

# **A Tutorial for Bayesian Integrative Factor Models**

Mavis Liang

2025-01-10

# Table of contents

<b>Welcome</b>	<b>3</b>
Quick start . . . . .	3
A summary of the models . . . . .	3
<b>1 Package Installation</b>	<b>4</b>
1.1 Stack FA, Ind FA, and BMSFA . . . . .	4
1.2 PFA . . . . .	4
1.3 MOM-SS . . . . .	5
1.4 SUFA . . . . .	5
1.5 Tetris . . . . .	5
1.6 Other utility packages . . . . .	5
<b>2 Case study: nutrition data</b>	<b>6</b>
2.1 Loading and previewing the data . . . . .	6
2.2 Data preprocessing . . . . .	7
2.3 Model fitting . . . . .	10
2.4 Post processing . . . . .	11
2.5 Visualization . . . . .	20
2.6 Mean squared error (MSE) . . . . .	20
<b>3 Case study: gene expression data</b>	<b>25</b>
<b>References</b>	<b>26</b>

# Welcome

This is the tutorial to guide statisticians to use Bayesian integrative factor models. Including Tetris (2023).

```
1 + 1
```

```
[1] 2
```

## Quick start

### Step 1: Prepare the package and data

A small simulated data can be downloaded here:

[RDS format](#)

### Step 2: Run BMSFA

### Step 3: Post-processing

### Step 4: Visualization

```
1 + 1
```

```
[1] 2
```

## A summary of the models

# 1 Package Installation

I have this separate section to introduce the package installation because installing some packages requires extra efforts.

## 1.1 Stack FA, Ind FA, and BMSFA

The Stack FA, Ind FA, and BMSFA models are implemented in the MSFA package, which is not available on CRAN. We can install it from GitHub using the `remotes` package. We use `sp_fa()` within the MSFA package to fit the Stack FA and Ind FA model, and `sp_msfa()` to fit the BMSFA model.

```
install.packages("remotes")
remotes::install_github("rdevito/MSFA")
library(MSFA)
```

## 1.2 PFA

PFA does not provide any downloadable R packages and we need to download the R scripts from their GitHub repository, put them in the same directory as the main script, and source them for use.

We only need the three files: `FBPFA-PFA.R`, `FBPFA-PFA with fixed latent dim.R`, and `PFA.cpp`, which can be found in <https://github.com/royarkaprava/Perturbed-factor-models>. The `FBPFA-PFA.R` file contains the full Bayesian inference algorithm for the PFA model, directly set the latent dimensions equal to the dimensions of the original data. The `FBPFA-PFA with fixed latent dim.R` file contains the same algorithm that requires to set numbers of common factors  $K$ . We also notice that two versions of the models are both `PFA()`, and some functions in the `FBPFA-PFA with fixed latent dim.R` file depend on the `FBPFA-PFA.R` file. Therefore, since we want to run the dimension reduction version of the model, we must source the `FBPFA-PFA.R` file first, and then source the `FBPFA-PFA with fixed latent dim.R` file.

```
# Suppose the files are in the same directory as the main script
source("FBPFA-PFA.R")
source("FBPFA-PFA with fixed latent dim.R")
```

## 1.3 MOM-SS

```
BiocManager::install("sparseMatrixStats") # Dependency
install.packages("mombf")

install.packages("devtools")

devtools::install_github("AleAviP/BFR.BE")
library(BFR.BE)
```

## 1.4 SUFA

To install SUFA on Linux, you need to also install extra dependencies like PROJ, sqlite3 and GDAL onto PATH. On Windows, you might need to do several updates, particularly the updates for terra. We can skip building the vignettes to save time, as it contains the computation of a large dataset.

```
devtools::install_github("noirritchandra/SUFA", build_vignettes = FALSE)
library(SUFA)
```

## 1.5 Tetris

Similar to PFA, Tetris does not provide any downloadable R packages and we need to download the R scripts from their GitHub repository, put them in the same directory as the main script, and source them for use. The R scripts can be found in <https://github.com/igrabski/tetris/tree/main>.

```
# Suppose the files are in the same directory as the main script
source("Tetris.R")
```

## 1.6 Other utility packages

```
library(tidyverse)
library(Matrix)#for the bdiag function
```

## 2 Case study: nutrition data

### 2.1 Loading and previewing the data

The data used in this section is from a large multi-site study investigating health and diet among Hispanic/Latino adults in the United States(2022). This data is not publicly available. Please contact the authors of the original study for access.

However, you can use this simulated data instead: [simulated\\_nutrition\\_data.rds](#). To read in the data, use `readRDS()`. The loaded object contains data in both `Y_mat` and `Y_list` formats, as well as covariates and other information. It is generated with the Scenario 4 in the manuscript (see [code](#)). Note that instead of having 6 studies, this simulated data has 12 studies.

```
load("./Data/dataLAT_projale2.rda")
```

The resulting object is a list of 6 data frames, each corresponding to a different study. Each data frame contains information about the nutritional intake of individuals, and the columns represent different nutrients. From Study 1 to Study 6, the number of individuals ( $N_s$ ) are 1364, 1517, 2210, 5184, 2478, and 959, respectively, and the number of nutrients ( $P$ ) are all 42.

```
# Check how many studies in the list
length(X_s2)
```

```
[1] 6
```

```
# Dimension of each study
lapply(X_s2, dim)
```

```
[[1]]
[1] 1364  42
```

```
[[2]]
[1] 1517  42
```

```
[[3]]  
[1] 2210 42
```

```
[[4]]  
[1] 5184 42
```

```
[[5]]  
[1] 2478 42
```

```
[[6]]  
[1] 959 42
```

