Final Project Report: Variational inference of polygenic risk score using empirical prior

Sagar Nandeshwar (260920948)

Abstract

In the project, I tried to explore the usefulness of empirical prior in PRS methods. I build VIPRS by implementing the updates of point estimates for the annotation weights w and the point estimates for the variance explained per annotation. I evaluated its performance using R-square methods on UK Biobank. The observed R-square values are low, which may be due to small sample size, selection of hyperparameters, biased initial values.

1 Introduction

1.1 Opportunity

Large-scale genome-wide association studies (GWASs) has driven the development of statistical methods for phenotype prediction. One of such techniques is the polygenic risk score (PRS) methods that formulate the task of polygenic prediction in terms of a multiple linear regression framework, where the goal is to infer the joint effect sizes of all genetic variants on the trait. These methods are used to estimate an individual's genetic risk for a particular trait or disease based on the combined effect of multiple genetic variants, hence they are a valuable tool for predicting genetic risk and understanding the genetic basis of complex traits and diseases.

1.2 Challenges

PRS methods when paired with modern GWAS sample sizes, which consists of hundreds of thousands of individuals, with high dimensional data possess several computational and statistical challenges. The closed-form update equations for some of the variational parameters involve terms that relate to the LD between the focal variant and all other variants in the genome. This computationally prohibitive to compute for millions of variants and for hundreds of EM iterations. Furthermore, most individual-level GWAS data sources are protected for privacy concerns.

1.3 Related Work

One of the common methods in these analyses is single nucleotide polymorphisms (SNPs), measured by either genotyping arrays or imputed using reference haplotypes. Byesian models are other PRS models that incorporate prior knowledge such as probability distributions over the genetic causal architecture of complex traits. However, a major limitation of Bayesian methods is that their scalability is extremely slow and inefficient with its inference techniques.

2 Methodology

2.1 Overview

In this project, I implemented Variational inference of polygenic risk score (VIPRS) which is an is an efficient Bayesian inference algorithm utilizes variational inference to approximate the posterior for the effect sizes for high-dimensional multiple regression applied in the context of polygenic risk score (PRS) problem. Here, I have implemented the updates of point estimates for the annotation weights w and the point estimates for the variance explained per annotation k, updating the Empirical Bayes algorithm (Expectation-Maximization (EM)).

2.2 Theory

For P SNPs and and N individuals, let X denotes P $\times$ N genotype matrix and $y \in R^{Nx1}$ the phenotype vector. Let $a_{j,k} \in 0, 1$ be the binary indicator for whether SNP j is in annotation k, i.e., $j \in C_k$, where C_k is the set of SNPs that are in annotation k. Therefore, for K annotations, we have a P $\times$ K binary annotation matrix A.

Let $w \in R^{kx1}$ be the unknown annotation weights, $\sigma_k^2$the variance explained by SNP j if it is in annotation k, and β and s be the effect size and binary indicator variables of causal SNPs, respectively.

Together, we have the following data generative process.

$$\pi_j = \frac{1}{1 + exp(-\sum_k a_{jk} w_k)} = \sigma(a_j w)$$

$$p(s_j|\pi_j) = Bernoulli(s_j|\pi_j) = \pi_j^{s_j}(1 - \pi_j)^{1-s_j}$$

$$p(\beta_j|s_j = 1, a_{j,k} = 1) = N(\beta_j; 0, \sigma_k^2)$$

Likelihood:

$$p(y|X, \beta) = N(X(s.\beta), \sigma_\epsilon^2 I)$$

Mean-field variational distribution to approximate the true posterior for $p(\beta, s|X, y)$:

$$q(\beta, s) = \Pi_j q(\beta_j|s_j)q(s_j)$$

2.3 Algorithm

The EM algorithm of Bayesian PRS with spike-and-slab prior

E-step: For each SNP j given the estimates for all other SNPs i != j, do:

$$q^*(\beta_j|s_j = 1) = N(\mu^*_{\beta_j}, 1/\tau^*_{\beta_j})$$

$$\tau^*_{\beta_j} = x'_j x_j \tau_\epsilon + \sum_{kwherea_{j,k}=1} \tau_k$$

$$\mu^*_{\beta_j} = N\frac{\tau_\epsilon}{\tau^*_{\beta_j}}(y' x_j - \sum_{i!=j} \gamma^*_i \mu^*_{\beta_i} r_{ij})$$

$$u_j = ln(\frac{\pi}{1-\pi}) + \frac{1}{2}ln(\frac{\tau_\beta}{\tau^*_{\beta_j}}) + \frac{\tau^*_{\beta_j}}{2}{\mu^*_{\beta_j}}^2$$

$$q^*(s_j = 1) = \frac{1}{1 + exp(-u_j)} == \gamma^*_j$$

M-step: update hyperparameters:

$$\tau_\epsilon = 1$$

$$\tau_k^{-1} = \sum_{j \in C_k} \gamma^*_j({\mu^*_{\beta_j}}^2 + {\tau^*_{\beta_j}}^{-1}) / \sum_{j \in C_k} \gamma^*_j$$

$$w = w - n(\nabla w) where \nabla w = (\gamma^* - \pi)A$$

2.4 Evaluation

I use R-squared to evaluate the performance of trained model performs on the corresponding testing fold. R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variable(s) included in the model. R-squared values range from 0 to 1, with a value of 1 indicating that the model explains 100% of the variation in the dependent variable and a value of 0 indicating that the model explains none of the variation.

