Reproducibility in the First Real Application (Image data)

This file contains instructions for reproducing the results and figures contained in the first real application. The codes are in "code" folder. We assume the working directory has been appropriately set.

Step 1: "Step1_Preprocessing_&_Step3_FigureS9.ipynb"

Preprocess raw data. We download the CIFAR-100 data set by conducting the following code in Python:
(X_train, y_train), (X_test, y_test) = keras.datasets.cifar100.load_data(label_mode="fine")

In this step, we focused on nine classes—chair, chimpanzee, dolphin, lion, maple, oak, oranges, flatfish, and shark—and each class has 500 training images. We used UMAP to reduce the dimension of images to 10. Subsequently, we assigned these 4,500 images to 30 sources with each source having 150 images, and the 30 sources further form three groups with each group having 10 sources, where group 1 contains three classes chair, chimpanzee, and dolphin, group 2 contains three classes lion, maple, and oak, and group 3 contains three classes oranges, flatfish, and shark.

The output of this step is given by "rawData.csv".

In the "rawData.csv":

- Column1 - Column10: Input data of the TreeTC model (4,500 * 10)
- Column11: Corresponding indices of each image in the original data

Step 2: "Step2_TreeTC_mcmc.R"

Apply the TreeTC model to the preprocessed data, conduct the MCMC posterior sampling, and select the iteration following the big node number rule. The output of this step is "Real_result_iteration.RData".

In the "Real_result_iteration.RData":

- q0: Root node of the tree (Normal R6 Class)
- tssb: Tree structure (TssbMCMC R6 Class)
- testData: Input data for TreeTC model
- numOfSubjData: Number of data for each source
- etaNormal: Shrinkage parameter for node variances
- 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 / ggFinal3: gg: Tree structures (igraph) of each source group for the iteration "tssbNum"
- subFinalClIds1 / subFinalClIds2 / subFinalClIds3: Observational-level clustering membership

of each source group for the iteration "tssbNum"

- subFinalCenters1 / subFinalCenters2 / subFinalCenters3: 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_FigureS9.R" and "Step1_Preprocessing&Step3_FigureS9.ipynb"

Draw Figure S9 in the supplementary using the output of Step 2. Specifically, we conduct "Step3_FigureS9.R" to obtain the csv files "TopClIds/group1.csv," "TopClIds/group2.csv," and "TopClIds/group3.csv." Then we conduct "Step1_Preprocessing&Step3_FigureS9.ipynb" to draw Figure S9.