Efficient inference for genetic association studies with multiple outcomes

Hélène Ruffieux[*†] Anthony C. Davison[*] Jörg Hager[†] Irina Irincheeva[†]

Abstract

Combined inference for heterogeneous high-dimensional data is critical in modern biology, where clinical and various kinds of molecular data may be available from a single study. Classical genetic association studies regress a single clinical outcome on many genetic variants one by one, but there is an increasing demand for analysing jointly a large number of molecular outcomes and genetic variants in order to unravel functional interactions. Unfortunately, most existing approaches to joint modelling are either too simplistic to be powerful or are impracticable for computational reasons. Inspired by Richardson et al. (2010, Bayesian Statistics 9), we consider a sparse multivariate regression model that allows simultaneous selection of predictors and associated responses by borrowing information across responses with shared associations. As Markov chain Monte Carlo (MCMC) inference on such models can be prohibitively slow when the number of genetic variants exceeds a few thousand, we instead propose a variational Bayes approach which produces posterior information very close to that of MCMC inference, at a much reduced computational cost. Extensive numerical experiments show that our approach outperforms popular variable selection methods and tailored Bayesian procedures, dealing within hours with problems involving hundreds of thousands of genetic variants and tens to hundreds of clinical or molecular outcomes.

Key Words: Computational efficiency; Genetic variant; High-dimensional data; Molecular quantitative trait loci analysis; Sparse multivariate regression; Variable selection; Variational inference.

Preprint available on `arXiv:1609.03400`

[*]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[†]Nestlé Institute of Health Sciences SA, Lausanne, Switzerland