

Master Project:
Statistical analysis on genomic data
Mid-term presentation

William van Rooij

EPFL

12.04.19

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

Introduction

- ▶ 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 .
- ▶ $y_{n \times q} = x_{n \times p} \beta_{p \times q} + \epsilon_{n \times q}$, $\epsilon_t \sim \mathcal{N}(0, \tau_t^{-1} I_n)$
- ▶ 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

- ▶ $s = 1, \dots, p, t = 1, \dots, q,$
- ▶ $\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st} \mathcal{N}(0, \sigma^2 \tau_t^{-1}) + (1 - \gamma_{st}) \delta_0,$
(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:

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

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

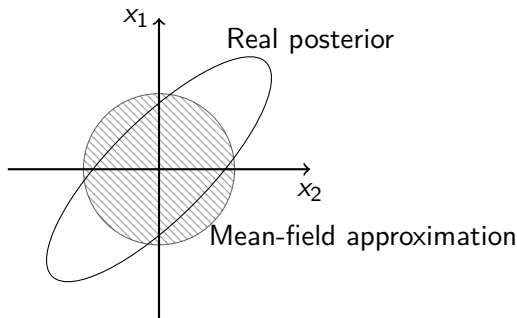
- ▶ Observed data \mathbf{y} , parameters θ , posterior distribution of parameters $p(\theta \mid \mathbf{y})$.
- ▶ Approximate the posterior density with a simpler density q , minimizing a "closeness" measure: the Kullback-Leibler divergence.
- ▶ $\text{KL}(q \parallel p) := \int q(\theta) \log \left(\frac{q(\theta)}{p(\theta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

- ▶ If we fix $q_l(\theta_l)$, $l \neq j$, the optimal for $q_j(\theta_j)$ verifies:
 $q_j^*(\theta_j) \propto \exp \{ \mathbb{E}_{-j} [\log p(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathbf{y})] \}$
- ▶ IN: $p(\mathbf{x}, \mathbf{z})$, data set \mathbf{x} , tolerance tol ,
OUT: $q(\mathbf{z}) = \prod q_j(\mathbf{z}_j)$.
INIT: $q_j(\mathbf{z}_j)$,
REPEAT:
 FOR: $j \in \{1, \dots, m\}$,
 SET: $q_j(\mathbf{z}_j) \propto \exp \{ \mathbb{E}_{-j} [\log p(\mathbf{z}_j \mid \mathbf{z}_{-j}, \mathbf{x})] \}$.
 COMPUTE:
 $ELBO^{old}(q) \leftarrow ELBO(q)$.
 $ELBO(q) = \mathbb{E} [\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E} [\log q(\mathbf{z})]$.
UNTIL: $|ELBO(q) - ELBO^{old}(q)| < tol$.
RETURN: $q(\mathbf{z})$.

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

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

"Bayesian model averaging"

- ▶ 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: $p(M_k) = 1/K$, $\forall k = 1, \dots, K$.

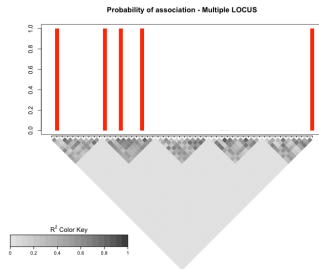
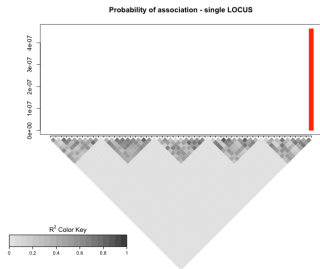
Implementation

- ▶ 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 ("BMA").
- ▶ The function yields probabilities of association between SNPs and traits.

