Package 'GFM'

August 11, 2023

Type Package

Title Generalized Factor Model

Version 1.2.1

Date 2023-08-10

License GPL-3

Author Wei Liu [aut, cre],
Huazhen Lin [aut],
Shurong Zheng [aut],
Jin Liu [aut],
Jinyu Nie [aut]

Maintainer Wei Liu <LiuWeideng@gmail.com>

Description Generalized factor model is implemented for ultra-high dimensional data with mixed-type variables.
Two algorithms, variational EM and alternate maximization, are designed to implement the generalized factor model,
respectively. The factor matrix and loading matrix together with the number of factors can be well estimated.
This model can be employed in social and behavioral sciences, economy and finance, and genomics,
to extract interpretable nonlinear factors. More details can be referred to
Wei Liu, Huazhen Lin, Shurong Zheng and Jin Liu. (2021) <doi:10.1080/01621459.2021.1999818>.

URL https://github.com/feiyoung/GFM

BugReports https://github.com/feiyoung/GFM/issues

Depends doSNOW, parallel, R (>= 3.5.0)

Imports MASS, stats,irlba, Rcpp, methods

Suggests knitr, rmarkdown

LinkingTo Rcpp, RcppArmadillo

VignetteBuilder knitr

Encoding UTF-8

RoxygenNote 7.1.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2023-08-11 06:22:04 UTC

R topics documented:

`chooseFacNumber` *Choose the Number of factors for Generalized Factor Models*

Description

This function is designed to chooose the number of factors for a generalized factor model.

Usage

```
chooseFacNumber(XList, types, q_set = 2: 10,
        select_method = c("SVR", "IC"),offset=FALSE,
        dc_eps=1e-4, maxIter=30, verbose = TRUE, parallelList=NULL)
```

Arguments

`XList`	a list consisting of matrices with the same rows n, and different columns (p1,p2, ..., p_d),observational mixed data matrix list, d is the types of variables, p_j is the dimension of varibles with the j-th type.
`types`	a d-dimensional character vector, specify the type of variables. For example, `types=c('gaussian','poisson', 'binomial')`, implies the components of `XList` are matrices with continuous, count and binomial values, respectively.
`q_set`	a positive integer vector, specify the candidates of factor number q, (optional) default as `c(2:10)` according to Bai (2013).
`select_method`	a string, specify the method to choose the number of factors. Two methods are supported: the singular value ratio (SVR) and information criterion (IC) based methods, default as 'SVR'. Empirically, 'SVR' is much faster than 'IC', especially for high-dimensional large-scale data.
`offset`	a logical value, whether add an offset term (the total counts for each row in the count component of XList) when there are Poisson variables.

`dc_eps`	positive real number, specify the tolerance of varing quantity of objective function in the algorithm. Optional parameter with default as `1e-4`.
`maxIter`	a positive integer, specify the times of iteration. Optional parameter with default as 50.
`verbose`	a logical value, specify whether ouput the information in iteration process, (optional) default as TRUE.
`parallelList`	a list with two components: (1) parallel: a logical value with TRUE or FALSE, indicates wheter to use prallel computating. Optional parameter with default as FALSE. (2)ncores: a positive integer, specify the number of cores when parallel computing is used. This argument plays its role if only `select_method='IC'`.

Value

return an integer value, the estimated number of factors.

Note

nothing

Author(s)

Liu Wei

References

Liu, W., Lin, H., Zheng, S., & Liu, J. (2021). Generalized factor model for ultra-high dimensional correlated variables with mixed types. Journal of the American Statistical Association, (just-accepted), 1-42.

See Also

nothing

Examples

```
## mix of normal and Poisson

dat <- gendata(seed=1, n=60, p=60, type='norm_pois', q=2, rho=2)
## we set maxIter=2 for example.
hq <- chooseFacNumber(dat$XList, dat$types, verbose = FALSE, maxIter=2)
```

`Factorm` *Factor Analysis Model*

Description

Factor analysis to extract latent linear factor and estimate loadings.

