Package ‘ICmiss’

Type Package

 Title An Imputation-Consistency Algorithm for High-Dimensional Missing Data Problems and Beyond

Version 1.0.0

Date 2017-10-03

Author Bochao Jia, Faming Liang

Maintainer Bochao Jia <jbc409@ufl.edu>

Depends R (>= 3.0.2)

Imports mvtnorm, equSA, huge, ncvreg

Description Missing data are frequently encountered in high-dimensional data analysis, but they are usually difficult to deal with using standard algorithms, such as the EM algorithm and its variants. This package provides a general algorithm, the so-called imputation-consistency (IC) algorithm, for high-dimensional missing data problems. This package has also extended the applications of the IC algorithm to random coefficient models.

License GPL-2

LazyLoad true

NeedsCompilation yes

Repository CRAN

Date/Publication 2017-09-29 00:19:24 UTC

RoxygenNote 6.0.1

R topics documented:

ICmiss-package .................................................. 2
EyeICC ............................................................ 3
eye_norm ......................................................... 4
GraphIC ........................................................... 5
RCDat .............................................................. 6
RCLM ............................................................... 6
RegICC ............................................................. 7
SimGraDat ........................................................ 8
SimRegDat ......................................................... 9
yeast .............................................................. 10
YeastIC ............................................................ 11

Index .......................................................... 13
Description

Missing data are frequently encountered in high-dimensional data analysis, but they are usually difficult to deal with using standard algorithms, such as the EM algorithm and its variants. This package provides a general algorithm, the so-called imputation-consistency (IC) algorithm, for treating high-dimensional missing data problems. A variant of the IC algorithm, the so-called imputation-conditional consistency (ICC) algorithm, has also provided in the package.

Details

| Package: | ICmiss |
| Type: | Package |
| Version: | 1.0.0 |
| Date: | 2017-10-03 |
| License: | GPL-2 |

This package illustrates the use of the IC/ICC algorithms in three modules:

The first module is to apply the IC algorithm to learning high-dimensional Gaussian Graphical Models (GGMs) in presence of missing data with a simulated dataset SimGradDat(n, p, \ldots) and Yeast cell example YeastIC(data, \ldots).

The second module is to apply the ICC algorithm to variable selection for high-dimensional linear regression in presence of missing data. The simulation study covers both cases, the covariates are mutually independent and generally dependent, with the code SimRegDat(n, p, \ldots). The real data example is for Bardet-Biedl syndrome (Scheetz et al., 2006) with the dataset available in the R package flare.

The third module is to apply the ICC algorithm to random coefficient models, where the random coefficients are treated as missing data. A simulated dataset data(RCDat) is included in the package, which can be used in RCLM(RCDat).

Author(s)

Bochao Jia, Faming Liang
Maintainer: Bochao Jia<jbc409@ufl.edu>

References


Examples

```r
#library(ICmiss)
#result <- SimRegDat(n = 100, p = 200, type = "dep", rate = 0.1)
#RegICC(result$x, result$y, result$coef, type = "dep", iteration = 30, warm = 20)
```

---

**EyeICC**  
*Variable selection for Bardet-Biedl syndrome data with missing observations.*

**Description**

The imputation-conditional consistency (ICC) algorithm is used to select variables for the Bardet-Biedl syndrome data with missing observations: We first randomly delete a specified percentage of observations and then apply the ICC algorithm for variable selection.

**Usage**

```
EyeICC(x, y, rate = 0.05, alpha1 = 0.1, alpha2 = 0.1, iteration = 30, warm = 20)
```

**Arguments**

- **x**  
a \(n\times p\) covariates matrix.
- **y**  
a \(n\times 1\) responses.
- **rate**  
Missing rate, the default value is 0.05.
- **alpha1**  
The significance level of correlation screening in the \(\psi\)-learning algorithm, see [equSA](#). In general, a high significance level of correlation screening will lead to a slightly large separator set, which reduces the risk of missing important variables in the conditioning set. In general, including a few false variables in the conditioning set will not hurt much the accuracy of the \(\psi\)-partial correlation coefficient, the default value is 0.1.
- **alpha2**  
The significance level of \(\psi\)-partial correlation coefficient screening for estimating the adjacency matrix, see [equSA](#), the default value is 0.1.
- **iteration**  
The number of total iterations, the default value is 30.
- **warm**  
The number of burn-in iterations, the default value is 20.

**Value**

- **topVar**  
Variables ranked by the frequency of appearance in the last few iterations.

**Author(s)**

Bochao Jia (<jbc409@ufl.edu>) and Faming Liang
References


Examples

```r
#library(ICmiss)
data(eye_norm)
#EyeICC(eye_norm$x, eye_norm$y, rate = 0.05, alpha1 = 0.1, alpha2 = 0.1)
```

**eye_norm**  
Example dataset for high-dimensional variable selection by the ICC algorithm.

Description

Gene expression data from the microarray experiments of mammalian-eye tissue samples of Scheetz et al. (2006). It should be used in EyeICC(x, y...).

x  a nxp gene expression data.

y  The expression level of gene TRIM32.

