Simulation Study - AntMAN and DP

Kevin Korsurat

2023-06-17

```
### Function: Simulating the data based on the scenario
f_data_sim <- function(sim_seed, scenario_index){

  ### place for storing result.
  actual_clus <- NULL
  dat <- NULL

  set.seed(sim_seed)

  if(! scenario_index %in% 1:4){
    warning("invalid scenario. we have only 4 scenarios")
  } else {
    if(scenario_index == 1){
      actual_clus <- sample(1:2, 500, replace = TRUE)
      dat <- rnorm(500, c(-5, 5)[actual_clus])
    } else if(scenario_index == 2){
      actual_clus <- sample(1:5, 500, replace = TRUE)
      dat <- rnorm(500, (c(0, 7.5, 15, 25, 35))[actual_clus])
    } else if(scenario_index == 3){
      actual_clus <- sample(1:2, 500, replace = TRUE)
      dat <- rnorm(500, c(-5, 5)[actual_clus], 3)
    } else {
      actual_clus <- sample(1:5, 500, replace = TRUE)
      dat <- rnorm(500, (c(0, 7.5, 15, 25, 35)[actual_clus])/2, 1)
    }
  }

  ### return the simulated data
  result <- data.frame(actual_clus, dat)
  return(result)
}
```

Hyperparameter choosing (SFDMM)

I have chosen the set of hyperparameters based on the sensitivity analysis. Based on the sensitivity analysis, the model works well if we choose something that looks like a noninformative prior.

- $\mathrm{K}_{\max} = 10$
- $\sigma_0^2 = 100$
- $a_\sigma = b_\sigma = 0.01$
- $\xi = 1$

- $a_\theta = b_\theta = 1$
- the number of launch step is 10

Then, I will test this set of the hyperparameter on all cases for both raw and scaled dataset.

Other models

AntMAN

- The default hyperparameters are $\mu_0 = 0, \lambda = 1, a_\sigma = 3, b_\sigma = 2$. However, I will set $a_\sigma = b_\sigma = 0.1$ instead to let this model to similar to SFDMM as much as possible.

Dirichlet Process

- The hyperparameter for this model is also the same as `AntMAN`. ($\mu_0 = 0, \lambda = 1, a_\sigma = 3, b_\sigma = 2$). So, I will set $a_\sigma = b_\sigma = 0.1$ instead to let this model to similar to SFDMM as much as possible.

Here is the result for the raw data.

```
### Raw data
for(i in 1:4){
  dat_sim <- f_data_sim(345324, i)
  dat_y <- dat_sim$dat

  print(paste0("============== Scenario ", i, " (Raw Data) =============="))

  ### AntMAN
  AntMAN_MCMC <- AM_mcmc_parameters(niter = 10000, burnin = 5000, thin = 1,
                                    verbose = 1, output = c("CI", "K"),
                                    parallel = FALSE, output_dir = NULL)
  data_hyper <- AM_mix_hyperparams_uninorm(m0 = 0, k0 = 1, nu0 = 0.1, sig02 = 0.1)
  cluster_hyper <- AM_mix_weights_prior_gamma(a = 1, b = 1)
  AntMAN_mod <- AntMAN::AM_mcmc_fit(y = dat_y, initial_clustering = rep(1, 500),
                                    mix_kernel_hyperparams = data_hyper,
                                    mix_weight_prior = cluster_hyper,
                                    mcmc_parameters = AntMAN_MCMC)
  AntMAN_method <- as.numeric(salso(AM_clustering(AntMAN_mod), maxNClusters = 10))
  table("AntMAN" = AntMAN_method, "Actual" = dat_sim$actual_clus) %>% print()

  ### SFDMM
  model <- SFDMM_model(iter = 10000, K_max = 10, init_assign = rep(0, 500),
                       y = dat_y, a0 = 0.01, b0 = 0.01, mu0 = 0, s20 = 100,
                       xi0 = 1, a_theta = 1, b_theta = 1, launch_iter = 10,
                       print_iter = 10001)
  table("SFDMM" = salso(model$iter_assign[-(1:5000), ]),
        "Actual" = dat_sim$actual_clus) %>% print()

  ### DP
  dp_mod <- DirichletProcessGaussian(as.matrix(dat_y),
                                     g0Priors = c(0, 1, 0.01, 0.01), alphaPriors = c(1, 1))
  dp_fit <- Fit(dp_mod, 10000, updatePrior = FALSE, progressBar = TRUE)
  dp_clus <- matrix(NA, nrow = 5000, ncol = 500)
```

```
  for(i in 1:5000){
    dp_clus[i, ] <- dp_fit$labelsChain[[(5000 + i)]]
  }

  table("DP" = salso(dp_clus),
        "Actual" = dat_sim$actual_clus) %>% print()
}
```

```
## [1] "============== Scenario 1 (Raw Data) =============="
##       Actual
## AntMAN   1   2
##      1 256   0
##      2   0 244
##      Actual
## SFDMM   1   2
##     1 256   0
##     2   0 244
##   |                                                     |
##    Actual
## DP    1   2
##   1 256   0
##   2   0 244
## [1] "============== Scenario 2 (Raw Data) =============="
##       Actual
## AntMAN   1   2   3   4   5
##      1 112   0   0   0   0
##      2   0   1  99   0   0
##      3   1  96   0   0   0
##      4   0   0   0  90 101
##      Actual
## SFDMM   1   2   3   4   5
##     1 113   0   0   0   0
##     2   0   0  99   0   0
##     3   0  97   0   0   0
##     4   0   0   0  90   0
##     5   0   0   0   0 101
##   |                                                     |
##    Actual
## DP    1   2   3   4   5
##   1 112   0   0   0   0
##   2   0   1  99   0   0
##   3   1  96   0   0   0
##   4   0   0   0  90 101
## [1] "============== Scenario 3 (Raw Data) =============="
##       Actual
## AntMAN   1   2
##      1 250  17
##      2   6 227
##      Actual
## SFDMM   1   2
##     1 250  18
##     2   6 226
##   |                                                     |
```

```
##    Actual
## DP    1   2
##   1 250  17
##   2   6 227
## [1] "============== Scenario 4 (Raw Data) =============="
##       Actual
## AntMAN   1   2   3   4   5
##      1  99   0   0   0   0
##      2  14  97  99   1   0
##      3   0   0   0  89 101
##      Actual
## SFDMM   1   2   3   4   5
##     1 107   4   0   0   0
##     2   0   4  99   1   0
##     3   6  89   0   0   0
##     4   0   0   0  88   0
##     5   0   0   0   1 101
##   |                                                         |
##    Actual
## DP    1   2   3   4   5
##   1  99   0   0   0   0
##   2  14  97  99   1   0
##   3   0   0   0  89 101
```

