

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

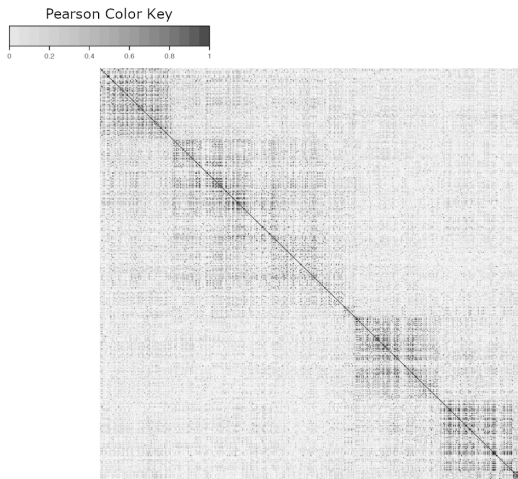
- Introduction
- Hierarchical model
- Variational inference
- Methods
- Simulations
- Conclusion

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.

Introduction

SNP correlation pattern



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

Hierarchical model

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

Variational Inference

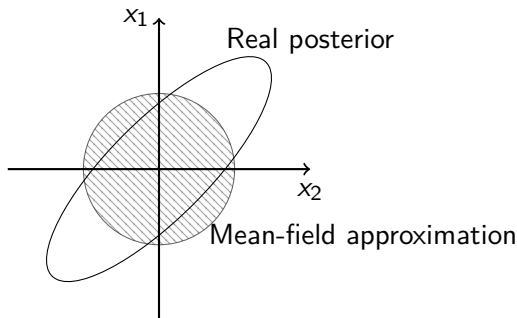
- Evidence lower bound (ELBO):
 $\mathcal{L}(q) = \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathbf{y})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})].$
- $\text{KL}(q \parallel p) = \log(p) - \mathcal{L}(q).$
- Minimizing KL is equivalent to maximizing ELBO.

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 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 $q(\boldsymbol{\theta})$

Problem statement

- High multimodality of $\mathcal{L}(q)$,
- mean-field independence assumption,
- reverse Kullback–Leibler divergence optimization,
- \Rightarrow variational inference underestimates posterior variances
- \Rightarrow tends to concentrate mass on a single mode.
- Two possibilities:
 - 1 Simulated annealing
 - 2 Weighted average

Averaged LOCUS

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

Averaged LOCUS

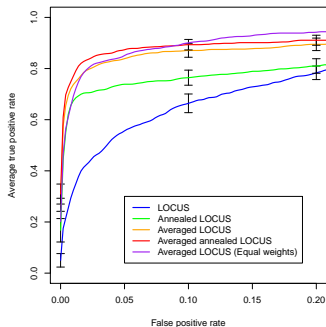
- Denote M_k , $k = 1, \dots, K$ the models yielded by the local optimums.
- $p(M_k | \mathbf{y}) = \frac{p(\mathbf{y} | M_k)p(M_k)}{\sum_{j=1}^K p(\mathbf{y} | M_j)p(M_j)}$,
- $\mathcal{L}(q)$ serves as an approximation of $\log p(\mathbf{y} | M_k)$, as $\text{KL}(q \parallel p) = \log p(\mathbf{y})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

Simulations

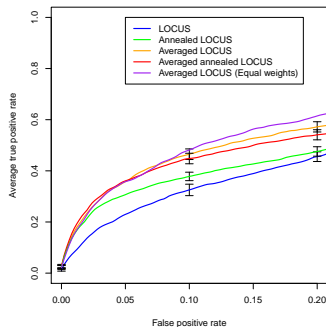
- $n = 300$ observations,
- $p = 500$ SNPs, with p_0 associated SNPs,
- $q = 1$ trait,
- 100 random initialisations,
- autocorrelation between the SNPs is between 0.95 and 0.99, in blocks of ten SNPs,
- we can specify the maximum proportion of response variance explained by the SNPs.
- We used 50 replications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

ROC curves comparison

ROC Curves comparison, $p_0 = 15$, Max Tot. PVE = 0.8

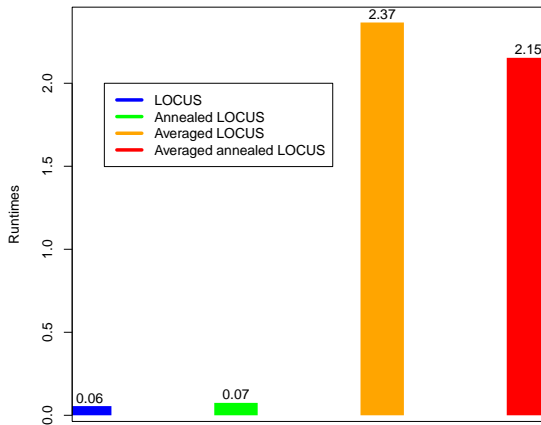


ROC Curves comparison, $p_0 = 50$, Max Tot. PVE = 0.8



Runtimes

Running times of the four methods (in seconds)



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 the averaged LOCUS.

Conclusion

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

Thank you

Thank you for your time.