# Package ‘MultiRFM’

December 3, 2025

**Type** Package

**Title** High-Dimensional Multi-Study Robust Factor Model

**Version** 1.1.0

**Date** 2025-11-28

**Author** Wei Liu [aut, cre],
Xiaolu Jiang [aut]

**Maintainer** Wei Liu <liuweideng@gmail.com>

**Description** We introduce a high-dimensional multi-study robust factor model, which learns latent features and accounts for the heterogeneity among source. It could be used for analyzing heterogeneous RNA sequencing data. More details can be referred to Jiang et al. (2025) <doi:10.48550/arXiv.2506.18478>.

**License** GPL-3

**Depends** R (>= 3.5.0)

**Imports** MASS, irlba, LaplacesDemon, mixtools, mvtnorm, Rcpp (>= 1.0.8.3)

**URL** https://github.com/feiyoung/MultiRFM

**BugReports** https://github.com/feiyoung/MultiRFM/issues

**Encoding** UTF-8

**Suggests** knitr, rmarkdown

**LinkingTo** Rcpp, RcppArmadillo

**VignetteBuilder** knitr

**RoxygenNote** 7.3.2

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2025-12-03 20:50:07 UTC

## Contents

## gendata_simu_multi *Generate Simulated Multi-Study Factor Analysis Data*

### Description

Generate simulated data for multi-study factor analysis under different error distributions. The data follows a factor model with common factors (shared across studies) and study-specific factors (unique to each study), plus noise.

### Usage

```
gendata_simu_multi(
  seed = 1,
  nvec = c(100, 300),
  p = 50,
  q = 3,
  qs = rep(2, length(nvec)),
  err.type = c("gaussian", "mvt", "exp", "t", "mixnorm", "pareto"),
  rho = c(1, 1),
  sigma2_eps = 0.1,
  nu = 1
)
```

### Arguments

`seed` Integer, default = 1. Random seed for reproducibility of simulated data.

`nvec` Numeric vector (length >= 2). Sample sizes of each study (e.g., ‘c(150, 200)‘ for 2 studies with 150 and 200 samples).

`p` Integer, default = 50. Number of variables (features) in the data.

`q` Integer, default = 3. Number of common factors (shared across all studies).

`qs` Numeric vector with length equal to ‘length(nvec)‘, default = ‘rep(2, length(nvec))‘. Number of study-specific factors for each study (e.g., ‘c(2,2)‘ for 2 studies each with 2 specific factors).

`err.type` Character, default = "gaussian". Error distribution type, one of: - "gaussian": Gaussian (normal) distribution;

- "mvt": Multivariate t-distribution;

- "exp": Exponential distribution (centered to mean 0);

- "t": Univariate t-distribution (independent across variables);

- "mixnorm": Mixture of two normal distributions;

- "pareto": Pareto distribution (centered to mean 0).

`rho` Numeric vector of length 2, default = ‘c(1,1)‘. Scaling factors for: - ‘rho1‘: Common factor loadings (matrix ‘A0‘); - ‘rho2‘: Study-specific factor loadings (matrix list ‘Blist0‘).

`sigma2_eps` Numeric, default = 0.1. Variance of the error term (controls noise level).

`nu` Integer, default = 1. Degrees of freedom for t-distribution ("mvt" or "t" ‘err.type‘). Ignored for other error distributions.

## Details

The simulated data follows the multi-study factor model:

Xs = mu0s + Fs x A0 + Hs x B0s + epsilons

True parameters (‘A0‘, ‘Blist0‘, ‘mu0‘) are generated with orthogonal constraints to ensure identifiability.

## Value

A list containing the simulated data and true parameter values (for model evaluation):

- `Xlist`: List of matrices. Each element is a data matrix (ns × p) for study s, where ns = ‘nvec[s]‘ (sample size of study s), p = number of variables.
- `mu0`: Matrix (p × S). True mean vector for each variable (row) in each study (column), where S = ‘length(nvec)‘ (number of studies).
- `A0`: Matrix (p × q). True common factor loadings (shared across all studies) — constructed as the first q columns of an orthogonal matrix (‘A1‘) generated internally. This is the "ground truth" that modeling functions (e.g., MultiRFM) aim to estimate.
- `Blist0`: List of matrices. Each element is a true study-specific factor loadings matrix (p × qs[s]) for study s. Constructed from orthogonal matrices (similar to ‘A0‘) and scaled by ‘rho[2]‘. Another "ground truth" for model evaluation.
- `Flist`: List of matrices. Each element is a true common factor score matrix (ns × q) for study s, generated from a standard normal distribution. These are the latent common factor values used to generate ‘Xlist‘.
- `Hlist`: List of matrices. Each element is a true study-specific factor score matrix (ns × qs[s]) for study s, generated from a standard normal distribution. Latent specific factor values used to generate ‘Xlist‘.
- q: Integer. Number of common factors used for data generation (same as input ‘q‘, for reference).
- qs: Numeric vector. Number of study-specific factors used for data generation (same as input ‘qs‘, for reference).

