Figure 1 (right plot t-dist) Simulation Code

Overview

This doc shows reproducible simulation code to generate the Figure 1 (right plot t-dist) result in manuscript *Community Detection with Heterogeneous Block Covariance Model*. Each value in Figure 1 is the average of 100 simulation results, which was executed by leveraging the George Washington University cloud computing platform. If the purpose is to test the code in local computing environment (i.e. personal laptop), we recommend to use few (<5) simulations to execute the code base.

Install packages

Install the HBCM package from github and other helper packages.

```
# install helper package
packages_helper <- c(
  'devtools', "parallel",
  "foreach", "doParallel",
  "kernlab", "matrixcalc"
  )
install.packages(packages_helper, repos = "http://cran.us.r-project.org")

# install hbcm package
devtools::install_github("xiangli2pro/hbcm")

# load all required packages
lapply(c('hbcm', packages_helper), require, char = TRUE)

# set random seed to reproduce the result
set.seed(2022)
```

Setup simulation parameters

Figure 1 (right plot t-dist) in the manuscript experiments t-distribution with different degree of freedom (df) . For the demo purpose, we set `df=5`. Meanwhile, we need to redefine the sample generating function, as the default function 'data_gen()' assumes normal distribution instead of t-distribution.

```
df <- 5
sample_gen <- function(n, p, mu, omega, labels, hlambda, hsigma) {
  alpha <- MASS::mvrnorm(n, mu, omega)
  x <- matrix(rep(0, n * p), nrow = n)
  for (i in 1:n) {
    for (j in 1:p) {
      x[i, j] <- hlambda[j] * (alpha[i, labels[j]]) + hsigma[j] * rt(1, df)
    }
  }

  list(
    x = x, alpha = alpha,
    hlambda = hlambda, hsigma = hsigma
  )
}
```

Other parameters do not change.

```
# data size and cluster number
centers <- 3 # number of Classes
n <- 1000 # number of Observations
p <- 1000 # number of Genes

# mean vector of normal distribution
mu <- rep(0, centers)

# class-level covariance matrix
off_diag <- 0.5
omega <- diag(rep(1, centers))
for (i in 1:centers) {
  for (j in 1:centers) {
    if (i != j) {
      omega[i, j] <- off_diag
    }
  }
```

```
}

# equally distributed class
ppi <- rep(1 / centers, centers)
labels <- sample(c(1:centers), size = p, replace = TRUE, prob = ppi)

# take hlambda=1 and hsigma=6
hlambda <- rep(1, p)
hsigma <- rep(6, p)
```

Generate a list of simulation data sets. Note that the manuscript used 100 simulations, but for the test purpose and limited computing resource, we recommend to use fewer simulations.

```
# set up the number of simulation data
size <- 1

# generate data
data_list <- lapply(
  c(1:size),
  function(i) sample_gen(n, p, mu, omega, labels, hlambda, hsigma)
  )

# save data
# save(data_list, labels, file = 'sim_data_npk.rda')
```

Estimate clusters

First use spectral clustering model to get initial estimation of the clusters, then plug in the initial estimation to HBCM model to perform the estimation.

Execute `?hbcm::heterogbcm` to see the details of input and output of the HBCM function.

```
# get the list of simulation data
X_list <- list(data_list[[1]]$x)

# apply spectral clustering for each simulation data in the list
system.time(
  spec_labels <- lapply(
    X_list,
    function(x) rSpecc(abs(cor(x)), centers = centers)$.Data),
  gcFirst = FALSE
```

```
)
```

```
# set up parallel environment
registerDoParallel(detectCores())

# apply HBCM for each simulation data in the list
system.time(
  hbcm_res <- foreach(
    m = c(1:size), .errorhandling = "pass",
    .packages = c("MASS", "Matrix", "matrixcalc", "kernlab", "RSpectra")
  ) %dopar%
    hbcm::heterogbcm(
      scale(X_list[[m]], center = TRUE, scale = FALSE), # standardize the data
      centers = centers, # set number of clusters
      tol = 1e-3, # iteration stop criteria
      iter = 100, # max iteration steps
      iter_init = 3, # iteration steps for estimation of hsigma and hlambda
      labels = spec_labels[[m]], # initial label estimation
      verbose = FALSE # whether or not to print out the iteration message
    ),
  gcFirst = FALSE
)

# save data
# save(
#   spec_labels, hbcm_res,
#   labels,
#   file = 'sim_cluster_npk.rda'
# )
```

