MegaPRS in Math

Tuesday, January 23, 2024 3:00 PM

[Improved genetic prediction of complex traits from individual-level data or summary statistics | Nature Communications]

Method

Suppose n individuals and m snps, X (size n*m), Xj column contains genotypes for SNP j, Y denotes phenotype.

Xj and Y been standardized, N(0,1), the error term is also N(0,1)

Consider the prediction of X on Y, by linear regression

$$\mathbf{E}[\mathbf{Y}] = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \ldots + \mathbf{X}_m\beta_m = \mathbf{X}\boldsymbol{\beta} \quad (1)$$

β is the effect size of SNP j, and since Xj and Y are standardized, so we have $h_j^2 = \beta_j^2$

Heritability

$$h^2 = \frac{Var(X\beta)}{Var(Y)}$$

The heritability model takes form:

$$E[h_j^2] = a_{j1}\tau_1 + a_{j2}\tau_2 + \ldots + a_{jK}\tau_K \quad (2)$$

where the a_{jk} are pre-specified SNP annotations, while the parameters τ_k are estimated from the data.

So if we take $h^2 = \tau_1 I_j(f_j(1-f_j)^{0.75})$, then the variance of β_j is $\tau_1 I_j(f_j(1-f_j)^{0.75}) \times h^2$, by SumHer

$$\boldsymbol{\beta_j} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\tau_1 I_j(f_j(1-f_j))^{0.75}} \times \boldsymbol{h^2})$$

$$\hat{Y} = \sqrt{h^2}X\beta + \sqrt{1-h^2}\,\epsilon$$

Calculation of heritability

Suppose I have X and Y as individual data, we can calculate h2 by the following formulas:

1. $r = {}^{XjY}/_{n}$, this is the correlation between X and Y
2. $S = n\frac{r^2}{1-r^2}$, this is the chi-square test statistics for SNP j.
3. For SumHer, it calculates τ by the following formula:

$$\mathbf{E}[S_j] \approx 1 + n\Sigma_l c_{jl}^2(a_{l1}\tau_1 + a_{l2}\tau_2 + \ldots + a_{lK}\tau_K) \quad (3)$$

   a. S was calculated by step 2. C_{jl}^2 is taken from the corrleation matrix of SNPs j and l.
   b. We can take a linear regression of S and Correlaation, and the coefficient is τ
4. $h^2 = \tau I_j(f(1-f))^{0.75}$,Where I is the LD matrix, and f is the MAF.

Suppose we are using Summary Statistics instead of individual data, the first step will be updated:
We have genotype data and Z-score.

$$\mathbf{S} = \boldsymbol{Z^2}$$

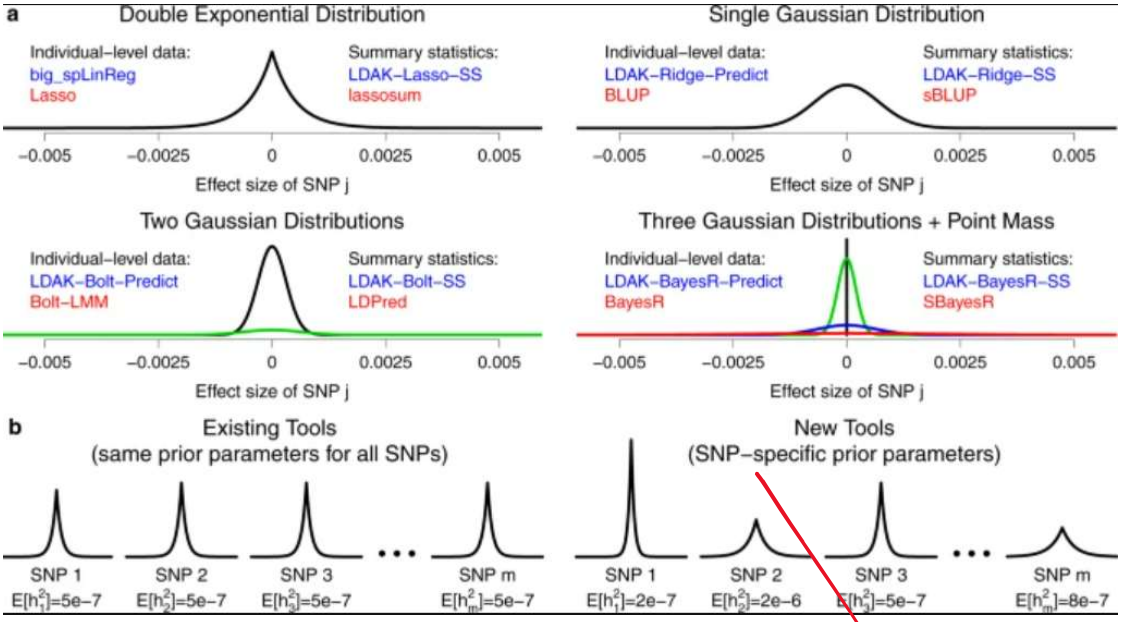
we can skip calculating the correlation between X and Y and directly get S = Zˆ2.

