R2GUESS: a GPU-based R package for sparse Bayesian variable selection

L. Bottolo[b], **M. Chadeau-Hyam**[b], **B. Liquet**[a], **S. Richardson**[a] and **H. Saadi**[b]

[a]MRC Biostatistics Unit
Institute of Public Health
University Forvie Site, Cambridge, UK
benoit.liquet@isped.u-bordeaux2.fr

[b]Department of Epidemiology and Biostatistics
Imperial College London
St Mary's Campus, London, UK
h.saadi@imperial.ac.uk

Mots clefs : Biology, Genomics, Bayesian, Variable Selection, GPU, High Dimensional Statistics.

Recent advances in high throughput "omics" technologies have given rise to a wealth of novel high dimensional data, ranging from thousands to hundreds of thousand variables, each demonstrating complex correlation structures. These data comprise genetic, epigenetic and transcriptomic profile which have shown a great potential in measuring the abundance of biologically relevant molecules over a whole biological system. The analysis of such complex data raises serious statistical challenges relating to the fact that the number of predictors exceeds the number of observations ("large p, small n" scenario).

Alongside multiple testing correction strategies, variable selection approaches are well suited to handle this situation, and we propose here a Bayesian implementation of this kind of approaches. As such, the method seeks for the best combination of covariates to predict the (possibly multivariate) outcome. The Bayesian framework it is based on allows for the construction of parsimonious regression models, adopting prior specifications that translate expected sparsity of the underlying biology, and therefore facilitating results interpretation.

R2GUESS is an R package that interfaces a C++ implementation of a fully Bayesian Variable Selection approach for multivariate linear regression . Using latest computational advancement, it can run on GPU (Graphical Processing Unit), and in its current form it enables the analysis of hundreds of thousands of predictors measured in thousands of individuals simultaneously. The efficient exploration of the 2^n dimensional space is possible thanks to the use of an Evolutionary Monte Carlo sampling scheme comprising a large portfolio of local and global moves. R2GUESS also provides refined numerical and graphical output facilitating post-processing and subsequent interpretation of the extensive output produced by the GUESS algorithm. Performances of the model and interpretability of its results are illustrated with examples from several omics platforms.

References

[1] Bottolo, L., Richardson, S., (2010). Evolutionary Stochastic Search for Bayesian Model Exploration. *Bayesian Analysis*, **5**, 583-618.
[2] Bottolo, L., Chadeau-Hyam, M., Hastie D.I., Langley, S.R., Petretto, E., Tiret, L., Tregouet, D., Richardson, S. (2011). ESS++: a C++ objected-oriented algorithm for Bayesian stochastic

search model exploration. *Bioinformatics*, **27**, 587-588.