

Averaged Variational Inference for Hierarchical Modelling of Genetic Association

Master thesis

William van Rooij

École Polytechnique Fédérale de Lausanne

09.07.19

- ▶ Introduction
- ▶ Hierarchical model
- ▶ Variational inference
- ▶ Methods
- ▶ Simulations
- ▶ Conclusion

Introduction

- ▶ Estimate association between genetic variants and diseases or phenotypes.
- ▶ The most common genetic variants are single nucleotide polymorphisms (SNPs).
- ▶ Not many observations compared to the number of parameters, i.e., small n , large p situation.
- ▶ Traditional techniques do not apply, so we need to find an alternative.

Hierarchical model

- ▶ We introduce $X = (X_1, \dots, X_p)$, and $y = (y_1, \dots, y_q)$.
- ▶ A SNP X_s and a trait y_t , SNPs are strongly correlated.
- ▶ Estimate the association between SNP s and trait t .
- ▶ $\mathbf{y}_{n \times q} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times q} + \boldsymbol{\epsilon}_{n \times q}$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \tau_t^{-1} \mathbf{I}_n)$
- ▶ y is a response matrix, x are candidate predictors.
- ▶ Each response y_t is linearly related with the predictors and has a residual precision $\tau_t \sim \text{Gamma}(\eta_t, \kappa_t)$.

Hierarchical model II

- ▶ For all $s = 1, \dots, p$, $t = 1, \dots, q$,
- ▶ $\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st} \mathcal{N}(0, \sigma^2 \tau_t^{-1}) + (1 - \gamma_{st}) \delta_0$,
- ▶ $\gamma_{st} \mid \omega_s \sim \text{Bernoulli}(\omega_s)$,
- ▶ $\omega_s \sim \text{Beta}(a_s, b_s)$,
- ▶ $\tau_t \sim$ and $\sigma^{-2} \sim$ have Gamma priors,
- ▶ a_s, b_s chosen to enforce sparsity. We define p^* the expected number of predictors involved in the model. Then, e.g.:

$$a_s \equiv 1, \quad b_s \equiv q(p - p^*)/p^*$$

Hierarchical model III

- ▶ Markov Chain Monte Carlo algorithms (MCMC) are the usual way to approximate inference in relatively small datasets.
- ▶ small n , large p , large q .
- ▶ MCMC gets time consuming, computational cost of operations increases with the number of parameters.
- ▶ Number of iterations needed increases with the number of parameters.
- ▶ Variational inference is an alternative to MCMC.

Variational inference

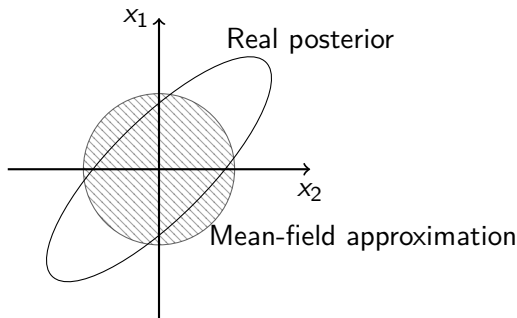
- ▶ Observed data \mathbf{y} , parameters θ , posterior distribution of parameters $p(\theta \mid \mathbf{y})$.
- ▶ Approximate the posterior density with a simpler density q , minimizing a "closeness" measure: the Kullback-Leibler divergence.
- ▶ $\text{KL}(q \parallel p) := \int q(\theta) \log \left(\frac{q(\theta)}{p(\theta \mid \mathbf{y})} \right) d\theta$.
- ▶ Evidence lower bound (ELBO):
 $\mathcal{L}(q) = \mathbb{E}_q [\log p(\theta, \mathbf{y})] - \mathbb{E}_q [\log q(\theta)]$.
- ▶ $\text{KL}(q \parallel p) = \log(p) - \mathcal{L}(q)$.
- ▶ Minimizing KL is equivalent to maximizing ELBO.

Mean-field approximation

- ▶ We assume independence for most of the parameters:

$$q(\theta) = \left\{ \prod_{s=1}^p \prod_{t=1}^q q(\beta_{st}, \gamma_{st}) \right\} \left\{ \prod_{s=1}^p q(\omega_s) \right\} \left\{ \prod_{t=1}^q q(\tau_t) \right\} q(\sigma^{-2}).$$

- ▶ The mean-field approximation does not represent the correlations between parameters.



