The R package bigstatsr: Memory- and Computation-Efficient Tools for Big Matrices

useR!2017 lightning talk

Florian Privé (@privefl)

July 6, 2017
About

I'm a PhD Student (2016-2019) in Predictive Human Genetics in Grenoble.

Disease ~ DNA mutations
Very large genotype matrices

- currently: 15K x 300K, celiac disease
- soon: 500K x 800K, UK Biobank
Problem I had

Celiac disease dataset as standard R matrix

32GB

RAM

8GB
Solution I found

Celiac disease dataset as standard R matrix

32GB

RAM

8GB

Disk storage space

500GB

big.matrix

R package bigmemory

Michael J. Kane, John Emerson, Stephen Weston (2013).
Similar accessor as R matrices

CAUTION

THIS MACHINE HAS NO BRAIN
USE YOUR OWN
Split-(par)Apply-Combine Strategy

Apply standard R functions to big matrices (in parallel)

strategy coined by Hadley Wickham (2011)
Similar accessor as Rcpp matrices

In Rcpp we trust
Partial Singular Value Decomposition

15K x 100K big matrix, 6 cores, K = 10, 1 min (vs 2h in base R)

based on R package RSpectra
Sparse linear models: biglasso

Other functions

- matrix operations (Split-Apply-Combine strategy)
- association of each variable with an output (RcppArmadillo)
- plotting functions (ggplot2)
- read from text files
- others..
I'm now able to run algorithms on 100GB of data
# R Packages

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigmemory</td>
<td><code>big.matrix</code> object</td>
</tr>
<tr>
<td>bigstatsr</td>
<td>Statistical functions for <code>big.matrix</code> objects to be used by any field</td>
</tr>
<tr>
<td>bigsnpr</td>
<td>Specific functions for my field of research</td>
</tr>
</tbody>
</table>

Paper in preparation: "Efficient management and analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr".
Contributors are welcomed!
Thanks!

Package's website: https://privefl.github.io/bigstatsr/

Twitter and GitHub: @privefl

Presentation available online: https://goo.gl/cv7L5s

Slides created via the R package xaringan.