# 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

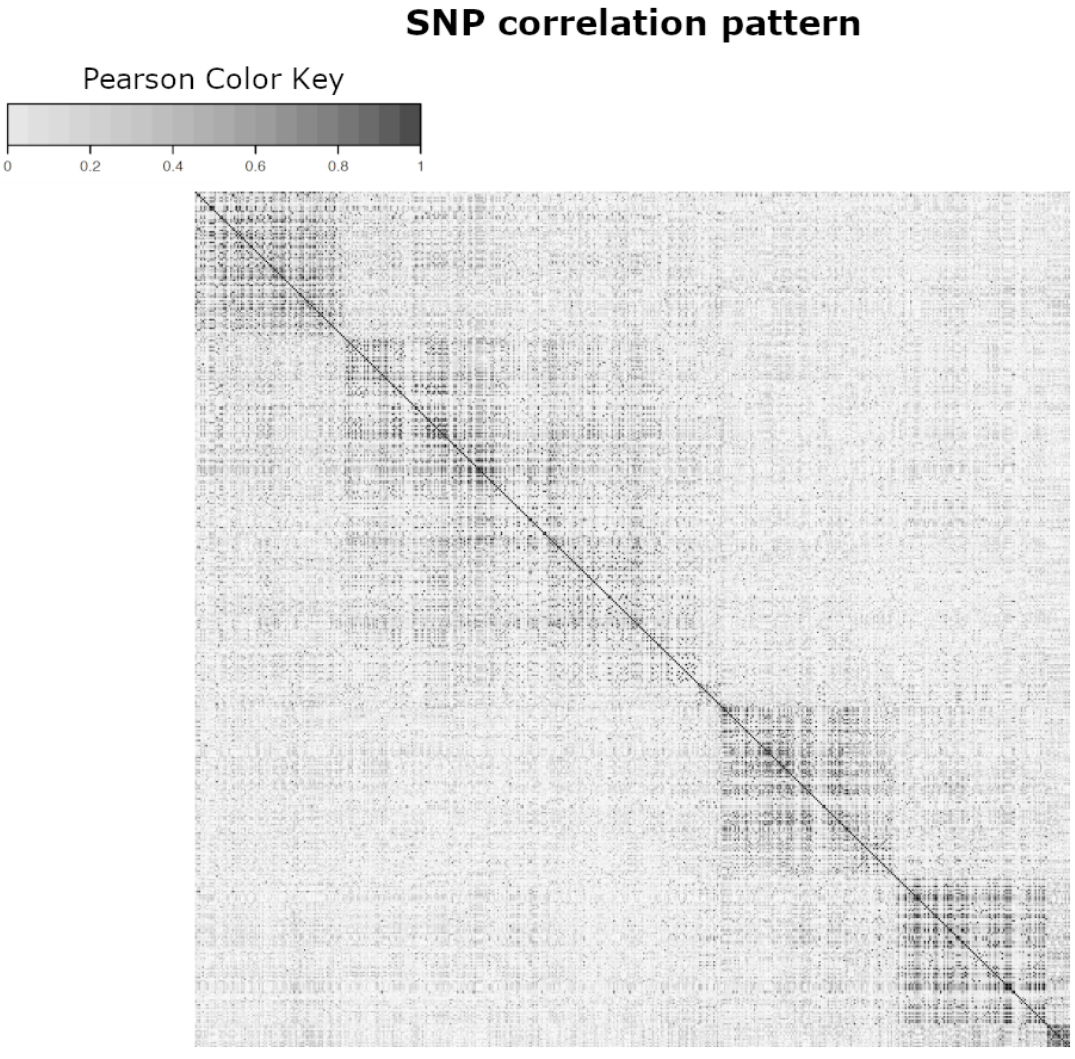
EPFL

- Introduction
- Hierarchical model
- Variational inference
- Averaged LOCUS
- Simulations
- Conclusion

# Introduction

- Estimate associations between genetic variants and diseases or phenotypes.
- The most common genetic variants are single nucleotide polymorphisms (SNPs).
- SNPs are strongly correlated with a block pattern.
- 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



# Hierarchical model

- We introduce $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$,
- A SNP $\boldsymbol{X}_s$ and a trait $\boldsymbol{y}_t$,
- Estimate the association between SNP $s$ and trait $t$,
- For all $s = 1, \ldots, p$, $t = 1, \ldots, q$,
- $\boldsymbol{y}_t \mid \beta_t, \tau_t \sim \mathcal{N}(\boldsymbol{X}\beta_t, \tau_t^{-1}\boldsymbol{I}_n)$,
- $\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st}\, \mathcal{N}(0, \sigma^2\tau_t^{-1}) + (1 - \gamma_{st})\, \delta_0$,
- $\gamma_{st} \mid \omega_s \sim \text{Bernoulli}(\omega_s)$,
- $\omega_s \sim \text{Beta}(a_s, b_s)$,
- $a_s, b_s$ chosen to enforce sparsity,
- $\tau_t$ and $\sigma^{-2}$ have Gamma priors.