Usage

```
Factorm(X, q=NULL)
```

Arguments

`X`	a n-by-p matrix, the observed data
`q`	an integer between 1 and p or `NULL`, default as `NULL` and automatically choose q by the eigenvalue ratio method.

Value

return a list with class named `fac`, including following components:

`hH`	a n-by-q matrix, the extracted lantent factor matrix.
`hB`	a p-by-q matrix, the estimated loading matrix.
`q`	an integer between 1 and p, the number of factor extracted.
`sigma2vec`	a p-dimensional vector, the estimated variance for each error term in model.
`propvar`	a positive number between 0 and 1, the explained propotion of cummulative variance by the q factors.
`egvalues`	a n-dimensional(n<=p) or p-dimensional(p<n) vector, the eigenvalues of sample covariance matrix.

Note

nothing

Author(s)

Liu Wei

References

Fan, J., Xue, L., and Yao, J. (2017). Sufficient forecasting using factor models. Journal of Econometrics.

See Also

`gfm`.

Examples

```
dat <- gendata(n = 300, p = 500)
res <- Factorm(dat$X)
measurefun(res$hH, dat$H0) # the smallest canonical correlation
```

---

gendata *Generate simulated data*

Description

Generate simulated data from high dimensional genelized nonlinear factor model.

Usage

```
gendata(seed = 1, n = 300, p = 50,
            type =  c('homonorm', 'heternorm',
            'pois', 'bino', 'norm_pois', 'pois_bino', 'npb'),
                   q = 6, rho = 1, n_bin=1)
```

Arguments

seed	a nonnegative integer, the random seed, default as 1.
n	a positive integer, the sample size.
p	an positive integer, the variable dimension.
type	a character, specify the variables types for generated data, default as 'homonorm', representing the homogeneous gaussian variables.
q	a positive integer, the number of factors.
rho	a positive number, controlling the magnitude of loading matrix.
n_bin	a positive integer, specify the number of trails for the binomial variables when type is set to one of 'bino', 'pois_bino' and 'npb'.

Details

This function provides a variaty of mix of different variable types, in which 'homonorm' represents the generated data with only homogenous normal variables; 'heternorm' represents the data with only heterogenous normal variables; 'pois' means the data with only poisson variables; 'bino' means the data with only binomial variables; 'norm_pois' means the mix of normal and poisson variables; 'pois_bino' represents the mix of poisson and binomial variables; and 'npb' means the most complex mix of normal, poisson and binomial variables.

Value

return a list including two components:

X	a n-by-p matrix, the observed data matrix.
XList	a list consisting of the above observed data matrices with the same rows n (observations), and different columns (p1,p2, ..., p_d) and p columns in total, where d is the types of variables, pj is the dimension of varibles with the j-th type.
H0	a n-by-q matrix, the true lantent factor matrix.
B0	a p-by-q matrix, the true loading matrix, the last pzero rows are vectors of zeros.
mu0	a p-dimensional vector, the true intercept terms.

Note

nothing

Author(s)

Wei Liu

See Also

Factorm; gfm.

Examples

```
dat <- gendata(n=300, p = 500)
str(dat)
```

---

gfm *Generalized Factor Model*

Description

This function is to implement the generalized factor model.

