Approach

Overall Research Design

Specific Aim 1: We will use mean field variational methods to provide analytic approximations to the posterior distribution for a Bayesian model with sparsity-inducing priors for polygenic risk scores.

Introduction

Current inference methods for Bayesian polygenic risk score models use sampling-based strategies like Markov chain monte carlo and, thus,
pose significant computational burdens for modern human genetics studies with large sample sizes and high-dimensional measurements. The *objective* of Specific Aim 1 is to use computationally efficient and scalable variational inference methods for posterior inference in polygenic risk score models. Variational inference uses an analytic approximation to the posterior distribution to estimate quantities of interest. To achieve this objective, we will use the polygenic risk score model developed Privé, Arbel, and Vilhjálmsson (2020). However, instead of using the sampling-based inference methods of Privé, Arbel, and Vilhjálmsson (2020), we will apply the variational inference methods of Ray, Szabo, and Clara (2020) and Yang, Pati, and Bhattacharya (2020) for approximate inferences from the posterior distribution. The *rationale* is that our variational inference-based strategy will diminish computing time and memory requirements while maintaining predictive ability of the sampling-based strategy of Privé, Arbel, and Vilhjálmsson (2020). Moreover, with the need to model polygenic risk scores from studies with sample sizes nearing one million subjects, each of which has thousands of phenotype measurements, current sampling-based strategies for posterior inference are inadequate. Variational methods, on the other hand, are computationally efficient and scalable to big data sets. We will compare our method's performance - in terms of predictive ability and computing resource requirements - against that of Privé, Arbel, and Vilhjálmsson (2020). Upon completion of Specific Aim 1, it is our *expectation* that we will have created a computationally scalable and efficient method for constructing polygenic risk scores. Our open source implementation ensures transparency in our research and provides a valuable analytic tool to human genetics researchers.

Research Design

Study Data

We will use imputed genotype and phenotype data from the UK Biobank Study (Bycroft et al. 2018). The UK Biobank study enrolled approximately 500,000 UK adults. Each subject has tens of thousands of phenotypic measurements. The UK Biobank Study shares protected individual-level data with investigators around the world through its data sharing agreement.

For polygenic risk score construction, we will restrict our genetic markers to those available in the Hapmap3 Study, like Privé, Arbel, and Vilhjálmsson (2020) and Ge et al. (2019; Consortium et al. 2010). This will afford us 1,117,493 across the genome. We will restrict the UK Biobank subject set to those used in principal components analysis, who are unrelated and passed quality control filters (Privé, Arbel, and Vilhjálmsson 2020; Bycroft et al. 2018). This will leave us with 362,320 subjects (Privé, Arbel, and Vilhjálmsson 2020).

Statistical modeling

We will use the Bayesian statistical models in LDpred (Vilhjálmsson et al. 2015) and LDpred2 (Privé, Arbel, and Vilhjálmsson 2020). LDpred assumes that the SNP (main) effects follow a distribution that is a mixture of a normal distribution and a point mass at zero. In mathematical notation,

$$\beta_j \sim \begin{cases} \text{Normal}(0, \frac{h^2}{Mp}) & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

where β_j is the SNP effect for SNP with index j, M is the number of SNPs across the genome, p is the proportion of SNPs that are causal for the disease, and h^2 is the SNP heritability of the disease (Vilhjálmsson et al. 2015).

Statistical inference methods

Because Bayesian statistical models almost always have intractible posterior distributions, researchers are often forced to construct elaborate sampling-based strategies that rely on Markov chain monte carlo or the related method called Gibbs sampling. However, recent advances in the mathematics of variational inference have provided alternatives for posterior inference in mathematically intractible Bayesian models. Mean field variational inference for Bayesian statistical models imposes a simplifying assumption and results in a

Assessing predictive performance

Following procedures of Privé, Arbel, and Vilhjálmsson (2020), we will assign 10,000 subjects to a "validation" set of subjects. We will use the validation set to tune model hyperparameters and to estimate genome-wide SNP-SNP correlations. With the remaining 352,320 subjects, we will randomly assign 300,000 for use in our genome-wide association studies, which are a prerequisite for our polygenic risk score calculations (Privé, Arbel, and Vilhjálmsson 2020). The remaining 52,320 subjects that are in neither the GWAS cohort nor the validation set, are assigned to the "test" set (Privé, Arbel, and Vilhjálmsson 2020). We will use the test set to evaluate the performance of our polygenic scores. To compare our proposed method with those from Privé, Arbel, and Vilhjálmsson (2020) and Ge et al. (2019), we will compute the area under the receiver operating characteristic curve for all methods. The receiver operating characteristic curve plots the performance of a classifier, like our polygenic risk scores used to classify subjects as disease cases or controls, across a range of classification thresholds. The area under the receiver operating characteristic curve is one measure of the method's predictive performance. We will follow the detailed procedure described by Privé, Arbel, and Vilhjálmsson (2020) by sampling 10,000 bootstrap replicates of the test set subjects and computing the area under the receiver operating characteristic curve for each bootstrap replicate. With the resulting 10,000 areas, we will report the mean, the 2.5 percentile, and the 97.5 percentile. Privé et al. (2018) have implemented this strategy in the user-friendly R package, `bigstatsr`.