Let's take a look at the first few rows of the first data frame to get an idea of the data structure.

```
x_s2[[1]][1:5, 1:5]
```

	Animal Protein (g)	Vegetable Protein (g)	Cholesterol (mg)	SCSFA	MCSFA
1	28.9560	14.7440	256.761	0.2665	0.939
2	33.6675	8.9710	104.217	0.2180	0.520
3	70.0000	31.0635	207.902	0.9845	1.692
4	20.6700	13.8240	148.921	0.0625	0.239
5	15.4250	10.5550	65.060	0.0090	0.033

We note that the data we have available is different from the original data (cite). The original data is a collection of 12 studies, and there are known covariates for each individuals, like the one we simulated in the previous section. However, for the purpose of this case study, the data we used are collapsed into 6 studies, and only the nutritional intake data are available.

## 2.2 Data preprocessing

Some individuals have missing values for all nutrients, thus we will remove these individuals from the data. Also, there are some nutrition intake are less than zero, for which we will replace with 0. We then apply a log transformation to the data.

We first count how many NA values and negative values are in each study.

```
count_na_and_negatives <- function(df) {
  # Count NA values
  na_count <- sum(is.na(df))
  # Count negative values
  negative_count <- sum(df < 0, na.rm = TRUE)

  # Print counts
  cat("Number of NAs:", na_count, "\n")
  cat("Number of negative values:", negative_count, "\n")
}
invisible(lapply(X_s2, count_na_and_negatives))
```

```
Number of NAs: 1344
Number of negative values: 0
Number of NAs: 1344
Number of negative values: 1
Number of NAs: 1344
Number of negative values: 0
Number of NAs: 1344
Number of negative values: 2
Number of NAs: 1344
Number of negative values: 1
Number of NAs: 1344
Number of negative values: 0
```

We will define a function to process the data, which removes rows where all values are NA. We also define a function that replaces negative values with 0, and applies a log transformation to the data.

```
process_study_data <- function(df) {
  # Remove rows where all values are NA
  cleaned_df <- df[!apply(df, 1, function(row) all(is.na(row))), , drop = FALSE]
  # Count remaining rows
  remaining_rows <- nrow(cleaned_df)
  # Print results for the study
  cat("Remaining rows:", remaining_rows, "\n")
  return(cleaned_df)
}
Y_list <- lapply(X_s2, process_study_data)
```

```
Remaining rows: 1332
```



```
Remaining rows: 1485
Remaining rows: 2178
Remaining rows: 5152
Remaining rows: 2446
Remaining rows: 927
```

The numbers of individuals in each study left for analysis ( $N_s$ ) are 1332, 1485, 2178, 5152, 2446, and 927, respectively.

```
# Replace negative values with 0, then log(x+0.01) + 0.01
replace_negatives <- function(df) {
  # Replace negative values with 0
  df[df < 0] <- 0
  # Apply log transformation. Add 0.01 to avoid log(0).
  transformed_df <- log(df + 0.01)
  return(transformed_df)
}

Y_list <- lapply(Y_list, replace_negatives)
```

```
# Check the processed data
invisible(lapply(Y_list, count_na_and_negatives))
```

```
Number of NAs: 0
Number of negative values: 11910
Number of NAs: 0
Number of negative values: 11222
Number of NAs: 0
Number of negative values: 15006
Number of NAs: 0
Number of negative values: 36230
Number of NAs: 0
Number of negative values: 19349
Number of NAs: 0
Number of negative values: 6707
```

Now we don't have any NA values or negative values in the data.

The assumptions for factor models require that each variable has a mean of 0. Therefore, for each study, we will center the data for each column. We note that for some model (Stack FA, Ind FA, BMSFA, and Tetris), this step is handled internally, and for MOM-SS, the random intercepts are estimated, so we do not need to center the data.

```

Y_list_scaled <- lapply(
  Y_list, function(x) scale(x, center = TRUE, scale = FALSE)
)
Y_mat_scaled <- Y_list_scaled %>% do.call(rbind, .) %>% as.matrix()

```

## 2.3 Model fitting

We recommend running model fitting and post-processing in a high-performance computing environment, as the model fitting process can be computationally intensive. PFA and Tetris are particularly computationally expensive, where PFA requires more than 10 hours to run, and Tetris requires more than 4 days to run. Other models can be completed in half an hour. We recommend at least 5GB of memory for running the models and post-processing.

For each model, we over-specify the numbers of factors to .. [to be written].

```

# Stack FA
Y_mat = Y_list %>% do.call(rbind, .) %>% as.matrix()
fit_stackFA <- MSFA::sp_fa(Y_mat, k = 6, scaling = FALSE, centering = TRUE,
                          control = list(nrun = 10000, burn = 800))

# Ind FA
fit_indFA <-
  lapply(1:6, function(s){
    j_s = c(8, 8, 8, 8, 8, 8)
    MSFA::sp_fa(Y_list[[s]], k = j_s[s], scaling = FALSE, centering = TRUE,
                control = list(nrun = 10000, burn = 8000))
  })

# PFA
N_s <- sapply(Y_list, nrow)
fit_PFA <- PFA(Y=t(Y_mat_scaled),
               latentdim = 6,
               grpind = rep(1:6,
                           times = N_s),
               Thin = 5,
               Total_itr = 10000, burn = 8000)

# MOM-SS
Y_mat = Y_list %>% do.call(rbind, .) %>% as.matrix()
# Construct the membership matrix
N_s <- sapply(Y_list, nrow)
M_list <- list()

