Version 1.0

Programming with Big Data in R

Speaking Serial R with a Parallel Accent

Package Examples and Demonstrations

pbdR Core Team

Speaking Serial R with a Parallel Accent

pbdR Package Examples and Demonstrations

Drew Schmidt
Remote Data Analysis and Visualization Center,
University of Tennessee, Knoxville

Wei-Chen Chen
Computer Science and Mathematics Division,
Oak Ridge National Laboratory

Pragneskumar Patel
Remote Data Analysis and Visualization Center,
University of Tennessee, Knoxville

George Ostrouchov
Computer Science and Mathematics Division,
Oak Ridge National Laboratory

Version 1.0

This publication was typeset using LaTeX. The illustrations were created using the **ggplot2** package (Wickham, 2009), except for Figure **??**, which was created in Microsoft Powerpoint.

Contents

List of Figures

List of Tables

Acknowledgements

Schmidt, Ostrouchov, and Patel were supported in part by the project "NICS Remote Data Analysis and Visualization Center" funded by the Office of Cyberinfrastructure of the U.S. National Science Foundation under Award No. ARRA-NSF-OCI-0906324 for NICS-RDAV center. Chen and Ostrouchov were supported in part by the project "Visual Data Exploration and Analysis of Ultra-large Climate Data" funded by U.S. DOE Office of Science under Contract No. DE-AC05-00OR22725.

This work used resources of National Institute for Computational Sciences at the University of Tennessee, Knoxville, which is supported by the Office of Cyberinfrastructure of the U.S. National Science Foundation under Award No. ARRA-NSF-OCI-0906324 for NICS-RDAV center. This work also used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work used resources of the Newton HPC Program at the University of Tennessee, Knoxville.

We also thank Brian D. Ripley, Kurt Hornik, Uwe Ligges, and Simon Urbanek from the R Core Team for discussing package release issues and helping us solve portability problems on different platforms.

Warning: This document is written to explain the main functions of **pbdDEMO** (Schmidt *et al.*, 2013), version 0.1-0. Every effort will be made to ensure future versions are consistent with these instructions, but features in later versions may not be explained in this document.

Information about the functionality of this package, and any changes in future versions can be found on website: "Programming with Big Data in R" at `http://r-pbd.org/`.

Part I

Applications

1 Model-Based Clustering

1.1 Introduction

Model-based clustering is an unsupervised learning technique and mainly based on finite mixture models to fit the data, cluster the data, and infer the data (Fraley and Raftery, 2002; Melnykov and Maitra, 2010). The major application of model-based clustering focuses on Gaussian mixture models. For example, $\boldsymbol{X}_n$ is a random p-dimensional observation from the Gaussian mixture model with K components which has a density

$$f(\boldsymbol{X}_n; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \eta_k \phi_p(\boldsymbol{X}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1.1}$$

where $\phi_p(\cdot; \cdot, \cdot)$ is a p-dimensional Gaussian density,

$$\boldsymbol{\Theta} = \{\eta_1, \eta_2, \ldots, \eta_{K-1}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_K\},$$

is the parameter space, η_k's are mixing proportion, $\boldsymbol{\mu}_k$'s are centers of components, and $\boldsymbol{\Sigma}_k$'s are dispersion of components.

Suppose a data set $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ has N observations from the Equation (1.1), then the log likelihood is

$$\log L(\boldsymbol{\Theta}; \boldsymbol{X}) = \sum_{n=1}^{N} \log f(\boldsymbol{X}_n; \boldsymbol{\Theta}) \tag{1.2}$$

which is usually solved by the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). After the EM algorithm is converged, let $\hat{\boldsymbol{\Theta}}$ is maximum likelihood estimator of Equation (1.2), then the maximum posterior probability

$$\underset{k}{\operatorname{argmax}} \frac{\hat{\eta}_k \phi_p(\boldsymbol{X}_n; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{f(\boldsymbol{X}_n; \hat{\boldsymbol{\Theta}})}$$

for all $n = 1, 2, \ldots, N$ indicates the membership of the data set $\boldsymbol{X}$.

mclust (Fraley *et al.*, 1999) and **EMCluster** (Chen *et al.*, 2012a) are two main R packages implementing the EM algorithm for the model-based clustering. The **mclust** has several selections

on models, and the **EMCluster** implements the most complicated model (dispersions are all unstructured) in a more efficient way, several initializations, and semi-supervised learning. Both are assuming small N and tiny p and only run in serial with sufficient memory.