Result

```
# cluster estimated by spectral cluster for the first simulation data
spec_labels[[1]]
```

```
  [1] 2 2 3 3 3 2 2 3 2 1 3 3 3 2 2 3 3 3 1 1 1 2 2 3 2 1 2 1 2 1 2 3 3 2 3 2 1
 [38] 1 1 1 3 2 2 2 1 3 1 2 2 1 3 3 1 3 2 2 3 2 2 3 3 2 1 1 3 1 2 3 3 2 2 1 1 1
 [75] 2 3 2 2 2 1 3 1 1 1 2 2 3 3 2 1 1 3 2 2 2 3 3 2 2 1 1 3 1 3 2 3 1 3 2 1 1
[112] 2 2 3 1 2 3 2 2 2 3 2 3 2 2 1 2 2 3 1 2 2 3 2 1 3 1 1 2 1 2 1 3 3 2 2 2 1
[149] 2 2 2 1 1 3 1 3 1 3 1 1 3 1 3 3 1 2 3 1 3 1 1 1 1 3 1 3 3 3 2 3 1 3 2 2 2
```

```
 [186] 2 2 2 1 3 3 1 3 1 1 3 2 2 1 1 2 2 3 2 3 1 2 2 1 1 3 2 1 3 3 2 2 1 1 3 1 1
 [223] 3 1 3 2 3 1 2 2 3 1 2 3 2 2 1 1 3 3 2 2 1 3 3 3 3 2 1 2 2 3 3 3 2 2 3 1 1
 [260] 3 1 1 1 2 2 2 3 3 3 3 2 2 1 2 1 1 3 3 3 3 1 2 2 3 2 3 1 1 2 3 1 2 2 3 1 2
 [297] 3 3 1 1 2 2 1 2 3 2 3 2 1 2 3 2 3 2 3 2 3 3 3 1 3 2 3 3 3 3 3 3 3 2 1 3 3
 [334] 1 1 1 1 1 3 1 3 1 2 2 2 3 3 1 3 1 3 3 3 1 2 2 1 3 1 3 3 2 3 1 2 2 1 2 3 3
 [371] 3 2 2 3 3 1 1 2 2 3 3 2 1 1 3 1 1 3 3 3 3 2 3 2 2 3 2 1 2 2 2 2 1 1 3 2 3
 [408] 1 2 3 2 1 1 3 2 3 3 2 2 2 2 2 3 1 2 2 1 1 1 1 1 1 3 2 3 3 1 1 1 2 3 3 2 2
 [445] 1 2 1 1 3 3 1 2 1 3 1 2 2 1 3 1 2 1 1 1 3 1 2 1 2 3 2 3 3 1 3 1 2 1 3 1 3
 [482] 1 2 1 1 1 2 2 2 1 3 1 3 2 1 1 2 1 2 3 1 1 3 1 2 2 3 2 3 2 1 2 3 1 1 1 3 1
 [519] 3 2 3 3 1 2 2 1 2 1 1 1 1 2 2 3 2 1 3 2 1 2 2 2 3 2 3 3 3 3 3 2 2 3 1 2 1
 [556] 3 2 3 2 3 3 1 2 3 1 3 2 3 1 1 2 3 3 3 3 2 3 1 2 1 2 3 3 2 1 1 2 2 3 2 3 3
 [593] 1 2 3 3 3 1 2 3 3 2 1 1 1 2 2 2 2 3 1 1 1 1 2 1 3 1 3 3 3 1 1 3 3 2 2 1 3
 [630] 1 2 2 1 1 3 2 3 2 3 1 1 3 2 2 1 2 2 2 3 2 2 3 3 3 1 2 1 3 1 1 3 1 1 1 1 2
 [667] 3 3 1 2 2 3 1 3 2 3 2 2 3 3 2 1 3 3 3 3 2 3 1 3 2 3 3 2 1 3 1 2 3 3 2 2 3
