods

- SetTrace(paralist, num_of_ iterations):** set parameters to be traced.
- paralist:** a list of parameters to be traced.
 - num_of_ iterations:** the maximum number of traced iterations.
- Run(burnin, iter, thinning):** run MCMC iterations.
- burnin:** number of burn in iterations.
 - iter:** number of MCMC iterations.
 - thinning:** thinning parameter.
- Resume():** resume from an interrupted call to run method.
- Parameters(paralist):** retrieve a selected list of model parameters from last MCMC iteration.
- paralist:** a list of parameters to be traced.
- GetTrace():** retrieve all traced iterations. Returns a list with all the parameters set using the method `SetTrace()`. See description of `snapshotreference` method for a description of the parameters.

References

- Manrique-Vallier, D. and Reiter, J.P. (2013), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros", JCGS.
- Manrique-Vallier, D. and Reiter, J.P. (2014), "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros", Survey Methodology.

Examples

```
require(NPBayesImpute)
data('NYexample')

#create the model
model <- CreateModel(X,MCZ,100,200000,0.25,0.25)
#run 1 burnin, 20 mcmc iterations and thin every 2 iterations.
#MCMC should be much longer when fitting real data
model$Run(1,20,2)

#retrieve parameters from the final iteration
result <- model$snapshot

#convert ImputedX matrix to dataframe, using proper factors/names etc.
IX <- GetDataFrame(result$ImputedX,X)
#View(IX)
```

X

Example dataframe for input categorical data with missing values.

Description

Example dataframe for input categorical data with missing values.

Index

*Topic **classes**

Rcpp_Lcm-class, [6](#)

*Topic **package**

NPBayesImpute-package, [2](#)

C++Object, [6](#)

CreateModel, [3](#), [6](#)

envRefClass, [6](#)

GetDataFrame, [4](#), [7](#)

GetMCZ, [5](#)

Lcm, [6](#)

MCZ, [6](#)

NPBayesImpute (NPBayesImpute-package), [2](#)

NPBayesImpute-package, [2](#)

Rcpp_Lcm, [6](#)

Rcpp_Lcm-class, [3](#), [6](#), [6](#)

X, [8](#)