Bayesian multiple logistic regression improves loci prioritization and finemapping in case-control GWAS

Saikat Banerjee[1], Lingyao Zeng[2], Heribert Schunkert[2] and Johannes Söding[*,1]

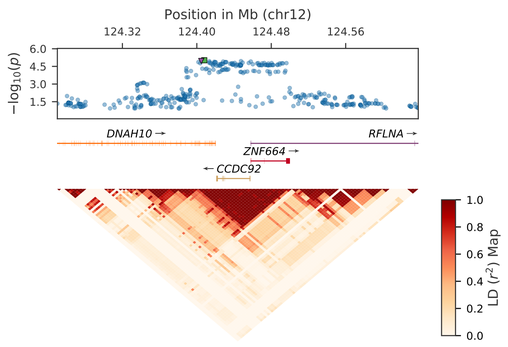
[1]Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany

[2]German Heart Centre, 80636 Munich, Germany

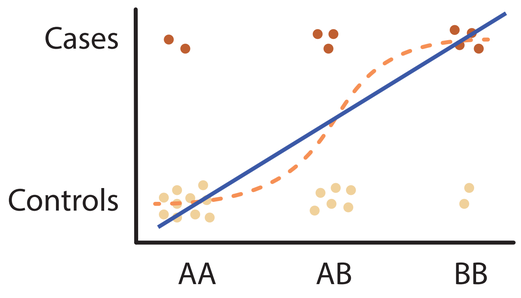
saikat.banerjee@mpibpc.mpg.de, soeding@mpibpc.mpg.de

https://github.com/soedinglab/b-lore

1. Post-GWAS analyses using multiple regression could not utilize the benefits of logistic model



Linkage disequilibrium ☹



Linear model ☹

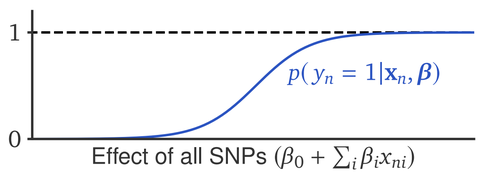
Challenges for using multiple logistic regression:

- The integration for the maximum likelihood cannot be solved analytically.
- MCMC sampling is computationally intractable.
- Solutions using Laplace and linear approximations essentially makes it a linear model.
- Metaanalysis requires knowledge of LD structure of the population.
- Limited to application on single genome-wide significant locus (cannot prioritize loci).

2. B-LORE uses logistic model and sparsity-inducing priors

Probability of n^{th} individual with genotype $\mathbf{x}_n$ to be diseased:

$$p\left(y_n = 1 \mid \mathbf{x}_n, \boldsymbol{\beta}\right) = \frac{\exp\left(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_n\right)}{1 + \exp\left(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_n\right)}$$



Prior on effect sizes given hyperparameters π and σ, $p\left(\beta_i \mid \pi, \sigma\right)$

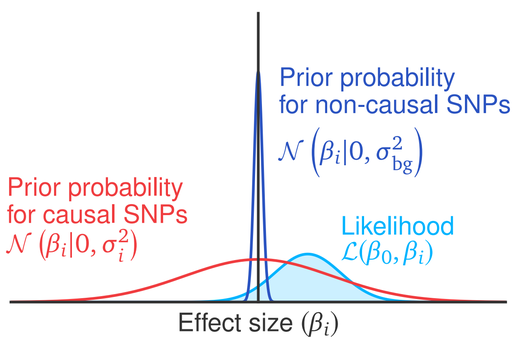
$$= \overbrace{\pi\,\mathcal{N}\left(\beta_i \mid 0, \sigma^2\right)}^{\text{Causal}} + \overbrace{(1-\pi)\,\delta_0}^{\text{Non-causal}}$$

$$= \sum_{z_i=0,1} \pi^{z_i}(1-\pi)^{(1-z_i)}\mathcal{N}\left(\beta_i \mid \mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\mathbf{z},\mathbf{i}})\right)$$

$$= \sum_{z_i=0,1} p\left(\mathbf{z} \mid \pi\right)\mathcal{N}\left(\beta_i \mid \mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\mathbf{z},\mathbf{i}})\right)$$

where, $\sigma^2_{\mathbf{z},i} = z_i\sigma^2$

$z_i \in \{0,1\} \Rightarrow$ Indicator variable of causality



- $z_i = 1$ SNP i is causal
- $z_i = 0$ SNP i is non-causal

3. We introduce the quasi-Laplace approximation

Evidence approximation: maximizing the marginal likelihood

$$\mathrm{m}\mathcal{L}(\pi, \sigma) := p\left(\mathbf{y} \mid \mathbf{x}, \pi, \sigma\right) = \sum_{\mathbf{z}} p\left(\mathbf{z} \mid \pi\right) \int p\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}\right)\mathcal{N}\left(\boldsymbol{\beta} \mid \mathbf{0}, \mathrm{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right) d\boldsymbol{\beta} \rightarrow \max$$

