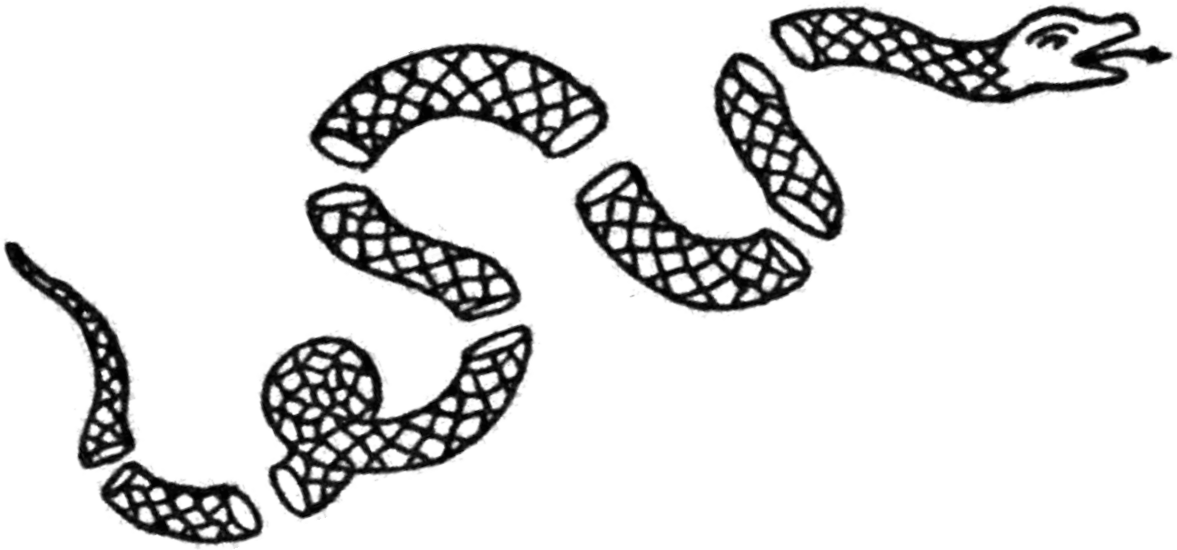
Version 0.3-0



Programming with Big Data in R

Speaking Serial R with a Parallel Accent (Ver. 0.3-0)

Package Examples and Demonstrations

pbdR Core Team

Speaking Serial R with a Parallel Accent (Ver. 0.3-0)

pbdR Package Examples and Demonstrations

June 12, 2014

Drew Schmidt
National Institute for Computational Sciences
University of Tennessee

Wei-Chen Chen
Department of Ecology and Evolutionary Biology
University of Tennessee

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

Pragneshkumar Patel
National Institute for Computational Sciences
University of Tennessee

Version 0.3-0

This publication was typeset using LaTeX. Illustrations were created using the **ggplot2** package (Wickham, 2009), native R functions, and Microsoft Powerpoint.

Contents

List of Figures ii
List of Tables iii
Acknowledgements iv
Disclaimer 1

1 Pairwise Distance and Comparisons **2**
1.1 Introduction 2
1.2 Distributed Distance and Comparisons 3
1.3 Hierarchical Clustering 4
1.4 Neighbor Joining Tree 4
1.5 Exercises 5

References **6**

Index **7**

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 was supported in part by the Department of Ecology and Evolutionary Biology at the University of Tennessee, Knoxville, and a grant from the National Science Foundation (MCB-1120370.)

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.

Disclaimer

Warning: The findings and conclusions in this article have not been formally disseminated by the U.S. Department of Energy and should not be construed to represent any determination or policy of University, Agency and National Laboratory.

This document is written to explain the main functions of **pbdDEMO** (Schmidt *et al.*, 2013), version 0.3-0. 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/`.

1

Pairwise Distance and Comparisons

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

—**John Tukey**

1.1 Introduction

Distance method is not only a fundamental tool in geometry, but also appears in statistics and other applied disciplines. For example, least square method in regression can be simply derived and computed via Euclidean distance. The resulting line is an approximate answer in terms of minimum total distance to all observations. Distance is also related to a similarity measure of two observations describing relationship of the two. Usually, the smaller of distance the closer of relation. For example, the higher probability (probability is a measure) of one virus evolving to a mutant means the smaller distance (related closely) of two viruses as described in Chapter **??**. Further, distance method is simple to apply on clustering problems and easy to visualize data structures such as K-means algorithm a special case of model-based clustering introduced in Chapter **??**. For instance, the observations of the same group are more similar in characteristics with each other than those between different groups.

Potentially, computing distance of several observations involves half of pairwise comparisons if distance is symmetric, and involves all pairwise comparisons if distance is not symmetric. Also, if number of observations is small, then most of distance methods can be compute efficient within one core. For moderate number of observations or complex distance systems, the computing can be parallelized wisely in several levels. For example, one may utilize multiple threads or co-processors to archive performance gains. For large number of observations, the computing is not trivial if data are distributed across cores. Further, the dimension of resulting distance array may be much larger the number of observations and can only be held distributed across cores. Note that for some models or iterative algorithms, it is not wise to dump the distance array into disk since that decreases performance due to overhead cost for I/O. For example, one may utilize distributed parallelization to avoid these restrictions.

In the context of **pbdR**, we focus on distributed methods and abstract computing of distance to allow user-defined comparison (dissimilarity) functions of any two observations. We briefly introduce issues and methods of distributed distance and comparisons first, and followed by demonstration of hierarchical clusterings on the `iris` dataset of Chapter **??**. This example can be done using exists distance function in R. Further, we provide a biological application of building phylogenetic trees on the *Pony 524* dataset of Chapter **??** utilizing evolutionary models to compute probability distance. This example demonstrate how user-defined function can be defined and used to obtain special distance.

