Package 'MMGFM'

September 3, 2024

Type Package

Title Multi-Study Multi-Modality Generalized Factor Model

Version 1.1.0

Date 2024-08-25

Author Wei Liu [aut, cre],
Qingzhi Zhong [aut]

Maintainer Wei Liu <liuwei8@scu.edu.cn>

Description We introduce a generalized factor model designed to jointly analyze high-dimensional multi-modality data from multiple studies by extracting study-shared and specified factors. Our factor models account for heterogeneous noises and overdispersion among modality variables with augmented covariates. We propose an efficient and speedy variational estimation procedure for estimating model parameters, along with a novel criterion for selecting the optimal number of factors. More details can be referred to Liu et al. (2024) <doi:10.48550/arXiv.2408.10542>.

License GPL-3

Depends irlba, R (>= 3.5.0)

Imports MASS, stats, GFM, MultiCOAP, Rcpp (>= 1.0.10)

Suggests knitr, rmarkdown

LinkingTo Rcpp, RcppArmadillo

Encoding UTF-8

RoxygenNote 7.3.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2024-09-03 15:50:10 UTC

Contents

gendata_mmgfm *Generate simulated data*

Description

Generate simulated data from MMGFM models

Usage

```
gendata_mmgfm(
  seed = 1,
  nvec = c(300, 200),
  pveclist = list(gaussian = c(50, 150), poisson = c(50), binomial = c(100, 60)),
  q = 6,
  d = 3,
  qs = rep(2, length(nvec)),
  rho = rep(1, length(pveclist)),
  rho_z = 1,
  sigmavec = rep(0.5, length(pveclist)),
  n_bin = 1,
  sigma_eps = 1,
  heter_error = FALSE
)
```

Arguments

`seed`	a postive integer, the random seed for reproducibility of data generation process.
`nvec`	a vector with postive integers, specify the sample size in each study/source.
`pveclist`	a named list, specify the number of modalities for each type and variable dimension in each type of modatlity.
`q`	a postive integer, specify the number of study-shared factors.
`d`	a postive integer, specify the dimension of covariate matrix.
`qs`	a vector with postive integers, specify the number of study-specified factors.
`rho`	a numeric vector with `length(pveclist)` and positive elements, specify the signal strength of loading matrices for each modality type.
`rho_z`	a positive real, specify the signal strength of covariates.
`sigmavec`	a positive real vector with `length(pveclist)`, specify the variance of study-specified and modality variable-shared factors; default as 0.5 for each element.
`n_bin`	a positive integer, specify the number of trails when generate Binomial modality matrix; default as 1.
`sigma_eps`	a positive real, the variance of overdispersion error; default as 1.
`heter_error`	a logical value, whether to generate the heterogeneous error; default as FALSE.

Value

return a list including the following components:

- hbeta - a M-length list composed by the estimated regression coefficient matrix for each modality;
- hA - a M-length list composed by the loading matrix corresponding to study-shared factors for each modality;
- hB - a S-length list composed by a M-length loading matrix list corresponding to study-specified factors for each study;
- hF - a S-length list composed by the posterior estimation of study-shared factor matrix for each study;
- hH - a S-length list composed by the posterior estimation of study-specified factor matrix for each study;
- hSigma - a S-length list composed by the estimated posterior variance of the study-shared factor;
- hPhi - a S-length list composed by the estimated posterior variance of study-specified factor;
- hv - a S-length list composed by a M-length vector list corresponding to the posterior estimation of study-specified and modality variable-shared factor for each study and modality;
- hzeta - the estimated posterior variance for study-specified and modality variable-shared factor;
- hsigma2 - the estimated variance for study-specified and modality variable-shared factor;
- hinvLambda - a S-length list composed by a M-length vector list corresponding to the inverse of the estimated variances of error;
- S - the approximated posterior covariance for each row of F;
- ELBO - the ELBO value when algorithm stops;
- ELBO_seq - the sequence of ELBO values.
- time_use - the running time in model fitting of SpaCOAP;

