Master Project: Statistical analysis on genomic data

Mid-term presentation

William van Rooij

EPFL

12.04.19

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

- $\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st}\mathcal{N}(0, \sigma^2\tau_t^{-1}) + (1 - \gamma_{st})\delta_0$,
  (spike and slab)
- $\gamma_{st} \mid \omega_s \sim \text{Bernoulli}(\omega_s)$,
- $\omega_s \sim \text{Beta}(a_s, b_s)$,
- a_s, b_s chosen to enforce sparsity. We choose 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 large compared to n.
- MCMC gets time consuming, computation 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

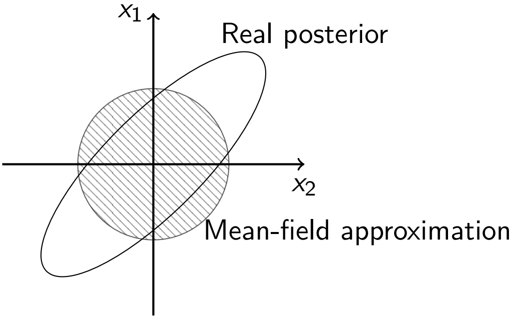
- Observed data $\boldsymbol{y}$, parameters $\boldsymbol{\theta}$, posterior distribution of parameters $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
- Approximate the posterior density with a simpler density q, minimizing a "closeness" measure: the Kullback-Leibler divergence.
- $\mathrm{KL}(q \parallel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

Mean-field approximation

- We assume independence for some 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 compute the correlations between parameters.



Parameters distributions

- $\beta_{st} \mid \gamma_{st} = 1, \boldsymbol{y} \sim \mathcal{N}\left(\mu_{\beta,st}, \sigma^2_{\beta,st}\right)$,
- $\beta_{st} \mid \gamma_{st} = 0, \boldsymbol{y} \sim \delta_0$,
- $\gamma_{st} \mid \boldsymbol{y} \sim \text{Bernoulli}(\gamma^{(1)}_{st})$,

- $\sigma^{-2}_{\beta,st} = \tau^{(1)}_t \left\{\|\boldsymbol{X}_s\|^2 + (\sigma^{-2})^{(1)}\right\}$,
- $\mu_{\beta,st} = \sigma^2_{\beta,st}\tau^{(1)}_t \boldsymbol{X}^T_s \left\{\boldsymbol{y}_t - \sum_{j=1,j\neq s}^{p} \gamma^{(1)}_{jt}\mu_{\beta,jt}\boldsymbol{X}_j\right\}$,

Coordinate ascent variational inference - CAVI

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

Coordinate ascent variational inference - CAVI II

- $\mathcal{L}(q)$ is guaranteed to augment at every iteration.
- 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.

"Bayesian model averaging"

- 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 $p(\boldsymbol{y} \mid M_k)$, 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: $p(M_k) = 1/K$, $\forall k = 1, \ldots, K$.

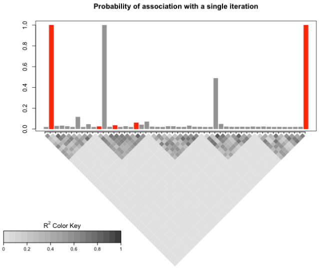
Implementation

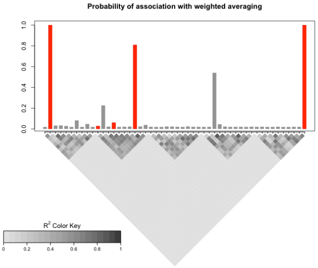
- Generate SNPs, traits, and dependences.
- Find the optimums $q^*(\boldsymbol{\theta})$ with different initial parameters, drawn at random.
- Generate the ELBOs and use them as weights in the weighted average ("BMA").
- The function yields the probabilities of association between SNPs and traits.

