Non-linear Genetic Effects for Complex Traits

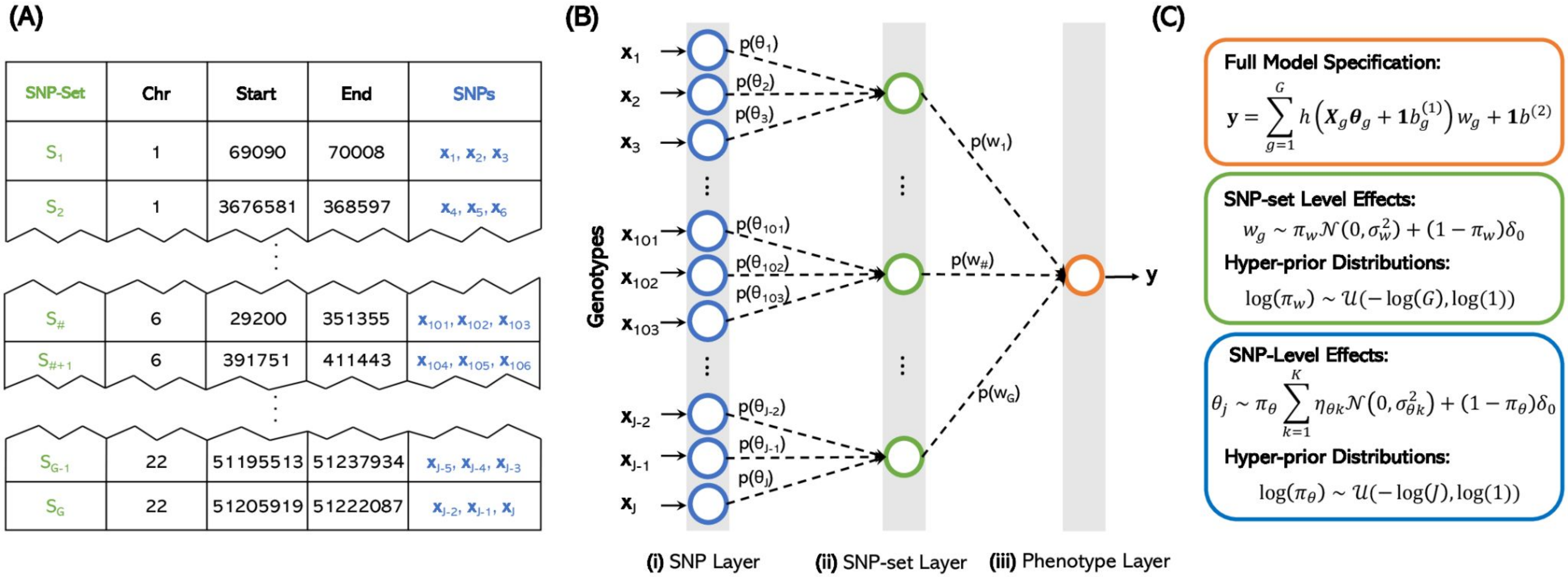
Yu-Wei Chen, Zhengtong Liu, Jacob Wallin

https://github.com/Zhengtong-Liu/CS-M226-Project

Introduction

- Polygenic Risk Score
    - Common method of evaluating association between genotypes and phenotypes
    - Captures linear relationship
- BANN (Biologically Annotated Neural Networks)
    - Machine learning model that can capture non-linear relationships
    - Interpretable statistical output
- VIPRS (Variational Inference of Polygenic Risk Scores)
    - Very efficient, uses VI instead of MCMC for posterior inference

Biologically Annotated Neural Networks (BANNs)



(A)

SNP-Set	Chr	Start	End	SNPs
S_1	1	69090	70008	x_1, x_2, x_3
S_2	1	3676581	368597	x_4, x_5, x_6
⋮				
$S_\#$	6	29200	351355	$x_{101}, x_{102}, x_{103}$
$S_{\#+1}$	6	391751	411443	$x_{104}, x_{105}, x_{106}$
⋮				
S_{G-1}	22	51195513	51237934	$x_{J-5}, x_{J-4}, x_{J-3}$
S_G	22	51205919	51222087	x_{J-2}, x_{J-1}, x_J

(B)

(C)

