Package ‘CMGFM’

Type Package
Title Covariate-Augmented Generalized Factor Model
Version 1.1
Date 2024-06-21
Author Wei Liu [aut, cre],
      Jiakun Jiang [aut],
      Dewei Xiang [aut],
      Xuancheng Zhou [aut]
Maintainer Wei Liu <LiuWeideng@gmail.com>
Description Covariate-augmented generalized factor model is designed to account for cross-modal heterogeneity, capture nonlinear dependencies among the data, incorporate additional information, and provide excellent interpretability while maintaining high computational efficiency.
BugReports https://github.com/feiyoung/CMGFM/issues
License GPL-3
Depends irlba, R (>= 3.5.0)
Imports MASS, stats, GFM, Rcpp (>= 1.0.10)
Suggests knitr, rmarkdown
LinkingTo Rcpp, RcppArmadillo
VignetteBuilder knitr
Encoding UTF-8
RoxygenNote 7.3.1
NeedsCompilation yes
Repository CRAN
Date/Publication 2024-06-25 15:00:05 UTC

Contents

  CMGFM ................................................................. 2
  gendata_cmgfm .................................................... 4
  MSVR ................................................................. 5
**CMGFM**

**Fit the CMGFM model**

**Description**

Fit the covariate-augumented generalized factor model

**Usage**

```r
CMGFM(
  XList,
  Z,
  types,
  numvarmat,
  q = 15,
  Alist = NULL,
  init = c("LFM", "GFM", "random"),
  maxIter = 30,
  epsELBO = 1e-08,
  verbose = TRUE,
  add_IC_iter = FALSE,
  seed = 1
)
```

**Arguments**

- **XList**: a list consisting of multiple matrices in which each matrix has the same type of values, i.e., continuous, or count, or binomial/binary values.
- **Z**: a matrix, the fixed-dimensional covariate matrix with control variables.
- **types**: a string vector, specify the variable type in each matrix in XList;
- **numvarmat**: a length(types)-by-d matrix, specify the number of variables in modalities that belong to the same type.
- **q**: an optional string, specify the number of factors; default as 15.
- **Alist**: an optional vector, the offset for each unit; default as full-zero vector.
- **init**: an optional character, specify the method in initialization.
- **maxIter**: the maximum iteration of the VEM algorithm. The default is 30.
- **epsELBO**: an optional positive value, tolerance of relative variation rate of the evidence lower bound value, default as '1e-8'.
- **verbose**: a logical value, whether output the information in iteration.
- **add_IC_iter**: a logical value, add the identifiability condition in iterative algorithm or add it after algorithm converges; default as FALSE.
- **seed**: an integer, set the random seed in initialization, default as 1;
Details

None

Value

return a list including the following components:

- $\beta_f$ - the estimated regression coefficient vector for each modality;
- $B_f$ - the estimated loading matrix for each modality;
- $M$ - the estimated modality-shared factor matrix;
- $X_f$ - the estimated modality-specified factor vector;
- $S$ - the estimated covariance matrix of modality-shared latent factors;
- $\Omega_m$ - the posterior variance of modality-specified latent factors;
- $m_f$ - the estimated intercept vector for each modality;
- $\Sigma_m$ - the variance of modality-specified factors;
- $\text{invLambd}_f$ - the inverse of the estimated variances of error for each modality.
- $\text{ELBO}$ - the ELBO value when algorithm stops;
- $\text{ELBO}_\text{seq}$ - the sequence of ELBO values.
- $\text{time_use}$ - the running time in model fitting;

References

None

See Also

None

Examples

```r
pveclist <- list('gaussian'=c(50, 150), 'poisson'=c(50, 150),
  'binomial'=c(100,60))
q <- 6
simgavec <- rep(1,3)
pvec <- unlist(pveclist)
datlist <- gendata_cmgfm(pveclist = pveclist, seed = 1, n = 300,d = 3,
  q = q, rho = rep(1,length(pveclist)), rho_z=0.2,
  simgavec=simgavec, sigma_eps=1)
XList <- datlist$XList
Z <- datlist$Z
numvarmat <- datlist$numvarmat
types <- datlist$types
rlist <- CMGFM(XList, Z, types=types, numvarmat, q=q)
str(rlist)
```
**gendata_cmgfm**

