Package 'IROmiss'

February 19, 2020

Type Package

Title Imputation Regularized Optimization Algorithm

Version 1.0.2

Date 2020-02-19

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 Regularized Optimization (IRO) algorithm, for high-dimensional missing data problems. You can refer to Liang, F., Jia, B., Xue, J., Li, Q. and Luo, Y. (2018) at <arXiv:1802.02251> for detail.

License GPL-2

LazyLoad true

NeedsCompilation yes

Repository CRAN

Date/Publication 2020-02-19 05:10:02 UTC

RoxygenNote 6.0.1

Author Bochao Jia [aut, ctb, cre, cph],
Faming Liang [ctb]

Maintainer Bochao Jia <jbc409@ufl.edu>

R topics documented:

`IROmiss-package` *Imputation Regularized Optimization Algorithm*

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 regularized optimization (IRO) algorithm, for treating high-dimensional missing data problems. A variant of the IRO algorithm, the so-called imputation conditional regularized optimization (ICRO) algorithm, has also been provided in the package.

Details

Package:	IROmiss
Type:	Package
Version:	1.0.2
Date:	2020-02-19
License:	GPL-2

This package illustrates the use of the IRO/ICRO algorithms in three modules:

The first module is to apply the IRO algorithm to learning high-dimensional Gaussian Graphical Models (GGMs) in presence of missing data with a simulated dataset `SimGraDat(n,p,...)` and Yeast cell example `YeastIRO(data,...)`.

The second module is to apply the ICRO algorithm to varisable 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,...)`. 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 ICRO algorithm to random coefficient linear models, where the random coefficients are treated as missing data. We can generate a dataset for the random coefficient linear models with `SimRCLM(I,J,...)` and a simulated dataset `data(RCDat)` is included in the package, which can be used in `RCLM(I,J,RCDat,...)` for estimate the random coefficents.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.<doi:10.1080/01621459.2015.1012391>

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.<doi:10.1093/biomet/asn036>

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B. <arXiv:1802.02251>

Jia, B., Xu, S., Xiao, G., Lamba, V., Liang, F. (2017) Inference of Genetic Networks from Next Generation Sequencing Data. Biometrics.

Examples

```
library(IROmiss)
p <- 200
beta <- rep(0,p)
beta[1:5] <- c(1, 2, -1.5, -2.5, 5)
result <- SimRegDat(n = 100, p = 200, coef = beta, data.type = "indep",
miss.type="MCAR", rate = 0.05)
RegICRO(result$x, result$y, result$coef, type = "indep", iteration = 30, warm = 20)
```

---

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

Description

Normalized Gene expression data from the microarray experiments of mammalian-eye tissue samples of Scheetz et al. (2006). It should be used in `EyeICRO(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

T. Scheetz, k. Kim, R. Swiderski, A. Philp, T. Braun, K. Knudtson, A. Dorrance, G. DiBona, J. Huang, T. Casavant, V. Sheffield, E. Stone .Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences of the United States of America, 2006.

`GraphIRO` *Learning high-dimensional Gaussian Graphical Models with Missing Observations.*

Description

The imputation regularized optimization (IRO) algorithm for learning high-dimensional Gaussian Graphical Models with simulated incomplete data.

Usage

```
GraphIRO(data, A, alpha1 = 0.05, alpha2 = 0.05, alpha3 = 0.05, iteration = 30, warm = 10)
```

Arguments

`data`	nxp Dataset with missing values.
`A`	True adjacency matrix for evaluating the performance of the IRO algorithm.
`alpha1`	The significance level of correlation screening in the ψ-learning algorithm, see R package **equSA** for detail. 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 0.05.
`alpha3`	The significance level of integrative ψ-partial correlation coefficient screening for estimating the adjacency matrix of IRO_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 10.

Value

`RecPre`	The output of Recall and Precision values for the IRO algorithm.
`Adj`	pxp Estimated adjacency matrix by our IRO algorithm.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

```
library(IROmiss)
library(huge)
result <- SimGraDat(n = 200, p = 100, type = "band", rate = 0.1)
Est <- GraphIRO(result$data, result$A, iteration = 20, warm = 10)
## plot network by our estimated adjacency matrix.
huge.plot(Est$Adj)
## plot the Recall-Precision curve.
plot(Est$RecPre[,1], Est$RecPre[,2], type="l", xlab="Recall", ylab="Precision")
```

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

Description

The dataset is generated using the default settings. The Number of customers I=100 and each customer responds to J=10 items. For the parameters, the true coefficient β is $(\beta_0, \beta_1, \beta_2, \beta_3) = (1, 2, 1.5, 1)$ and the true value of σ^2 is 0.25. The first column of the dataset denote the response $\mathbf{y}$. The dataset should be used in `RCLM(I,J,RCDat,...)`.

RCDat A simulated dataset.

Usage

```
data(RCDat)
```

Format

matrix

References

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

RCLM — *Fit Random Coefficient Linear Models*

Description

An extension of the ICRO algorithm for Bayesian Computation. It can be used to fit a Random Coefficient Linear Models and estimate the coefficients β and σ^2.

