Averaged Variational Inference for Hierarchical Modelling of Genetic Association

William van Rooij

Master thesis supervised by Hélène Ruffieux and Anthony Davison

École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

1. Introduction

In **genome-wide association studies** (GWAS), we estimate the associations between **single nucleotide polymorphisms** (SNPs) and a phenotype or **trait**. It is a *small n, large p* situation where

- n is the number of observations,
- p is the number of SNPs.

In a *small n, large p* situation, traditional Bayesian inference methods such as **Markov Chain Monte Carlo** (MCMC) algorithms often scale poorly. An alternative is to use **variational inference** [Blei et al., 2017].

Moreover, SNPs data are often highly correlated with a block structure, which complicates inference and interpretation. Here, we seek to enhance regression approaches in such difficult settings.

2. Hierarchical Regression Model

Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ represent the SNPs, and $\boldsymbol{y}$ represent the trait observed.

$\boldsymbol{y}$ is linearly related with the predictors $\boldsymbol{X}$ and has residual precision τ. We suppose that they follow the following hierarchical model,

$$\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{\beta}, \tau &\sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}_n\right), \\
\beta_s \mid \gamma_s, \tau, \sigma^{-2} &\sim \gamma_s\, \mathcal{N}\left(0, \sigma^2\tau^{-1}\right) + (1-\gamma_s)\, \delta_0, \\
\gamma_s \mid \omega_s &\sim \text{Bernoulli}\left(\omega_s\right), \\
\omega_s &\sim \text{Beta}\left(a_s, b_s\right),
\end{aligned}$$

for all $s = 1, \ldots, p$, where δ_0 is the Dirac distribution, and a_s and b_s are chosen to enforce sparsity [Ruffieux et al., 2017]. $\boldsymbol{\gamma}$ is a binary vector that that indicates which SNP is associated with the trait, i.e.,

$$\gamma_s = 1 \iff \text{SNP } s \text{ is associated with the trait.}$$

3. Variational Inference

Given a family of densities $\mathcal{D}$ over the parameters $\boldsymbol{\theta}$, variational inference approximates the density of interest $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ by a distribution $q \in \mathcal{D}$ that minimizes the "reverse" **Kullback–Leibler divergence**,

$$\text{KL}(q \parallel p) = \int q(\boldsymbol{\theta}) \log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\right\}.$$

We introduce the **evidence lower bound**,

$$\mathcal{L}(q) = \mathbb{E}_q\left[\log p(\boldsymbol{\theta}, \boldsymbol{y})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\theta})\right] = \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}}\text{d}\boldsymbol{\theta},$$

where $\mathbb{E}_q$ represents the expectation with respect to q.

The lower bound verifies

$$\text{KL}(q \parallel p) = \log(p) - \mathcal{L}(q).$$

As the Kullbak–Leibler divergence can be difficult to minimize, variational inference maximizes the lower bound instead.

When applied to highly correlated data, variational inference **underestimates** posterior variances,this is a consequence of:

- the **high multimodality** of the lower bound $\mathcal{L}(q)$,
- the **mean-field independence** assumption,
- the **reverse Kullback–Leibler** divergence optimisation,

and results in a tendency for the approximation to concentrate mass on a single mode and report a set of predictors with very high confidence.

To better handle the multimodality, we build a method consisting of averaging over multiple runs with different parameter initialisations with weights equal to the posterior model probability.

4. Averaged LOCUS method

Assume that the data $\boldsymbol{y}$ has been obtained from a one of K models $M_k, \quad k = 1, \ldots, K$.

We want to estimate the associations between the SNPs and the trait.

