Reproducibility in the Overlapping Simulation Case

This file contains instructions for reproducing the results and figures in the overlapping simulation case. The codes are in "code" folder. We assume the working directory has been appropriately set.

Step 1: "Step1_Simulation_data.R"

Generate the data in the simulation study. The output of this step is "Sim_data.RData".

Step 2: "Step2_TreeTC_mcmc.R"

Implement the TreeTC model on the data from Step 1, conduct the MCMC posterior inference, and select the iteration following the big node number rule. The output of this step is "Sim_result_iteration.RData".

In the "Sim_result_iteration.RData":

- q0: Root node of the tree (Normal Class)
- tssb: Tree structure (TssbMCMC Class)
- ll_UnPost: Unnormalized posterior
- clIds: Observational-level clustering membership of all the data as well as of each source group for each iteration after burn-in
- numOfTrees: Maximal number of source groups ("L" in the model)
- tssbNum: Selected iteration index under the big node number rule
- finalSubjAssignments: Source group membership for the iteration "tssbNum"
- finalClIds: Observational-level clustering membership of all the data for the iteration "tssbNum"
- groups_subj: Group numbers (distinct values of the vector "finalSubjAssignments")
- ggFinal1 / ggFinal2: gg: Tree structures (igraph) of each source group for the iteration "tssbNum"
- subFinalClIds1 / subFinalClIds2: Observational-level clustering membership of each source group for the iteration "tssbNum"
- subFinalCenters1 / subFinalCenters2: Observational-level cluster centers of each source group for the iteration "tssbNum"
- finalSigma: Variance of root node for the iteration "tssbNum"
- finalDrift: Drift parameter for the iteration "tssbNum"

Step 3: "Step3_baseline_gmm_cam.R"

Implement two baseline methods (hierarchical clustering and k-centroids) and the Gaussian mixture model to conduct the clustering at the observational level. Implement CAM to conduct the clustering at both the source and observational level. The output of this step is "sim_hc.csv," "sim_kc.csv," "sim_gmm.csv," "sim_cam_sub.csv," "sim_cam_obs.csv."

- sim_hc.csv: Clustering result of hierarchical clustering
- sim_kc.csv: Clustering result of k-centroids
- sim_gmm.csv: Clustering result of Gaussian mixture model
- sim_cam_sub.csv: Clustering result of CAM, the source-level clustering indices

- sim_cam_obs.csv: Clustering result of CAM, the observation-level clustering indices

Step 4: "Step4_FigureS8.R"

Draw Figure S8 in the supplementary using the output of Step 2 and 3.