Averaged Variational Inference for Hierarchical Modelling of Genetic Association

Master thesis

William van Rooij
supervised by Hélène Ruffieux and Anthony Davison

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.
- $\boldsymbol{y}_{n \times q} = \boldsymbol{X}_{n \times p} \boldsymbol{\beta}_{p \times q} + \boldsymbol{\epsilon}_{n \times q}$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \tau_t^{-1} \boldsymbol{I}_n)$
- $\boldsymbol{y}$ is a response matrix, $\boldsymbol{X}$ 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 $s = 1, \dots, p$, $t = 1, \dots, q$,
- $\boldsymbol{y}_t \mid \beta_t, \tau_t \ \sim \ \mathcal{N}(\boldsymbol{X}\beta_t, \tau_t^{-1}\boldsymbol{I}_n)$,
- $\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \ \sim \ \gamma_{st}\, \mathcal{N}(0, \sigma^2\tau_t^{-1}) + (1 - \gamma_{st})\, \delta_0$,
- $\gamma_{st} \mid \omega_s \sim \text{Bernoulli}(\omega_s)$,
- $\omega_s \sim \text{Beta}(a_s, b_s)$,
- a_s, b_s chosen to enforce sparsity,
- τ_t and σ^{-2} have Gamma priors.

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 as an alternative to MCMC.

Variational Inference

- Observed data $\boldsymbol{y}$, parameters $\boldsymbol{\theta}$, posterior distribution of parameters $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
- Approximate the posterior density with a simpler density q, minimizing a ”closeness” measure: the reverse Kullback–Leibler divergence.
- $\mathrm{KL}(q \parallel p) := \int q(\boldsymbol{\theta}) \log \left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right) \mathrm{d}\boldsymbol{\theta}$.

Variational Inference II

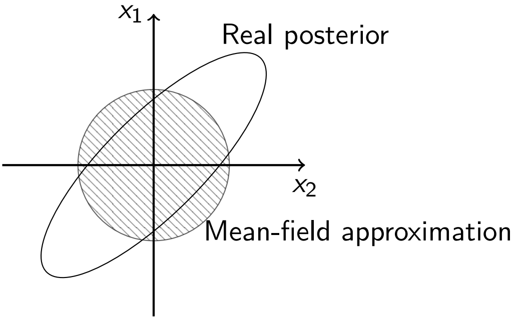
- Evidence lower bound (ELBO): $\mathcal{L}(q) = \mathbb{E}_q\left[\log p(\boldsymbol{\theta}, \boldsymbol{y})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\theta})\right]$.
- $\mathrm{KL}(q \parallel p) = \log(p) - \mathcal{L}(q)$.
- Minimizing KL is equivalent to maximizing ELBO.

Mean-field approximation

- We assume independence for most of the parameters:

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

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



Coordinate ascent variational inference algorithm

Algorithm 1: Coordinate ascent variational inference

```
input      : $p(\boldsymbol{y}, \boldsymbol{\theta})$, dataset $y$, tolerance $\varepsilon$
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 $q_j(\theta_j) \propto \exp\{\mathbb{E}_{-j}[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})]\}$
    $\mathcal{L}^{\text{old}}(q) \leftarrow \mathcal{L}(q)$
    $\mathcal{L}(q) \leftarrow \mathbb{E}[\log p(\boldsymbol{\theta}, \boldsymbol{y})] - \mathbb{E}[\log q(\boldsymbol{\theta})]$
until $|\mathcal{L}^{\text{old}}(q) - \mathcal{L}(q)| < \varepsilon$
return $q(\boldsymbol{\theta})$
```

EPFL

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

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

Averaged LOCUS

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

Averaged LOCUS

- Generate SNPs, traits, and associations.
- Find the optimums $q^*(\boldsymbol{\theta})$ with different initial parameters, drawn at random.
- Generate the ELBOs and use them as weights in the weighted average (Averaged LOCUS).
- $\mathbb{E}\left[\gamma_{st} \mid \boldsymbol{y}\right] = \sum_{k=1}^{K} \mathbb{E}\left[\gamma_{st} \mid M_k, \boldsymbol{y}\right] p(M_k \mid \boldsymbol{y})$
- 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.
- $p_T(\boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{y}, \boldsymbol{\theta})^{1/T}$,
- $\mathcal{L}_T(q_T) = \int q_T(\boldsymbol{\theta}) \log p(\boldsymbol{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - T \int q_T(\boldsymbol{\theta}) \log q_T(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$

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

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.