Usage

data(eye_norm)

Format

A list containing the matrix x and response matrix y

References

**Description**

The imputation-consistency (IC) algorithm for learning high-dimensional Gaussian Graphical Models with simulated incomplete data.

**Usage**

```r
GraphIC(data, A, alpha1 = 0.05, alpha2 = 0.05, alpha3 = 0.05, iteration = 30, warm = 20)
```

**Arguments**

- `data`: $n \times p$ Dataset with missing values.
- `A`: True adjacency matrix for evaluating the performance of the IC algorithm.
- `alpha1`: The significance level of correlation screening in the $\psi$-learning algorithm, see `equSA`. In general, a high significance level of correlation screening will lead to a slightly large separator set, which reduces the risk of missing important variables in the conditioning set. In general, including a few false variables in the conditioning set will not hurt much the accuracy of the $\psi$-partial correlation coefficient, the default value is 0.05.
- `alpha2`: The significance level of $\psi$-partial correlation coefficient screening for estimating the adjacency matrix, see `equSA`, the default value is 0.05.
- `alpha3`: The significance level of integrative $\psi$-partial correlation coefficient screening for estimating the adjacency matrix of IC_Ave method, the default value is 0.05.
- `iteration`: The number of total iterations, the default value is 30.
- `warm`: The number of burn-in iterations, the default value is 20.

**Value**

- `RecPre`: The output of Recall and Precision values for the IC algorithm.
- `Adj`: $p \times p$ Estimated adjacency matrix by our IC algorithm.

**Author(s)**

Bochao Jia<jbc49@ufl.edu> and Faming Liang

**References**


Examples

```r
#library(icmiss)
#library(huge)
#result <- SimGraDat(n = 100, p = 50, type = "band", rate = 0.1)
#Est <- GraphIC(result$data, result$A, alpha1 = 0.05, alpha2 = 0.05, alpha3 = 0.05, iteration = 10, warm = 5)
#huge.plot(Est$Adj) ## plot network by our estimated adjacency matrix.
#plot(Est$RecPre[,1], Est$RecPre[,2], type="1", xlab="Recall", ylab="Precision") ## plot the Recall-Precision curve.
```

**RCDat**  
A simulated dataset for random coefficient models.

**Description**
Number of customers I=100 and each customer responds to J=10 items. The first column is for responses. It should be used in RCLM(RCDat).

**RCDat**  
A simulated dataset.

**Usage**

```r
data(RCDat)
```

**Format**

matrix

**References**

**RCLM**  
Random Coefficient Models

**Description**
An extension of the ICC algorithm for Bayesian Computation.

**Usage**

```r
RCLM(Data, iteration = 10000, warm = 100)
```

**Arguments**

- **Data**  
  A simulated dataset. The first column is the response and the rest is for explanatory variables.

- **iteration**  
  The number of total iterations, the default value is 10000.

- **warm**  
  The number of burn-in iterations, the default value is 100.
Value

- **path**: The traces of estimated coefficients vs. iterations.
- **coef**: The mean of estimated coefficients.

Author(s)

Bochao Jia<jbc409@ufl.edu> and Faming Liang

References


Examples

```r
library(icmiss)
data(RCDat)
RCLM(RCDat, iteration = 1000, warm = 100)
```

RegICC

*Variable selection for high-dimensional Regression with Missing Data.*

Description

Application of the imputation-conditional consistency (ICC) algorithm for high-dimensional variable selection in presence of missing data.

Usage

```r
RegICC(x, y, coef, type = "indep", alpha1 = 0.1, alpha2 = 0.05, iteration = 30, warm = 20)
```

Arguments

- **x**: \(n \times p\) covariates matrix.
- **y**: \(n \times 1\) responses.
- **coef**: \(p \times 1\) coefficients for generating responses from the covariates matrix.
- **type**: When `type="indep"`, the case with independent covariates, or `type="dep"`, the case with dependent covariates, the default type is "indep".
- **alpha1**: The significance level of correlation screening in the \(\psi\)-learning algorithm, see `equSA`. In general, a high significance level of correlation screening will lead to a slightly large separator set, which reduces the risk of missing important variables in the conditioning set. In general, including a few false variables in the conditioning set will not hurt much the accuracy of the \(\psi\)-partial correlation coefficient, the default value is 0.1.
alpha2  The significance level of $\psi$-partial correlation coefficient screening for estimating the adjacency matrix, see equSA, the default value is 0.05.
iteration  The number of total iterations, the default value is 30.
warm  The number of burn-in iterations, the default value is 20.

Value
Var  Selected variables and their estimated coefficients by our ICC algorithm.
table  The summarized table for evaluating the performance of IC (ICC) algorithm. 'bias' denotes Euclidean distance between estimated coefficients and true coefficients; 'fsr' denotes false selection rate and 'nsr' denotes negative selection rate. The smaller the measurements are, the better the performance is.