1.2 Distributed Distance and Comparisons

Suppose x and y are two observations and $d(x, y)$ is a distance or a comparison of x and y. Note that x, y, and $d(\cdot, \cdot)$ could be very generic as long as they are well defined. Although, it is efficient to compute a distance of any two observations in R via `dist()` serially, it becomes non-trivial to compute distance of distributed observations in parallel.

The potential problems include:

(P1) Communication must be evoked between processors when any two observations are not located within the same processor.

(P2) The resulting distance matrix may be too big to held in one processor as data size increased even only a half (lower triangular matrix is stored as row-major in a vector.)

(P3) Compute all comparisons may be too time consuming even for small data sets.

Distributed situations of observations and computed results (distance matrix) are categorized next.

(C1) Both observations and distance matrix are in one node and may both be in serial or in parallel within the node, typically via OpenMP (OpenMP ARB, 1997).

(C2) Observations are in common in all processors and distance matrix is distributed across nodes.

(C3) Observations are distributed across nodes and distance matrix is in common in all nodes.

(C4) Both observations and distance matrix are distributed across nodes.

Here, we may presume the distribution method is GBD row-major matrix (or row-block major) as introduced in Section **??** since most of native R functions can be extend and reused in such a way.

Note that the `dist()` only supports a few distance methods and assume distance is symmetric by definition. However, in practice, a more general measure may not be necessarily symmetric of two observations. i.e. $d(x, y) \neq d(y, x)$. In some cases, $d(x, x) \neq 0$ and the distance may also be dependent on other measurements or conditions.

1.3 Hierarchical Clustering

Hierarchical clustering is a popular statistical tools in fundamental multivariate statistics and is heavily based on distance matrix to classify data. Several algorithms are proposed to build dendrograms or trees, and prune branches of the resulting trees to identify possible subgroups. The basic function `hclust()` takes a dissimilarity structure as produced by `dist()` and returns a tree object can be visualized. The method option "average" linkage is equivalent to UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method (Sokal and Michener, 1985) one of popular methods in ecology for classification.

For example, the `iris` dataset used in Chapter **??** can be clustered in hierarchical clustering. First, we distribute 150 observations in four cores and compute Euclidean distances in four dimensional space ('Sepal.Length', 'Sepal.Width', 'Petal.Length', and 'Petal.Width'). Note that the distance may not be meaningful to the data, but preserve some (dis-) similarity of the observations. We compute the dissimilarity matrix in distributed manners via a utility function `comm.dist()` of **pbdMPI** (Chen *et al.*, 2012) and store the result in a common matrix across all cores. We based on the matrix to perform a UPGMA clustering. The example in SPMD can be found in demo via

```
### At the shell prompt, run the demo with 4 processors by
### (Use Rscript.exe for windows system)
mpiexec -np 4 Rscript -e "demo(dist_iris,'pbdDEMO',ask=F,echo=F)"
```

and it returns a dendrogram as Figure **??**.

1.4 Neighbor Joining Tree

In some sense, Figure **??** is a rooted tree and the "average" method as well as UPGMA assumes a constant rate of evolution (molecular clock hypothesis). However, these assumption may not be appropriate to most sequence evolutionary topics where a gene tree should be more suitable to interpret relation of sequences and species. We introduce a common approach in evolution biology and build a evolutionary tree for *Pony 524* dataset. We select JC69 evolutionary model (Jukes and Cantor, 1969) as a probability measure to compute for distance (evolution time) of 146 EIAV sequences and use a neighbor joining tree (Saitou and Nei, 1987) to build an unrooted tree.

The purpose is to design a wrapper function, says `my.dist(x, y)`, that takes a pairs of observations `x` and `y` as inputs, and returns a user-defined distance of given data. The utility function `comm.pairwise()` of **pbdMPI** (Chen *et al.*, 2012) is more flexible than `comm.dist()`. Through the options `pairid.gbd` and `FUN = my.dist`, the function can evaluate `my.dist()` on the given dataset `X` in row major blocks. For *Pony 524*, the `X` is the DNA sequences and `my.dist()` is a wrapper of `phyclust.edist`.

The example in SPMD can be found in demo via

```
### At the shell prompt, run the demo with 4 processors by
### (Use Rscript.exe for windows system)
```

```
mpiexec -np 4 Rscript -e "demo(dist_pony,'pbdDEMO',ask=F,echo=F)"
```

and it returns a neighbor-joining tree as Figure **??**.

1.5 Exercises

1-1 What are potential limitations of distance approaches?

1-2 Prove that clustering based on Euclidean distance is equivalent to that clustering based on multivariate Normal distributions with identity variance covariance matrices.

1-3 Prove that the "average" method of `hclust()` is equivalent to the UPGMA method.

1-4 Given n observations or taxa, find total numbers of possible rooted and unrooted trees analytically, $(2n-5)!!$ and $(2n-3)!!$.

1-5 As number of observations increases, the data and the distance matrix are both distributed as the category (C4). State potential problems of implementations and minimum costs of communications.

References

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

Jukes T, Cantor C (1969). *Evolution of Protein Molecules*. New York: Academic Press.

OpenMP ARB (1997). "OpenMP." URL http://www.openmp.org/.

Saitou N, Nei M (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, **4**, 406–425.

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.

Sokal R, Michener C (1985). "A statistical method for evaluating systematic relationships." *University of Kansas Science Bulletin*, **38**, 1409–1438.

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

Code
- `comm.dist()`, 4
- `comm.pairwise()`, 4
- `dist()`, 3
- `hclust()`, 4
- `phyclust.edist()`, 4

Data
- iris, 3
- Pony 524, 3, 4

Library
- OpenMP, 3

UPGMA, 4