Full Model Specification:

$$\mathbf{y} = \sum_{g=1}^{G} h\left(\boldsymbol{X}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)}\right) w_g + \mathbf{1}b^{(2)}$$

SNP-set Level Effects:

$$w_g \sim \pi_w \mathcal{N}(0, \sigma_w^2) + (1 - \pi_w)\delta_0$$

Hyper-prior Distributions:

$$\log(\pi_w) \sim \mathcal{U}(-\log(G), \log(1))$$

SNP-Level Effects:

$$\theta_j \sim \pi_\theta \sum_{k=1}^{K} \eta_{\theta k} \mathcal{N}\left(0, \sigma_{\theta k}^2\right) + (1 - \pi_\theta)\delta_0$$

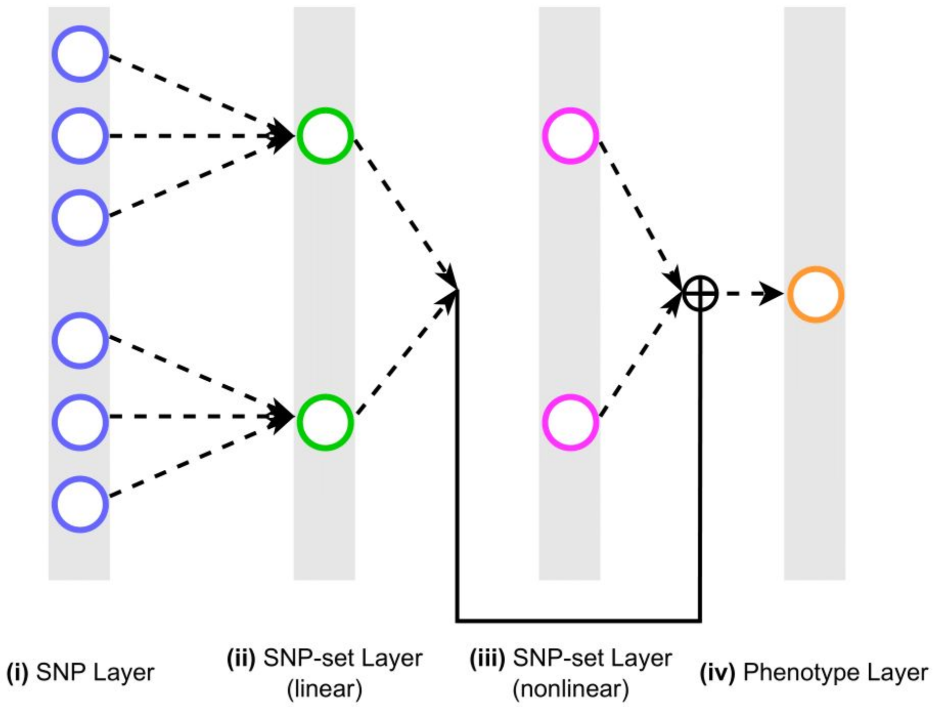
Hyper-prior Distributions:

$$\log(\pi_\theta) \sim \mathcal{U}(-\log(J), \log(1))$$

P. Demetci, W. Cheng, G. Darnell, X. Zhou, S. Ramachandran, and L. Crawford, “Multi-scale inference of genetic trait architecture using biologically annotated neural networks,” PLOS Genetics, vol. 17, no. 8, pp. 1–53, 08 2021.

Kernel BANNs

(A)



(B)

Modified Model:

$$\boldsymbol{y} = \sum_{g=1}^{c}(\boldsymbol{X}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g^{(1)} + \sum_{g=1}^{c} h(\boldsymbol{X}_g)\boldsymbol{w}_g^{(2)} + \mathbf{1}b^{(2)}$$

SNP Layer and SNP-set Level (linear):