```

```

    for(s in 1:6){
      M_list[[s]] <- matrix(1, nrow = N_s[s], ncol = 1)
    }
M <- as.matrix(bdiag(M_list))
fit_MOMSS <- BFR.BE::BFR.BE.EM.CV(x = Y_mat, v = NULL,
                                b = M, q = 6, scaling = FALSE)

# SUFA
fit_SUFA <- SUFA::fit_SUFA(Y_list_scaled, qmax=6, nrun = 10000)

# BMSFA
fit_BMSFA <- MSFA::sp_msfa(Y_list, k = 6, j_s = c(2, 2, 2, 2, 2, 2),
                           outputlevel = 1, scaling = FALSE,
                           centering = TRUE,
                           control = list(nrun = 10000, burn = 8000))

```

Fitting Tetris requires a 3-step process. First, we run `tetris()` to draw posterior samples of the model parameters, including  $\mathcal{T}$ . Then we run `choose.A()` to choose the best  $\mathcal{T}$  based on the posterior samples. Finally, we run `tetris()` again with the chosen  $\mathcal{T}$  to obtain the final model. Hyperparameters  $\alpha_{\mathcal{T}}$  are set to 1.25 times the number of studies.

```

# Tetris
set_alpha <- ceiling(1.25*6)
fit_Tetris <- tetris(Y_list, alpha=set_alpha, beta=1, nprint = 200,
                    nrun=10000, burn=8000)
big_T <- choose.A(fit_Tetris, alpha_IBP=set_alpha, S=6)
run_fixed <- tetris(Y_list, alpha=set_alpha, beta=1,
                  fixed=TRUE, A_fixed=big_T, nprint = 200,
                  nrun=10000, burn=8000)

```

## 2.4 Post processing

Post processing includes calculating the point estimates of the factor loadings and covariance matrix from the posterior samples, and determining the number of factors for each model.

For methods of MOM-SS, SUFA, and Tetris, the number of factors is determined internally in the algorithms, therefore, the output of the models are final results.

For MOM-SS,  $\Phi$  is directly obtained with its post-processed common loadings in the fitted output. The common covariance is calculated with  $\Phi\Phi^\top$ . The marginal covariance matrix

$\Sigma_{\text{marginal}}$  is calculated by adding the estimated study-specific error covariances to the common variance. The study-specific intercepts  $\alpha$  and the coefficients for the known covariates  $B$  are also extracted from the fitted output.

```
post_MOMSS <- function(fit, version = 2){ # version 1: M, version 2: Mpost
  est_Phi <- fit$M
  if (version==2){est_Phi <- fit$Mpost}
  est_SigmaPhi <- tcrossprod(est_Phi)

  # Marginal covariance
  S <- dim(fit$sigma)[2]
  est_PsiList <- est_SigmaMarginal <- list()
  for(s in 1:S){
    est_PsiList[[s]] <- fit$sigma[,s]
    est_SigmaMarginal[[s]] <- est_SigmaPhi + diag(fit$sigma[,s])
  }
  # last S columns of fit$Theta are the study-specific intercepts
  est_alphas <- fit$Theta[, (dim(fit$Theta)[2]-S+1):dim(fit$Theta)[2]]
  # The rest are coefficients for the known covariates
  est_B <- fit$Theta[, 1:(dim(fit$Theta)[2]-S)]

  return(list(Phi = est_Phi, SigmaPhi = est_SigmaPhi, Psi = est_PsiList, alpha = est_alphas,
             SigmaMarginal = est_SigmaMarginal))
}
res_MOMSS <- post_MOMSS(fit_MOMSS)
saveRDS(res_MOMSS, "Data/Rnutrition_MOMSS.rds")
```

For SUFA, the shared and study-specific loading matrices, as well as the common and marginal covariance can be conveniently obtained via the `lam.est.all()`, `SUFA_shared_covmat()` and `sufa_marginal_covs()` functions. Error covariance is obtained by taking averages of the “residuals” fitted output. And the study-specific covariance matrices are calculated by subtracting the common covariance from the marginal covariance. Note that in the definition of SUFA, common covariance is  $\Phi\Phi^\top + \Sigma$ .

```
post_SUFA <- function(fit){
  all <- dim(fit$Lambda)[3]
  burnin <- floor(all * 0.8) # We will use the last 20% samples
  # shared and study-specific loading matrices
  loadings <- lam.est.all(fit, burn = burnin)
  # Obtain common covariance matrix and loading from fitting
  est_Phi <- loadings$Shared
  est_SigmaPhi <- SUFA_shared_covmat(fit, burn = burnin)
```

```

est_Psi <- diag(colMeans(fit$residuals))
# Study-specific loadings
est_LambdaList <- loadings$Study_specific

# Obtain study-specific covariance matrices
S <- length(fit$A)
marginal_cov <- sufa_marginal_covs(fit, burn = burnin)
est_SigmaLambdaList <- list()
for (s in 1:S) {
  est_SigmaLambdaList[[s]] <- marginal_cov[, , s] - est_SigmaPhi
}

return(list(SigmaPhi = est_SigmaPhi, Phi = est_Phi,
           SigmaLambdaList = est_SigmaLambdaList,
           LambdaList = est_LambdaList,
           Psi = est_Psi,
           SigmaMarginal = lapply(1:S, function(s) marginal_cov[, , s])
))
}
res_SUFA <- post_SUFA(fit_SUFA)
saveRDS(res_SUFA, "Data/Rnutrition_SUFA.rds")

