atlasqtl: an R package for variable selection in sparse regression with hierarchically-related responses

Hélène Ruffieux

Statistical studies aimed at identifying associations between a high-dimensional vector of candidate predictors and large numbers of correlated outcomes multiply, prompted by the proliferation of technologies capable of measuring large volumes of information, whether on health and lifestyle parameters, molecular entities or even galaxies. As well as growing in dimension, the datasets collected are also growing in complexity, which calls for elaborate and flexible modelling approaches. When coupled with robust and efficient inference, Bayesian hierarchical modelling is a powerful framework for describing intricate dependencies at the scale required by current applications, while conveying uncertainty in a coherent fashion. In this note, we present the R package `atlasqtl`, which implements a scalable hierarchical framework for variable selection in regression problems with high-dimensional predictor and response spaces.

Model and inference

The model consists of a series of hierarchically-related spike-and-slab regressions that permit borrowing information across a large number of responses. A graphical representation is provided in Figure 1 — the full prior specification is given in Ruffieux et al. (2020). Each pair of predictor X_s and response y_t has a corresponding regression coefficient β_{st} and spike-and-slab binary latent variable γ_{st}, from which posterior probabilities of association, $\text{pr}(\gamma_{st} = 1 \mid y)$, can be obtained and employed to implement Bayesian false discovery rate control. The model is also tailored to the detection of *hotspots*, namely, predictors associated with several responses: the top-level hierarchy entails a probit submodel which involves a response-specific contribution to the spike-and-slab probability of association, via ζ_t, and a predictor-specific modulation of this contribution, via θ_s. The latter parameter also acts as the propensity of predictor X_s to be a hotspot and is assigned a horseshoe prior; such a global-local prior specification permits adapting to the overall problem sparsity (via the global scale σ_0), while flexibly capturing hotspot effects (via the Cauchy tail of the local scale λ_s). Joint inference requires special attention as the binary latent matrix $\Gamma = \{\gamma_{st}\}$ creates a discrete search space of dimension $2^{p \times q}$. To overcome this complication, `atlasqtl` implements a variational inference algorithm based on a structured mean-field factorisation and efficient batch updates. The algorithm is augmented with a simulated annealing scheme that improves the exploration of highly multimodal spaces by introducing so-called "temperature" parameters controlling the degree of separation of the modes of a series of "heated" distributions (Ruffieux et al. (2020)).

`atlasqtl` can be employed in any sparse multiple-response regression setting. Hereafter we discuss a use case in the context of expression quantitative trait locus (eQTL) analysis, in which *hotspot genetic variants*, controlling many gene expression traits at once, may be responsible for important functional mechanisms underlying specific disease endpoints — such studies aim to clarify the genetic architecture of diseases by estimating associations between up to a few million candidate predictors (genetic variants) and several thousand responses (molecular traits, such as genomic, proteomic, metabolomic levels).

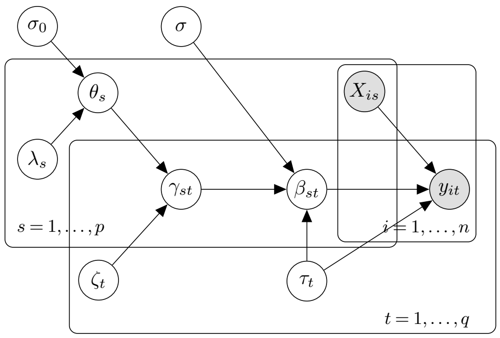

Figure 1: *Graphical representation of the model. The shaded nodes are observed, the others are inferred.*

Installation

The `atlasqtl` package is written in R with C++ subroutines. It can be installed following the instructions provided at https://github.com/hruffieux/atlasqtl#installation. Note that the GSL library must be installed prior to installing the package.

Data and reproducibility

The material used in this example is available at https://github.com/hruffieux/software_corner_ibs_bulletin for replication purposes (Rmarkdown document). It involves synthetic genotyping data generated using our in-house R package `echoseq`; these data emulate real genotypes (single nucleotide polymorphisms, "SNPs") in the *LYZ* gene region. The gene expressionn traits consist of monocyte levels before and after immune stimulation (resting, IFNg, LPS2h, LPS24h), as well as B-cell levels. To each of these transcriptomic datasets corresponds a separate eQTL problem.

Analysis

We illustrate the main package functionalities on the five eQTL problems (corresponding datasets stored in the list `list_data`). `atlasqtl` requires the specification of just two hyperparameters via the elicitation of a prior mean `mu_t` and variance `v_t` for the number of predictors associated with each response. The sensitivity of inference to these choices should be evaluated on a case-by-case basis; it is however very limited when a permutation-based FDR threshold is employed, where the "null-case" permutation runs use the same hyperparameter specifications. The remaining model parameters are inferred by variational inference. In particular, the horseshoe prior on the hotspot propensity circumvents ad-hoc choices of top-level variance parameters (for which inference is prone to spurious effects). This specification also has desirable multiplicity adjustment properties in large-response settings, see Ruffieux et al. (2020).

