Statistica Sinica Preprint No: SS-2021-0386	
Title	Statistical Inference for Genetic Relatedness Based on High-Dimensional Logistic Regression
Manuscript ID	SS-2021-0386
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0386
Complete List of Authors	Rong Ma, Zijian Guo, T. Tony Cai and Hongzhe Li
Corresponding Authors	Hongzhe Li
E-mails	hongzhe@upenn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

STATISTICAL INFERENCE FOR GENETIC RELATEDNESS BASED ON HIGH-DIMENSIONAL LOGISTIC REGRESSION

Rong Ma[1], Zijian Guo[2], T. Tony Cai[3] and Hongzhe Li[4]

Department of Statistics[1]

Stanford University

Department of Statistics[2]

Rutgers University

Department of Statistics and Data Science[3]

Department of Biostatistics, Epidemiology and Informatics[4]

University of Pennsylvania

Abstract: This paper studies the problem of statistical inference for genetic relatedness between binary traits based on individual-level genome-wide association data. Specifically, under the high-dimensional logistic regression models, we define parameters characterizing the cross-trait genetic correlation, the genetic covariance and the trait-specific genetic variance. A novel weighted debiasing method is developed for the logistic Lasso estimator and computationally efficient debiased estimators are proposed. The rates of convergence for these es-

timators are studied and their asymptotic normality is established under mild conditions. Moreover, we construct confidence intervals and statistical tests for these parameters, and provide theoretical justifications for the methods, including the coverage probability and expected length of the confidence intervals, as well as the size and power of the proposed tests. Numerical studies are conducted under both model generated data and simulated genetic data to show the superiority of the proposed methods. By analyzing a real data set on autoimmune diseases, we demonstrate its ability to obtain novel insights about the shared genetic architecture between ten pediatric autoimmune diseases.

Key words and phrases: Confidence interval; debiasing methods; functional estimation; genetic correlation; hypothesis testing.

1. Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants or single nucleotide polymorphisms (SNPs) that are associated with various complex phenotypes. Among them, many variants were found to be associated with multiple complex traits, reflecting the pleiotropic action of genes or correlation between causal loci in two traits. Understanding the shared genetic architecture among different traits can potentially lead to further insights into the biological etiology of diseases and inform therapeutic interventions (Van Rheenen et al., 2019).

Various definitions of genetic relatedness or correlation have been proposed in different contexts to characterize quantitatively the shared genetic associations between complex traits based on GWAS data. Understanding of genetic relatedness between complex traits help identify new trait-associated variants (Turley et al., 2018), improve genetic risk prediction (Maier et al., 2015) and assist inference on causality (O'Connor and Price, 2018). Comparing to the traditional approaches from family studies, where measurements of both traits are required for the same individuals, methods based on GWAS enjoy the advantages of increased sample sizes and reduced risk of confounding or ascertainment biases, and have greater potential for large-scale analysis involving multiple traits (Zhang et al., 2020).

Bivariate linear mixed-effects models have been widely applied to estimate the genetic covariance and genetic correlation between two traits from individual-level GWAS data (Lee et al., 2011, 2012; Vattikuti et al., 2012; Lee et al., 2013). These models decompose the phenotypic variance into genetic and residual variance components, and define the genetic correlation to be the correlation between the two trait-specific random generic effects. However, the mixed-effect model approach requires knowledge about the genetic relationship matrix, which is commonly approximated by the genetic relationship across the set of all genotyped variants (Yang et al.,

2010). More recently, computationally efficient methods based on the cross-trait linkage disequilibrium (LD) score regression were developed (Bulik-Sullivan et al., 2015; Ning et al., 2020) to estimate genetic correlation using GWAS summary statistics over a large set of SNPs. This approach relies on the classical asymptotics that does not take into account the high-dimensionality of the SNPs compared to the sample sizes, and might lead to inaccurate inference results (Zhao and Zhu, 2019a). Some other approaches such as Shi et al. (2017), Lu et al. (2017) and Guo et al. (2021) aim to explore differences in local genetic correlations between traits through genome partitioning based on genomic annotations. Weissbrod et al. (2018) noticed that many of the existing methods are primarily geared toward quantitative traits, and direct application of these methods to data sets with binary outcomes may suffer from reduced statistical power; they proposed a mixed effects model to estimate the genetic correlation between binary traits.

In this study, we take a high-dimensional regression approach with fixed genetic effects for identifying trait-associated genetic variants and quantifying the genetic relatedness between two traits. An important advantage of multiple regression over the simple univariate regression is its potential of identifying more trait-associated variants (Wu et al., 2009). Existing studies of heritability or co-heritability within the high-dimensional regression

framework include, for example, Bonnet et al. (2015); Janson et al. (2017); Verzelen and Gassiat (2018); Guo et al. (2019); Zhao and Zhu (2019a,b); Guo et al. (2021). Under the linear regression model, Guo et al. (2019) proposed bias-corrected estimators for the genetic covariance and correlation parameters based on individual-level GWAS data and Zhao and Zhu (2019a) proposed consistent estimators for polygenic risk score and genetic correlation based on GWAS summary statistics. However, these papers focus on the genetic relatedness between continuous traits, and do not provide inference procedures such as statistical tests.