```

For Tetris, the common loading matrix  $\Phi$  can be obtained through the `getLambda()` function. The common covariance matrix is calculated as  $\Phi\Phi^\top$ . The study-specific loading matrices  $\Lambda_s$  are obtained by multiplying the common loading matrix with the study-specific matrices  $T_s - P$ . The study-specific covariance matrices are calculated as  $\Lambda_s\Lambda_s^\top$ . The marginal covariance matrix is calculated as  $\Lambda T \Lambda^\top + \Psi$ .

```

post_Tetris <- function(fit){
  # Estimated common covariance
  A <- fit$A[[1]]
  Lambda <- getLambda(fit,A)
  S <- dim(A)[1]
  est_Phi <- as.matrix(Lambda[,colSums(A)==S])
  est_SigmaPhi <- tcrossprod(est_Phi)
  # Estimated study-specific covariance
  P = diag((colSums(A) == S)*1)
  T_s <- list()
  est_LambdaList <- list()
  for(s in 1:S){
    T_s[[s]] <- diag(A[s,])
    Lambda_s <- Lambda %*% (T_s[[s]] - P)
  }
}

```

```

Lambda_s <- Lambda_s[,-which(colSums(Lambda_s == 0) == nrow(Lambda_s))]
Lambda_s <- matrix(Lambda_s, nrow=nrow(Lambda))
est_LambdaList[[s]] <- Lambda_s}
est_SigmaLambdaList <- lapply(1:S, function(s){
  tcrossprod(est_LambdaList[[s]])})

# Estimated marginal covariance
Psi <- list()
est_SigmaMarginal <- lapply(1:S, function(s){
  Psi[[s]] <- diag(Reduce("+", fit$Psi[[s]])/length(fit$Psi[[s]]))
  Sigma_s <- Lambda %*% T_s[[s]] %*% t(Lambda) + Psi[[s]]
})

return(list(Phi = est_Phi, SigmaPhi = est_SigmaPhi,
           LambdaList = est_LambdaList, SigmaLambdaList = est_SigmaLambdaList,
           Psi = Psi, T_s = T_s,
           SigmaMarginal = est_SigmaMarginal))
}
res_Tetris <- post_Tetris(run_fixed)
saveRDS(res_Tetris, "Data/Rnutrition_Tetris.rds")

```

For PFA, the post-processing is a little bit tricky. According to the original paper, the common loading matrix  $\Phi$  can be just the average of the estimated loadings from the posterior samples. However, we still apply OP as in BMSFA to get the common loadings, and the loadings in PFA should be the product of the estimated loadings and the square root of the variance of the factors. Therefore, we carry out multiplication for  $\Phi$  for each posterior sample, and then OP the sequence. We also carry out multiplication for  $Q_s$ ,  $\Sigma_\Phi$ ,  $\Sigma_s$ ,  $\Sigma_{\Lambda_s}$ , and  $\Psi$  as in the definition of PFA, and then take the average of the results.

After that, the number of factors is determined by counting the number of columns in  $\Phi$  that have all loadings less than  $10^{-3}$ .

```

post_PFA <- function(fit){
  p <- nrow(fit$Loading[[1]])
  k <- ncol(fit$Loading[[1]])
  npost <- length(fit$Loading)
  Q_list <- fit$Pertmat
  S <- dim(Q_list[[1]])[2]
  posteriorPhis <- array(0, dim = c(p, k, npost))

  #---estimated common loadings and shared variance---
  # Element-wise multiplication

```

```

for(i in 1:npost){
  posteriorPhis[,i] <- fit$Loading[[i]] %*% diag(fit$Latentsigma[[i]])
}
# Varimax for common loadings
est_Phi <- MSFA::sp_OP(posteriorPhis, itermax = 10, trace = FALSE)$Phi

# Estimated shared covariance, study-specific covariance matrix, and SigmaMarginal
sharevar <- list()
est_SigmaLambdaList <- list()
est_SigmaMarginal <- list()
est_Psi <- list()
for(s in 1:S){ # Loop over each study
  post_SigmaLambda_s <- list()
  post_SigmaMarginal_s <- list()
  Psi <- list()
  for(i in 1:npost){ # Loop over each posterior sample
    sharevar[[i]] <- fit$Loading[[i]]%*%diag(fit$Latentsigma[[i]]^2)%*%t(fit$Loading[[i]]) +
      diag(fit$Errorsigma[[i]]^2) # Get the shared variance
    Q_temp_inv <- solve(
      matrix(Q_list[[i]][, s], p, p)
    )
    post_SigmaMarginal_s[[i]] <- Q_temp_inv%*%sharevar[[i]]%*%t(Q_temp_inv)
    post_SigmaLambda_s[[i]] <- post_SigmaMarginal_s[[i]] - sharevar[[i]]
    Psi[[i]] <- diag(fit$Errorsigma[[i]]^2)
  }
  est_SigmaMarginal[[s]] <- Reduce('+', post_SigmaMarginal_s)/length(post_SigmaMarginal_s)
  est_SigmaLambdaList[[s]] <- Reduce('+', post_SigmaLambda_s)/length(post_SigmaLambda_s)
}
est_Psi <- Reduce('+', Psi)/length(Psi)
est_SigmaPhi <- Reduce('+', sharevar)/length(sharevar)
est_Q <- Reduce('+', Q_list)/length(Q_list)
est_Q_list <- lapply(1:S, function(s) matrix(est_Q[, s], p, p))

# Return the results
return(list(Phi = est_Phi, SigmaPhi = est_SigmaPhi, Psi = est_Psi, Q = est_Q_list,
           SigmaLambdaList = est_SigmaLambdaList,
           SigmaMarginal = est_SigmaMarginal))
}
res_PFA <- post_PFA(fit_PFA)
saveRDS(res_PFA, "Data/Rnutrition_PFA.rds")