The code below runs the five `atlasqtl` analyses in parallel. The annealing schedule (initial temperature, grid size and type of spacing) can be supplied via the argument `anneal`; here we use a default geometric schedule on the inverse temperature, with initial temperature of 2 and grid size of 10. The argument `add_collinear_back` is a boolean indicating whether or not to include the final posterior summaries for all possible collinear variables in X (which are removed during the run to ensure the stability of inference). A number of additional settings (e.g., tolerance, maximum number of iterations, checkpointing, etc.) can be chosen as detailed on the help page of the function, which is accessed by running `?atlasqtl` in the console.

```r
require(atlasqtl)
```

```r
mu_t <- 1; v_t <- 4 # hyperparameter specification

vec_type <- names(list_data) # names of the eQTL problems
                             # (resting, IFNg, LPS2h,
                             # LPS24h monocytes and B-cells).

n_cpus <- 4 # nb of CPUs for parallel execution - please
            # change this number according to your setup.

list_out <- parallel::mclapply(vec_type, function(type) {

  snps <- list_data[[type]]$snps
  expr <- list_data[[type]]$expr

  atlasqtl(Y = expr, X = snps,
           p0 = c(mu_t, v_t),
           add_collinear_back = TRUE)
}, mc.cores = n_cpus)

names(list_out) <- vec_type
```

The object returned by `atlasqtl` contains posterior quantities, which can be employed to assess:

- pairwise associations between each pair of SNP and trait: using the variational posterior probabilities (PPIs) stored in `gam_vb` ($p \times q$ matrix) and the variational posterior means of the regression estimates stored in `beta_vb` ($p \times q$ matrix);
- hotspot propensities: using the variational posterior mean of θ_s stored in `theta_vb` (vector of length p).

The custom `print.atlasqtl`, `summary.atlasqtl` and `plot.atlasqtl` S3 functions provide further information about the run and a summary of the above posterior quantities; e.g., for the LPS2h-monocyte eQTL analysis:

```r
print(list_out$lps2h)
```

```
## *******************************************************
## Successful convergence after 84 iterations, using a
## tolerance of 0.1 on the absolute changes in the ELBO.
## *******************************************************
##
## Geometric annealing on the inverse temperature was
## applied for the first 10 iterations, with initial
## temperature of 2 (default).
##
## Number of samples: 260;
## Number of (non-redundant) candidate predictors: 100;
## Number of responses: 1751;
## Prior expectation for the number of predictors
## associated with each response: 1 (sd: 2).
##
## The posterior quantities inferred by ATLASQTL can
## be accessed as list elements from the `atlasqtl` S3
## object, and a summary can obtained using the
## `summary` function.
```

The summary can optionnaly be based on a supplied Bayesian FDR threshold (here 5%); a $p \times q$ matrix of FDR estimates can also be directly obtained using the `assign_bFDR` function.

```r
thres_fdr <- 0.05
summary(list_out$lps2h, thres = thres_fdr,
        fdr_adjust = TRUE, full_summary = FALSE)
```

```
## *******************************************************
## * ATLASQTL: posterior summary for variable selection *
## *******************************************************
##
## Using FDR adjustment of 0.05:
## ----------------------------
##
## Nb of pairwise (predictor-response) associations: 491
##
## Nb of predictors associated with > 1 response (active
## predictors): 28
##
## Hotspot sizes (nb of responses associated with each
## active predictors):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    1.00   17.54    3.00  241.00
##
## Top hotspots:
## snp_99 (size 241), snp_54 (size 189), snp_43 (size 16)
## snp_15 (size 7), snp_49 (size 7), snp_44 (size 4)
```

The Manhattan-type plot indicates the position of the hotspots, as estimated in the five different eQTL analyses, using the Bayesian FDR threshold of 0.05 on the PPIs; alternatively, thanks to the scalability of the variational implementation, permutation-based Bayesian FDR can be obtained as an alternative (Ruffieux et al. (2017) and Ruffieux et al. (2020)).

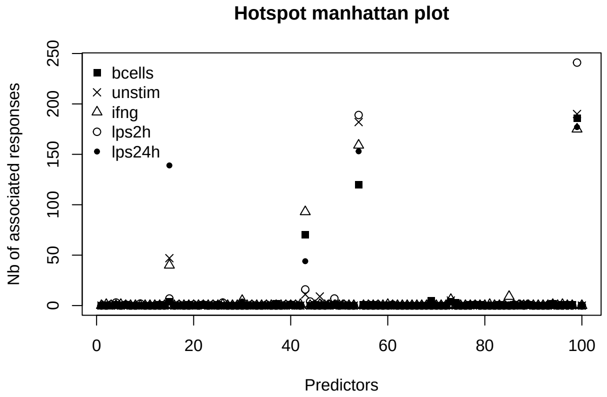

Figure 2: *Hotspot sizes, as obtained using the S3 function plot.atlasqtl.*

Encoding predictor-level information

We recently proposed a new approach, `epispot` (Ruffieux et al. (2021)) based on the same hierarchical framework, to encode and leverage information on the probability of candidate predictors to be involved in associations. Specifically, the effects of a (possibly high-dimensional) predictor-level vector of covariates are estimated using a secondary spike-and-slab regression at the top of the model hierarchy. The R package is available at https://github.com/hruffieux/epispot.

References

Ruffieux, H., A. C. Davison, J. Hager, J. Inshaw, B. Fairfax, S. Richardson, and L. Bottolo. 2020. "A Global-Local Approach for Detecting Hotspots in Multiple Response Regression." *The Annals of Applied Statistics* 14: 905–28.

Ruffieux, H., A. C. Davison, J. Hager, and I. Irincheeva. 2017. "Efficient Inference for Genetic Association Studies with Multiple Outcomes." *Biostatistics* 18: 618–36.

Ruffieux, H., B. Fairfax, I. Nassiri, E. Vigorito, C. Wallace, S. Richardson, and L. Bottolo. 2021. "EPISPOT: an epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies." *The American Journal of Human Genetics* 108: 983–1000.