 [704] 3 1 1 1 2 3 1 2 2 3 2 1 3 1 1 3 1 1 1 2 2 2 2 1 3 1 3 1 3 2 1 1 2 1 2 1 1
 [741] 2 1 3 2 3 2 1 1 3 2 3 3 1 3 3 2 2 2 1 3 1 3 1 1 3 3 2 2 1 2 3 1 3 2 1 1 2
 [778] 1 2 2 3 2 3 3 1 1 3 2 1 3 1 2 2 2 3 3 2 1 1 2 3 3 3 3 2 2 1 2 1 2 2 1 1 3
 [815] 1 2 3 1 3 2 2 2 2 1 3 3 2 3 2 1 3 1 1 3 3 1 1 2 1 2 2 3 3 3 1 2 1 3 1 2 1
 [852] 2 2 1 2 2 3 3 1 1 1 3 2 3 3 3 3 3 3 2 1 1 1 2 1 1 3 2 3 3 2 3 2 3 3 2 3 2
 [889] 2 3 1 1 2 1 3 3 1 3 2 1 3 3 2 3 3 3 1 1 1 1 1 1 3 1 2 2 1 1 3 1 3 1 2 2 2
 [926] 2 1 1 3 2 2 2 2 2 1 1 1 2 1 2 1 3 2 1 2 1 1 1 2 1 3 2 3 3 3 3 3 2 2 2 1 2 2
 [963] 1 2 2 1 3 1 1 2 2 2 2 1 1 3 3 3 3 3 1 1 2 3 3 2 1 1 2 3 2 3 3 1 1 1 3 3 1
[1000] 3
```

```
# cluster estimated by HBCM for the first simulation data
hbcm_res[[1]]$cluster
```

```
  [1] 1 2 3 2 3 2 3 3 2 1 3 3 3 2 2 3 3 2 2 1 1 2 2 3 3 1 2 1 2 1 2 2 3 2 3 3 1
 [38] 2 1 1 2 2 2 2 1 3 1 2 2 1 3 3 1 3 2 2 3 2 2 3 2 2 2 1 3 1 2 2 3 2 1 1 1 1
 [75] 2 3 2 2 2 2 2 1 1 1 3 2 3 3 2 1 3 3 2 2 2 3 3 2 2 1 1 3 1 1 2 3 1 3 2 1 1
[112] 2 2 3 1 2 3 2 2 2 3 2 3 3 2 1 2 2 3 1 2 2 3 2 1 3 1 1 2 1 2 1 3 3 1 2 2 1
[149] 1 2 2 1 1 1 1 3 1 3 1 1 3 1 3 3 1 2 3 1 3 1 1 2 1 3 1 3 1 2 2 3 1 3 2 2 2
[186] 2 2 2 1 3 3 1 3 1 1 1 2 2 1 1 2 3 3 2 3 1 2 2 1 3 1 2 1 3 3 2 2 1 1 3 1 1
[223] 3 1 3 2 3 1 2 2 3 3 2 1 2 2 1 1 3 3 2 2 1 3 3 3 3 2 1 2 2 3 3 3 2 2 3 1 1
[260] 3 1 1 2 2 2 2 3 3 3 3 2 2 1 2 1 1 3 3 3 3 1 2 2 3 2 2 1 3 2 1 1 2 2 3 3 2
[297] 3 3 1 1 2 2 1 2 3 2 3 2 1 2 3 1 3 3 3 2 1 1 3 3 3 2 3 3 3 3 3 3 3 2 1 3 3
[334] 1 1 1 1 1 3 1 3 1 2 2 2 3 3 1 3 1 2 3 3 2 2 2 1 3 1 3 3 2 3 1 2 2 1 2 3 1
[371] 3 2 2 2 3 1 1 2 2 3 3 2 1 1 3 1 2 3 3 3 3 2 2 2 1 3 2 2 2 2 2 2 1 1 3 2 3