Expected Outcomes, Potential Problems & Alternative Strategies

Our expected outcomes from Specific Aim 1 include a new statistical method for polygenic risk scores. Unlike existing methods, we expect that our method will be computationally efficient and scalable to data sets with millions of subjects and thousands of traits.

One possible problem lies in our use of mean field variational inference instead of other variational inference approaches. Mean field variational inference is equivalent to α-variational inference with $\alpha = 1$ (Yang, Pati, and Bhattacharya 2020). Should our mean field variational inference method underperform in predictive ability, we will pivot to using other values of α in the $(0, 1]$ interval (Yang, Pati, and Bhattacharya 2020). We will then assess performance in terms of predictive ability as a function of α.

Specific Aim 2: We will develop a Bayesian statistical model for polygenic risk scores based on SNP effect estimates and estimates for SNP-SNP interaction effects

Introduction

One possible reason for the modest predictive performance of current polygenic risk scores is their collective failure to account for SNP-SNP interactions in their statistical modeling. This oversight is partially due to the computing resources that are needed to accommodate not only SNP main effects, but the very large number of possible genome-wide SNP-SNP interactions. The *objective* of Specific Aim 2 is to develop polygenic risk score statistical methods that model both SNP main effects *and* SNP-SNP interactions.

While the shear number of SNP-SNP interactions across the genome may make it too computationally costly to use sampling-based inference methods like LDPred2 (Privé, Arbel, and Vilhjálmsson 2020), we anticipate that the gains in efficiency from use of variational inference will make it computationally feasible for us to incorporate modeling of SNP-SNP interactions into our polygenic risk scores.

Upon completion of Specific Aim 2, it is our *expectation* that we will have created a computationally scalable and efficient method for constructing polygenic risk scores that models both SNP main effects and SNP-SNP interactions. We expect that our modeling of SNP-SNP interactions will lead to improved predictive performance of our method relative to current standard methods, such as LDpred2 (Privé, Arbel, and Vilhjálmsson 2020) and PRS-CS (Ge et al. 2019).

Research Design

Study Data

As in Specific Aim 1, we will use data from 362,320 UK Biobank subjects at 1,117,493 genetic markers. We will examine dozens of diseases for every subject and will ensure that we consider traits across the spectrum of SNP heritability values and traits with distinct patterns of genetic architectures.

Statistical modeling

Statistical inference methods

Assessing predictive performance

Like in our strategy for Specific Aim 1, we will measure predictive performance through area under the receiver operating characteristic curve. For Specific Aim 2, we need to quantify the anticipated gains in predictive performance from modeling
SNP-SNP interactions. To do this, we will compare areas under the curve for our polygenic risk scores that omit SNP-SNP interactions (*i.e.*, those from Specific Aim 1) to those that model SNP-SNP interactions for every disease of interest. We expect to see improved performances for the

polygenic risk scores that model SNP-SNP interactions, and we expect that the size of the performance improvement to be greater for those traits with greater SNP heritability values.

Expected Outcomes, Potential Problems & Alternative Strategies

Our expected outcomes from Specific Aim 2 include a new polygenic risk score statistical method that accounts for not only SNP main effects, but also models SNP-SNP interactions. With this more comprehensive modeling of genetic effects, we expect that our method with SNP-SNP interactions will outperform, in terms of predictive ability, current state-of-the-art polygenic risk scores, since none of them model SNP-SNP interactions.

Potential problems include the possibility that our modeling of SNP-SNP interactions doesn't improve predictive performance over that of polygenic risk scores that model only SNP main effects. Should our initial studies on a limited set of diseases from the UK Biobank not provide evidence that our modeling of SNP-SNP interactions improves predictive performance, we will expand our study to examine more diseases in the UK Biobank.

References

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, et al. 2018. "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–9.

Consortium, International HapMap 3 et al. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. 2019. "Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors." *Nature Communications* 10 (1): 1776.

Privé, Florian, Julyan Arbel, and Bjarni J Vilhjálmsson. 2020. "LDpred2: Better, Faster, Stronger." *Bioinformatics* 36 (22-23): 5424–31.

Privé, Florian, Hugues Aschard, Andrey Ziyatdinov, and Michael GB Blum. 2018. "Efficient Analysis of Large-Scale Genome-Wide Data with Two r Packages: Bigstatsr and Bigsnpr." *Bioinformatics* 34 (16): 2781–87.

Ray, Kolyan, Botond Szabo, and Gabriel Clara. 2020. "Spike and Slab Variational Bayes for High Dimensional Logistic Regression." *Advances in Neural Information Processing Systems* 33: 14423–34.

Vilhjálmsson, Bjarni J, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, et al. 2015. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *The American Journal of Human Genetics* 97 (4): 576–92.

Yang, Yun, Debdeep Pati, and Anirban Bhattacharya. 2020. "α-Variational Inference with Statistical Guarantees." *Annals of Statistics*.