Prediction Model for β

Y in fact is based on Xβ and ϵ, and ϵ is normally distributed.

$$\boldsymbol{Y} \sim \boldsymbol{N}(\mathrm{X}\beta, \boldsymbol{\sigma}_{\epsilon}^{2})$$

We have h^2 already, so we need to estimate β. The newest methods estimate β by using Bayes Inference. Different methods would have different prior distributions.



1. Double Exponential Distribution

$$\beta_j \sim \mathrm{DE}(\lambda / \mathrm{E}[h^2{}_j]^{0.5})$$

2. Single Gaussian Distribution (Ridge Regression)

$$\beta_j \sim \mathrm{N}(0, \mathrm{E}[h^2{}_j]) \text{ and } \beta_j \sim \mathrm{N}(0, v\mathrm{E}[h^2{}_j])$$

Respectively for using Individual data or Summary Statistics.

3. Two Gaussian Distribution (Bolt-lmm)

Predict and LDAK-Bolt-SS use a mixture of two Gaussian distributions, $\beta_j \sim p\ \mathrm{N}(0, (1\text{-}f_2)/p\ \mathrm{E}[h^2{}_j]) + (1\text{-}p)\ \mathrm{N}(0, f_2/(1\text{-}p)\ \mathrm{E}[h^2{}_j])$. I

4. Three Gaussian Distributions + Point Mass

LDAK-BayesR-Predict and LDAK-BayesR-SS use a mixture of a point mass at zero and three Gaussian distributions, $\beta_j \sim \pi_1 \delta_0 + \pi_2\ \mathrm{N}(0, s\mathrm{E}[h^2{}_j]/100) + \pi_3\ \mathrm{N}(0, s\mathrm{E}[h^2{}_j]/10) + \pi_4\ \mathrm{N}(0, s\mathrm{E}[h^2{}_j])$, where $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$ and $s = (\pi_2/100 + \pi_3/10 + \pi_4)^{-1}$.

Calculation for β

Assuming we have $\hat{\beta}$ from Summary Statistics ($\beta = nZ$) as $\mathrm{P}(\hat{\beta})$, and we get the prior distribution of β as $P(\beta)$

For Bayesian Inference:

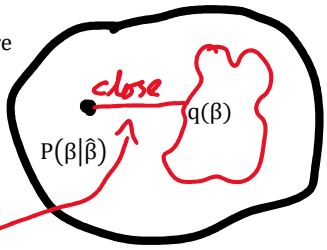
$$\mathrm{P}(\beta|\hat{\beta}) = \frac{\mathrm{P}(\hat{\beta}|\beta)\mathrm{P}(\beta)}{\mathrm{P}(\hat{\beta})}$$

$$\text{Posterior} = \frac{\text{MLE} \times \text{Prior}}{\text{Observe}}$$

So we need to calculate the MLE of β and $\hat{\beta}$, for example by Gradient Descent MLE, while these methods are slow in general. And posterior has no closed form, so it's not feasible to compute $\mathrm{P}(\beta|\hat{\beta})$ directly.

$$\text{Posterior} = \frac{\text{MLE} \quad \text{Prior}}{\text{Observe}}$$

So we need to calculate the MLE of β and $\hat{\beta}$, for example by Gradient Descent MLE, while these methods are slow in general. And posterior has no closed form, so it's not feasible to compute $P(\beta|\hat{\beta})$ directly.



Using Variational Bayes

Our goal becomes to find a distribution q(β) that approximates the posterior distribution $P(\beta|\hat{\beta})$

- Variational Bayes (VB): L is KL divergence

$$L\big(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x})\big) = \mathrm{KL}\big(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})\big)$$

$$\mathrm{KL}(q(\beta)|\,|P(\beta|\hat{\beta})) = \int q(\beta)\, log\, \frac{q(\beta)}{P(\beta|\hat{\beta})} d\beta$$

$$= \int q(\beta)\, log\, \frac{q(\beta)P(\hat{\beta})}{P(\beta,\hat{\beta})} d\beta$$

$$= -\int \mathbf{q(\beta)}\, \boldsymbol{log}\, \frac{\mathbf{P(\beta,\hat{\beta})}}{\mathbf{q(\beta)}} \boldsymbol{d\beta} + logP(\hat{\beta})$$

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

The first part is called ELBO (evidence lower bound), and since we need to minimize KL divergence, we can drop the second term $logP(\hat{\beta})$.