R square can be computed from test summary statistics as follows: Formula

$$R^2 = 1 - \frac{SS_{res}}{SS_{tol}}$$

$$SS_{res} = \frac{1}{N_{test}}(y_{test} - X\hat{\beta})'(y_{test} - X\hat{\beta})$$

$$= \frac{1}{N_{test}} y'_{test} y_{test} - 2\frac{y'_{test} X_{test}}{N_{test}}\hat{\beta} + \hat{\beta}'\frac{1}{N_{test}} X_t est' X_t est \hat{\beta}$$

$$= 1 - 2\tilde{\beta}'_{test}\hat{\beta} + \hat{\beta}' R \hat{\beta}$$

$$SS_{tot} = \frac{1}{N_{test}}(y_{test} - \bar{y}_{test})'(y_{test} - \bar{y}_{test}) = 1$$

2.5 Evidence lower bound

I have also implemented the L_{ELBO} for the model, but due to computational limitation, I am unable to run it for the whole data.

$$L_{ELBO} = E_{q(\beta,s)}[lnp(y|\beta, s)] + E_{q(\beta)}[lnp(\beta|s)] + E_{q(s)}[lnp(s|\pi)] - E_{q(\beta,s)}[lnq(\beta|s)] - E_{q(s)}[lnq(s)]$$

2.6 Data-set

2.6.1 Linkage-Disequilibrium matrices

I used pre-computed LD matrices from the UK Biobank dataset from Zenodo. These matrices record the SNP correlations in a random sample of 50000 individuals from the White British

cohort in the UK Biobank dataset. I used the matrix for autosomal chromosome chr21 and ch22.

2.6.2 Annotations

The annotations folder stores the annotations for SNPs on chromosome 21 and 22. For the purpose of this project I have only used binary annotations.

2.6.3 Summary statistics data

I use the training and testing summary statistics data computed for human standing height on chromosome 21 and 22 from 337,205 White British individuals from the UK Biobank.

2.7 Data Processing

I use magenpy, a python library for loading, manipulating, and simulating with genotype data. I created GWADataLoader object for training and testing data per fold, and then creates several numpy arrays to store per SNPs data and per annotations data. The use of numpy methods has significantly speed up the algorithm. I have also limits the γ value it $0.01, 0.99$), in order to stabalize the algorithm.

2.8 Simulation

For each of the two LD matrix (ch21 and ch21) For each of the five fold

1. Use mergepy GWADataLoader to
   (a) Load the LD matrix
   (b) Load the Training marginal beta
   (c) Load the Annotation matrix

   mergy simultaneous harmonize data all the three dataset
2. Use numpy to create array for
   (a) per snps data
      i. $\mu^*_{\beta_j}$
      ii. $\tau^*_{\beta_j}$
      iii. τ_β
      iv. π
      v. γ^*_j
   (b) per annotation data
      i. τ_k
      ii. w
3. Run the EM algorithm for 10 iterations with
   (a) Learning rate = 0.001

(b) $\tau_\epsilon = 1, 0$

4. Use mergepy to load testing marginal beta
5. match the testing Marginal beta with training Marginal beta
6. Compute the R square

