Statistica Sinica **33** (2023), 1-22
doi:https://doi.org/10.5705/ss.202021.0060

HIGH-DIMENSIONAL ASYMPTOTIC BEHAVIOR OF INFERENCE BASED ON GWAS SUMMARY STATISTIC

Jiming Jiang[1], Wei Jiang[2], Debashis Paul[1], Yiliang Zhang[2] and Hongyu Zhao[2]

[1] *University of California, Davis and* [2] *Yale University*

Abstract: We study the high-dimensional asymptotic behavior of inferences based on summary statistics that are widely used in genome-wide association studies (GWAS) under model misspecification. The high dimensionality is in the sense that the number of single-nucleotide polymorphisms (SNPs) under consideration may be much larger than the sample size. The model misspecification is in the sense that the number of causal SNPs may be much smaller than the total number of SNPs under consideration. Specifically, we establish two parameters of genetic interest, namely, the consistency and asymptotic normality of the estimators of the heritability and genetic covariance. Our theoretical results are supported by the findings of empirical studies involving simulated and real data.

Key words and phrases: Asymptotic normality, Bernoulli, consistency, genetic covariance, heritability, martingale, model misspecification, random matrix theory.

1. Introduction

Over the past 15 years, genome-wide association studies (GWAS) have identified tens of thousands of single-nucleotide polymorphisms (SNPs) associated with complex human traits and diseases (Buniello et al. (2019)). In addition to the success in finding risk loci, estimations of heritability and genetic covariance based on collected GWAS data also provide insights into the genetic basis of complex traits/diseases (Tenesa and Haley (2013); van Rheenen et al. (2019)). Heritability is the proportion of phenotypic variance due to genetic effects, and genetic covariance is the covariance of genetic effects contributing to two phenotypes. Methods based on the linear mixed model (LMM) and the restricted maximum likelihood (REML) algorithm have been developed to estimate these two quantities of significant genetic interest (Yang et al. (2010); Lee et al. (2012)). Compared with traditional family-based approaches for estimating these two quantities, these methods do not need to collect related samples and can use large GWAS samples for estimation. Moreover, they do not require the

Corresponding author: Jiming Jiang, Department of Statistics, University of California, Davis, CA 95616, USA. E-mail: jimjiang@ucdavis.edu.