Here is the result for the scaled data.

```
### Scaled data
for(i in 1:4){
  dat_sim <- f_data_sim(34120, i)
  dat_y <- as.numeric(scale(dat_sim$dat))

  print(paste0("============== Scenario ", i, " (Raw Data) =============="))

  ### AntMAN
  AntMAN_MCMC <- AM_mcmc_parameters(niter = 10000, burnin = 5000, thin = 1,
                                    verbose = 1, output = c("CI", "K"),
                                    parallel = FALSE, output_dir = NULL)
  data_hyper <- AM_mix_hyperparams_uninorm(m0 = 0, k0 = 1, nu0 = 0.1, sig02 = 0.1)
  cluster_hyper <- AM_mix_weights_prior_gamma(a = 1, b = 1)
  AntMAN_mod <- AntMAN::AM_mcmc_fit(y = dat_y, initial_clustering = rep(1, 500),
                                    mix_kernel_hyperparams = data_hyper,
                                    mix_weight_prior = cluster_hyper,
                                    mcmc_parameters = AntMAN_MCMC)
  AntMAN_method <- as.numeric(salso(AM_clustering(AntMAN_mod), maxNClusters = 10))
  table("AntMAN" = AntMAN_method, "Actual" = dat_sim$actual_clus) %>% print()

  ### SFDMM
  model <- SFDMM_model(iter = 10000, K_max = 10, init_assign = rep(0, 500),
                       y = dat_y, a0 = 0.01, b0 = 0.01, mu0 = 0, s20 = 100,
                       xi0 = 1, a_theta = 1, b_theta = 1, launch_iter = 10,
                       print_iter = 10001)
  table("SFDMM" = salso(model$iter_assign[-(1:5000), ]),
        "Actual" = dat_sim$actual_clus) %>% print()
```

```r
  ### DP
  dp_mod <- DirichletProcessGaussian(as.matrix(dat_y),
                                     g0Priors = c(0, 1, 0.01, 0.01), alphaPriors = c(1, 1))
  dp_fit <- Fit(dp_mod, 10000, updatePrior = FALSE, progressBar = TRUE)
  dp_clus <- matrix(NA, nrow = 5000, ncol = 500)
  for(i in 1:5000){
    dp_clus[i, ] <- dp_fit$labelsChain[[(5000 + i)]]
  }

  table("DP" = salso(dp_clus),
        "Actual" = dat_sim$actual_clus) %>% print()
}
```

```
## [1] "============== Scenario 1 (Raw Data) =============="
##       Actual
## AntMAN   1   2
##      1 264   0
##      2   0 236
##      Actual
## SFDMM   1   2
##     1 264   0
##     2   0 236
##   |                                                     |
##    Actual
## DP    1   2
##   1 264   0
##   2   0 236
## [1] "============== Scenario 2 (Raw Data) =============="
##       Actual
## AntMAN   1   2   3   4   5
##      1  96  94   0   0   0
##      2   0   0 106   0   0
##      3   0   0   0  88 116
##      Actual
## SFDMM   1   2   3   4   5
##     1  96   0   0   0   0
##     2   0   0 106   0   0
##     3   0   0   0  88   0
##     4   0  94   0   0   0
##     5   0   0   0   0 116
##   |                                                     |
##    Actual
## DP    1   2   3   4   5
##   1  96  94   0   0   0
##   2   0   0 106   0   0
##   3   0   0   0  88   0
##   4   0   0   0   0 116
## [1] "============== Scenario 3 (Raw Data) =============="
##       Actual
## AntMAN   1   2
##      1 264 236
##      Actual
## SFDMM   1   2
```

```
##     1 249  12
##     2  15 224
##   |                                                                 |
##    Actual
## DP    1   2
##   1 249  12
##   2  15 224
## [1] "============== Scenario 4 (Raw Data) =============="
##       Actual
## AntMAN   1   2   3   4   5
##      1  96  92  12   0   0
##      2   0   2  94   0   0
##      3   0   0   0  80   0
##      4   0   0   0   8 116
##      Actual
## SFDMM   1   2   3   4   5
##     1  96   6   0   0   0
##     2   0   6 103   0   0
##     3   0   0   0  83   0
##     4   0  82   3   0   0
##     5   0   0   0   5 116
##   |                                                                 |
##    Actual
## DP    1   2   3   4   5
##   1  96  92  11   0   0
##   2   0   2  94   0   0
##   3   0   0   1  80   0
##   4   0   0   0   8 116
```