```

```

# columns have all loadings less than 10^-3
fun_neighbour <- function(Phi, threshold = 1e-3) {
  return(
    sum(apply(Phi, 2, function(x) {
      sum(abs(x) <= threshold) < length(x)}
    ))
  )
}

# The post_PFA(fit_PFA) object
res_PFA <- readRDS("Data/Rnutrition_PFA.rds")
Phi_PFA <- res_PFA$Phi
K_PFA <- fun_neighbour(Phi_PFA)
K_PFA

```

```
[1] 6
```

The estimated  $K$  for PFA is 6.

Next, we show how to process the output of Stack FA, Ind FA, and BMSFA. While point estimates for the loadings and covariances can be obtained in the similar way as in PFA, the number of factors is determined with eigen value decompositions of the covariance matrix. Once the numbers of factors are determined, we have to run the models again with the correct number of factors, and then extract the final results.

Therefore, for Stack FA, we first extract the point estimates of the  $\Phi$  by applying OP to the posterior samples of the loadings. The common covariance matrix is calculated as  $\Phi\Phi^\top$ . The marginal covariance matrix is calculated as the average of its posterior samples.

```

post_stackFA <- function(fit, S){
  est_Phi <- MSFA::sp_OP(fit$Lambda, trace=FALSE)$Phi
  est_SigmaPhi <- tcrossprod(est_Phi)
  est_SigmaMarginal <- lapply(1:S, function(s)
    apply(fit$Sigma, c(1, 2), mean)
  )
  Psi_chain <- list()
  for(i in 1:dim(fit$Sigma)[3]){
    Psi_chain[[i]] <- fit$Sigma[, , i] - tcrossprod(fit$Lambda[, , i])
  }
  est_Psi <- Reduce('+', Psi_chain)/length(Psi_chain)
  return(list(Phi = est_Phi, SigmaPhi = est_SigmaPhi, Psi = est_Psi,
    SigmaMarginal = est_SigmaMarginal))
}

```



```

}
res_stackFA <- post_stackFA(fit_stackFA, S=6)
saveRDS(res_stackFA, "Data/Rnutrition_StackFA.rds")

```

After that, we determine the number of factors by eigen value decomposition of the common covariance matrix. We then run the model again with the correct number of factors, and extract the final results.

```

fun_eigen <- function(Sig_mean) {
  val_eigen <- eigen(Sig_mean)$values
  prop_var <- val_eigen/sum(val_eigen)
  choose_K <- length(which(prop_var > 0.05))
  return(choose_K)
}
res_stackFA <- readRDS("Data/Rnutrition_StackFA.rds")
SigmaPhi_StackFA <- res_stackFA$SigmaPhi
K_StackFA <- fun_eigen(SigmaPhi_StackFA)
K_StackFA

```

```
[1] 4
```

The estimated  $K$  for Stack FA is 4.

Then we re-run the model with the correct number of factors, and extract the final results.

```

fit_stackFA_2 <- MSFA::sp_fa(Y_mat_scaled, k = K_StackFA, scaling = FALSE, centering = TRUE,
                           control = list(nrun = 10000, burn = 8000))
res_stackFA_2 <- post_stackFA(fit_stackFA_2, S=6)
saveRDS(res_stackFA_2, "Data/Rnutrition_StackFA_2.rds")

```

We repeat this process for Ind FA and BMSFA.

```

# Ind FA
post_indFA <- function(fit){
  # Estimated study-specific covariance and loading
  S <- length(fit_list)
  est_LambdaList <- lapply(1:S, function(s){
    MSFA::sp_OP(fit_list[[s]]$Lambda, trace=FALSE)$Phi
  })
  est_SigmaLambdaList <- lapply(est_LambdaList, function(x) tcrossprod(x))
}

```

```

# Marginal covariance matrices
est_SigmaMarginal <- lapply(1:S, function(s) {
  fit <- fit_list[[s]]
  apply(fit$Sigma, c(1, 2), mean)
})

Psi <- list()
for(s in 1:S){
  Psi_chain <- list()
  for(i in 1:dim(fit_list[[1]]$Sigma)[3]){
    Psi_chain[[i]] <- fit_list[[s]]$Sigma[, , i] - tcrossprod(fit_list[[s]]$Lambda[, , i])
  }
  Psi[[s]] <- Reduce('+', Psi_chain)/length(Psi_chain)
}

return(list(LambdaList = est_LambdaList, SigmaLambdaList = est_SigmaLambdaList, Psi = Psi,
           SigmaMarginal = est_SigmaMarginal))
}

res_indFA <- post_indFA(fit_indFA)
saveRDS(res_indFA, "Data/Rnutrition_IndFA.rds")

# BMSFA
post_BMSFA <- function(fit){
  # Common covariance matrix and loading
  est_Phi <- sp_OP(fit$Phi, trace=FALSE)$Phi
  est_SigmaPhi <- tcrossprod(est_Phi)

  # Study-specific covariance matrices and loadings
  est_LambdaList <- lapply(fit$Lambda, function(x) sp_OP(x, trace=FALSE)$Phi)
  est_SigmaLambdaList <- lapply(est_LambdaList, function(x) tcrossprod(x))

  # Marginal covariance matrices
  S <- length(est_SigmaLambdaList)
  # Get point estimate of each Psi_s
  est_PsiList <- lapply(1:S, function(s) {
    apply(fit$psi[[s]], c(1, 2), mean)
  })
  est_margin_cov <- lapply(1:S, function(s) {
    est_SigmaPhi + est_SigmaLambdaList[[s]] + diag(est_PsiList[[s]] %>% as.vector())
  })
}

