Supplementary Material for A Resampling based Clustering Algorithm for Replicated Gene Expression Data

Han Li, Chun Li, Jie Hu, Xiaodan Fan

This package contains five files:

1) SuppMaterial.pdf: this file, including instructions for running RC.r
2) syn_gen.r: R code for generating synthetic data in the paper
3) syn_data: includes three subfolders, s1, s2, s3 include synthetic sample data in simulation study 1, 2, 3, respectively
4) RC.r: R code for resampling based clustering using quasi-MCMC algorithm
5) comparison.pdf: simulation studies comparing RC and MCLUST using only the mean profile in cases with/without outliers

For RC, we first use K-means clustering to partition the genes into 10 clusters (based on the mean profile), and then set the initial values of the parameters of mixture model equal their component mean/variance. We set μ_0 be the mean of overall expression data, κ=10, η and ζ be the half of the median of the overall variance across time points.

To run RC.r, we use the command line argument, by specifying

- dataName: the prefix of output file
- fileName: the file route for input data
- R: the number of replicates
- BurnIn: the number of burn-in iterations, usually 1000-2000 is enough for convergence
- ITER: the number of total iterations, usually 3000-5000 is enough to summarize the result
- flag: if flag=1, the input data will be normalized to have mean 0, standard deviation 1; if flag=0, the input data will not be normalized

Th input data is a matrix with N rows and $J \times R$ columns, with each row representing a gene. Denote y_{ijr} as the expression level of the i-th gene at the j-th experiment condition in the r-th replicate. The replicated data is formatted in the following way.

$$\text{i-th row: } y_{i11} \cdots y_{i1R} \; y_{i21} \cdots y_{i2R} \cdots y_{iJ1} \cdots y_{iJR}$$

RC requires all the experiment conditions have R replicates. For varying number of replicates, one can change the code for sampling the replicate acoordingly. Here is an example for running RC. First change the working directory to where this file locates. Next make sure that the local R program has installed R packages “MCMCpack” and “mcclust”. We will use the data stored in syn_data/s1/case1.txt as input. Set dataName=s1, fileName=syn_data/s1/case1.txt, R=4, BurnIn=500, ITER=1000, flag=0. To run RC.r, type the following command.

```
R - -vanilla - -slave - -args s1 syn_data/s1/case1.txt 4 500 1000 0 < RC.r
```

The output is saved in the working directory, including two files, one is s1_clust.txt, which is the clustering result for all genes with each row representing the cluster index of the gene; the other is s1_res.txt, whose first row is the number of inferred clusters, the second row is the adjusted Rand index.

For the yeast galactose data, we download the data from http://expression.washington.edu/publications/kayee/yeunggb2003/. For the Drosophila Notch signaling pathway data, we obtain the data in the CRAN R package “DIRECT”. Considering the authorship, we do not include these two data in this package.