Usage

```
gfm(XList, types, q=10, offset=FALSE, dc_eps=1e-4, maxIter=30,
              verbose = TRUE, algorithm=c("VEM", "AM"))
```

Arguments

XList	a list consisting of matrices with the same rows n, and different columns (p1,p2, ..., p_d),observational mixed data matrix list, d is the types of variables, p_j is the dimension of varibles with the j-th type.
types	a d-dimensional character vector, specify the type of variables. For example, types=c('gaussian','poisson', 'binomial'), implies the components of XList are matrices with continuous, count and binomial values, respectively.
q	a positive integer or empty, specify the number of factors, defualt as 10.
offset	a logical value, whether add an offset term (the total counts for each row in the count component of XList) when there are Poisson variables.
dc_eps	a positive real, specify the relative tolerance of objective function in the algorithm. Optional parameter with default as 1e-4.
maxIter	a positive integer, specify the times of iteration. Optional parameter with default as 30.
verbose	a logical value with TRUE or FALSE, specify whether ouput the information in iteration process, (optional) default as TRUE.
algorithm	a string, specify the algorithm to be used for fitting model. Now it supports two algorithms: variational EM (VEM) and alternate maximization (AM) algorithm, default as VEM. Empirically, we observed that VEM is more robust than AM to the high noise data.

Details

This function also has the MATLAB version at https://github.com/feiyoung/MGFM/blob/master/gfm.m.

Value

return a list with class name 'gfm' and including following components,

hH	a n*q matrix, the estimated factor matrix.
hB	a p*q matrix, the estimated loading matrix.
hmu	a p-dimensional vector, the estimated intercept terms.
obj	a real number, the value of objective function when the convergence achieves.
q	an integer, the used or estimated factor number.
history	a list including the following 7 components: (1)dB: the varied quantity of B in each iteration; (2)dH: the varied quantity of H in each iteration; (3)dc: the varied quantity of the objective function in each iteration; (4)c: the objective value in each iteration; (5) realIter: the real iterations to converge; (6)maxIter: the tolerance of maximum iterations; (7)elapsedTime: the elapsed time.

Note

nothing

Author(s)

Liu Wei

References

Liu, W., Lin, H., Zheng, S., & Liu, J. (2021). Generalized factor model for ultra-high dimensional correlated variables with mixed types. Journal of the American Statistical Association, (just-accepted), 1-42.

Bai, J. and Liao, Y. (2013). Statistical inferences using large esti- mated covariances for panel data and factor models.

See Also

nothing

Examples

```
## mix of normal and Poisson

dat <- gendata(seed=1, n=60, p=60, type='norm_pois', q=2, rho=2)
## we set maxIter=2 for example.
gfm2 <- gfm(dat$XList,  dat$types,  q=2, verbose = FALSE, maxIter=2)
measurefun(gfm2$hH, dat$H0, type='ccor')
measurefun(gfm2$hB, dat$B0, type='ccor')
```

---

measurefun *Assess the performance of an estimator on a matrix*

Description

Evaluate the smallest cononical correlation (ccor) coefficients or trace statistic between two matrices, where a larger ccor or trace statistic is better.

Usage

```
measurefun(hH, H, type=c('trace_statistic','ccor'))
```

Arguments

hH	a n-by-q matrix, the estimated matrix.
H	a n-by-q matrix, the true matrix.
type	a character taking value within `c('trace_statistic','ccor')`, default as 'trace_statistic'.

Value

return a real number.

Note

nothing

Author(s)

Liu Wei

Examples

```
dat <- gendata(n = 100, p = 200, q=2, rho=3)
res <- Factorm(dat$XList[[1]])
measurefun(res$hB, dat$B0)
```

`overdispersedGFM` *Overdispersed Generalized Factor Model*

Description

This function is to implement the overdispersed generalized factor model.

Usage

```
overdispersedGFM(XList, types, q, offset=FALSE, epsELBO=1e-5,
                 maxIter=30, verbose=TRUE)
```

Arguments

`XList`	a list consisting of matrices with the same rows n, and different columns (p1,p2, ..., p_d),observational mixed data matrix list, d is the types of variables, p_j is the dimension of varibles with the j-th type.
`types`	a d-dimensional character vector, specify the type of variables. For example, `types=c('gaussian','poisson', 'binomial')`, implies the components of `XList` are matrices with continuous, count and binomial values, respectively.
`q`	a positive integer or empty, specify the number of factors.
`offset`	a logical value, whether add an offset term (the total counts for each row in the count component of XList) when there are Poisson variables.
`epsELBO`	a positive real, specify the relative tolerance of ELBO function in the algorithm. Optional parameter with default as `1e-5`.
`maxIter`	a positive integer, specify the times of iteration. Optional parameter with default as 30.
`verbose`	a logical value with TRUE or FALSE, specify whether ouput the information in iteration process, (optional) default as TRUE.