Author(s)
Bochao Jia<jbc4PY@ufl.edu> and Faming Liang

References

Examples
library(ICmiss)
result <- SimRegDat(n = 100, p = 50, type = "indep", rate = 0.1)
RegICC(result$x, result$y, result$coef, type = "indep", iteration = 10, warm = 5)

SimGraDat  Simulate Incomplete Data for Gaussian Graphical Models

Description
Simulate incomplete data with a band structure, which can be used in GraphIC(data,...) for estimating the structure of the Gaussian graphical network.

Usage
SimGraDat(n = 200, p = 100, type = "band", rate = 0.1)

Arguments
n  Number of observations, default of 200.
p  Number of covariates, default of 100.
type  type="band" which denotes the band structure, see equSA.
rate  Missing rate, the default value is 0.1.
**SimRegDat**

**Value**
- **data** \( n \times p \) Gaussian distributed data with missing.
- **A** \( p \times p \) adjacency matrix used for generating data.

**Author(s)**
Bochao Jia<jbc409@ufl.edu> and Faming Liang

**References**

**Examples**

```r
library(ICmiss)
SimGraDat(n = 200, p = 100, type = "band", rate = 0.1)
```

**SimRegDat**

*Simulate Incomplete Data for High-Dimensional Linear Regression.*

**Description**
Simulate incomplete data for high-dimensional linear regression with dependent or independent covariates RegICC(x,y...).

**Usage**

```r
SimRegDat(n = 100, p = 200, type = "indep", rate = 0.1)
```

**Arguments**
- **n** Number of observations, default of 100.
- **p** Number of covariates, default of 200.
- **type** When **type**="indep", it simulates the data with independent covariates, or **type**="dep", it simulates the data with dependent covariates, the default type is "indep".
- **rate** Missing rate, the default value is 0.1.

**Value**
- **x** \( n \times p \) covariates matrix.
- **y** \( n \times I \) responses.
- **coef** \( p \times I \) coefficients for generating responses from the covariates matrix.
Author(s)

Bochao Jia <jbc409@ufl.edu> and Faming Liang

References


Examples

library(ICmiss)
SimRegDat(n = 100, p = 200, type = "dep", rate = 0.1)

---

**yeast**  
Example dataset for learning Gaussian Graphical Models by the IC Algorithm

Description

Genomic expression patterns in the yeast Saccharomyces cerevisiae responding to diverse environmental changes. The whole dataset consists of 173 samples collected under different environmental settings, and is available at http://genome-www.stanford.edu/yeast-stress/. It should be used in YeastIC(data,...).

Usage

data(yeast)

Format

yeast  a n x p Yeast Cell expression data.

References

YeastIC

Learning gene regulatory networks for Yeast Cell Expression Data.

Description

An Imputation Consistency (IC) algorithm for learning gene regulatory networks with missing data. The dataset is collected from the yeast Saccharomyces cerevisiae responding to diverse environmental changes and is available at http://genome-www.stanford.edu/yeast-stress/.

Usage

YeastIC(data, alpha1 = 0.05, alpha2 = 0.01, alpha3 = 0.01, iteration = 30, warm = 20)

Arguments

data n x p

Yeast Cell expression data.

alpha1 The significance level of correlation screening in the ψ-learning algorithm, see equSA. In general, a high significance level of correlation screening will lead to a slightly large separator set, which reduces the risk of missing important variables in the conditioning set. In general, including a few false variables in the conditioning set will not hurt much the accuracy of the ψ-partial correlation coefficient, the default value is 0.05.

alpha2 The significance level of ψ-partial correlation coefficient screening for estimating the adjacency matrix, see equSA, the default value is 0.01.

alpha3 The significance level of integrative ψ-partial correlation coefficient screening for estimating the adjacency matrix of IC_Ave method, the default value is 0.01.

iteration The number of total iterations, the default value is 30.

warm The number of burn-in iterations, the default value is 20.

Value

A p x p Estimated adjacency matrix for network construction.

Author(s)

Bochao Jia<jbc409@ufl.edu> and Faming Liang

References


Examples

```r
#library(ICmiss)
#library(huge)
data(yeast)
A <- YeastIC(yeast, alpha1 = 0.05, alpha2 = 0.01, alpha3 = 0.01, iteration = 30, warm = 20)
huge.plot(A)  ## plot gene regulatory network by our estimated adjacency matrix.
```
Index

*Topic **EyeICC**
  EyeICC, 3

*Topic **GraphIC**
  GraphIC, 5

*Topic **RCLM**
  RCLM, 6

*Topic **RegICC**
  RegICC, 7

*Topic **SimGraDat**
  SimGraDat, 8

*Topic **SimRegDat**
  SimRegDat, 9

*Topic **YeastIC**
  YeastIC, 11

*Topic **datasets**
  eye_norm, 4
  RCDat, 6
  yeast, 10

*Topic **package**
  ICmiss-package, 2

eye_norm, 4
EyeICC, 3

GraphIC, 5

ICmiss-package, 2

RCDat, 6
RCLM, 6
RegICC, 7

SimGraDat, 8
SimRegDat, 9

yeast, 10
YeastIC, 11