Leveraging Probabilistic Programming for Flexible Inference in High-Dimensional Statistical Genetics Models

Viraj Chhajed[1,2], Richard Border[3,4], Noah Zaitlen[3,5]

[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA [2] Henry Samueli School of Engineering, UCLA
[3] Department of Neurology, David Geffen School of Medicine, UCLA [4] Department of Computer Science, UCLA
[5] Department of Human Genetics, UCLA

Abstract

Many scientific hypotheses can be formulated in terms of specific parameters in the context of a generative model. However, performing inference about these parameters in large datasets is often a challenging task, frequently requiring bespoke software implementations for every question of interest. This means that much of researchers' time is spent developing methods for inference rather than addressing questions of direct scientific consequence. In this project, we introduce a novel approach to streamline hypothesis interrogation using NumPyro, a Probabilistic Programming Language (PPL), and efficient Bayesian inference techniques. Our method builds on existing techniques, starting with GWAS and Linear Mixed Effects Model and can be extended to non-infinitesimal and bivariate models, which are particularly relevant for high-dimensional genotype data. By employing PPLs and Hamiltonian Monte Carlo (HMC), we hope to demonstrate that researchers can greatly reduce the time and effort involved in methods development. This approach holds the potential to accelerate scientific discovery, allowing researchers to focus more on hypothesis generation and exploration.

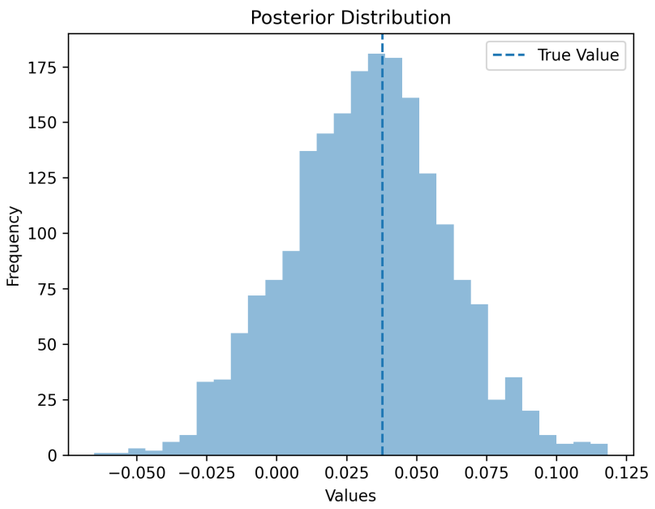
Background

Genotypic Data in Statistical Genetics Research:

- Often characterized by high dimensionality
- Number of observed data points (i.e., genetic markers) is much smaller than the number of parameters.
- Exact inference on such data is extremely challenging.
- Many models are simple extensions of each other. Custom models and software implementations lead to limited interoperability and duplicated efforts.
- Need for a more efficient and unified approach to inference.

PPLs and Bayesian Inference

Probabilistic Programming Languages (PPLs) offer a unified approach to modelling and inference where users can define their models and utilize inference techniques, such as MCMC/HMC, across different models, which can foster interoperability, code reusability, and collaboration.



Bayesian inference allows us to update our beliefs about an unknown parameter by combining prior knowledge with new data in a principled way, using Bayes' theorem given by:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

Where:

- $P(\theta|D)$: is the posterior distribution
- $P(D|\theta)$: is the conditional probability of observing the data D given θ
- $P(\theta)$: is the prior probability distribution
- $P(D)$: is the marginal likelihood

Simulation framework

The vector f contains minor allele frequencies (MAFs) for each SNP, drawn from a uniform distribution:

$$f_j \sim \text{Uniform}(0, 0.5)$$

The matrix X is a $N \times M$ matrix, where each entry X_{ij} represents the number of minor alleles for SNP j in individual i, generated using binomial distribution with corresponding MAFs:

$$X_{ij} \sim \text{Binomial}(2, f_j)$$

Finally the phenotype vector Y is calculated as the product of X and the effect vector β, with added error vector, ε, given by:

$$\beta_i \sim \mathcal{N}(0, \sqrt{\frac{\sigma_g}{M}})$$

$$\varepsilon_i \sim \mathcal{N}(0, \sqrt{\sigma_e})$$

$$Y = X\beta + \varepsilon$$

OLS and Bayesian GWAS: Theory

Ordinary Least Squares (OLS) Regression:

- A frequentist approach to linear regression.
- Estimates the coefficients of the linear model by minimizing the sum of squared residuals.
- Assumptions: The errors are independently and identically distributed (IID) with constant variance, and they are normally distributed.
- Objective function to minimize:

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \beta^T \cdot X_i\right)^2$$

Bayesian Regression:

- A probabilistic approach to linear regression. Represents uncertainty in parameter estimates using probability distributions.

$$\text{Y} \sim \mathcal{N}(X\beta, \sigma^2 I)$$

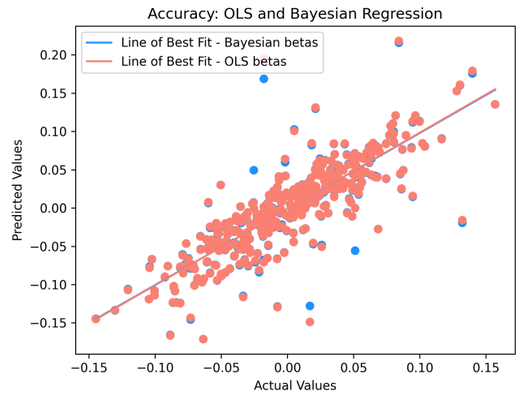
- Assumptions: Similar to OLS, but allows for incorporation of prior knowledge about the parameters.
- Objective: Find the posterior distribution of the parameters given the data and prior knowledge.
- Posterior likelihood of parameters given data (using Bayes' theorem):

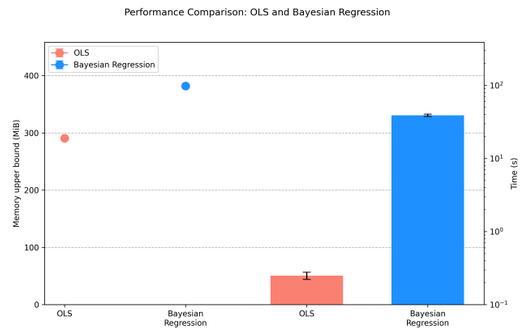
$$P(\beta|X, Y) \propto P(Y|X, \beta) \cdot P(\beta)$$

Hamiltonian Monte Carlo (HMC)

- HMC is a Markov Chain Monte Carlo (MCMC) algorithm that uses Hamiltonian dynamics to efficiently sample from complex probability distributions.
- HMC generates proposals that efficiently explore the target distribution, reducing random walk behavior and leading to faster convergence.

OLS and Bayesian GWAS: Implementation





Conclusion and Future Directions

- Using NumPyro (PPL) we streamlined hypothesis interrogation, saving time on methods development and enabling focused hypothesis exploration.
- Both OLS and Bayesian regression show similar accuracy, but Bayesian regression with HMC has much higher memory footprint and time complexity.
- Therefore future efforts will involve use of JAX and TFP to leverage GPUs to accelerate computations. Sharding mechanism will also be implemented to address the memory constraints of GPUs.