Examples

```
q <- 3; qsvec<-rep(2,3)
nvec <- c(100, 120, 100)
pveclist <-  list('gaussian'=rep(150, 1),'poisson'=rep(50, 2),'binomial'=rep(60, 2))
datlist <- gendata_mmgfm(seed = 1,  nvec = nvec, pveclist =pveclist,
                         q = q,  d= 3,qs = qsvec,  rho = rep(3,length(pveclist)), rho_z=0.5,
                            sigmavec=rep(0.5, length(pveclist)),  sigma_eps=1)
```

MMGFM *Fit the high-dimensional multi-study multi-modality covariate-augmented generalized factor model*

Description

Fit the high-dimensional multi-study multi-modality covariate-augmented generalized factor model via variational inference.

Usage

```
MMGFM(
  XList,
  ZList,
  numvarmat,
  tauList = NULL,
  q = 15,
  qsvec = rep(2, length(XList)),
  init = c("MSFRVI", "random", "LFM"),
  epsELBO = 1e-12,
  maxIter = 30,
  verbose = TRUE,
  seed = 1
)
```

Arguments

`XList` a S-length list with each component a m-length list composed by a combined modality matrix of the same type modalities, which is the observed matrix from each source/study and each modality, where m is the number of modality types.

`ZList` a S-length list with each component a matrix that is the covariate matrix from each study.

`numvarmat` a m-by-T matrix with rownames modality types that specifies the variable number for each modality of each modality type, where m is the number of modality types, T is the maximum number of modalities for one of modality types .

`tauList` an optional S-length list with each component a m-length list correponding the offset term for each combined modality of each study; default as full-zero matrix.

`q` an optional string, specify the number of study-shared factors; default as 15.

`qsvec` a integer vector with length S, specify the number of study-specifed factors; default as 2.

`init` an optional string, specify the initialization method, supporting "MSFRVI", "random" and "LFM", default as "MSFRVI".

`epsELBO` an optional positive vlaue, tolerance of relative variation rate of the envidence lower bound value, defualt as '1e-5'.

`maxIter` the maximum iteration of the VEM algorithm. The default is 30.

`verbose` a logical value, whether output the information in iteration.

`seed` an optional integer, specify the random seed for reproducibility in initialization.

Details

If `init="MSFRVI"`, it will use the results from multi-study linear factor model in MultiCOAP package as initial values; If `init="LFM"`, it will use the results from linear factor model by combing data from all studies as initials.

Value

return a list including the following components:

- `hbeta` - a M-length list composed by the estimated regression coefficient matrix for each modality;
- `hA` - a M-length list composed by the loading matrix corresponding to study-shared factors for each modality;
- `hB` - a S-length list composed by a M-length loading matrix list corresponding to study-specified factors for each study;
- `hF` - a S-length list composed by the posterior estimation of study-shared factor matrix for each study;
- `hH` - a S-length list composed by the posterior estimation of study-specified factor matrix for each study;
- `hSigma` - a S-length list composed by the estimated posterior variance of the study-shared factor;
- `hPhi` - a S-length list composed by the estimated posterior variance of study-specified factor;
- `hv` - a S-length list composed by a M-length vector list corresponding to the posterior estimation of study-specified and modality variable-shared factor for each study and modality;
- `hzeta` - the estimated posterior variance for study-specified and modality variable-shared factor;
- `hsigma2` - the estimated variance for study-specified and modality variable-shared factor;
- `hinvLambda` - a S-length list composed by a M-length vector list corresponding to the inverse of the estimated variances of error;
- `S` - the approximated posterior covariance for each row of F;
- `ELBO` - the ELBO value when algorithm stops;
- `ELBO_seq` - the sequence of ELBO values.
- `time_use` - the running time in model fitting of SpaCOAP;