```

```

    return(list(Phi = est_Phi, SigmaPhi = est_SigmaPhi,
               LambdaList = est_LambdaList, SigmaLambdaList = est_SigmaLambdaList,
               PsiList = est_PsiList,
               SigmaMarginal = est_margin_cov))
  }
res_BMSFA <- post_BMSFA(fit_BMSFA)
saveRDS(res_BMSFA, "Data/Rnutrition_BMSFA.rds")

```

```

SigmaLambda_IndFA <- readRDS("Data/Rnutrition_IndFA.rds")$SigmaLambda
Js_IndFA <- lapply(SigmaLambda_IndFA, fun_eigen)
Js_IndFA %>% unlist()

```

```
[1] 4 5 5 5 4 4
```

```

SigmaPhi_BMSFA <- readRDS("Data/Rnutrition_BMSFA.rds")$SigmaPhi
K_BMSFA <- fun_eigen(SigmaPhi_BMSFA)
SigmaLambda_BMSFA <- readRDS("Data/Rnutrition_BMSFA.rds")$SigmaLambda
Js_BMSFA <- lapply(SigmaLambda_BMSFA, fun_eigen)
K_BMSFA %>% unlist()

```

```
[1] 4
```

```
Js_BMSFA %>% unlist()
```

```
[1] 2 2 2 2 2 2
```

The estimated  $J_s$  for Ind FA are 4, 5, 5, 5, 4, and 4. The estimated  $K$  for BMSFA is 4 and the estimated  $J_s$  are 2, 2, 2, 2, 2, and 2

Then we re-run the models with the correct number of factors, and extract the final results.

```

# Ind FA
fit_indFA_2 <-
  lapply(1:6, function(s){
    j_s = c(4, 5, 5, 5, 4, 4)
    MSFA::sp_fa(Y_list[[s]], k = j_s[s], scaling = FALSE, centering = TRUE,
                control = list(nrun = 10000, burn = 8000))
  })
res_indFA_2 <- post_indFA(fit_indFA_2)
saveRDS(res_indFA_2, "Data/Rnutrition_IndFA_2.rds")

```

```
# BMSFA
fit_BMSFA_2 <- MSFA::sp_msfa(Y_list, k = 4, j_s = c(2, 2, 2, 2, 2, 2),
                             outputlevel = 1, scaling = FALSE,
                             centering = TRUE,
                             control = list(nrun = 10000, burn = 8000))
res_BMSFA_2 <- post_BMSFA(fit_BMSFA_2)
saveRDS(res_BMSFA_2, "Data/Rnutrition_BMSFA_2.rds")
```

Now the final results are obtained and you can get the saved files mentioned above from the [GitHub repository](#).

## 2.5 Visualization

We can make some heatmap for the loadings. Please see the figures in the paper.

## 2.6 Mean squared error (MSE)

We can assess the goodness-of-fit of the models by reconstructing data and calculating the reconstruction errors. With estimated loadings and error covariances, we can estimate the factor scores for a new observation,  $\hat{\mathbf{f}}_{is,(new)}$  and  $\hat{\mathbf{l}}_{is,(new)}$ , derived by adapting the Bartlett method?, ?. Then we can use the estimated factor scores, together with the estimated loadings, to reconstruct  $\hat{\mathbf{y}}_{is,(new)}$  and calculate the reconstruction error which represents the fit of the models. To be specific, suppose that we have the multivariate data of a new observation from a specific study,  $\mathbf{y}_{is,(new)}$ , we can estimate its factor score and reconstruct its  $\hat{\mathbf{y}}_{is,(new)}$  in the following ways:

- Stack FA:  $\hat{\mathbf{f}}_{is,(new)} = (\hat{\Phi}^\top \hat{\Psi}^{-1} \hat{\Phi})^{-1} \hat{\Phi}^\top \hat{\Psi}^{-1} \mathbf{y}_{is,(new)}$ , and  $\hat{\mathbf{y}}_{is,(new)} = \hat{\Phi} \hat{\mathbf{f}}_{is,(new)}$ .
- Ind FA:  $\hat{\mathbf{l}}_{is,(new)} = (\hat{\Lambda}_s^\top \hat{\Psi}_s^{-1} \hat{\Lambda}_s)^{-1} \hat{\Lambda}_s^\top \hat{\Psi}_s^{-1} \mathbf{y}_{is,(new)}$ , and  $\hat{\mathbf{y}}_{is,(new)} = \hat{\Lambda}_s \hat{\mathbf{l}}_{is,(new)}$ .
- PFA:  $\hat{\mathbf{f}}_{is,(new)} = (\hat{\Phi}^\top \hat{\Psi}^{-1} \hat{\Phi})^{-1} \hat{\Phi}^\top \hat{\Psi}^{-1} \hat{Q}_s \mathbf{y}_{is,(new)}$ , and  $\hat{\mathbf{y}}_{is,(new)} = \hat{Q}_s^{-1} \hat{\Phi} \hat{\mathbf{f}}_{is,(new)}$ .
- MOM-SS:  $\hat{\mathbf{f}}_{is,(new)} = (\hat{\Phi}^\top \hat{\Psi}_s^{-1} \hat{\Phi})^{-1} \hat{\Phi}^\top \hat{\Psi}_s^{-1} (\mathbf{y}_{is,(new)} - \hat{\mathbf{s}} - \hat{\mathbf{x}}_{is,(new)})$ , and  $\hat{\mathbf{y}}_{is,(new)} = \hat{\mathbf{s}} + \hat{\mathbf{x}}_{is,(new)} + \hat{\Phi} \hat{\mathbf{f}}_{is,(new)}$ .
- SUFA: let  $\Omega = [\hat{\Phi}, \hat{\Phi} \hat{A}_s]$ , then  $\begin{bmatrix} \hat{\mathbf{f}}_{is,(new)} \\ \hat{\mathbf{l}}_{is,(new)} \end{bmatrix} = (\Omega^\top \hat{\Psi}^{-1} \Omega)^{-1} \Omega^\top \hat{\Psi}^{-1} \mathbf{y}_{is,(new)}$ , and  $\hat{\mathbf{y}}_{is,(new)} = \hat{\Phi} \hat{\mathbf{f}}_{is,(new)} + \hat{\Phi} \hat{A}_s \hat{\mathbf{l}}_{is,(new)}$ .