# Hierarchical model

- Markov Chain Monte Carlo algorithms (MCMC) are the usual way to approximate inference in relatively small datasets.
- This is a small $n$, large $p$, large $q$ situation.
- 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 inference.

# 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.
- $\mathrm{KL}(q \parallel p) := \int q(\boldsymbol{\theta}) \log \left\{ \dfrac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right\} \mathrm{d}\boldsymbol{\theta}.$
- $\mathrm{KL}$ difficult to minimize.

# Variational Inference

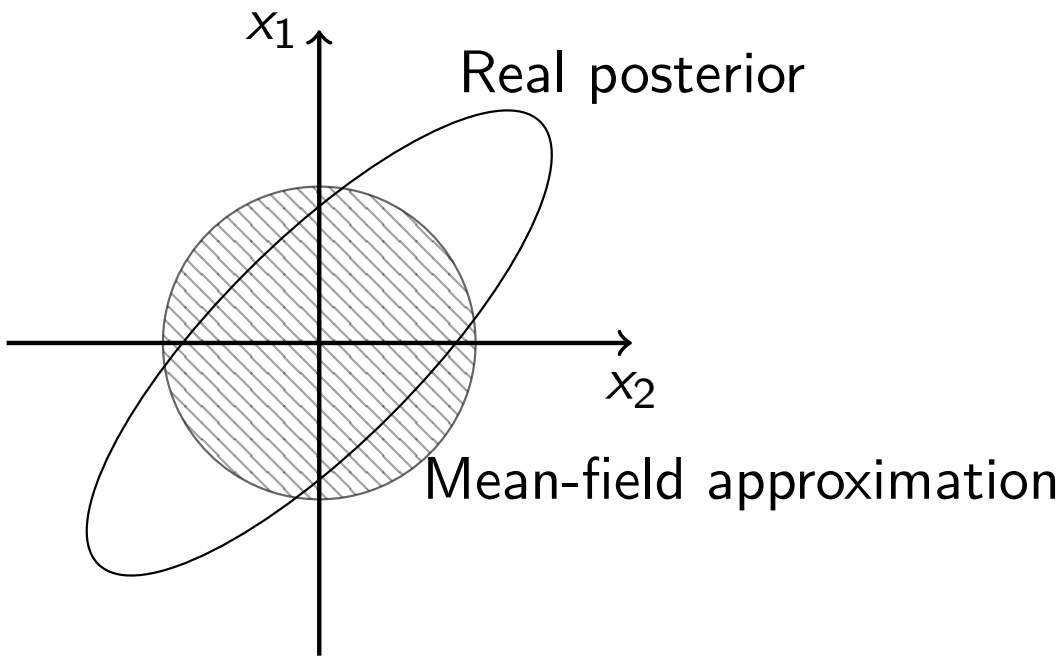
- Evidence lower bound (ELBO):
  $\mathcal{L}(q) = \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{y})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\theta})\right]$.
- $\mathrm{KL}(q \parallel p) = \log(p) - \mathcal{L}(q)$.
- Minimizing KL is equivalent to maximizing ELBO.
- It is easier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

# 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** $j \in \{1, \ldots, J\}$ **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** $q(\boldsymbol{\theta})$