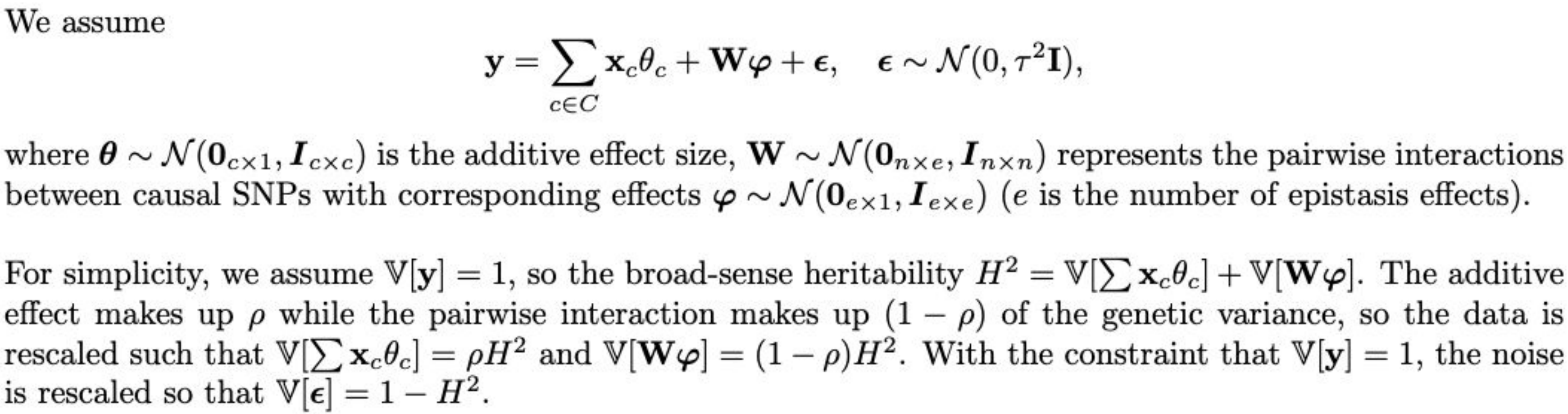
$$\boldsymbol{y}_{linear} = \sum_{g=1}^{c}(\boldsymbol{X}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g^{(1)} + \mathbf{1}b^{(2)}$$

SNP-set Level (nonlinear):

$$\boldsymbol{y}_{nonlinear} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{linear}$$

$$\boldsymbol{y}_{nonlinear} = \sum_{g=1}^{c} h(\boldsymbol{X}_g)\boldsymbol{w}_g^{(2)}$$

Simulation Design



We assume

$$\mathbf{y} = \sum_{c \in C} \mathbf{x}_c \theta_c + \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \tau^2 \mathbf{I}),$$

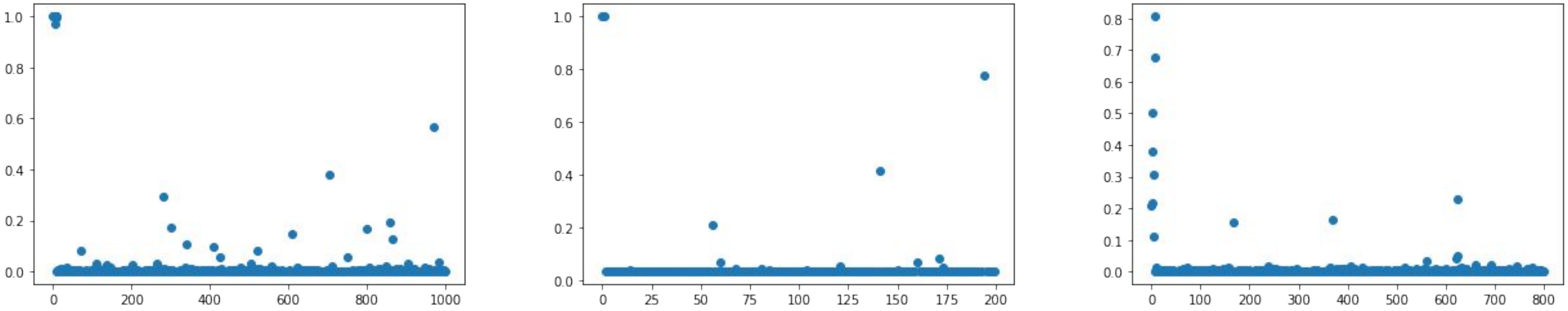
where $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_{c\times 1}, \boldsymbol{I}_{c\times c})$ is the additive effect size, $\mathbf{W} \sim \mathcal{N}(\mathbf{0}_{n\times e}, \boldsymbol{I}_{n\times n})$ represents the pairwise interactions between causal SNPs with corresponding effects $\boldsymbol{\varphi} \sim \mathcal{N}(\mathbf{0}_{e\times 1}, \boldsymbol{I}_{e\times e})$ (e is the number of epistasis effects).