Results

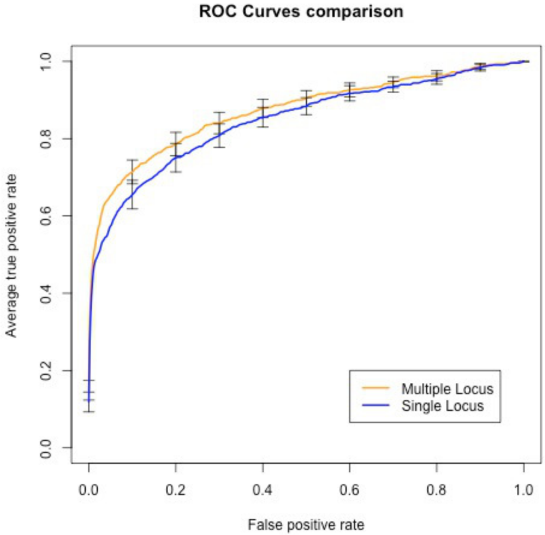
- $n = 300$ observations,
- $p = 500$ SNPs, with p_0 active SNPs per trait,
- $q = 1$ trait,
- 100 random initialisations,
- correlation between the SNPs is between 0.95 and 0.99, in blocks of ten SNPs,
- we can specify the maximum variance explained by

Weighted averaging with $p_0 = 5$

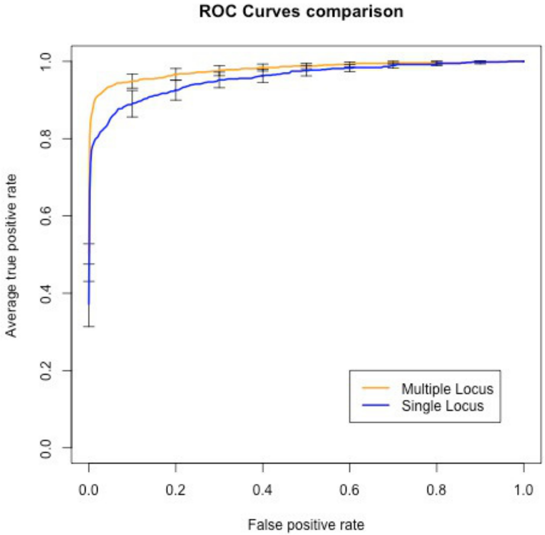




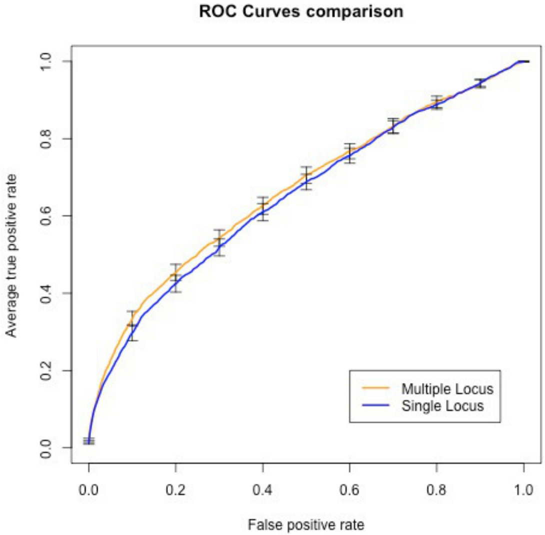
ROC curves comparison, $p_0 = 15$, max var.$= 0.5$



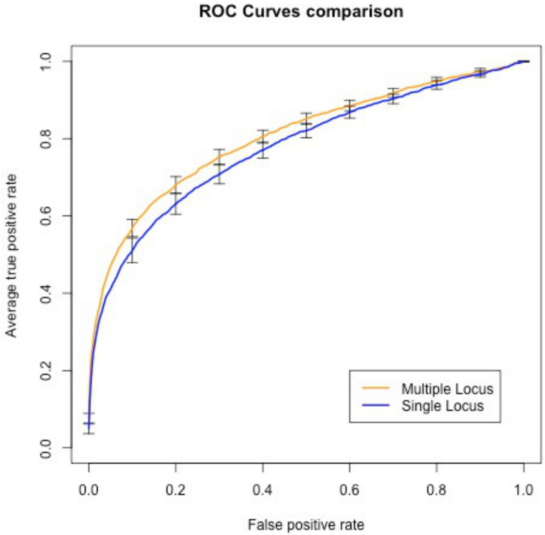
ROC curves comparison, $p_0 = 15$, max var.$= 0.8$



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



ROC curves comparison, $p_0 = 50$, max var.$= 0.8$



EPFL

- The paralleled version is not necessarily more time consuming.
- The difference is bigger when phenotypic variance is better explained from the SNPs.
- The difference is bigger with fewer active SNPs.

Next steps

- Optimize code,
- Comparison with annealing for strong correlations,
- Do we find the right modes?