Report on:
A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits

GAO, Bingsong
21108544

December 2024

1 Overview of the Paper

1.1 Introduction

The paper by Mingxuan Cai et al. introduces a novel framework for constructing polygenic risk scores (PRSs) that are more accurate across different populations. This is particularly significant because previous PRSs have been less accurate in non-European populations due to genetic differences. The major challenge addressed in this paper is improving the prediction accuracy in under-represented populations by leveraging the genetic correlation across populations.

1.2 Major Challenges

The major challenges of transferable genetic studies arise from three aspects:

- Underrepresentation of Critical SNPs: Genetic studies often overlook single-nucleotide polymorphisms (SNPs) that play crucial roles in non-European populations. This occurs when these SNPs are either absent or exhibit very low allele frequencies among individuals of European descent. Consequently, the genetic diversity and specific risk factors pertinent to non-European populations may not be adequately captured in genome-wide association studies (GWASs).
- Variability in SNP Effect Sizes: There is considerable variability in the impact of the same SNP on a particular phenotype across different populations. This variability underscores the limitations of directly applying GWAS findings and the resultant PRSs to populations other than those in which they were discovered. The differing effect sizes can diminish the predictive power of PRSs when used in non-European populations.

- Population-Specific Linkage Disequilibrium (LD) Patterns: The patterns of linkage disequilibrium, which describe the non-random association of alleles at different loci, vary significantly among populations. These variations can lead to biases when extrapolating PRSs developed from one population to another, as the genetic correlations that influence risk prediction may not be consistent across different ancestral groups.

1.3 How the Proposed Method Addresses the Challenges

The authors propose a cross-population analysis framework, XPA (individual-level) and XPASS (summary-level), which incorporates trans-ancestry genetic correlation to improve risk prediction in non-European populations. The framework can also integrate population-specific effects to further refine PRS construction. Innovations in data structure and algorithm design allow for substantial savings in computational time and memory usage, making the framework efficient for handling biobank-scale data.

2 Simulations to Illustrate the Key Result

I have implemented the XPA in R, using simulated data. The simulations demonstrate the XPA can improve the accuracy of prediction by using data from auxiliary populations if the genetic correlation is non-zero.

Due to the limited computational power of my personal computer, set the number of target populations $n_1 = 100$. Set $p = 10$, $c_1 = 2$, $c_2 = 4$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\sigma_\epsilon = 0.5$, $\sigma_\xi = 1$, $\omega_1 = (1, 1)^T$, $\omega_2 = (2, 2, 2, 2)^T$. Let the number of target populations n_2 be $100, 200, 300, 400, 500, 600, 700$. Simulate every entry of Z_1 and $Z_2 \sim N(0, 1)$, every entry of G_1 and $G_2 \sim Bernoulli(\frac{1}{2})$. Simulate $y_1 = Z_1\omega_1 + X_1\beta_1 + \epsilon$, $y_2 = Z_2\omega_2 + X_2\beta_2 + \xi$. Then I implemented XPA to compute the averaged prediction R^2 from 10 replications. The results are shown in Figure 1.

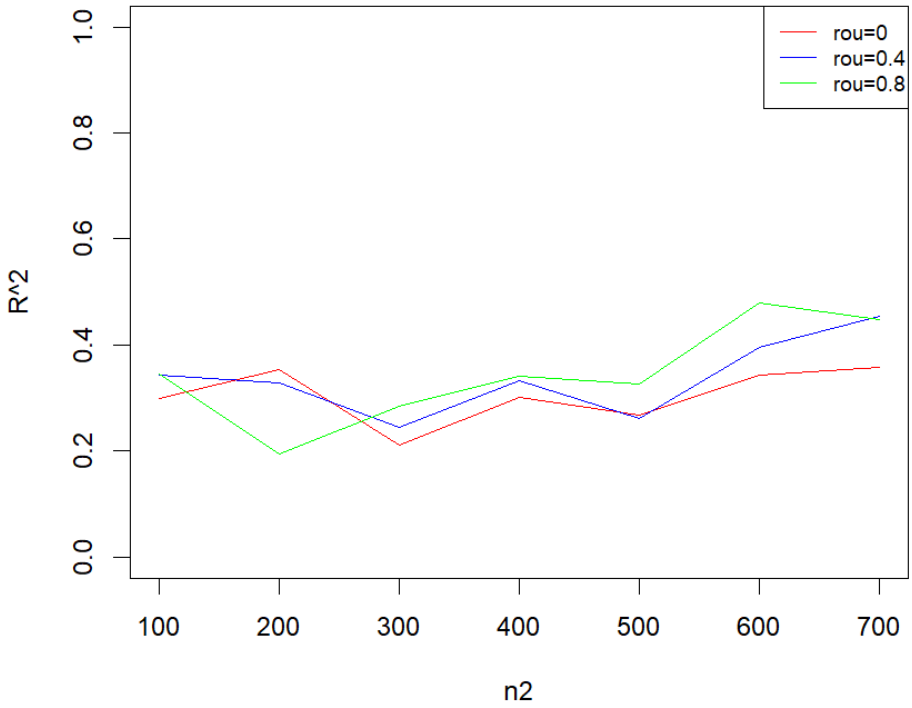

Figure 1: Comparison of prediction accuracy for different genetic correlations.

When the genetic correlation $\rho = 0$, as n_2 increases, R^2 does not change significantly. When $\rho = 0.4$, R^2 becomes larger slightly. When $\rho = 0.8$, R^2 increases more significantly.

In conclusion, auxiliary populations can help improve the prediction accuracy, especially when the genetic correlation is large. Auxiliary populations with larger size can be more helpful.