Coordinate ascent variational inference algorithm

Algorithm 1: Coordinate ascent variational inference

input : $p(\mathbf{y}, \boldsymbol{\theta})$, dataset \mathbf{y} , tolerance ε

output : $q(\boldsymbol{\theta}) = \prod_{j=1}^J q_j(\theta_j)$

initialize: the parameters of each $q(\theta_j)$

repeat

for $j \in \{1, \dots, J\}$ **do**

set $q_j(\theta_j) \propto \exp \{ \mathbb{E}_{-j} [\log p(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathbf{y})] \}$

$\mathcal{L}^{\text{old}}(q) \leftarrow \mathcal{L}(q)$

$\mathcal{L}(q) \leftarrow \mathbb{E} [\log p(\boldsymbol{\theta}, \mathbf{y})] - \mathbb{E} [\log q(\boldsymbol{\theta})]$

until $|\mathcal{L}^{\text{old}}(q) - \mathcal{L}(q)| < \varepsilon$

return $q(\boldsymbol{\theta})$

Coordinate ascent variational inference algorithm II

- ▶ $\mathcal{L}(q)$ is guaranteed to increase at every iteration.
- ▶ We assume there exists a best model and we want to find it
- ▶ CAVI yields a local optimum, depending on the initialization of the parameters.
- ▶ Another possible solution is annealing, which consists of "heating" the distribution to have only a global maximum.
- ▶ Annealing yields a unique model, so averaging might better represent the incertitude.

Parameters posterior distributions

- ▶ $\beta_{st} \mid \gamma_{st} = 1, \mathbf{y} \sim \mathcal{N}(\mu_{\beta,st}, \sigma_{\beta,st}^2),$
- ▶ $\beta_{st} \mid \gamma_{st} = 0, \mathbf{y} \sim \delta_0,$
- ▶ $\gamma_{st} \mid \mathbf{y} \sim \text{Bernoulli}(\gamma_{st}^{(1)}),$
- ▶ $\omega_s \mid \mathbf{y} \sim \text{Beta}(a_s^*, b_s^*),$
- ▶ $\tau_t \mid \mathbf{y} \sim \text{Gamma}(\eta_t^*, \kappa_t^*),$
- ▶ $\sigma^{-2} \mid \mathbf{y} \sim \text{Gamma}(\lambda^*, \nu^*),$

Averaged LOCUS

- ▶ Denote M_k , $k = 1, \dots, K$ the models yielded by the local optimums.
- ▶ $p(\gamma_{st} \mid \mathbf{y}) = \sum_{k=1}^K p(\gamma_{st} \mid M_k) p(M_k \mid \mathbf{y})$,
- ▶ $p(M_k \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_k) p(M_k)}{\sum_{j=1}^K p(\mathbf{y} \mid M_j) p(M_j)}$,
- ▶ $\mathcal{L}(q)$ serves as an approximation of $\log p(\mathbf{y} \mid M_k)$, as $\text{KL}(q \parallel p) = \log p(\mathbf{y}) - \mathcal{L}(q)$.
- ▶ $p(M_k)$ is the prior probability of the models, we consider them to be equiprobable: $p(M_k) = 1/K$, $\forall k = 1, \dots, K$.

Averaged LOCUS

- ▶ Generate SNPs, traits, and associations.
- ▶ Find the optimums $q^*(\theta)$ with different initial parameters, drawn at random.
- ▶ Generate the ELBOs and use them as weights in the weighted average (Averaged LOCUS).
- ▶ $\mathbb{E}[\gamma_{st} | \mathbf{y}] = \sum_{k=1}^K \mathbb{E}[\gamma_{st} | M_k, \mathbf{y}] p(M_k | \mathbf{y})$
- ▶ The function yields probabilities of association between SNPs and traits.

Annealed LOCUS

- ▶ Temperature T , “smoothing” the density of interest, and gets lower until initial density is reached.
- ▶ $p_T(\mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y}, \boldsymbol{\theta})^{1/T}$,
- ▶ $\mathcal{L}_T(q_T) = \int q_T(\boldsymbol{\theta}) \log p(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} - T \int q_T(\boldsymbol{\theta}) \log q_T(\boldsymbol{\theta}) d\boldsymbol{\theta}$

$$\begin{aligned}\mathcal{L}_T(q) &= \mathbb{E}_j [\mathbb{E}_{-j} \{\log p(\mathbf{y}, \boldsymbol{\theta})\} - T \log q_T(\theta_j)] + \text{const}, \\ &= T \mathbb{E}_j \left[\log \left\{ \frac{p_{T,-j}(\mathbf{y}, \theta_j)}{q_T(\theta_j)} \right\} \right] + \text{const} .\end{aligned}$$

Annealed LOCUS II

- ▶ Geometric spacing,

$$T_l = (1 + \Delta)^{l-1}, \quad \Delta = T_L^{1/(L-1)} - 1,$$

- ▶ harmonic spacing,

$$T_l = 1 + \Delta(l - 1), \quad \Delta = \frac{T_L - 1}{L - 1},$$

- ▶ linear spacing,

$$T_l^{-1} = T_L^{-1} + \Delta(L - l), \quad \Delta = \frac{1 - T_L^{-1}}{L - 1}.$$

- ▶ We can also combine annealing with the Averaged LOCUS method, which we call Averaged annealed LOCUS.

