Averaged Variational Inference for Hierarchical Modelling of Genetic Association

Master thesis

William van Rooij

École Polythecnique Fédérale de Lausanne

12.04.19

EPFL

- Introduction
- Variational inference
- Mean-field approximation
- Implementation
- Results
- Next steps

Introduction

- We introduce $X = (X_1, \ldots, X_p)$, and $y = (y_1, \ldots, y_q)$.
- A SNP X_s and a trait y_t, SNPs are strongly correlated.
- Estimate the association between SNP s and trait t.
- $y_{n\times q} = x_{n\times p}\beta_{p\times q} + \epsilon_{n\times q}$, $\epsilon_t \sim \mathcal{N}(0, \tau_t^{-1} I_n)$
- y is a response matrix, x are candidate predictors.
- Each response y_t is linearly related with the predictors and has a residual precision $\tau_t \sim \text{Gamma}(\eta_t, \kappa_t)$.

Introduction II

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

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

Introduction III

- Markov Chain Monte Carlo algorithms (MCMC) are the usual way to approximate inference in relatively small datasets.
- p and q are large compared to n.
- 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

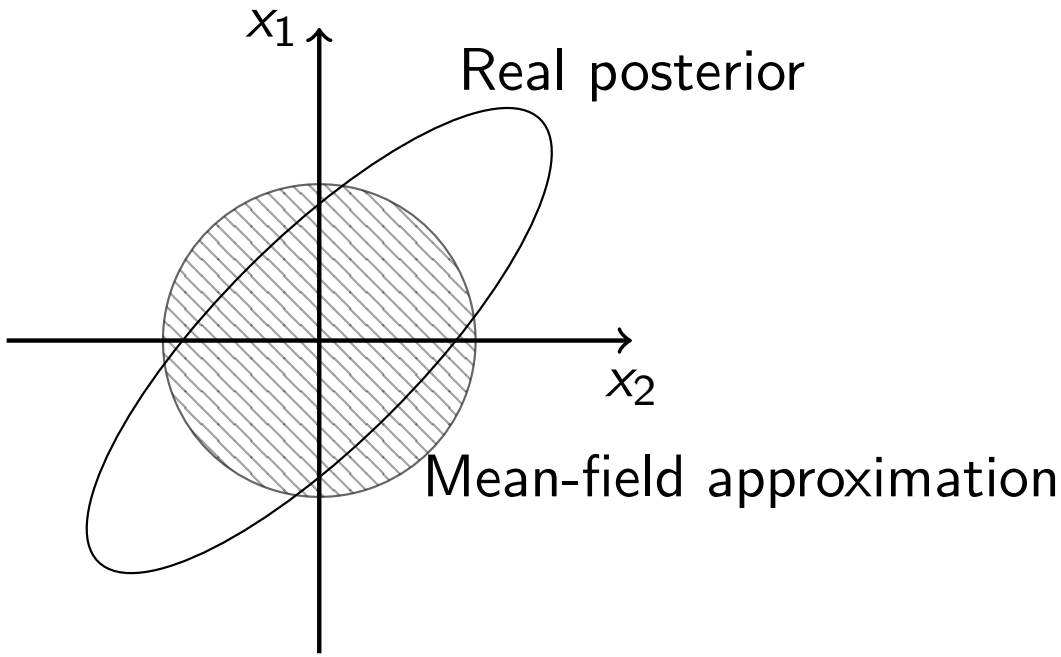
- Observed data $\boldsymbol{y}$, parameters $\boldsymbol{\theta}$, posterior distribution of parameters $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
- Approximate the posterior density with a simpler density q, minimizing a "closeness" measure: the Kullback-Leibler divergence.
- $\mathrm{KL}(q \,\|\, p) := \int q(\boldsymbol{\theta}) \log \left(\dfrac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right) \mathrm{d}\boldsymbol{\theta}$.
- Evidence lower bound (ELBO):
  $\mathcal{L}(q) = \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{y}) \right] - \mathbb{E}_q \left[\log q(\boldsymbol{\theta}) \right]$.
- $\mathrm{KL}(q \,\|\, p) = \log(p) - \mathcal{L}(q)$.
- 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 - CAVI

Algorithm 1: Coordinate ascent variational inference

input : $p(\boldsymbol{y}, \boldsymbol{\theta})$, dataset 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 $j \in \{1, \ldots, J\}$ **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 $q(\boldsymbol{\theta})$

Coordinate ascent variational inference - CAVI 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 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).
- The function yields probabilities of association between SNPs and traits.

Annealed LOCUS & Averaged annealed LOCUS

- Temperature T,
- $p_T(\boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{y}, \boldsymbol{\theta})^{1/T}$,