We perform a weighted average accounting for the likelihood that the data corresponds to each model, i.e., for all $s = 1, \ldots, p$,

$$\mathbb{E}\left[\gamma_s \mid \boldsymbol{y}\right] = \sum_{k=1}^{K} \mathbb{E}\left[\gamma_s \mid M_k, \boldsymbol{y}\right]\ p(M_k \mid \boldsymbol{y}).$$

The posterior probability for model M_k is

$$p(M_k \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid M_k)\ p(M_k)}{\sum_{j=1}^{K} p(\boldsymbol{y} \mid M_j)\ p(M_j)} \approx \frac{\exp\left\{\mathcal{L}(q)\right\}\ p(M_k)}{\sum_{j=1}^{K} \exp\left\{\mathcal{L}(q)\right\}\ p(M_j)}$$

where $p(\boldsymbol{y} \mid M_k)$ is the likelihood under model M_k, approximated by $\exp\left\{\mathcal{L}(q)\right\}$, and $p(M_k)$ is the prior probability of model M_k.

5. Performance

We compare the performance of variable selection of five methods,

- classical variational algorithm "LOCUS", [Ruffieux et al., 2017],
- averaged variational algorithm "averaged LOCUS",
- their annealing augmented equivalents "annealed LOCUS" and "averaged annealed LOCUS",
- averaged variational algorithm with equal weights "averaged LOCUS (Equal weights)"

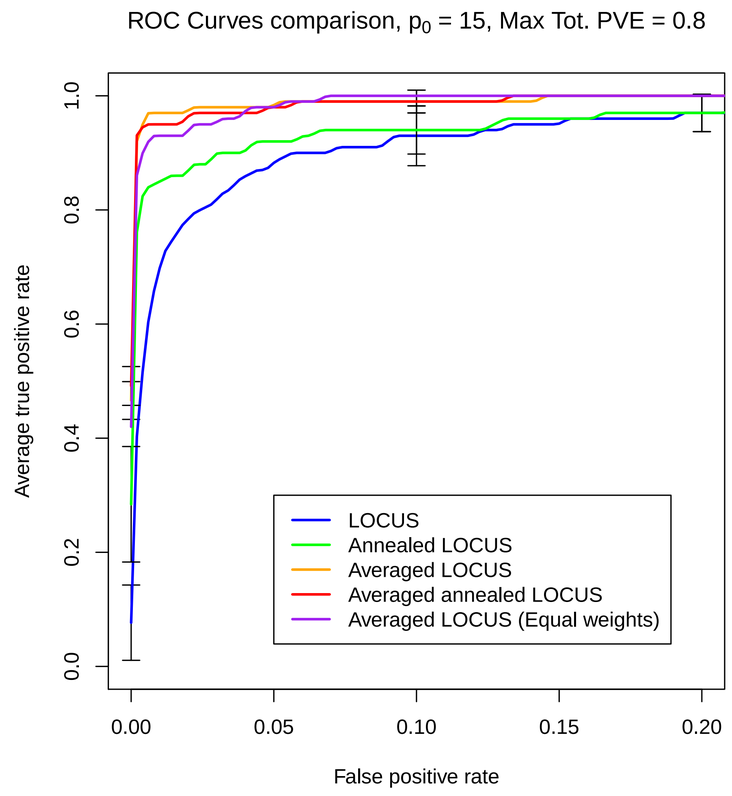

Figure 1: *Comparison of ROC curves between LOCUS, averaged LOCUS, their respective annealed versions, and averaged LOCUS with equal weights, colored in blue, orange, green, red, and purple respectively. The data involve* 300 *observations,* 15 *SNPs out of* 500 *SNPs are associated with the trait. The response variance explained by the SNPs is below* 80%.

The three averaging methods have similar variable selection performance, which is better than the standard LOCUS method. The annealing step only improves the performance of variable selection in the non-averaged cases.

6. Conclusion

The averaged variational inference better handles the multimodality of the posterior and better conveys the uncertainty due to the strong correlation structure in the data.

The averaged methods can be implemented in parallel, drastically diminishing the runtime.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877.

Ruffieux, H., Davison, A. C., Hager, J., and Irincheeva, I. (2017). Efficient inference for genetic association studies with multiple outcomes. *Biostatistics*, 18:618–636.