Note that the k-means algorithm (Forgy, 1965) equivalently assumes $\eta_1 = \eta_2 = \cdots = \eta_K \equiv 1/K$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_K \equiv \boldsymbol{I}$ in the Equation (1.1) where $\boldsymbol{I}$ is an identity matrix. The k-means algorithm is a restricted Gaussian mixture model such that it can be implemented in a simplified way of the EM algorithm. However, the cluster results are always unrealistic, unable to infer, and sometimes seriously with high classification errors.

1.1.1 Parallel Model-Based Clustering

pmclust (Chen and Ostrouchov, 2012) is an R package for parallel model-based clustering based on mixture Gaussian models with unstructured dispersions. The package assumes data are distributed on several machines, therefore, some gathering and reducing are necessary at some stages of EM algorithm. An expectation-gathering-maximization (EGM) algorithm (Chen *et al.*, 2013) is established for minimizing communication and data moving between machines. There are four variants of EGM-like algorithms implemented in **pmclust** include EM, AECM (Meng and van Dyk, 1997) , APECM (Chen and Maitra, 2011) , and APECMa (Chen *et al.*, 2013) . The variants are trying to yield better convergent rate and less computing time than the original EM algorithm. A simple k-means algorithm is also implemented in **pmclust**.

The **pmclust** is the first pbdR application, and the first R package in SPMD to analyze distributed data in Terabytes level. Originally, it is designed for analyzing Climate simulation outputs (CAM5) as discussed in Section **??**, and is a product for the project "Visual Data Exploration and Analysis of Ultra-large Climate Data" supported by U.S. DOE Office of Science.

The **pmclust** initially depended on **Rmpi**, but designed in SPMD approach rather than in master/workers approach even before pbdR. Later, it migrates to use **pbdMPI** (Chen *et al.*, 2012b) for performance issues of larger machines. So, by default, the package assumes data are stored in SPMD row-major matrix format. Currently, the package also fully utilizes **pbdSLAP** (Chen *et al.*, 2012c), **pbdBASE** (Schmidt *et al.*, 2012a), and **pbdDMAT** (Schmidt *et al.*, 2012b) to implement algorithms in `ddmatrix` format. Table 1.1 lists the current implementations.

Table 1.1: Parallel Mode-Based Clustering Algorithms in **pmclust**

Algorithm	SPMD	`ddmatrix`
EM	yes	no
AECM	yes	no
APECM	yes	no
APECMa	yes	no
k-means	yes	yes

Based on **pmclust** version 0.1-4

1.2 The *Iris* Dataset

The `iris` (Fisher, 1936) dataset in R is a tiny dataset for 50 iris flowers from each of three species of iris which are *Iris setosa*, *Iris versicolor*, and *Iris virginica*. The dataset has in total 150 rows and five columns including four features (sepal length and width, petal length and width) and class of species. We take the first four columns of `iris` to form $\boldsymbol{X}$ matrix where each row can be classified in three groups by the true id (the fifth column of `iris`) for supervised learning, or clustered in three groups by algorithms for unsupervised learning. Note that the dimension of $\boldsymbol{X}$ is $N = 150$ by $p = 4$.

From the supervised learning point view, the empirical estimation for $\boldsymbol{\Theta}$ from data will be the best description for the data assuming the true model is Gaussian mixture. The demo code `iris_overlap` in **pbdDEMO** quickly suggests the overlap level of three iris species. It can be obtained by

R Code

```
R> demo(iris_overlap, 'pbdDEMO', ask = F, echo = F)
```

which utilize `overlap` function of **MixSim** (Melnykov *et al.*, 2012). The output is

R Output

```
R> (ret <- overlap(ETA, MU, S))
$OmegaMap
             [,1]          [,2]        [,3]
[1,] 1.000000e+00 7.201413e-08 0.00000000
[2,] 1.158418e-07 1.000000e+00 0.02302315
[3,] 0.000000e+00 2.629446e-02 1.00000000

$BarOmega
[1] 0.01643926

$MaxOmega
[1] 0.0493176

$rcMax
[1] 2 3

R> (levels(iris[, 5]))
[1] "setosa"     "versicolor" "virginica"
```

The `OmegaMap` is a map of pair-wised overlap of three species where 1, 2, 3 are *Iris setosa*, *Iris versicolor*, and *Iris virginica*, respectively. The outputs also indicate that the averaged pair-wised overlap (`BarOmega`) is about 1.6%, and the maximum pair-wised overlap (`MaxOmega`) is about 4.9% among these three iris species. Also, the maximum occurs at 2 (*Iris versicolor*) and 3 (*Iris virginica*) indicating these two species are partly inseparable given these four features.

