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

Hélène Ruffieux, MRC Biostatistics Unit, University of Cambridge

Statistical studies aimed at identifying associations between a high-dimensional vector of candidate predictors and a collection of correlated outcomes have multiplied, 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 collected datasets 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 q responses y_t, $t = 1, \ldots, q$, while also accounting for p candidate predictors X_s, $s = 1, \ldots, p$, where both p and q may be larger than the number of samples n. Specifically, to each pair of predictor and response corresponds a regression coefficient β_{st} and a spike-and-slab binary latent variable γ_{st} from which posterior probabilities of association, $\text{pr}(\gamma_{st} = 1 \mid y)$, can employed for selection. The model is also tailored to the detection of *hotspots*, namely, predictors associated with multiple 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 — the corresponding global-local specification adapts to the overall problem sparsity (thanks to the global scale σ_0), while flexibly capturing the individual hotspot effects (thanks to the Cauchy tail of the local scale λ_s). A graphical representation of the model is provided in Figure 1 and its full specification is given in Ruffieux et al. (2020).

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` relies on variational inference with a structured mean-field factorisation and efficient batch updates. The algorithm is augmented with a simulated annealing scheme that improves the exploration of multimodal spaces by introducing a so-called "temperature" parameter controlling the degree of separation of the modes for a series of "heated" distributions (Ruffieux et al. (2020)).

`atlasqtl` can be employed in any sparse multiple-response regression setting with Gaussian errors. Hereafter we discuss a use case in the context of molecular quantitative trait locus (QTL) analysis, in which *hotspot genetic variants*, controlling many molecular 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 several thousand responses (the molecular traits, e.g., genomic, proteomic, metabolomic levels) and up to a few million candidate predictors (the genetic variants). We have previously shown that accounting jointly for all traits and genetic variants with `atlasqtl` substantially improves the detection of weak association effects over more conventional marginal screening approaches (Ruffieux et al. (2020)).

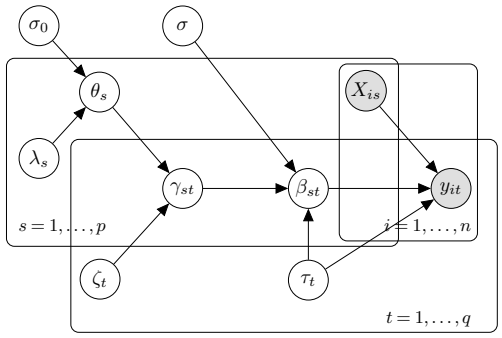

Figure 1: *Graphical representation of the model with n samples, p candidate predictors and q responses. The shaded nodes are observed, the others are inferred.*

Installation

The `atlasqtl` package is written in R with C++ subroutines. It can be installed from GitHub[1], following the provided instructions.

Data and reproducibility

The Rmarkdown document and data used to generate this note are publicly available for replication[2]. We restrict our illustration to the analysis of a small semi-synthetic dataset: it involves $q = 875$ traits capturing B-cell gene expression for $n = 100$ individuals ($n \times q$ matrix `Y`), and $p = 200$ candidate predictors generated with our in-house package `echoseq`[3] to emulate real genetic variants ($n \times p$ matrix `X`). A subset of the traits was modified so as to be associated with five hotspot predictors; these hotspots control 40, 49, 92, 119, 221 genes, respectively.

[1] https://github.com/hruffieux/atlasqtl
[2] https://github.com/hruffieux/software_corner_ibs_bulletin
[3] https://github.com/hruffieux/echoseq

Illustration: joint expression QTL analysis in B cells

The following code chunk runs `atlasqtl` jointly on all the candidate predictors and traits, using a default geometric schedule on the inverse temperature, with initial temperature of 2 and a grid of 10 temperatures.

```
require(atlasqtl)
```

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

obj_atlasqtl <- atlasqtl(Y = Y, X = X, p0 = c(mu_t, v_t),
                         user_seed = 1)
```

The analysis typically takes less than 10 seconds on a standard laptop[4].

`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 low when a permutation-based Bayesian false discovery rate (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 user can control additional settings — e.g., the annealing schedule, seed for random number generation, convergence tolerance threshold, checkpointing — as detailed on the help page of the function, which can be accessed by running `?atlasqtl` in the console.

Custom S3 functions are also available. For instance, the `print.atlasqtl` function provides basic information about the run:

```
obj_atlasqtl
```

```
## ******************************************************
## Successful convergence after 117 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: 100;
## Number of (non-redundant) candidate predictors: 200;
## Number of responses: 875;
## 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 object `obj_atlasqtl` returned by `atlasqtl` contains posterior quantities that can be employed to assess:

- associations between each pair of predictor and response: using the variational posterior probabilities of association 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 function `summary.atlasqtl` displays a variable selection summary, which can be based on a supplied Bayesian FDR threshold (here 5%):

```
thres_fdr <- 0.05
summary(obj_atlasqtl, thres = thres_fdr,
        fdr_adjust = TRUE, full_summary = FALSE)
```

```
## ******************************************************
## * ATLASQTL: posterior summary for variable selection *
## ******************************************************
##
## Using a 5% FDR control:
## -----------------------
##
## Nb of pairwise (predictor-response) associations: 499
##
## Nb of predictors associated with at least one response
## (active predictors): 14
##
## Hotspot sizes (nb of responses associated with each
## active predictor):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    1.50   35.64   38.00  209.00
##
## Top hotspots:
## snp_103 (size 209), snp_115 (size 132), snp_171 (size 81),
## snp_54 (size 44), snp_15 (size 20), snp_75 (size 4)
```

The user can also directly access the $p \times q$ matrix of FDR estimates for all predictor-response associations, using the function `assign_bFDR`. Alternatively, the scalablility of `atlasqtl` permits efficient permutation-based FDR estimation (see Ruffieux et al. (2017) or Ruffieux et al. (2020)).

Finally, the function `plot.atlasqtl` displays a Manhattan-type plot (Figure 2) with the hotspot positions and sizes after Bayesian FDR control:

```
sim_hs <- rowSums(pat) # simulated hotspot sizes (5 hotspots)
                       # pat is a p x q binary 'pattern matrix'
                       # locating the simulated associations
plot(obj_atlasqtl, thres = thres_fdr, fdr_adjust = TRUE,
     ylim_max = max(sim_hs))
points(which(sim_hs>0), sim_hs[sim_hs>0], pch = "x")
```

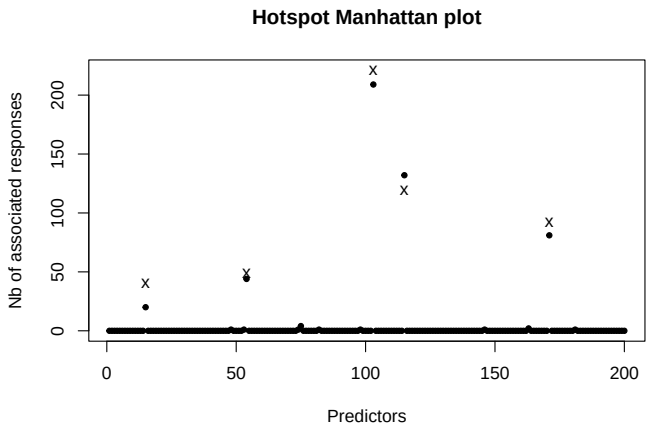

Figure 2: *Hotspot sizes, as obtained using the S3 function plot.atlasqtl. The crosses indicate the simulated sizes of the five hotspots.*

[4]Here a 2.8 GHz Quad-Core Intel Core i7 machine (serial execution)

Encoding predictor-level information

We recently proposed a new approach, called `epispot`, which extends the `atlasqtl` hierarchical framework to leverage information on the probability of candidate predictors to be involved in associations (Ruffieux et al. (2021)). Specifically, the effect of a (possibly high-dimensional) predictor-level vector of covariates is estimated using a secondary spike-and-slab regression at the top of the model hierarchy. The R package for this approach is also available on GitHub[5].

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.

[5] https://github.com/hruffieux/epispot