[408] 1 2 3 2 1 1 3 2 3 3 2 2 2 2 2 3 1 2 2 1 1 1 1 1 1 3 2 2 3 1 1 1 2 3 3 2 2
[445] 1 1 1 1 3 3 1 2 1 2 1 2 2 1 3 1 2 3 1 1 3 1 2 1 2 3 1 3 3 1 3 1 2 1 3 1 3
[482] 1 2 1 2 3 2 2 1 1 3 1 3 2 1 1 1 2 2 3 1 1 3 1 2 2 3 2 3 2 1 2 3 1 1 1 3 1
```

```
 [519] 3 2 3 3 1 2 2 1 2 1 1 2 1 2 3 1 2 1 3 2 1 2 3 2 3 2 3 3 3 3 3 2 3 3 1 2 1
 [556] 3 2 1 2 3 3 1 2 3 1 3 3 3 1 1 2 3 1 3 3 2 3 1 3 1 2 3 3 2 1 3 2 2 2 2 3 3
 [593] 3 3 2 3 3 2 2 3 3 2 3 1 2 2 2 1 2 3 1 1 1 1 2 1 3 1 3 1 2 1 1 3 3 2 2 2 3
 [630] 1 2 3 1 1 3 2 3 2 3 1 1 3 2 2 1 2 2 2 3 2 1 3 3 3 1 2 1 2 3 1 3 1 2 1 1 2
 [667] 3 3 2 2 3 3 1 2 2 2 2 2 3 3 2 1 2 3 2 3 2 2 1 3 2 2 3 2 1 2 1 2 3 3 2 2 3
 [704] 3 1 1 3 2 3 1 2 2 3 2 1 3 1 1 3 1 1 1 2 3 2 2 1 3 1 3 1 3 2 1 3 2 1 2 1 1
 [741] 2 1 3 1 3 1 1 1 3 2 3 3 2 3 3 2 2 3 1 3 2 1 1 1 3 3 2 2 1 3 3 1 3 2 1 1 2
 [778] 1 2 2 3 2 3 3 1 1 1 2 1 3 1 2 2 2 3 3 2 1 3 2 1 3 3 3 2 2 1 2 1 2 2 1 1 3
 [815] 1 2 2 1 3 2 1 2 2 1 3 3 2 3 2 1 3 1 1 3 1 1 1 2 1 2 2 1 3 3 1 3 3 3 1 2 1
 [852] 2 2 1 2 1 1 3 3 1 1 3 2 3 3 3 1 3 3 2 1 1 3 2 1 1 3 3 2 2 2 3 2 2 3 3 3 2
 [889] 2 3 1 1 2 1 3 3 1 3 2 2 3 3 2 3 3 3 1 3 1 1 1 1 3 1 2 2 3 1 3 1 3 1 2 2 1
 [926] 2 3 1 3 2 1 2 2 2 1 1 1 2 1 1 1 2 2 1 2 1 1 1 2 3 3 2 3 3 3 3 2 2 2 1 2 2
 [963] 1 2 2 1 1 1 3 2 2 2 2 1 1 3 3 3 3 3 1 1 2 3 3 2 1 1 2 3 2 3 3 1 1 1 3 3 1
[1000] 3
```

```r
# adjRand to evaluate how good is the spectral cluster estimation to truth
matchLabel(labels, spec_labels[[1]])$adjRand
```

```
[1] 0.5183668
```

```r
# adjRand to evaluate how good is the HBCM estimation to truth
matchLabel(labels, hbcm_res[[1]]$cluster)$adjRand
```

```
[1] 0.7725328
```