Details

Overdispersion is prevalent in practical applications, particularly in fields like biomedical and genomics studies. To address this practical demand, we propose an overdispersed generalized factor model (OverGFM) for performing high-dimensional nonlinear factor analysis on overdispersed mixed-type data.

Value

return a list with class name 'overdispersedGFM' and including following components,

`hH`	a n*q matrix, the estimated factor matrix.
`hB`	a p*q matrix, the estimated loading matrix.
`hmu`	a p-dimensional vector, the estimated intercept terms.
`obj`	a real number, the value of objective function when the convergence achieves.
`q`	an integer, the used or estimated factor number.
`history`	a list including the following 7 components: (1)dB: the varied quantity of B in each iteration; (2)dH: the varied quantity of H in each iteration; (3)dc: the varied quantity of the objective function in each iteration; (4)c: the objective value in each iteration; (5) realIter: the real iterations to converge; (6)maxIter: the tolerance of maximum iterations; (7)elapsedTime: the elapsed time.

Note

nothing

Author(s)

Liu Wei

See Also

nothing

Examples

```
## mix of normal and Poisson

dat <- gendata(seed=1, n=60, p=60, type='norm_pois', q=2, rho=2)
## we set maxIter=2 for example.
gfm2 <- overdispersedGFM(dat$XList,  dat$types,  q=2, verbose = FALSE, maxIter=2)
measurefun(gfm2$hH, dat$H0, type='ccor')
measurefun(gfm2$hB, dat$B0, type='ccor')
```

`OverGFMchooseFacNumber` *Choose the Number of factors for Overdispersed Generalized Factor Models*

Description

This function is designed to chooose the number of factors for the overdispersed generalized factor model by using the singular value ratio (SVR) based method.

Usage

```
OverGFMchooseFacNumber(XList, types, q_max=15,offset=FALSE, epsELBO=1e-4, maxIter=30,
                       verbose = TRUE, threshold= 1e-2)
```

Arguments

`XList`	a list consisting of matrices with the same rows n, and different columns (p1,p2, ..., p_d),observational mixed data matrix list, d is the types of variables, p_j is the dimension of varibles with the j-th type.
`types`	a d-dimensional character vector, specify the type of variables. For example, `types=c('gaussian','poisson', 'binomial')`, implies the components of `XList` are matrices with continuous, count and binomial values, respectively.
`q_max`	a positive integer, specify the upper bound of the number of factors, defualt as 15.
`offset`	a logical value, whether add an offset term (the total counts for each row in the count component of XList) when there are Poisson variables.
`epsELBO`	a positive real, specify the relative tolerance of ELBO function in the algorithm. Optional parameter with default as `1e-5`.
`maxIter`	a positive integer, specify the times of iteration. Optional parameter with default as 30.
`verbose`	a logical value with TRUE or FALSE, specify whether ouput the information in iteration process, (optional) default as TRUE.
`threshold`	a postive real, the threshold that is used to filter the small singular values in the SVR criterion.

Value

return an integer value, the estimated number of factors.

Note

nothing

Author(s)

Liu Wei

See Also

nothing

Examples

```
## mix of normal and Poisson

dat <- gendata(seed=1, n=60, p=60, type='norm_pois', q=2, rho=2)
## we set maxIter=2 for example.
hq <- OverGFMchooseFacNumber(dat$XList, dat$types, verbose = FALSE, maxIter=2)
```

Index