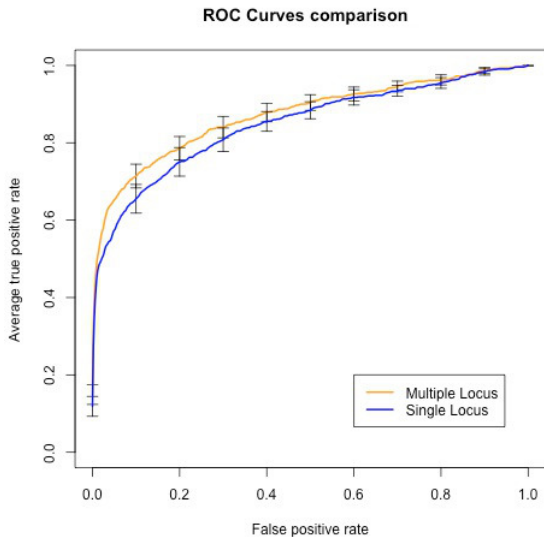
Results

- ▶ $n = 300$ 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.

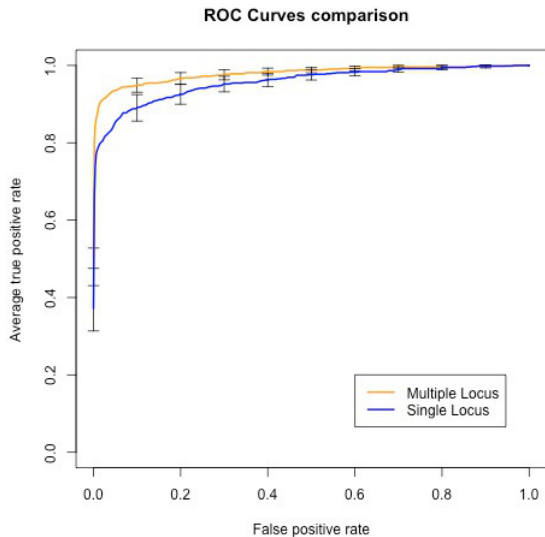
Weighted averaging with $p_0 = 5$, max var. = 0.5



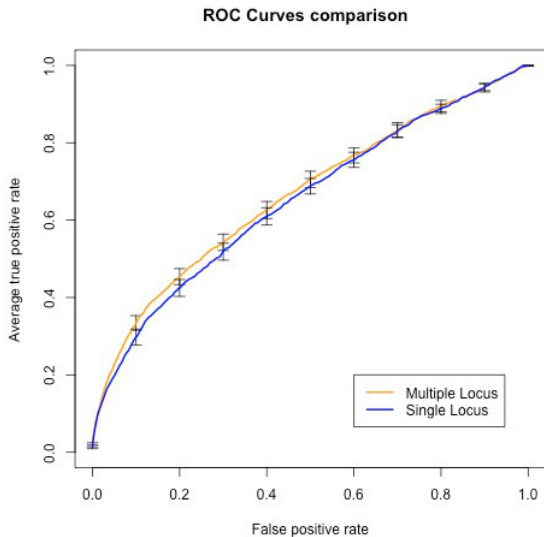
ROC curves comparison, $p_0 = 15$, max var.= 0.5



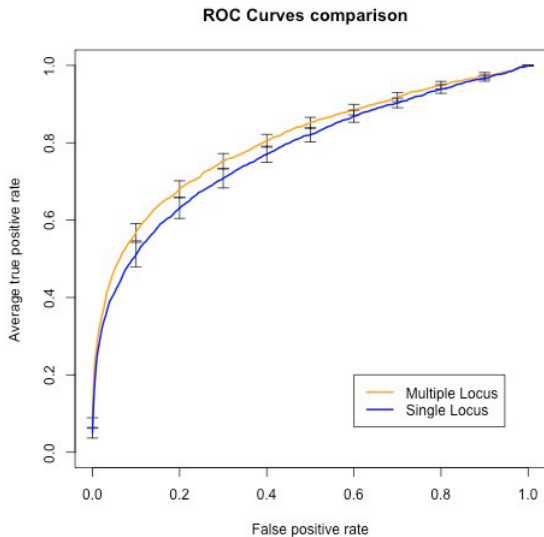
ROC curves comparison, $p_0 = 15$, max var.= 0.8



ROC curves comparison, $p_0 = 50$, max var.= 0.5



ROC curves comparison, $p_0 = 50$, 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.