- BMSFA: let  $\Omega = [\widehat{\Phi}, \widehat{\Lambda}_s]$ , then  $\begin{bmatrix} \widehat{\mathbf{f}}_{is,(new)} \\ \widehat{\mathbf{l}}_{is,(new)} \end{bmatrix} = (\Omega^\top \widehat{\Psi}_s^{-1} \Omega)^{-1} \Omega^\top \widehat{\Psi}_s^{-1} \mathbf{y}_{is,(new)}$ , and  $\widehat{\mathbf{y}}_{is,(new)} = \widehat{\Phi} \widehat{\mathbf{f}}_{is,(new)} + \widehat{\Lambda}_s \widehat{\mathbf{l}}_{is,(new)}$ .
- Tetris:  $\widehat{\mathbf{f}}_{is,(new)} = (\widehat{T}_s^\top (\widehat{\Phi}^*)^\top \widehat{\Psi}_s^{-1} \widehat{\Phi}^* \widehat{T}_s)^{-1} \widehat{T}_s^\top (\widehat{\Phi}^*)^\top \widehat{\Psi}_s^{-1} \mathbf{y}_{is,(new)}$ , and  $\widehat{\mathbf{y}}_{is,(new)} = \widehat{\Phi}^* \widehat{T}_s \widehat{\mathbf{f}}_{is,(new)}$ .

If we divide the whole data into training set and test set, we can obtain the **mean square difference (MSE)** between the true and estimated  $\mathbf{y}_{is,(new)}$  in the test set for all individuals in all studies with  $\frac{1}{P \sum_s N_s} \sum_s^S \sum_i^{N_s} \sum_p^P (\hat{y}_{isp,(new)} - y_{isp,(new)})^2$ .

In the following, we divide the nutrition data into training set (70%) and test set (30%) and calculate the MSE of each model. Note that  $\mathbf{y}_{new}$  should also be centered except for MOM-SS.

```
train_ratio <- 0.7
train_list <- list()
test_list <- list()

for (s in seq_along(Y_list)) {
  N_s <- nrow(Y_list[[s]]) # Number of rows in the study
  train_indices <- sample(1:N_s, size = floor(train_ratio * N_s), replace = FALSE)

  train_list[[s]] <- Y_list[[s]][train_indices, ]
  test_list[[s]] <- Y_list[[s]][-train_indices, ]
}
# Test data has to be centered
test_list <- lapply(test_list, as.matrix)
test_list_scaled <- lapply(
  test_list, function(x) scale(x, center = TRUE, scale = FALSE)
)
```

We fit the models in the training data (on high-performance computing clusters), and then we load the fitted models and calculate the MSE on the test set for each model. The numbers of factors we input are determined by the previous results.

```
fit_StackFA_train <- readRDS("Data/Rnutrition_stackFA_train.rds")
fit_IndFA_train <- readRDS("Data/Rnutrition_IndFA_train.rds")
fit_PFA_train <- readRDS("Data/Rnutrition_PFA_train.rds")
fit_MOMSS_train <- readRDS("Data/Rnutrition_MOMSS_train.rds")
fit_SUFA_train <- readRDS("Data/Rnutrition_SUFA_train.rds")
```

```
fit_BMSFA_train <- readRDS("Data/Rnutrition_BMSFA_train.rds")
fit_Tetris_train <- readRDS("Data/Rnutrition_Tetris_train.rds")
```