Generate simulated data from covariate-augmented generalized factor model

**Usage**

```r
gedata_cmgfm(
  seed = 1,
  n = 300,
  pveclist = list(gaussian = c(50, 150), poisson = c(50), binomial = c(100, 60)),
  q = 6,
  d = 3,
  rho = rep(1, length(pveclist)),
  rho_z = 1,
  sigmavec = rep(0.5, length(pveclist)),
  n_bin = 1,
  sigma_eps = 1,
  seed.para = 1
)
```

**Arguments**

- `seed` a positive integer, the random seed for reproducibility of data generation process.
- `n` a positive integer, specify the sample size.
- `pveclist` a named list, specify the number of modalities for each variable type and dimension of variables in each modality.
- `q` a positive integer, specify the number of modality-shared factors.
- `d` a positive integer, specify the dimension of covariate matrix.
- `rho` a numeric vector with length `length(pveclist)` and positive elements, specify the signal strength of loading matrix for each modality with the same variable type.
- `rho_z` a positive real, specify the signal strength of covariates.
- `sigmavec` a positive vector with length `length(pveclist)`, the variance of modality-specified latent factors.
- `n_bin` a positive integer, specify the number of trails in Binomial distribution.
- `sigma_eps` a positive real, the variance of overdispersion error.
- `seed.para` a positive integer, the random seed for reproducibility of data generation process by fixing the regression coefficient vector and loading matrices.

**Details**

None
Value

return a list including the following components:

- \( \mathbf{Xlist} \) - a list consisting of multiple matrices in which each matrix has the same type of values, i.e., continuous, or count, or binomial/binary values.
- \( \mathbf{Z} \) - a matrix, the fixed-dimensional covariate matrix with control variables;
- \( \mathbf{Alist} \) - the offset vector for each modality;
- \( \mathbf{B0list} \) - the true loading matrix for each modality;
- \( \mathbf{mu0} \) - the true intercept vector for each modality;
- \( \mathbf{U0} \) - the modality-specified factor vector;
- \( \mathbf{F0} \) - the modality-shared factor matrix;
- \( \mathbf{UpList} \) - the true intercept-loading matrix for each modality;
- \( \mathbf{beta} \) - the true regression coefficient vector for each modality;
- \( \mathbf{sigma_{eps}} \) - the standard deviation of error term;
- \( \mathbf{numvarmat} \) - a length(types)-by-d matrix, the number of variables in modalities that belong to the same type.

References

None

See Also

CMGFM

Examples

```r
n <- 300;
pveclist = list('gaussian'=c(50, 150), 'poisson'=c(50), 'binomial'=c(100, 60))
d <- 20; q <- 6;
datlist <- gendata_cmgfm(n=n, pveclist=pveclist, q=q, d=d)
str(datlist)
```

Description

Select the number of factors using maximum singular value ratio based method
Usage

\[
MSVR(\quad XList, \quad Z, \quad types, \quad numvarmat, \quad Alist = NULL, \quad q_{\text{max}} = 20, \quad threshold = 1e-05, \quad \ldots)
\]

Arguments

- **XList**: a list consisting of multiple matrices in which each matrix has the same type of values, i.e., continuous, or count, or binomial/binary values.
- **Z**: a matrix, the fixed-dimensional covariate matrix with control variables.
- **types**: a string vector, specify the variable type in each matrix in XList;
- **numvarmat**: a \(\text{length}(\text{types})\)-by-\(d\) matrix, specify the number of variables in modalities that belong to the same type.
- **Alist**: an optional vector, the offset for each unit; default as full-zero vector.
- **q_{\text{max}}**: an optional string, specify the maximum number of factors; default as 20.
- **threshold**: an optional positive value, a cutoff to filter the singular values that are smaller than it.
- **\ldots**: other arguments passed to CMGFM

Details

None

Value

return the estimated number of factors.

References

None

See Also

None
Examples

```r
pveclist <- list('gaussian'=c(50, 150), 'poisson'=c(50, 150),
                 'binomial'=c(100, 60))
q <- 6
sigmavec <- rep(1,3)
pvec <- unlist(pveclist)
datlist <- gendata_cmgfm(pveclist = pveclist, seed = 1, n = 300, d = 3,
                         q = q, rho = rep(1, length(pveclist)), rho_z=0.2,
                         sigmavec=sigmavec, sigma_eps=1)
XList <- datlist$XList
Z <- datlist$Z
numvarmat <- datlist$numvarmat
types <- datlist$types
hq <- MSVR(XList, Z, types=types, numvarmat, q_max=20)
print(c(q_true=q, q_est=hq))
```
Index

CMGFM, 2, 5

gendata_cmgfm, 4

MSVR, 5