For simplicity, we assume $\mathbb{V}[\mathbf{y}] = 1$, so the broad-sense heritability $H^2 = \mathbb{V}[\sum \mathbf{x}_c \theta_c] + \mathbb{V}[\mathbf{W}\boldsymbol{\varphi}]$. The additive effect makes up ρ while the pairwise interaction makes up $(1-\rho)$ of the genetic variance, so the data is rescaled such that $\mathbb{V}[\sum \mathbf{x}_c \theta_c] = \rho H^2$ and $\mathbb{V}[\mathbf{W}\boldsymbol{\varphi}] = (1-\rho)H^2$. With the constraint that $\mathbb{V}[\mathbf{y}] = 1$, the noise is rescaled so that $\mathbb{V}[\boldsymbol{\epsilon}] = 1 - H^2$.

Preliminary Results

- Phenotypic Variance Explained (PVE)
  - Ground truth: $H^2 = 0.6$
  - SNP Layer, SNP-set Layer (linear), SNP-set Layer (nonlinear): 0.48887597222414664, 0.4525421533848896, 0.1233089950726313

- Posterior Inclusion Probability (PIP)
  - Ground truth: First 10 SNPs are causal ones

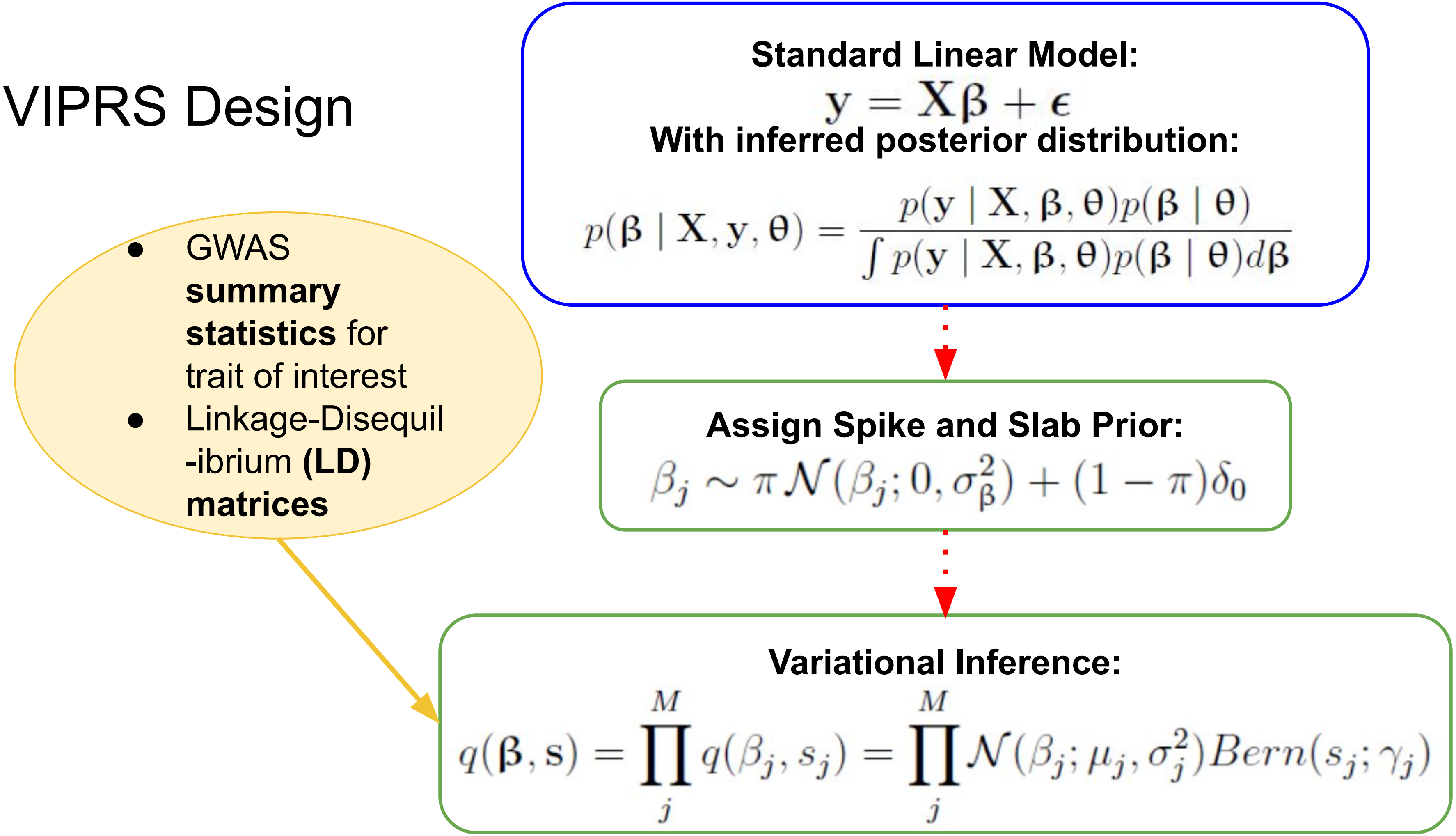


VIPRS

PRS methods that operate on GWAS summary statistics and sparse Bayesian methods have shown competitive predictive ability.