We used the derived MSE function to calculate the MSE for each model.

```
# Stack FA
Phi <- fit_StackFA_train$Phi
Psi <- fit_StackFA_train$Psi

mse_stackFA <- 1/(42 * sum(sapply(test_list, nrow))) * sum(
  sapply(1:6, function(s){
    scores <- test_list_scaled[[s]] %*% solve(Psi) %*% Phi %*% mnormt::pd.solve(signif(t(Phi),
    norm(test_list_scaled[[s]] - scores %*% t(Phi), "F")^2
  })
)
```

```
# Ind FA
LambdaList <- fit_IndFA_train$LambdaList
Psi <- fit_IndFA_train$Psi
mse_IndFA <- 1/(42 * sum(sapply(test_list, nrow))) * sum(
  sapply(1:6, function(s){
    scores <- test_list_scaled[[s]] %*% solve(Psi[[s]]) %*% LambdaList[[s]] %*% mnormt::pd.s
    norm(test_list_scaled[[s]] - scores %*% t(LambdaList[[s]]), "F")^2
  })
)
```

```
# PFA
Phi <- fit_PFA_train$Phi
Psi <- fit_PFA_train$Psi
Q_list <- fit_PFA_train$Q
mse_PFA <- 1/(42 * sum(sapply(test_list, nrow))) * sum(
  sapply(1:6, function(s){
    scores <- test_list_scaled[[s]] %*% t(Q_list[[s]]) %*% Phi %*% mnormt::pd.solve(signif(t
    Y_est <- scores %*% t(Phi) %*% t(solve(Q_list[[s]]))
    norm(test_list_scaled[[s]] - Y_est, "F")^2
  })
)
```

```
# MOM-SS
Phi <- fit_MOMSS_train$Phi
```

```

Psi <- fit_MOMSS_train$Psi %>% lapply(diag)
alpha <- lapply(1:6, function(s) {
  fit_MOMSS_train$alpha[,s]
})

mse_MOMSS <- 1/(42 * sum(sapply(test_list, nrow))) * sum(
  sapply(1:6, function(s){
    scores <- t(apply(test_list[[s]], 1, function(row) {row - alpha[[s]]})) %*% solve(Psi[[s]]
    Y_est <- t(apply(scores %*% t(Phi), 1, function(row) {row + alpha[[s]]}))
    norm(test_list[[s]] - Y_est, "F")^2
  })
)

```

```

# SUFA
Phi <- fit_SUFA_train$Phi
LambdaList <- fit_SUFA_train$LambdaList
Psi <- fit_SUFA_train$Psi
mse_SUFA <- 1/(42 * sum(sapply(test_list, nrow))) * sum(
  sapply(1:6, function(s){
    Omega <- cbind(Phi, LambdaList[[s]])
    scores <- test_list_scaled[[s]] %*% solve(Psi) %*% Omega %*% mnormt::pd.solve(signif(t(O
    norm(test_list_scaled[[s]] - scores %*% t(Omega), "F")^2
  })
)

```

```

# BMSFA
Phi <- fit_BMSFA_train$Phi
LambdaList <- fit_BMSFA_train$LambdaList
Psi <- fit_BMSFA_train$PsiList %>% lapply(as.vector) %>% lapply(diag)
mse_BMSFA <- 1/(42 * sum(sapply(test_list, nrow))) * sum(
  sapply(1:6, function(s){
    Omega <- cbind(Phi, LambdaList[[s]])
    scores <- test_list_scaled[[s]] %*% solve(Psi[[s]]) %*% Omega %*% mnormt::pd.solve(signi
    norm(test_list_scaled[[s]] - scores %*% t(Omega), "F")^2
  })
)

```

```

# Tetris
Phi <- fit_Tetris_train$Phi
LambdaList <- fit_Tetris_train$LambdaList

```

```

Marginal <- fit_Tetris_train$SigmaMarginal
SigmaPhi <- fit_Tetris_train$SigmaPhi
SigmaLambdaList <- fit_Tetris_train$SigmaLambdaList
Psi <- lapply(1:6, function(s) {
  Marginal[[s]] - SigmaPhi - SigmaLambdaList[[s]]
})
T_s_list <- fit_Tetris_train$T_s
mse_Tetris <- 1/(42 * sum(sapply(test_list, nrow))) * sum(
  sapply(1:6, function(s){
    Omega <- cbind(Phi, LambdaList[[s]])
    scores <- test_list_scaled[[s]] %*% solve(Psi[[s]]) %*% Omega %*% mnormt::pd.solve(signi

    norm(test_list_scaled[[s]] - scores %*% t(Omega), "F")^2
  })
)

```

We display the MSE of each model.

```

print(paste0("Stack FA: ", mse_stackFA %>% round(3)))
print(paste0("Ind FA: ", mse_IndFA %>% round(3)))
print(paste0("PFA: ", mse_PFA %>% round(3)))
print(paste0("MOM-SS: ", mse_MOMSS %>% round(3)))
print(paste0("SUFA: ", mse_SUFA %>% round(3)))
print(paste0("BMSFA: ", mse_BMSFA %>% round(3)))
print(paste0("Tetris: ", mse_Tetris %>% round(3)))

```

```

[1] "Stack FA: 0.48"
[1] "Ind FA: 0.464"
[1] "PFA: 0.669"
[1] "MOM-SS: 0.456"
[1] "SUFA: 0.421"
[1] "BMSFA: 0.436"
[1] "Tetris: 0.275"

```



### 3 Case study: gene expression data

```
library(curatedOvarianData)  
library(tidyverse)
```

```
#data(package="curatedOvarianData")  
data(GSE13876_eset)  
data(GSE26712_eset)  
data(GSE9891_eset)  
data(PMID17290060_eset)
```

## References

- De Vito, Roberta, Briana Stephenson, Daniela Sotres-Alvarez, Anna-Maria Siega-Riz, Josiemer Mattei, Maria Parpinel, Brandilyn A Peters, et al. 2022. “Shared and Ethnic Background Site-Specific Dietary Patterns in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).” *medRxiv*, 2022–06.
- Grabski, Isabella N, Roberta De Vito, Lorenzo Trippa, and Giovanni Parmigiani. 2023. “Bayesian Combinatorial MultiStudy Factor Analysis.” *The Annals of Applied Statistics* 17 (3): 2212.