Quasi-Laplace approximation:

$$p\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}\right)\mathcal{N}\left(\boldsymbol{\beta} \mid \mathbf{0}, \mathrm{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right) = \underbrace{p\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}\right)\mathcal{N}\left(\boldsymbol{\beta} \mid \mathbf{0}, \tilde{\sigma}^2\mathbb{I}\right)}_{\propto\, \mathcal{N}\left(\boldsymbol{\beta} \mid \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Lambda}}^{-1}\right)} \frac{\mathcal{N}\left(\boldsymbol{\beta} \mid \mathbf{0}, \mathrm{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right)}{\mathcal{N}\left(\boldsymbol{\beta} \mid \mathbf{0}, \tilde{\sigma}^2\mathbb{I}\right)}$$

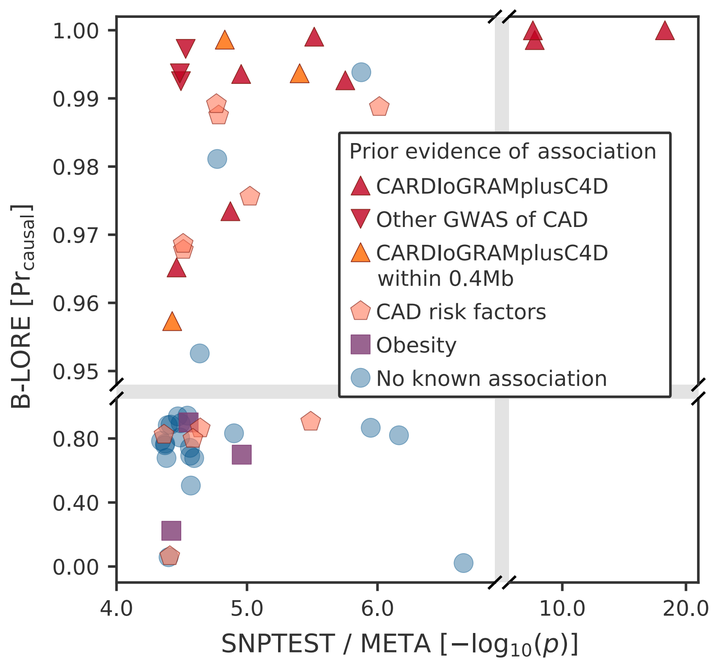
Benefits:

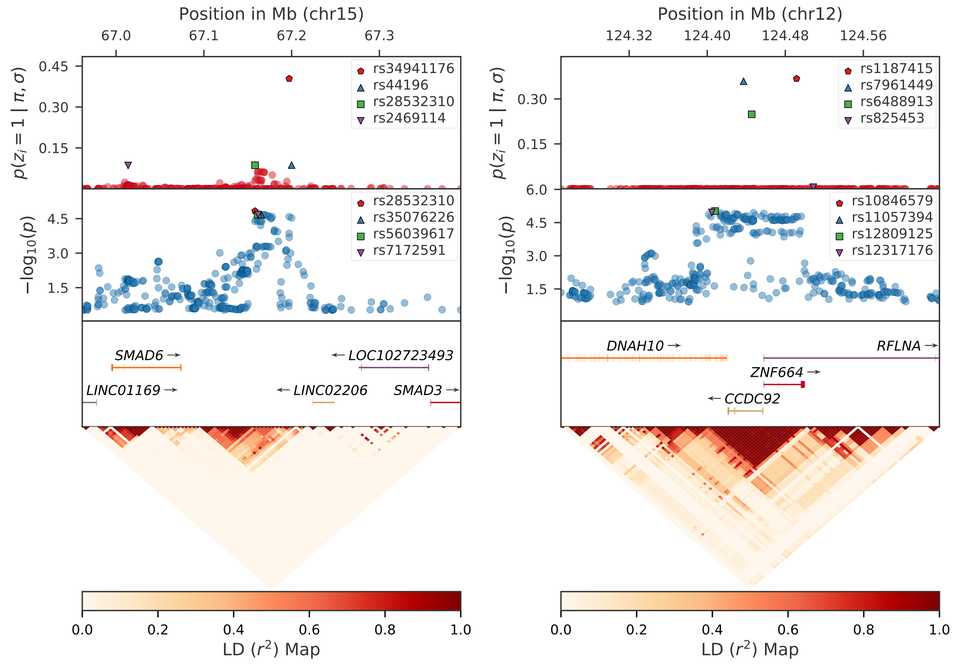
- The regularizer pulls the maximum of the regularized likelihood near to the mode of the integral, making it more accurate than Laplace approximation.
- Can be extended to multiple studies.
- Fast gradient-descent optimization.