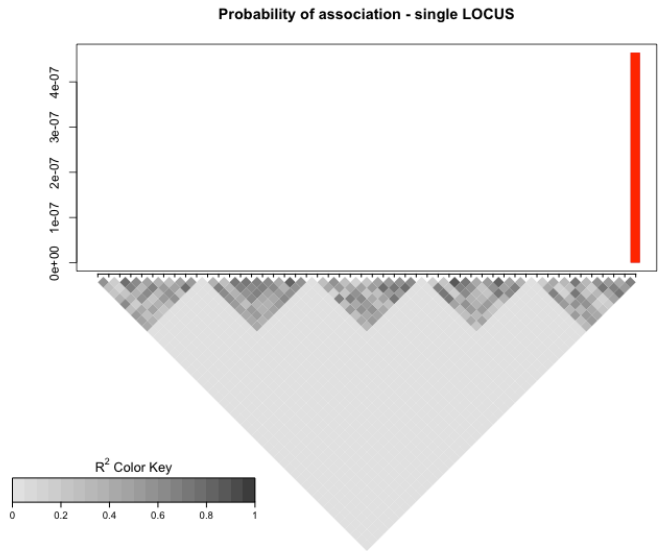
Meaned LOCUS

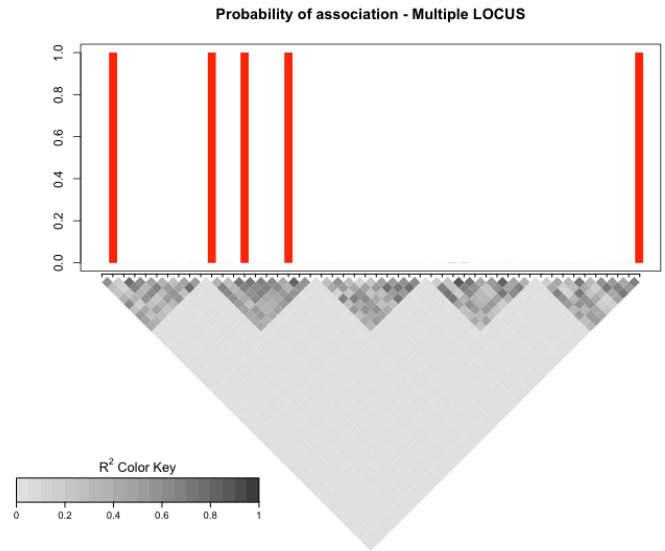
Expected weights averaged LOCUS

Results

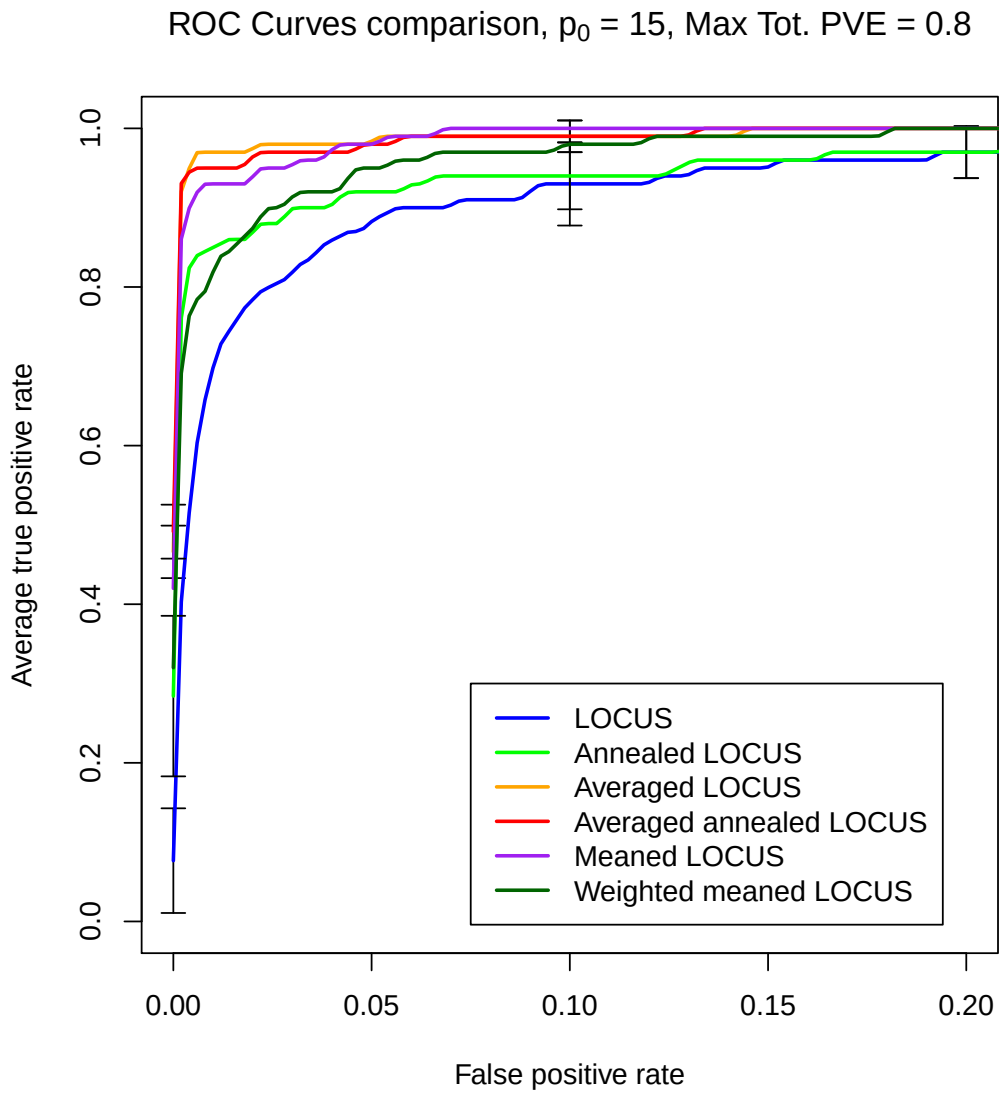
- $n = 300$ observations,
- $p = 500$ SNPs, with p_0 associated SNPs,
- $q = 1$ trait,
- 100 random initialisations,
- autocorrelation between the SNPs is between 0.95 and 0.99, in blocks of ten SNPs,
- we can specify the maximum proportion of response variance explained by the SNPs.
- We used 50 replications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

Weighted averaging with $p_0 = 5$, max var. $= 0.5$





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



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.

Next steps

- Optimization of the code, $\rightarrow$ ev. integration to R-package,
- Comparison with annealing and non-weighted averaging for strong correlations.
- Do we find the right modes? 2D visualisations (Rocková).
- Application to real data.

Thank you for your time.