Reproducibility in the Simulation Case I

This file contains instructions for reproducing the results and figures in the similation case I in Section 5. The codes are in "code" folder. Please set the working directory to the source file location ("code" folder).

The following code files can be directly run as their input data are generated by itself or in the "input\_data" and "result\_data" folders: Step1\_Generate\_data.R, Step2\_BIC\_analysis.R, Step3\_BUSbeta.R, Step4\_BUS.R, Step5\_ComBat.R, Step6\_BEclear.R, Step7\_Figure1.R, Step8\_Kmeans.R, Step9\_GMM.R and Step10\_MetaSparseKmeans.R.

**Data**

The simulated data in Simulation Case I are artificially generated by the assigned parameters and consist of 660 samples from 3 batches. The 660 samples have DNA methylation beta values for 1000 CpG sites and are assigned to 5 subgroups.

In the data generating procedure, we assign the true subgroup proportions in parameter specifications, which characterize the subgroup structures in each batch.

The generated data includes:

* "Y.RData": The raw simulated data list.
* "Z\_real.RData": The annotated subgroups list.
* "truemean\_by\_subgroup.RData": The true subgroup means list for each CpG site in each subgroup.

**Code**

**Step 1: "Step1\_Generate\_data.R"**

Generate simulated data. The outputs of this step are "Y.RData," "Z\_real.RData" and "truemean\_by\_subgroup.RData."

**Step 2: "Step2\_BIC\_analysis.R"**

Calculated the BIC values for subgroups value K ranging from one to ten, and draw Figure 1(d) in Section 5 to determine the number of subgroups for the following BUSbeta analysis. The output of this step is "Figure1(d).png" in "figures" folder.

**Step 3: "Step3\_BUSbeta.R"**

Implement BUSbeta on the data from Step 1, and conduct the MCMC posterior inference. Then compute the ARI values in each single batch and the overall ARI values for BUSbeta. The outputs of this step are "Y\_BUSbeta.RData" which saves the corrected data, "Z\_BUSbeta.RData" which saves the clustering labels and "ARI\_BUSbeta.csv" in "result\_data" folder.

In the "ARI\_BUSbeta.csv":

* "Batch1," "Batch2," "Batch3": samples in each batch.
* "Overall": samples combined from all batches together.
* "BUSbeta ARI": ARI values by BUSbeta.

**Step 4: "Step4\_BUS.R"**

Implement BUS model using the R package BUScorrect. Then compute the ARI values in each single batch and the overall ARI values for BUS. The outputs of this step are "Y\_BUS.RData" which saves the corrected data, "Z\_BUS.RData" which saves the clustering labels and "ARI\_BUS.csv" in "result\_data" folder.

In the "ARI\_BUS.csv":

* "Batch1," "Batch2," "Batch3": samples in each batch.
* "Overall": samples combined from all batches together.
* "BUS ARI": ARI values by BUS.

**Step 5: "Step5\_ComBat.R"**

Implement ComBat method using the R package sva. The output of this step is "Y\_ComBat.RData" which saves the corrected data in "result\_data" folder.

**Step 6: "Step6\_BEclear.R"**

Implement BEclear method using the R package BEclear. The output of this step is "Y\_BEclear.RData" which saves the corrected data in "result\_data" folder.

**Step 7: "Step7\_Figure1.R"**

Draw Figures 1(a)-(c) and 1(e)-(h) in Section 5 using the data from Step 1 and the corrected data of all methods. For BUSbeta and BUS, the clustering labels from Steps 3 and 4 are also used.

**Step 8: "Step8\_Kmeans.R"**

Implement Kmeans on the data from Step 1 and compute the ARI values in each single batch. The output of this step is "ARI\_Kmeans.csv" in "result\_data" folder.

In the "ARI\_Kmeans.csv":

* "Batch1," "Batch2," "Batch3": samples in each batch.
* "Overall": samples combined from all batches together.
* "Kmeans ARI": ARI values by Kmeans.

**Step 9: "Step9\_GMM.R"**

Implement GMM on the data from Step 1 using the R package mclust, and compute the ARI values in each single batch. The output of this step is "ARI\_GMM.csv" in "result\_data" folder.

In the "ARI\_GMM.csv":

* "Batch1," "Batch2," "Batch3": samples in each batch.
* "Overall": samples combined from all batches together.
* "GMM ARI": ARI values by GMM.

**Step 10: "Step10\_MetaSparseKmeans.R"**

Implement MetaSparseKmeans on the data from Step 1 using the R package MetaSparseKmeans, and compute the ARI values in each single batch and the overall ARI values. The output of this step is "ARI\_MetaSparseKmeans.csv" in "result\_data" folder.

In the "ARI\_ MetaSparseKmeans.csv":

* "Batch1," "Batch2," "Batch3": samples in each batch.
* "Overall": samples combined from all batches together.
* "MetaSparseKmeans ARI": ARI values by MetaSparseKmeans.