B-LORE schema

1. Two-step optimization at each cohort to estimate $\tilde{\sigma}$ and $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Lambda}})$.
2. Estimation of hyperparameters (π, σ).

5. Meta-analysis example: B-LORE discovers novel loci associated with coronary artery disease





Meta-analysis of 5 cohorts, Germal Myocardial Infarction Family Studies (GerMIFS I-V) – 6234 cases and 6848 controls.

4. Inference

Prediction of causality of each locus.

The probability for a locus to be causally associated with the disease is

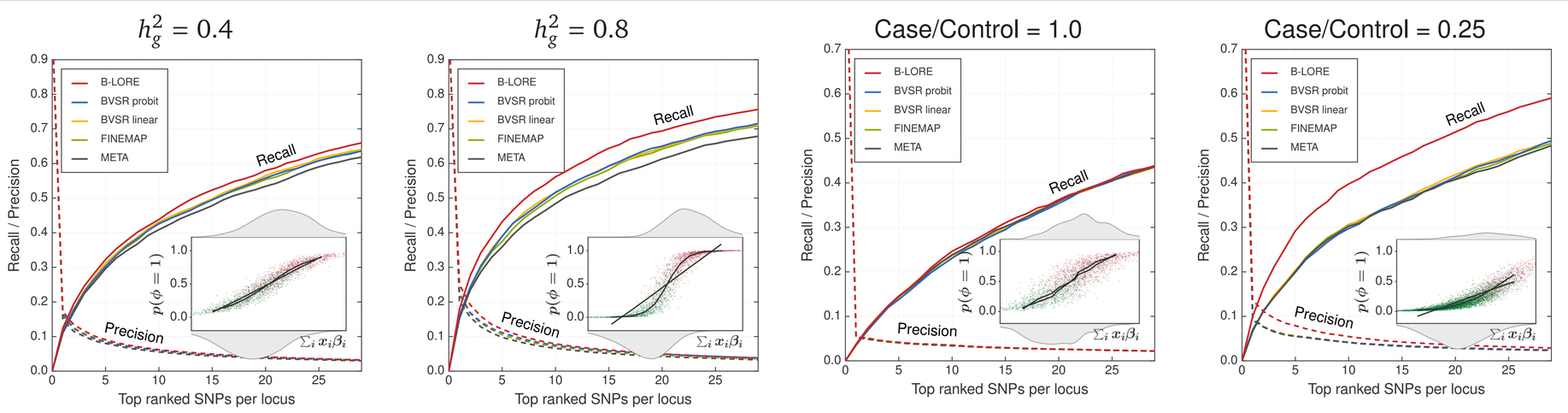
$$\begin{aligned}\mathrm{Pr}_{\text{causal}} &= p\left(\text{locus is causal} \mid \mathbf{y}, \mathbf{X}, \hat{\pi}, \hat{\sigma}\right) \\ &= 1 - p\left(\mathbf{z} = 0 \mid \mathbf{y}, \mathbf{X}, \hat{\pi}, \hat{\sigma}\right)\end{aligned}$$

Statistical finemapping of causal variants.

The posterior probability for SNP i to be causal is

$$p\left(z_i = 1 \mid \mathbf{y}, \mathbf{X}, \hat{\pi}, \hat{\sigma}\right)$$

6. Examples of non-linear regimes in case-control GWAS



7. References

1. Banerjee *et al.* PLOS Genet 2018, doi:10.1371/journal.pgen.1007856
2. Servin *et al.* PLOS Genet 2007, doi:10.1371/journal.pgen.0030114
3. Guan *et al.* Ann Appl Stat 2011, doi:10.1214/11-AOAS455
4. CARDIoGRAMplusC4D Nat Genet 2015, doi:10.1038/ng.3396

8. Acknowledgement

We thank Prof. Dr. Jeanette Erdmann for helpful discussions. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant 01ZX1313A-2014).

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
e:Med SYSTEMS MEDICINE