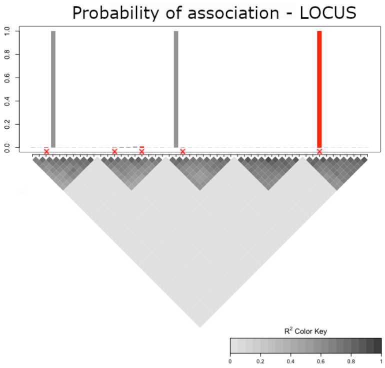
Averaged LOCUS with equal weights

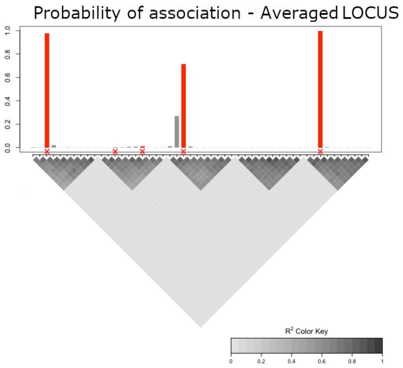
- Instead of using the lower bound as weights, we average over all the models with equal weights.
- $\mathbb{E}\left[\gamma_{st} \mid \boldsymbol{y}\right] = \sum_{k=1}^{K} \mathbb{E}\left[\gamma_{st} \mid M_k, \boldsymbol{y}\right] p(M_k \mid \boldsymbol{y})$

Results

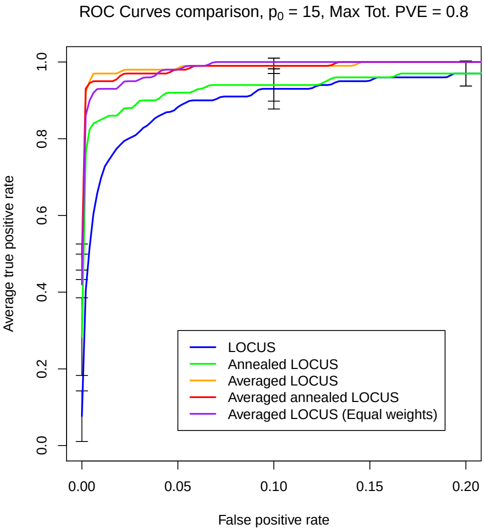
- $n = 300$ observations,
- $p = 500$ SNPs, with p_0 associated SNPs,
- $q = 1$ trait,
- 100 random initialisations,
- autocorrelation between the SNPs is between 0.95 and 0.99, in blocks of ten SNPs,
- we can specify the maximum proportion of response variance explained by the SNPs.
- We used 50 replications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

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

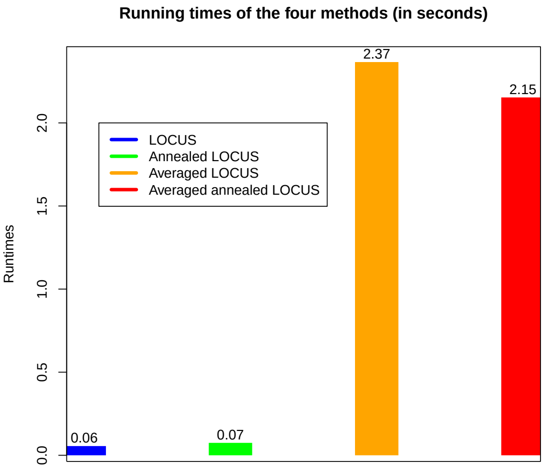




ROC curves comparison, $p_0 = 15$, max var.$= 0.8$



Runtimes


Running times of the four methods (in seconds)
Runtimes
0.0
0.5
1.0
1.5
2.0
0.06
0.07
2.37
2.15
LOCUS
Annealed LOCUS
Averaged LOCUS
Averaged annealed LOCUS

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, → ev. integration to R-package,
- Application to real data.

Thank you

Thank you for your time.