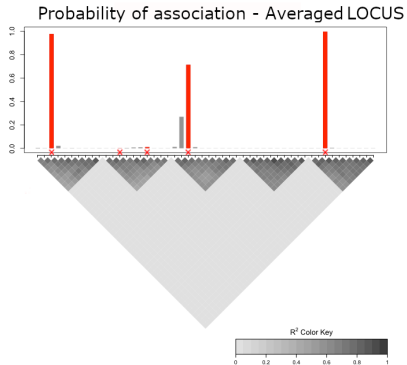
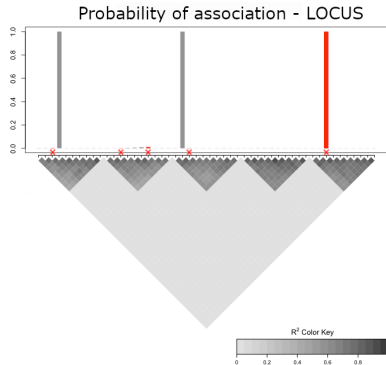
Averaged LOCUS with equal weights

- ▶ Instead of using the lower bound as weights, we average over all the models with equal weights.
- ▶ $\mathbb{E}[\gamma_{st} \mid \mathbf{y}] = \sum_{k=1}^K \mathbb{E}[\gamma_{st} \mid M_k, \mathbf{y}] p(M_k \mid \mathbf{y})$

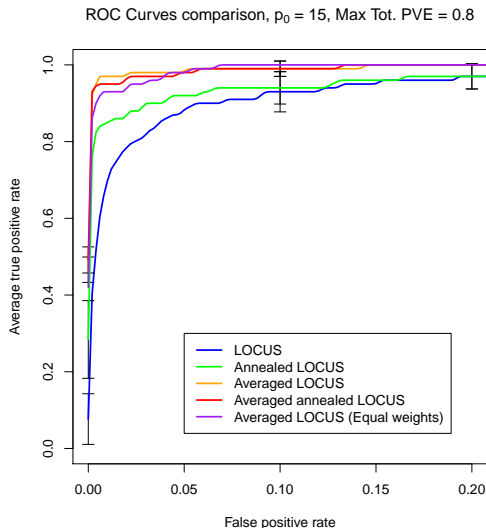
Results

- ▶ $n = 300$ observations,
- ▶ $p = 500$ SNPs, with p_0 associated SNPs,
- ▶ $q = 1$ trait,
- ▶ 100 random initialisations,
- ▶ autocorrelation between the SNPs is between 0.95 and 0.99, in blocks of ten SNPs,
- ▶ we can specify the maximum proportion of response variance explained by the SNPs.
- ▶ We used 50 replications to determine the ROC curves.

LOCUS VS. Averaged LOCUS with $p_0 = 5$, max var. = 0.5

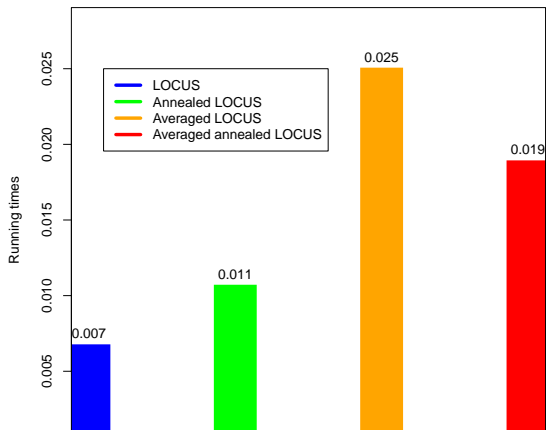


ROC curves comparison, $p_0 = 15$, max var.= 0.8



Runtimes

Running times of the four methods



Results

- ▶ Paralleled computation is possible.
- ▶ The difference is bigger when phenotypic variance is better explained from the SNPs.
- ▶ The difference is bigger with fewer active SNPs.

Conclusion

- ▶ On strong correlated structures, Averaged LOCUS performs better than LOCUS.
- ▶ The weights do not necessarily improve the performance.
- ▶ Simulated annealing improves the standard LOCUS, but less Averaged LOCUS.

Conclusion II

- ▶ Optimization of the code, \rightarrow ev. integration to R-package,
- ▶ Application to real data.

Thank you for your time.