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

F. Privé (a), H. Aschard (b) and M.G.B. Blum (a)

(a) Laboratoire TIMC-IMAG – Université Grenoble Alpes - CNRS – Faculté de Médecine - 38706 La Tronche cedex - France

florian.prive.21@gmail.com – florian.prive@univ-grenoble-alpes.fr – michael.blum@univ-grenoble-alpes.fr

(b) Département de Génomes et Génétique – Centre de Bioinformatique, Biostatistique et Biologie Intégrative – Institut Pasteur - 25-28 Rue du Dr Roux, 75015 Paris - France

hugues.aschard@pasteur.fr

Mots clefs : Statistics, Big Data, Memory-mapping, Parallelism.

Abstract

The *R* package **bigstatsr** (https://github.com/privefl/bigstatsr) provides functions for fast statistical analysis of large-scale data encoded as matrices (Privé et al. 2018). The package can handle matrices that are too large to fit in memory. The package **bigstatsr** is based on a similar format (called `FBM`) as the format `big.matrix` provided by the *R* package **bigmemory** (Kane, Emerson, and Weston 2013).

The package **bigstatsr** enables users with laptop to perform statistical analysis of several dozens of gigabytes of data. The package is fast and efficient because of four different reasons. First, **bigstatsr** is memory-efficient because it uses only small chunks of data at a time. Second, special care has been taken to implement effective algorithms. Third, `FBM` objects use memory-mapping, which provides efficient accesses to matrices. Finally, as matrices are stored on-disk, many processes can easily access them in parallel.

The main features currently available in **bigstatsr** are:

- partial singular value decomposition (SVD) via randomized projections (Lehoucq and Sorensen 1996),
- sparse linear and logistic regressions (Zeng and Breheny 2017),
- column-wise linear and logistic regressions tests,
- matrix operations,
- parallelization / apply.

References

Kane, Michael J, John W Emerson, and Stephen Weston. 2013. "Scalable Strategies for Computing with Massive Data." *Journal of Statistical Software* 55 (14): 1–19. doi:10.18637/jss.v055.i14.

Lehoucq, Rich Bruno, and D. C. Sorensen. 1996. "Deflation Techniques for an Implicitly Restarted Arnoldi Iteration." *SIAM Journal on Matrix Analysis and Applications* 17 (4). Society for Industrial; Applied Mathematics: 789–821. doi:10.1137/S0895479895281484.

Privé, Florian, Hugues Aschard, Andrey Ziyatdinov, and Michael G B Blum. 2018. "Efficient Analysis of Large-Scale Genome-Wide Data with Two R Packages: Bigstatsr and Bigsnpr." *Bioinformatics*, bty185. doi:10.1093/bioinformatics/bty185.

Zeng, Yaohui, and Patrick Breheny. 2017. "The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R," January. http://arxiv.org/abs/1701.05936.