Package 'MultiCOAP'

January 20, 2025

Type Package

Title High-Dimensional Covariate-Augmented Overdispersed Multi-Study Poisson Factor Model

Version 1.1

Date 2024-03-06

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

Maintainer Wei Liu <liuweideng@gmail.com>

Description We introduce factor models designed to jointly analyze high-dimensional count data from multiple studies by extracting study-shared and specified factors. Our factor models account for heterogeneous noises and overdispersion among counts 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 and the rank of regression coefficient matrix. More details can be referred to Liu et al. (2024) <doi:10.48550/arXiv.2402.15071>.

License GPL-3

Depends irlba, R (>= 3.5.0)

Imports MASS, Rcpp (>= 1.0.10)

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

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

LinkingTo Rcpp, RcppArmadillo

Encoding UTF-8

RoxygenNote 7.1.2

NeedsCompilation yes

Repository CRAN

Date/Publication 2024-03-07 10:40:05 UTC

Contents

`gendata_simu_multi2` *Generate simulated data*

Description

Generate simulated data from covariate-augmented Poisson factor models

Usage

```
gendata_simu_multi2(
  seed = 1,
  nvec = c(100, 300),
  a_interval = c(0, 1),
  p = 50,
  d = 3,
  q = 3,
  qs = rep(2, length(nvec)),
  rank0 = 3,
  rho = c(rhoA = 1, rhoB = 1, rhoZ = 1),
  sigma2_eps = 1,
  seed.beta = 1
)
```

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.
`a_interval`	a numeric vector with two elements, specify the range of offset term values in each study.
`p`	a postive integer, specify the dimension of count variables.
`d`	a postive integer, specify the dimension of covariate matrix.
`q`	a postive integer, specify the number of study-shared factors.
`qs`	a vector with postive integers, specify the number of study-specified factors.
`rank0`	a postive integer, specify the rank of the coefficient matrix.
`rho`	a numeric vector with length 3 and positive elements, specify the signal strength of regression coefficient and loading matrices, respectively.
`sigma2_eps`	a positive real, the variance of overdispersion error.
`seed.beta`	a postive integer, the random seed for fixing the regression coefficient matrix and loading matrix generation.

Details

None

Value

return a list including the following components: (1) Xlist, the list consisting of high-dimensional count matrices from multiple studies; (2) aList: the known normalization term (offset) for each study; (3) Zlist, the list consisting of covariate matrix; (4) bbeta0, the true regression coefficient matrix; (5) A0, the loading matrix of study-shared factors; (6) Blist, the list consisting of loading matrices of study-specified factors; (7)lambdavec, the variance vector of the random error vector; (8)Flist, the list composed by study-shared factor matrices; (9) Hlist, the list composed by study-specified factor matrices; (10) rank0, the rank of underlying regression coefficient matrix; (11) q, the number of study-shared factors; (12)qs, the numbers of study-specified factors.

References

None

See Also

None

Examples

```
seed <- 1; nvec <- c(100,300); p<- 300;
d <- 3; q<- 3; qs <- rep(2,2)
datlist <- gendata_simu_multi2(seed=seed, nvec=nvec, p=p, d=d, q=3, qs=qs)
str(datlist)
```

MSFRVI *Fit the multi-study covariate-augmented linear factor model via variational inference*

Description

Fit the multi-study covariate-augmented linear factor model via variational inference

Usage

```
MSFRVI(
  XList,
  ZList,
  q = 15,
  qs = rep(2, length(XList)),
  rank_use = NULL,
  aList = NULL,
  epsELBO = 1e-05,
  maxIter = 30,
```

```
  verbose = TRUE,
  seed = 1
)
```

Arguments

XList	A length-M list, where each component represents a matrix and is the observed response matrix from each source/study. Ideally, each matrix should be continuous.
ZList	a length-M list with each component a matrix that is the covariate matrix from each study.
q	an optional integer, specify the number of study-shared factors; default as 15.
qs	a integer vector with length M, specify the number of study-specifed factors; default as 2.
rank_use	an optional integer, specify the rank of the regression coefficient matrix; default as NULL, which means that rank is the dimension of covariates in Z.
aList	an optional length-M list with each component a vector, the normalization factors of each study; default as full-one vector.
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