For unsupervised learning point view, such as model-based clustering, suppose we were blinded to the true class ids or assuming the fifth column of $\boldsymbol{X}$ is unobserved, but only use the four features to form the model and cluster the data. Note that *Iris versicolor* and *Iris virginica* are

partly inseparable, so misclassification can happen at the overlap region. We validate the results by comparing the clustering ids to the true class ids using adjusted Rand index (Hubert and Arabie, 1985). The adjusted Rand index has values between 1 and -1 where 1 means perfect match otherwise less than 1. The function `RRand` in **MixSim** also provide the adjusted Rand index.

In order to show the unsupervised learning, we then use the `iris` in the following steps to show the scalability of **pmclust**. We first illustrate the `iris` in a serial code:

- decompose the $\boldsymbol{X}$ on principal components,
- project the $\boldsymbol{X}$ on the first two dimension with largest variation,
- visualize the $\boldsymbol{X}$ on the x-y plane,
- label $\boldsymbol{X}$ with true ids, and
- label $\boldsymbol{X}$ with estimated ids clustered by algorithms.

Then, we repeat these steps in SPMD code and in `ddmatrix` code to show the similarity of codes. This shows that **pmclust** can cluster data from very tiny dataset on single machines, but there is no difficulty to scale to very large dataset on supercomputers.

1.2.1 *Iris* in Serial Code

1.2.2 *Iris* in SPMD Code

1.2.3 *Iris* in `ddmatrix` Code

1.3 Exercises

1-1

References

Chen WC, Maitra R (2011). "Model-Based Clustering of Regression Time Series Data via APECM — an AECM Algorithm Sung to an Even Faster Beat." *Statistical Analysis and Data Mining*, **4**, 567–578.

Chen WC, Maitra R, Melnykov V (2012a). "EMCluster: EM Algorithm for Model-Based Clustering of Finite Mixture Gaussian Distribution." R Package, URL http://cran.r-project.org/package=EMCluster.

Chen WC, Ostrouchov G (2012). "pmclust: Parallel Model-Based Clustering." R Package, URL http://cran.r-project.org/package=pmclust.

Chen WC, Ostrouchov G, Pugmire D, Prabhat M, Wehner M (2013). "A Parallel EM Algorithm for Model-Based Clustering with Application to Explore Large Spatio-Temporal Data." *Technometrics*. (in revision).

Chen WC, Ostrouchov G, Schmidt D, Patel P, Yu H (2012b). "pbdMPI: Programming with Big Data – Interface to MPI." R Package, URL http://cran.r-project.org/package=pbdMPI.

Chen WC, Schmidt D, Ostrouchov G, Patel P (2012c). "pbdSLAP: Programming with Big Data – Scalable Linear Algebra Packages." R Package, URL http://cran.r-project.org/package=pbdSLAP.

Dempster A, Laird N, Rubin D (1977). "Maximum Likelihood for Incomplete Data via the EM Algorithm (with discussion)." *Jounal of the Royal Statistical Society, Series B*, **39**, 1–38.

Fisher R (1936). "The use of multiple measurements in taxonomic problems." *Annals of Eugenics*, **2**, 179–188.

Forgy E (1965). "Cluster analysis of multivariate data: efficiency vs. interpretability of classifications." *Biometrics*, **21**, 768–780.

Fraley C, Raftery A (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association*, **97**, 611–631.

Fraley C, Raftery A, Scrucca L (1999). "mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation." R Package, URL http://cran.r-project.org/package=mclust.

Hubert L, Arabie P (1985). "Comparing partitions." *Journal of Classification*, **2**, 193–218.

Melnykov V, Chen WC, Maitra R (2012). "MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms." *Journal of Statistical Software*, **51**(12), 1–25. URL http://www.jstatsoft.org/v51/i12/.

Melnykov V, Maitra R (2010). "Finite Mixture Models and Model-Based Clustering." *Statistics Surveys*, **4**, 80–116.

Meng X, van Dyk D (1997). "The EM Algorithm — an Old Folk-song Sung to a Fast New Tune (with discussion)." *Journal of the Royal Statistical Society B*, **59**, 511–567.

Schmidt D, Chen WC, Ostrouchov G, Patel P (2012a). "pbdBASE: Programming with Big Data – Core pbd Classes and Methods." R Package, URL http://cran.r-project.org/package=pbdBASE.

Schmidt D, Chen WC, Ostrouchov G, Patel P (2012b). "pbdDMAT: Programming with Big Data – Distributed Matrix Algebra Computation." R Package, URL http://cran.r-project.org/package=pbdDMAT.

Schmidt D, Chen WC, Ostrouchov G, Patel P (2013). "pbdDEMO: Programming with Big Data – Demonstrations of pbd Packages." R Package, URL http://cran.r-project.org/package=pbdDEMO.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

Index