Supplementary Material Item Descriptions

The supplementary material contains 4 subdirectories: **Code**, **Data**, **Graphs** and **Results**.

Code

core_functions.R: Contains all core functions of the EM algorithm of MCLUST-ME. (R file)

simulation_functions.R: Contains all functions required to run the simulations in the paper. (R file)

sim 1.R: Runs Simulation 1 in Section 7. (R file)

sim 2.R: Runs Simulation 2 in Section 7. (R file)

bic_simulation.R: Runs the BIC simulation in Section 8. (R file)

rnaseq_analysis.R: Performs cluster analysis in Section 9. (R file)

Data

rna_raw.RData: Raw RNA-seq data used in Section 9. (RData file)

rna_processed.RData: Processed RNA-seq data for Section 9. (RData file)

Graphs

Figure *.pdf, Figure *.png: Graphs used in the paper. (PDF and PNG files)

Results

res_sim1_0*.RData: Contains results from Simulation 1 with $p = 0.*$. (RData files)

res_sim2_0*.RData: Contains results from Simulation 2 with $p = 0.*$. (RData files)

bic1_multiseeds.RData: Contains results from Case 1 of BIC simulation. (RData file)

bic2_79.RData: Contains results from Case 2 of BIC simulation. (RData file)

bic3_51.RData: Contains results from Case 3 of BIC simulation. (RData file)

rna_*group.RData: Contains MCLUST-ME clustering result of RNA-seq data assuming $G = *$ groups. (RData files)

rna_cluster_res.RData: Contains MCLUST-ME clustering result of RNA-seq data with the optimal BIC value. (RData file)

Instructions on Running MCLUST-ME Clustering Algorithm

Data Preparation

The R function `mcmeVVV` requires the following arguments as input:

data: A dataframe or numerical matrix of dimension $n \times p$, with rows representing data points and columns representing responses.

z: Initial membership matrix of dimension $n \times G$, with rows representing data points and columns representing responses. The author recommends using function `hclust` from MCLUST package to obtain such a matrix. See the next section for more details.

err: An array of measurement error (ME) covariance matrices, of dimension $p \times p \times n$. Each $p \times p$ matrix represents the ME covariance of a data point.

d,itmax,lb: Control parameters. Default values recommended.

Running the Algorithm

Once the data matrix (`data`) and ME array (`err`) are available, and the number of clusters (`G`) is determined, run the following script to obtain an initial membership matrix (`z`):

```
# Obtain group labels
n = nrow(data)
hcTree = hc(data)
cl = hclass(hcTree, G)

# Convert labels into binary form
z = matrix(0,n,G)
for(i in 1:n){
for(j in 1:G){
     z[i,j] = ifelse(cl[i]==j,1,0)
    }
}
```

Next, run the following script to begin the clustering algorithm:

```
# Group data with MCLUST-ME:
res = mcmeVVV(data,z,err)
```

Extracting Results

To obtain the membership matrix and parameter estimates upon convergence, run the following script:

```
## Parameter estimates
par = res$parameters
# Center
cen = par$muhat
# Covariance
covar = par$sigmahat
# Mixing proportion
prop = par$tauhat

## Membership matrix
z.final = res$z
```