Mid-term presentation

William van Rooij

EPFL

12.04.19

- Introduction
- Variational inference
- Mean-field approximation
- LOCUS R
- Results
- Future

EPFL

Introduction

- Estimate the association between SNP s and trait t, $s = 1, \ldots, p$, $t = 1, \ldots, q$
- $y_{n \times q} = x_{n \times p} \beta_{p \times q} + \epsilon_{n \times q}$, $\epsilon \sim \mathcal{N}(0, \tau_t^{-1} I_n)$
- y is response matrix, x are candidate predictors.
- Each response y_t is linearly related with the predictors and has a residual precision $\tau_t \sim \text{Gamma}(\eta_t, \kappa_t)$.
- p and q really large compared n.

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)
- $\gamma_{st} \mid \omega_s \sim \text{Bernoulli}(\omega_s)$,
- $\omega_s \sim \text{Beta}(a_s, b_s)$,

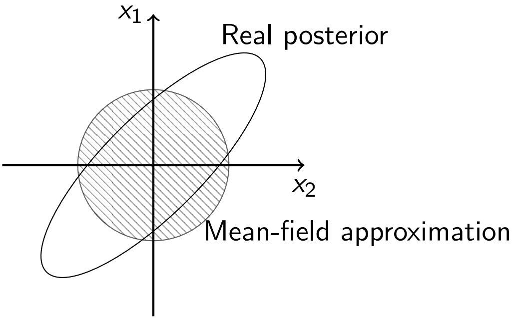
EPFL

Variational inference

- Observed data $\boldsymbol{y}$, parameters $\boldsymbol{\theta}$, posterior distribution of parameters $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
- Approximate 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): $\mathcal{L}(q) = \mathbb{E}\left[\log p(\boldsymbol{\theta}, \boldsymbol{y})\right] - \mathbb{E}\left[\log q(\boldsymbol{\theta})\right]$.
- $\mathrm{KL}(q \parallel p) = \log(p) - \mathcal{L}(q)$.
- Minimizing KL is equivalent to maximizing ELBO.

Mean-field approximation

- We assume $q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j)$, where $\{\theta_j\}$, $j = 1, \ldots, J$ is a decomposition of $\boldsymbol{\theta}$.
- The mean-field approximation does not compute the correlations between parameters.

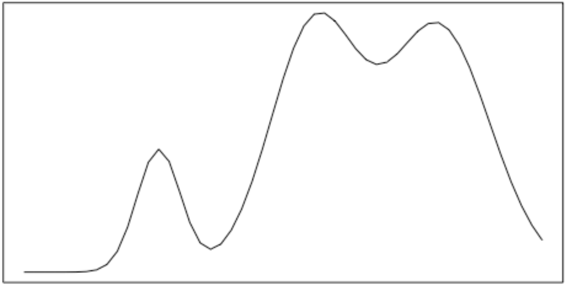


Coordinate ascent variational inference - CAVI

- If we fix $q_l(\theta_l)$, $l \neq j$, the optimal for $q_j(\theta_j)$ verifies: $q_j^*(\theta_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y}\right]\right\}$
- 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: $q(z)$.

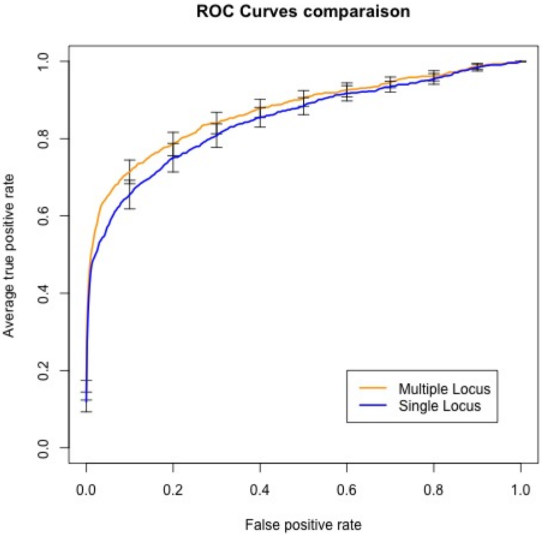
Coordinate ascent variational inference - CAVI II

- $\mathcal{L}(q)$ is guaranteed to augment at every iteration.
- CAVI yields a local optimum, depending on the initialization of the parameters.

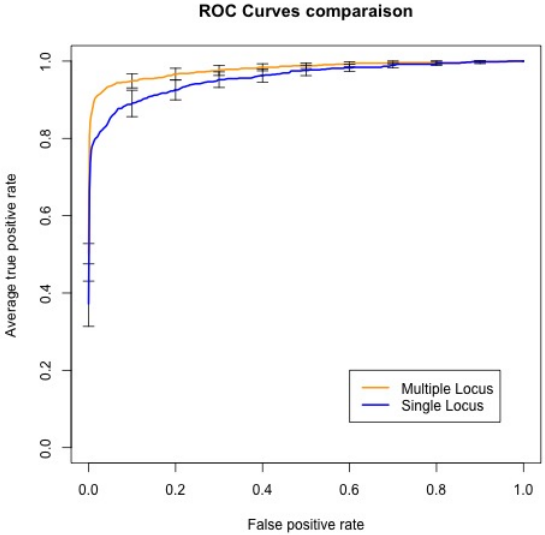


Bayesian model averaging

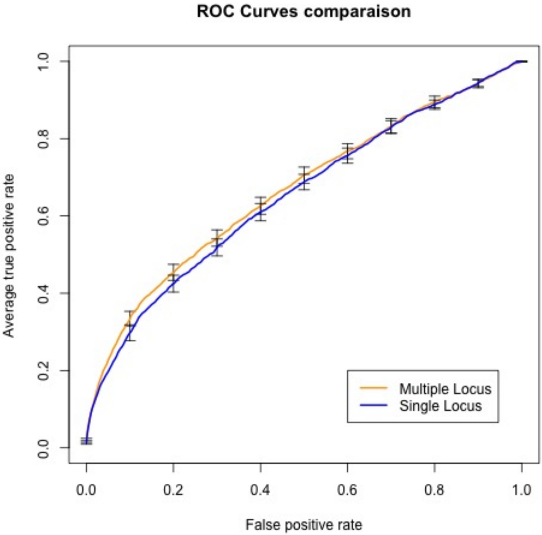
ROC curves comparison, $p_0 = 15$, max var.$= 0.5$



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



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



EPFL

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