# 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 averaging.

# Averaged LOCUS

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

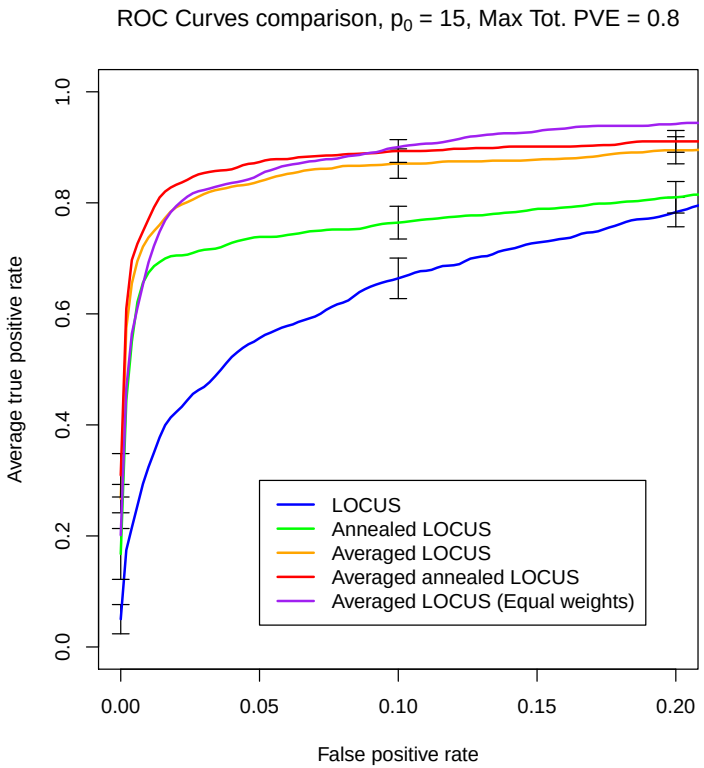
# Averaged LOCUS

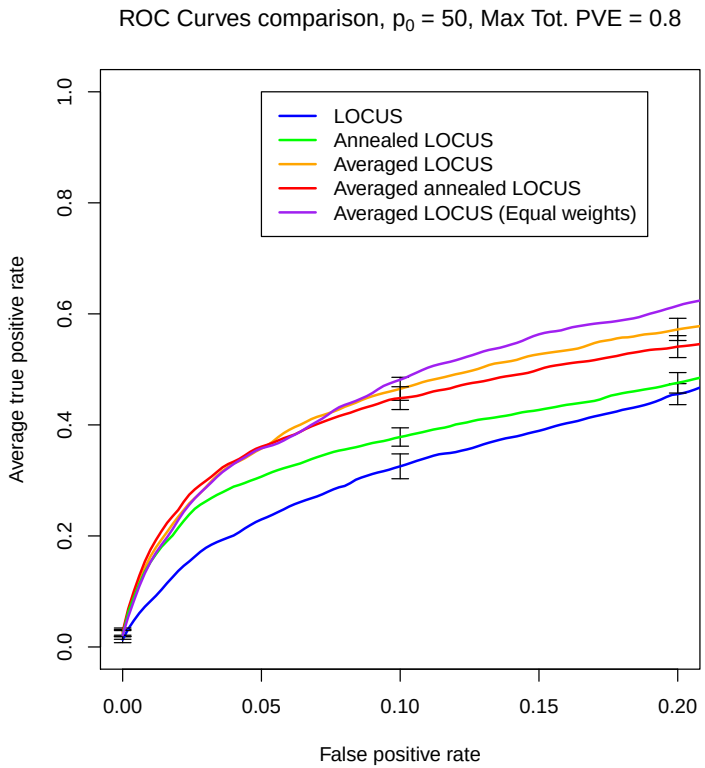
- Denote $M_k$, $k = 1, \ldots, K$ the models yielded by the local optimums.
- $p(M_k \mid \boldsymbol{y}) = \dfrac{p(\boldsymbol{y} \mid M_k) p(M_k)}{\sum_{j=1}^{K} p(\boldsymbol{y} \mid M_j) p(M_j)}$,
- $\mathcal{L}(q)$ serves as an approximation of $\log p(\boldsymbol{y} \mid M_k)$, as $\mathrm{KL}(q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

# 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 and running times

# ROC curves comparison



# Runtimes

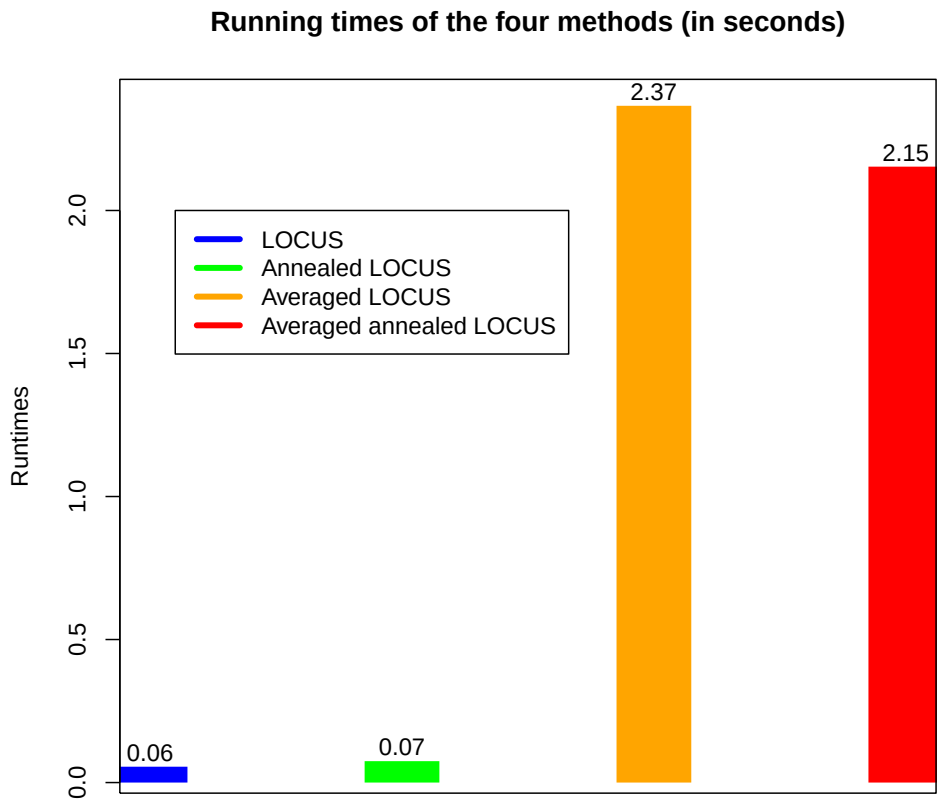



Figure: Running times of the different methods with $p_0 = 15$ and the max. variance explained by the SNPs is 80%.

EPFL

# 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.

# Parameters posterior distributions

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