Usage

```
RCLM(I=100, J=10, Data, iteration = 10000, warm = 100)
```

Arguments

`I`	Number of first subjects in the random coefficient linear model (RCLM).
`J`	Number of second subjects in the random coefficient linear model (RCLM).
`Data`	A simulated dataset. The first column is the response and the rest is for explanatory variables, see `RCDat` for detail.
`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 β and σ^2.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Penalized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

```
library(IROmiss)
data(RCDat)
RCLM(I=100, J=10, RCDat, iteration = 10000, warm = 1000)
```

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

Description

Application of the imputation conditional regularized optimization (ICRO) algorithm for high-dimensional variable selection in presence of missing data.

Usage

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

Arguments

`x`	nxp covariates matrix.
`y`	nx1 responses.
`coef`	A px1 vector of coefficients for the linear regression model. The intercept coefficient is default to 1.
`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 ψ-learning algorithm, see R package **equSA** for detail. 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 0.1.
`alpha2`	The significance level of ψ-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 ICRO algorithm.
`table`	The summarized table for evaluating the performance of ICRO 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<jbc409@ufl.edu> and Faming Liang

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

```
library(IROmiss)
p <- 200
beta <- rep(0,p)
beta[1:5] <- c(1, 2, -1.5, -2.5, 5)
result <- SimRegDat(n = 100, p = 200, coef = beta, data.type = "indep",
miss.type="MAR", rate = 0.05)
RegICRO(result$x, result$y, result$coef, type = "indep", iteration = 20, warm = 10)
```

SimGraDat *Simulate Incomplete Data for Gaussian Graphical Models*

Description

Simulate incomplete data with a band structure, which can be used in `GraphIRO(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, with precision matrix

$$C_{i,j} = \begin{cases} 0.5, & \text{if } |j-i| = 1, i = 2, ..., (p-1), \\ 0.25, & \text{if } |j-i| = 2, i = 3, ..., (p-2), \\ 1, & \text{if } i = j, i = 1, ..., p, \\ 0, & \text{otherwise.} \end{cases}$$

`rate`	Missing rate, the default value is 0.1.

Value

`data`	*n*x*p* Gaussian distributed data with missing.
`A`	*p*x*p* adjacency matrix used for generating data.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

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

`SimRCLM` *Simulate Dataset for Random Coefficient Linear Models*

Description

Simulate a dataset for random coefficient linear model, which can be used in `RCLM(I,J,RCDat,...)`.

Usage

```
SimRCLM(I=100, J=10, beta, sigma)
```

Arguments

`I`	Number of first subjects in the random coefficient linear model (RCLM).
`J`	Number of second subjects in the random coefficient linear model (RCLM).
`beta`	A 4x1 vector of random coefficients of the model, now only allows length 4.
`sigma`	The standard diviation for the noise term.

Value

`D`	A simulated data matrix for random coefficient models. The first column of the dataset denote the response $\mathbf{y}$. The dataset should be used in `RCLM(I,J,RCDat...)`.
`coef`	The mean of estimated coefficients β and σ^2.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Penalized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

```
library(IROmiss)
beta<-c(1,2,1.5,1)
sigma <- 0.5
D <- SimRCLM(I=100, J=10, beta, sigma)
RCLM(I=100, J=10, D, iteration = 10000, warm = 1000)
```

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

Description

Simulate incomplete data for high-dimensional linear regression with dependent or independent covariatesRegICRO(x,y...).

Usage

```
SimRegDat(n = 100, p = 200, coef, data.type = "indep",
miss.type="MCAR", rate = 0.1)
```

Arguments

`n`	Number of observations, default of 100.
`p`	Number of covariates, default of 200.
`coef`	A px1 vector of coefficients for the linear regression model. The intercept coefficient is default to 1.
`data.type`	When `data.type=="indep"`, it simulates the data with independent covariates, each covariate independently follow the normal distribution with mean 0 and variance 4. When `data.type=="dep"`, it simulates the data with dependent covariates with "band" dependent structure, see `SimGraDat` for detail. The default data type is "indep".
`miss.type`	`miss.type=="MCAR"` refer to the case of missing completely at random. when `miss.type=="MAR"`, the missing probability for each data point is proportional to the mean of its conditional normal distribution, the default missing type is "MCAR".
`rate`	Missing rate, the default value is 0.1.

Value

`x`	nxp covariates matrix.
`y`	nx1 responses.
`coef`	px1 vector of coefficients for the linear regression model.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

```
library(IROmiss)
p <- 200
beta <- rep(0,p)
beta[1:5] <- c(1, 2, -1.5, -2.5, 5)
```

```
SimRegDat(n = 100, p = 200, coef = beta, data.type = "dep",
miss.type="MAR", rate = 0.1)
```

---

yeast *Example dataset for learning Gaussian Graphical Models by the IRO 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://www-genome.stanford.edu/yeast_stress/. It should be used in `YeastIRO(data,...)`.

Usage

```
data(yeast)
```

Format

yeast a *n*x*p* Yeast Cell expression data.

References

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell, 11, 4241-4257.

YeastIRO *Learning gene regulatory networks for Yeast Cell Expression Data.*

Description

An Imputation Regularized Optimization (IRO) 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

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

Arguments

`data`	nxp Yeast Cell expression data.
`alpha1`	The significance level of correlation screening in the ψ-learning algorithm, see R package **equSA** for detail. 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 IRO_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`	pxp Estimated adjacency matrix for network construction.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

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

Index