$\mathbf{q(\beta)} = \mathrm{argmax}\big(\mathrm{ELBO}(q(\beta))\big)$

And q(β) is the distribution of β that we need.

Prediction

We have β and h^2, and we have the genotype data, we can then calculate PRS score:

$$\hat{Y} = \sqrt{h^2} X\beta + \sqrt{1-h^2}\,\epsilon$$

Discussion

About β

If we standardize X and Y into N(0,1), then β will be N(0,1)

$Z = {}^{\beta}/_{SE(\beta)} = \sqrt{n}\beta \; and \; SE(\beta) = 1/\sqrt{n}$

Because $SE = SD/\sqrt{n}$

Steps of MegaPRS

Correlation Matrix and High LD matrix

Date. 2024/1/28

MegaPRS each step, with input and output.

一、calculation of high-ld and correlation matrix.

1. ① Correlation Matrix.

Input, randomly pick 5000 individuals with bim/bed/fam.

	ind 1	0 1 2 0 1 2 ———
n	ind 2	0 1 2 1 0 2 ———
	...	———
	ind 5000	0 0 1 2 1 0 ———

m

② Calculation: Pearson's correlation's coefficient.

· SNP$_i$ and SNP$_j$.

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2 \sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$$

③ Output: m×m matrix.

Date.

II. ① High LD matrix.

Input: reference panel of N inds and M snps.

High LD region list.

Index	From	End	chr.
region 1	47999999	52000000	1
—	—	—	—
region 24	—	—	—

• We'll compute high LD of SNPs in these regions.

② Calculation.

• P_{AB}, P_{Ab}, P_{aB}, P_{ab}.

• D'_{ij} represents LD of SNP_i and SNP_j

✓ $D'_{ij} = \frac{P_{AB}P_{ab} - P_{Ab}P_{aB}}{\min[P_A P_a, P_B P_b]}$ — ChatGPT

$[-1, 1]$

$D^+_{ij} = P_{AB} - P_A P_B$ — Pozhen Note.

✓ $r^2 = \frac{D^2}{P_A(1-P_A)P_B(1-P_B)}$.

$[0, 1]$

③ Output: M×M matrix.

Date.

二. Prediction Model.

① Input: summary statistics
cor matrix
LD matrix.
→ model bayesR.

② Predicting by BayesR, h^2.

$$h^2 = \sum_i I_i (f_i(1-f_i))^{0.75}.$$

$I_i = \sum I_{ij}/N$, $f_i = \sum f_{ij}/N$.
and τ_i is set to be 0.1, 0.3, 0.5.

$\Rightarrow h_1^2$, h_2^2, h_3^2.

③ Predicting by BayesR, β.

Prior: $\beta_j \sim \pi_1 \delta_0 + \pi_2 N(0, s[h_j^2]/w) + \pi_3 N(0, s[h_j^2]/10) + \pi_4 N(0, s[h_j^2])$

$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$, $s = (\pi_2/w + \pi_3/10 + \pi_4)^{-1}$.
and π_2, π_3, π_4 take from $(0, 0.01, 0.05, 0.1, 0.2)$
can be replicated. $\Rightarrow$ 35 models.

$\Rightarrow 35 \times 3 = 105$ models.

Date.

③ Posterior $P(\beta|\hat{\beta})$.

Assuming we have $\hat{\beta}$ from summary statistics (standardised)
$P(\hat{\beta})$ represents it ($\delta=1$, $\hat{\beta}=\frac{1}{\sqrt{n}}z$)
and $P(\beta)$ as prior distribution from ②.

$$P(\beta|\hat{\beta}) = \frac{P(\hat{\beta}|\beta)P(\beta)}{P(\hat{\beta})}.$$

⇒ Using VB to approximate $P(\beta|\hat{\beta})$.

Assuming $q(\beta)$ approximates $P(\beta|\hat{\beta})$.

$$q(\beta) = \text{argmax}\left(+\int q(\beta)\log\frac{P(\beta,\hat{\beta})}{q(\beta)}\,d\beta\right)$$

三. ④ Calculating PRS.

Input, Model and genotype data.

$$\hat{Y} = \sqrt{h^2}X\beta + \sqrt{1-h^2}\,\varepsilon \sim N(0,1).$$

⑤. Comparison: correlation $(Y, \hat{Y})$.