Distributed Parallel Clustering in *R* with Large Data

Wei-Chen Chen*
Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA chenwc@ornl.gov

George Ostrouchov
Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA ostrouchovg@ornl.gov
and The University of Tennessee, Knoxville, Tennessee, USA

Clustering is an exploratory technique for partitioning multivariate data points into groups suggested by their density variations in the multivariate space. Clustering is important because exploratory visualization is very limited in high dimensions and the resulting partitions can lead to simplifications in model development.

We report on the *R* implementation of a set of distributed parallel clustering algorithms for fitting general Gaussian mixture models to large data using several variants of the EM algorithm. The resulting *R* package, *pmclust*, can be used when the data set is larger than the available memory on a single processor but fits into the collective memory of a large distributed cluster of processors. We describe the algorithms, the package, and its use of *pbdMPI*, a package for communication between processors that is needed to solve the large problem. In addition, we give a short overview of the *pbd* packages providing distributed parallel infrastructure in *R* (see r-pbd.org).

Key Words: parallel computing, big data, supercomputing, model-based clustering