None

Value

return a list including the following components: (1) F, a list composed by the posterior estimation of study-shared factor matrix for each study; (2) H, a list composed by the posterior estimation of study-specified factor matrix for each study; (3) Sf, a list consisting of the posterior estimation of covariance matrix of study-shared factors for each study; (4) Sh, a list consisting of the posterior estimation of covariance matrix of study-specified factors for each study; (5) A, the loading matrix corresponding to study-shared factors; (6) B, a list composed by the loading matrices corresponding to the study-specified factors; (7) bbeta, the estimated regression coefficient matrix; (8) invLambda, the inverse of the estimated variances of error; (9) ELBO: the ELBO value when algorithm stops; (7) ELBO_seq: the sequence of ELBO values. (11) qrlist, the number of factors and rank of regression coefficient matrix used in fitting; (12) time.use, the elapsed time for model fitting.

References

None

See Also

MultiCOAP

Examples

```
seed <- 1; nvec <- c(100,300); p<- 300;
d <- 3; q<- 3; qs <- rep(2,2)
datlist <- gendata_simu_multi2(seed=seed, nvec=nvec, p=p, d=d, q=3, qs=qs)
XList <- lapply(datlist$Xlist,  function(x) log(1+x))
fit_msfavi <- MSFRVI(XList, ZList = datlist$Zlist, q=3, qs=qs, rank_use = d)
str(fit_msfavi)
```

`MultiCOAP` *Fit the multi-study covariate-augmented overdispersed Poisson factor model via variational inference*

Description

Fit the high-dimensional multi-study covariate-augmented overdispersed Poisson factor model via variational inference.

Usage

```
MultiCOAP(
  XcList,
  ZList,
  q = 15,
  qs = rep(2, length(XcList)),
  rank_use = NULL,
  aList = NULL,
  init = c("MSFRVI", "LFM"),
  epsELBO = 1e-05,
  maxIter = 30,
  verbose = TRUE,
  seed = 1
)
```

Arguments

`XcList`	a length-M list with each component a count matrix, which is the observed count matrix from each source/study.
`ZList`	a length-M list with each component a matrix that is the covariate matrix from each study.
`q`	an optional integer, specify the number of study-shared factors; default as 15.
`qs`	a integer vector with length M, specify the number of study-specifed factors; default as 2.
`rank_use`	an optional integer, specify the rank of the regression coefficient matrix; default as NULL, which means that rank is the dimension of covariates in Z.
`aList`	an optional length-M list with each component a vector, the normalization factors of each study; default as full-one vector.

init	an optional string, specify the initialization method, 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 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: (1) F, a list composed by the posterior estimation of study-shared factor matrix for each study; (2) H, a list composed by the posterior estimation of study-specified factor matrix for each study; (3) Sf, a list consisting of the posterior estimation of covariance matrix of study-shared factors for each study; (4) Sh, a list consisting of the posterior estimation of covariance matrix of study-specified factors for each study; (5) A, the loading matrix corresponding to study-shared factors; (6) B, a list composed by the loading matrices corresponding to the study-specified factors; (7) bbeta, the estimated regression coefficient matrix; (8) invLambda, the inverse of the estimated variances of error; (9) ELBO: the ELBO value when algorithm stops; (7) ELBO_seq: the sequence of ELBO values. (11) qrlist, the number of factors and rank of regression coefficient matrix used in fitting; (12) time.use, the elapsed time for model fitting.

References

None

See Also

MSFRVI

Examples

```
seed <- 1; nvec <- c(100,300); p<- 300;
d <- 3; q<- 3; qs <- rep(2,2)
datlist <- gendata_simu_multi2(seed=seed, nvec=nvec, p=p, d=d, q=3, qs=qs)
fit_mcoap <- MultiCOAP(datlist$Xlist, ZList = datlist$Zlist, q=3, qs=qs, rank_use = d)
str(fit_mcoap)
```

Index