References

None

See Also

None

Examples

```
q <- 3; qsvec<-rep(2,3)
nvec <- c(100, 120, 100)
pveclist <-  list('gaussian'=rep(150, 1),'poisson'=rep(50, 2),'binomial'=rep(60, 2))
datlist <- gendata_mmgfm(seed = 1,  nvec = nvec, pveclist =pveclist,
                      q = q,  d= 3,qs = qsvec,  rho = rep(3,length(pveclist)), rho_z=0.5,
                         sigmavec=rep(0.5, length(pveclist)),  sigma_eps=1)
XList <- datlist$XList
ZList <- datlist$ZList
numvarmat <- datlist$numvarmat
### For illustration, we set maxIter=3. Set maxIter=50 when running formally
reslist1 <- MMGFM(XList, ZList=ZList, numvarmat, q=q, qsvec = qsvec, init='MSFRVI',maxIter = 3)
str(reslist1)
```

---

selectFac.MMGFM   *Select the number of study-shared and study-specified factors for MMGFM*

Description

Select the number of study-shared and study-specified factors for the high-dimensional multi-study multi-modality covariate-augmented generalized factor model.

Usage

```
selectFac.MMGFM(
  XList,
  ZList,
  numvarmat,
  q.max = 15,
  qsvec.max = rep(4, length(XList)),
  threshold.vec = c(0.01, 0.001),
  tauList = NULL,
  init = c("MSFRVI", "random", "LFM"),
  epsELBO = 1e-12,
  maxIter = 30,
  verbose = TRUE,
  seed = 1
)
```

Arguments

XList	a S-length list with each component a m-length list composed by a combined modality matrix of the same type modalities, which is the observed matrix from each source/study and each modality, where m is the number of modality types.
ZList	a S-length list with each component a matrix that is the covariate matrix from each study.

`numvarmat` a m-by-T matrix with rownames modality types that specifies the variable number for each modality of each modality type, where m is the number of modality types, T is the maximum number of modalities for one of modality types .

`q.max` an optional integer, specify the upper bound for the number of study-shared factors; default as 15.

`qsvec.max` an optional integer vector with length S, specify the upper bound for the number of study-specifed factors; default as 4 for each study.

`threshold.vec` an optional real vector with length 2, specify the threshold for the singular values of study-shared loading and study-specified loading matrices, respectively.

`tauList` an optional S-length list with each component a m-length list correponding the offset term for each combined modality of each study; default as full-zero matrix.

`init` an optional string, specify the initialization method, supporting "MSFRVI", "random" and "LFM", default as "MSFRVI".

`epsELBO` an optional positive vlaue, tolerance of relative variation rate of the envidence lower bound value, defualt as '1e-5'.

`maxIter` the maximum iteration of the VEM algorithm. The default is 30.

`verbose` a logical value, whether output the information in iteration.

`seed` an optional integer, specify the random seed for reproducibility in initialization.

Value

return a list with two components: q and qs.vec.

Examples

```
q <- 3; qsvec<-rep(2,3)
nvec <- c(100, 120, 100)
pveclist <-  list('gaussian'=rep(150, 1),'poisson'=rep(50, 2),'binomial'=rep(60, 2))
datlist <- gendata_mmgfm(seed = 1,  nvec = nvec, pveclist =pveclist,
                      q = q,  d= 3,qs = qsvec,  rho = rep(3,length(pveclist)), rho_z=0.5,
                         sigmavec=rep(0.5, length(pveclist)),  sigma_eps=1)
XList <- datlist$XList
ZList <- datlist$ZList
numvarmat <- datlist$numvarmat
### For illustration, we set maxIter=3. Set maxIter=50 when running formally
selectFac.MMGFM(XList, ZList=ZList, numvarmat, q.max=6, qsvec.max  = rep(4,3),
init='MSFRVI',maxIter = 3)
```

Index