Data and Reproducibility of the Article Robust Model-Based Clustering

Juan D. Gonzalez

26/03/2022

Data and Reproducibility of the Article Robust Model-Based Clustering

Summary

This document is for reproducing the analysis on real data and simulations of the article *Robust Model-Based* Clustering whose authors are Juan D. Gonzalez, Ricardo Maronna, Victor J. Yohai, and Ruben H. Zamar.

Firstly, you should install the RMBC package from CRAN, by typing

```
install.packages(RMBC)
```

Note: *For excecuting the following steps, the working directory should include the* `\utils` *folder, and the scripts*

- `fig3_and_Table6.R`,
- `main_simulation_clean_models.R`
- `main_simulation_contamimated_models.R`,
- `main_simulation_clean_and_small_n_models_balanced.R`,
- `main_simulation_small_n_contaminated_models_balanced.R`
- `table table_building_2022.R`

Reproducing the Figure 3 and Table 6 of the paper

Once you have already installed the RMBC package, the steps you should do in order to reproduce the results is to run in R,

```
source("Fig3_and_Table6.R")
```

This command will do the following steps

- Install the package "RMBC" from github. This packages implements the estimator developed in the paper and contains the real data set used to asses our procedure.
- Install the necessary packages from CRAN, namely
  - `tclust`
  - `RSKC`

  - `GSE`
  - `otrimle`
  - `mclust`
  - `mvtnorm`
  - `ktaucenters`
  - `combinat`

- Run the auxiliary routines that aid to plot the results, compute performance measures and give to the different estimators the same format.
  - `auxComputeClusters.R`
  - `auxPlotGroupsModelBasedCluster.R`
  - `performance_measures.R`

Finally, if success, the program is supposed to reproduce figure 3 and table 6 of the paper.

Reproducing the Simulation Results : Table 2 to 5 of the paper

In addition to this, in order to obtain the results of the simulation on contaminated data, that is, Table 2 of the paper, you can run

```
source("main_simulation_contamimated_models.R")
```

This procedure will display the values of the table on the R - console and print a `.tex` file with the table in the directory `tex_tables`.

The remainder simulation studies, namely clean data with small or big sample size and clean data with big samples can be obtained by running the analogous scripts

```
source("main_simulation_clean_models.R")
source("main_simulation_clean_and_small_n_models_balanced.R")
source("main_simulation_small_n_contaminated_models_balanced.R")
```

We want to warn that these last cases consist on Monte Carlo simulations, where four estimators are taken into account in each sample, so computing time can take several hours (the computing time for 500 replications took 38 hours in a regular personal computer). You can reduce the time by setting the variable from `nrep=500` to a lower value, say, `nrep=10`, obviously in that case, the results will not be the same that the table of the paper.

Note: The version of the packages used for carried out this simulation studies were

- `otrimle` (version 2.0)
- `mclust` (version 5.4.8)
- `tclust` (version 1.4.2)
- `RMBC` (version 0.1.0)