2.9 Result

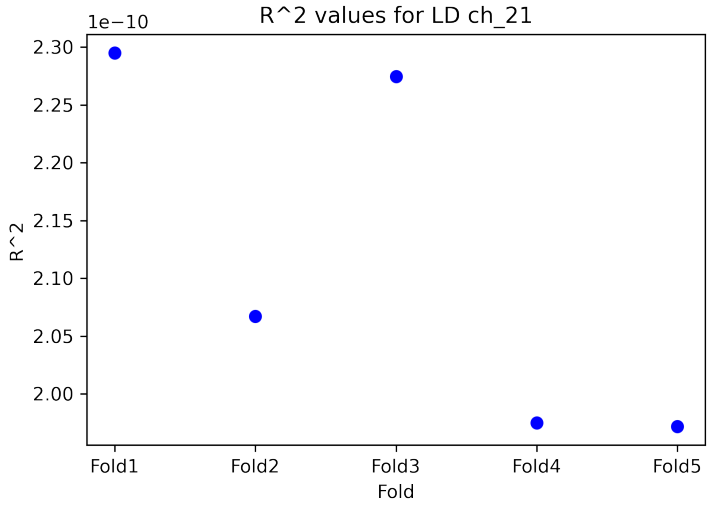

Figure 1: R^2 values with LD ch 21

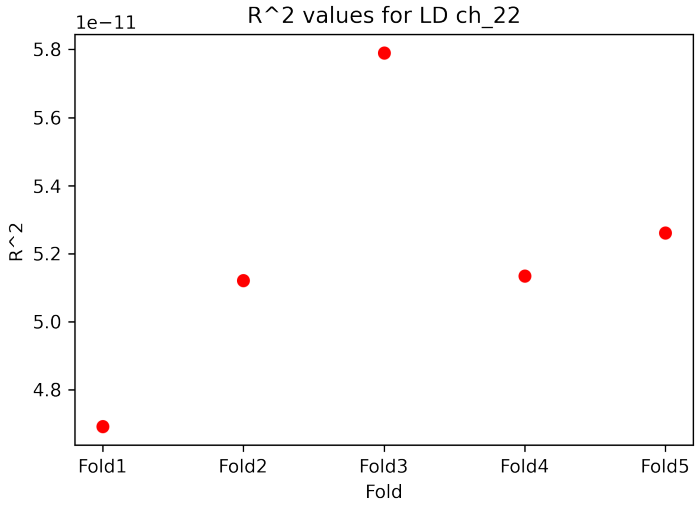

Figure 2: R^2 values with LD ch 22

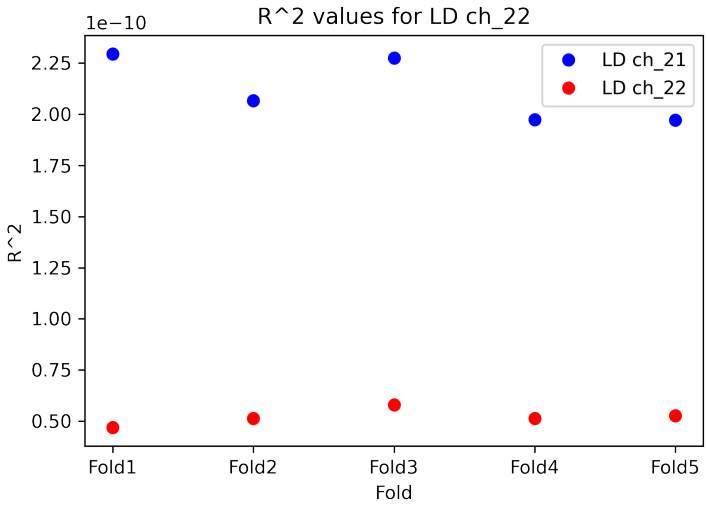

Figure 3: R^2 values

3 Discussion

3.1 Conclusion

I was able to implement the EM algorithm for VIPRS and performed 5-fold cross-validation by training and evaluating your model on the training and testing fold by computing standard error of the R-squared estimates for each method for both chromosomes ch21 and ch22 sets in an efficient way with the help of magenpy and numpy.

Within the same chromosomes group the r square values are very similar, however we see that the model performs slightly better when trained with LD of chromosomes ch21 than ch22. All the R square values observed are low. This may be due to small sample space, biased initial values or because of the hyperparameter selection.

3.2 Limitations

The method can be affected from population stratification, which occurs when there are systematic differences in allele frequencies between different populations. This can lead to spurious associations between genetic variants and traits or diseases.

Secondly, the method assumes that there is a linear relationship between the effect of individual genetic variants and the risk for the trait or disease. The actual relationship may be non-linear or depend on environmental factors or different genetic variants interactions.

The PRS methods heavily relies on genome-wide association studies (GWAS) to identify genetic variants associated with a particular trait or disease. However, GWAS studies often have limited sample sizes, which can lead to false positives or miss important genetic variants that contribute to the trait or disease.

3.3 Future work

We could try to run the model with different hyperparameters and initial values, and could also implement Evidence lower bound to observe the models performances over the time of training.

Second, the current spike-and-slab prior assumes that all genetic variants have a uniform prior probability of being causal and that the causal SNPs have equal expected contribution to the heritability. For the future work we could explore a more general and flexible Gaussian mixture prior.

Lastly, we could try to joint the model effect sizes from multiple ancestrally homogeneous populations within the same framework and evaluate the performance across different populations.

4 Codes

- code Folder: Contains the codes of the Final Project
- data Folder: Contains the datasets for the project
- instructions Folder: Contains the instruction for the project
- pics Folder: Contains pictures of results plot

5 Reference

Please note that some of the content in this paper is inspired and taken from Professor Li and TA's research paper.

[1] "Fast and accurate Bayesian polygenic risk modeling with variational inference"(2022) - Shadi Zabad, Simon Gravel, and Yue Li