1. **Markov Chain Monte Carlo (MCMC)** for posterior inference: computationally inefficient
2. **Variational Inference**:
   a. Propose a simple parametric distribution q(β, s)
   b. Optimize the parameters to match the true posterior as closely as possible
   c. Closeness between the true posterior and the proposed distribution is measured by the Kullback-Leibler (KL) divergence and maximizing ELBO (Evidence Lower BOund)

VIPRS Design



- GWAS **summary statistics** for trait of interest
- Linkage-Disequilibrium **(LD) matrices**

Standard Linear Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

With inferred posterior distribution:

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta} \mid \boldsymbol{\theta})}{\int p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta} \mid \boldsymbol{\theta}) d\boldsymbol{\beta}}$$

Assign Spike and Slab Prior:

$$\beta_j \sim \pi \mathcal{N}(\beta_j; 0, \sigma_{\beta}^2) + (1 - \pi)\delta_0$$

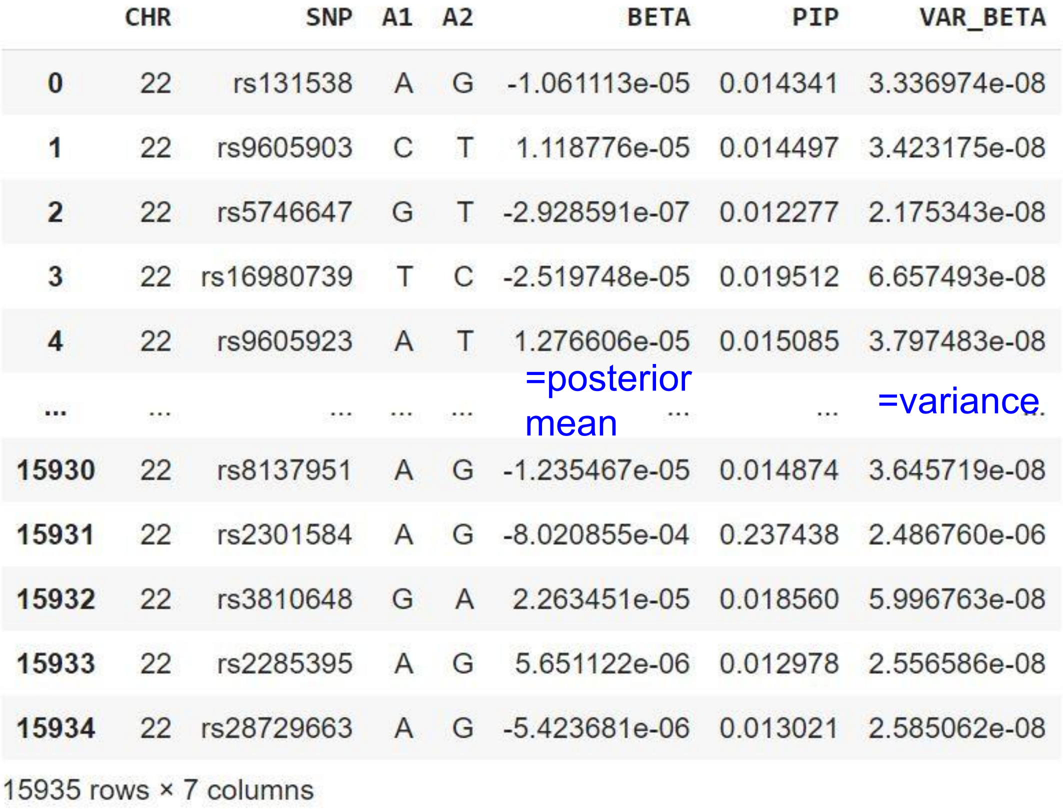
Variational Inference:

$$q(\boldsymbol{\beta}, \mathbf{s}) = \prod_{j}^{M} q(\beta_j, s_j) = \prod_{j}^{M} \mathcal{N}(\beta_j; \mu_j, \sigma_j^2) Bern(s_j; \gamma_j)$$

Shadi Zabad, Simon Gravel, and Yue Li. (2022) “Fast and Accurate Bayesian Polygenic Risk Modeling with Variational Inference,” bioRxiv 2022.05.10.491396

Example Results

- Examine posterior estimates for the model parameters



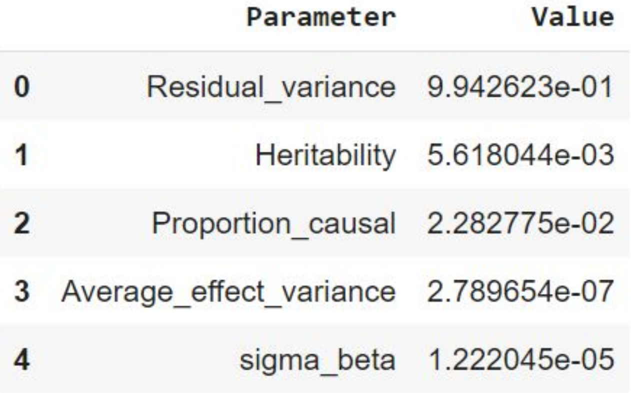
	CHR	SNP	A1	A2	BETA	PIP	VAR_BETA
0	22	rs131538	A	G	-1.061113e-05	0.014341	3.336974e-08
1	22	rs9605903	C	T	1.118776e-05	0.014497	3.423175e-08
2	22	rs5746647	G	T	-2.928591e-07	0.012277	2.175343e-08
3	22	rs16980739	T	C	-2.519748e-05	0.019512	6.657493e-08
4	22	rs9605923	A	T	1.276606e-05	0.015085	3.797483e-08
...	...	...	...	...	...	...	...
15930	22	rs8137951	A	G	-1.235467e-05	0.014874	3.645719e-08
15931	22	rs2301584	A	G	-8.020855e-04	0.237438	2.486760e-06
15932	22	rs3810648	G	A	2.263451e-05	0.018560	5.996763e-08
15933	22	rs2285395	A	G	5.651122e-06	0.012978	2.556586e-08
15934	22	rs28729663	A	G	-5.423681e-06	0.013021	2.585062e-08

15935 rows × 7 columns

=posterior mean

=variance

- Examine inferred hyperparameters



	Parameter	Value
0	Residual_variance	9.942623e-01
1	Heritability	5.618044e-03
2	Proportion_causal	2.282775e-02
3	Average_effect_variance	2.789654e-07
4	sigma_beta	1.222045e-05

Comparisons

<table>
<tr><th></th><th>BANNs</th><th>VIPRS</th></tr>
<tr><td rowspan="3"><b>Similarity</b></td><td colspan="2">Bayesian paradigm</td></tr>
<tr><td colspan="2">Same prior assumption</td></tr>
<tr><td colspan="2">Variational approximation</td></tr>
<tr><td rowspan="3"><b>Differences</b></td><td>Two layers, interpretable statistical output</td><td>One layer, computation efficiency</td></tr>
<tr><td>Relies on gene range information</td><td>GWAS summary statistics
Scalable and flexible</td></tr>
<tr><td>Worse performance on sparse data</td><td>Able to perform train-test split on data</td></tr>
</table>

Thanks!