## Author(s)

Wei Liu

## Examples

```
# Example 1: Gaussian error (2 studies, 100/200 samples, 50 variables)
set.seed(123)
sim_data <- gendata_simu_multi(
  seed = 123,
  nvec = c(100, 200),
  p = 50,
  q = 3,          # 3 common factors
  qs = c(2, 2),   # 2 specific factors per study
  err.type = "gaussian",
  rho = c(1, 1),
  sigma2_eps = 0.1
)
str(sim_data)  # Check structure of simulated data

# Extract true parameters for model evaluation
true_A <- sim_data$A0        # True common loadings
true_B1 <- sim_data$Blist0[[1]]  # True specific loadings (study 1)
```

## MultiRFM — *Fit the high-dimensional multi-study robust factor model*

### Description

Fit the high-dimensional multi-study robust factor model which learns latent features and accounts for the heterogeneity among source.

### Usage

```
MultiRFM(
  XList,
  q = 15,
  qs = rep(2, length(XList)),
  epsELBO = 1e-05,
  maxIter = 30,
  verbose = TRUE,
  seed = 1
)
```

### Arguments

`XList` A length-M list, where each component represents a matrix and is the

`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.

`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;default as 1.

## 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) mu,the mean of XList;(8) ELBO: the ELBO value when algorithm stops; (9) ELBO_seq: the sequence of ELBO values. (10) time_use, the elapsed time for model fitting.

## Examples

```
p <- 100
nvec <- c(150,200); qs <- c(2,2)
datList <- gendata_simu_multi(seed=1, nvec=nvec, p=p, q=3, qs=qs, rho=c(5,5),
       err.type='mvt', sigma2_eps = 1, nu=3)
XList <- datList$Xlist;
res <- MultiRFM(XList, q=3, qs= qs)
str(res)
```

---

# selectFac.MultiRFM *Select the number of factors*

---

## Description

Select the number of factors that are shared among studies q and thos that are specific to individual studies(qs).More details are in Section 3.1 of the article.

## Usage

```
selectFac.MultiRFM(
  XList,
  q_max = 15,
  qs_max = 4,
  method = c("SSVR", "CUP"),
  threshold = 1e-05,
  cup.upper = 0.95,
  epsELBO = 1e-05,
  maxIter = 30,
  verbose = TRUE,
  seed = 1
)
```

## Arguments

`XList` A length-M list, where each component represents a matrix and is the

`q_max` an optional integer, specify the maximum number of study-shared factors; default as 15.

`qs_max` an optional integer, specify the maximum number of study-specified factors; default as 4.

`method` an optional character, contains the methods of "SSVR" and "CUP", where ‘SSVR‘ is the sequential singular value ratio method while ‘CUP‘ is the criterion based on cumulative proportion of explained variance.

`threshold` the cutoff of the singular values, where the singular values less than this value will be removed.

`cup.upper` upper limit of the cumulative proportion of explained variance.

`epsELBO` an optional positive value, tolerance of relative variation rate of the evidence 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;default as 1.

## Details

None

## Value

return a list contains the following components:(1) q, the number of shared factors; (2) qs,the number of specified factors.

## Examples

```
p <- 100
nvec <- c(150,200); qs <- c(2,2)
datList <- gendata_simu_multi(seed=1, nvec=nvec, p=p, q=3, qs=qs, rho=c(5,5),
        err.type='mvt', sigma2_eps = 1, nu=3)
XList <- datList$Xlist;
## Set maxIter=5 for demonstration while set it to 30 in the formal run.
hqlist <- selectFac.MultiRFM(XList, q_max=6, qs_max= rep(4,2), maxIter = 5) #
str(hqlist)
```

# Index