This paper aims to address the following two questions concerning binary traits. How to define and study the genetic relatedness between two binary traits under the high-dimensional regression framework? How to perform valid statistical inference such as testing hypothesis or constructing confidence intervals for the genetic relatedness parameters? We address these questions in a principled way with rigorous statistical justifications.

To that end, for a pair of binary traits $(y, w) \in \{0, 1\}^2$, we consider the following high-dimensional logistic regression models

$$y|X \sim \text{Bernoulli}(\pi_y(X)), \qquad \log\left\{\frac{\pi_y(X)}{1-\pi_y(X)}\right\} = \alpha + X^\top \beta, \tag{1.1}$$

$$w|X \sim \text{Bernoulli}(\pi_w(X)), \qquad \log\left\{\frac{\pi_w(X)}{1-\pi_w(X)}\right\} = \zeta + X^\top \gamma, \tag{1.2}$$

where $\pi_y(X) = P(y = 1|X)$, $\pi_w(X) = P(w = 1|X)$, $X \in \mathbb{R}^p$ is a random vector of p genetic variants with population covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, $\beta, \gamma \in \mathbb{R}^p$ are the corresponding trait-specific regression coefficients, which are assumed to be sparse vectors, and $\alpha, \zeta \in \mathbb{R}$ are the trait-specific intercepts. The genetic covariance between the two traits is defined as the covariance between the log-odds ratios associated to the two traits, i.e., genetic covariance$(y, w) = \text{Cov}\left(\log\left\{\frac{\pi_y(X)}{1-\pi_y(X)}\right\}, \log\left\{\frac{\pi_w(X)}{1-\pi_w(X)}\right\}\right)$, which, by definition, admits the following expressions $\text{Cov}\left(\log\left\{\frac{\pi_y(X)}{1-\pi_y(X)}\right\}, \log\left\{\frac{\pi_w(X)}{1-\pi_w(X)}\right\}\right) = \text{Cov}(X^\top\beta, X^\top\gamma) = \beta^\top\Sigma\gamma$. Similarly, we define the genetic variance of the binary trait y as the variance of its associated log-odds, i.e., genetic variance$(y) = \text{Var}\left(\log\left\{\frac{\pi_y(X)}{1-\pi_y(X)}\right\}\right)$, which satisfies $\text{Var}\left(\log\left\{\frac{\pi_y(X)}{1-\pi_y(X)}\right\}\right) = \text{Var}(X^\top\beta) = \beta^\top\Sigma\beta$, and define the genetic variance of the trait w as $\text{Var}\left(\log\left\{\frac{\pi_w(X)}{1-\pi_w(X)}\right\}\right) = \text{Var}(X^\top\gamma) = \gamma^\top\Sigma\gamma$. Whenever both the genetic variances of y and w are nonzero, we can define the genetic correlation $R(y, w)$ between the two traits as the correlation between the associated log-odds ratios, that is, $\text{Corr}\left(\log\left\{\frac{\pi_y(X)}{1-\pi_y(X)}\right\}, \log\left\{\frac{\pi_w(X)}{1-\pi_w(X)}\right\}\right) = \frac{\beta^\top\Sigma\gamma}{\sqrt{\beta^\top\Sigma\beta\gamma^t op\Sigma\gamma}}$, and set $R(y, w) = 0$ whenever $\beta^\top\Sigma\beta \cdot \gamma^\top\Sigma\gamma = 0$.

The concept of covariance or correlation between two log-odds ratios is both statistically and empirically meaningful, and has been adopted by Wei and Higgins (2013) to account for the correlated outcomes in meta-

analysis, and by Bagos (2012) when the data are presented in the form of contingency tables. In our context, as parameters or functionals quantifying the conditional co-occurrence risk of two traits, the genetic covariance and correlation defined above characterize the shared effect size of the genetic variants by taking into account the true covariance structure of the variants.

This paper studies the problem of statistical inference for these genetic relatedness functionals based on individual-level GWAS data with binary outcomes. By carefully analyzing the logistic Lasso estimator, we develop a novel weighted debiasing method and propose computationally efficient debiased estimators for these functionals. We further study their rates of convergence and obtain their asymptotic normality under mild theoretical conditions. Moreover, confidence intervals and statistical tests for these functionals are constructed. We provide theoretical justifications for the methods, including the coverage probability and expected length of the confidence intervals, as well as the size and power of the proposed tests. Our results provide a rigorous statistical inference framework for studying the genetic relatedness between binary traits.

Throughout, for a symmetric matrix $A \in \mathbb{R}^{p\times p}$, $\lambda_i(A)$ stands for its i-th largest eigenvalue and $\lambda_{\max}(A) = \lambda_1(A)$, $\lambda_{\min}(A) = \lambda_p(A)$. For a smooth function $f(x)$ defined on $\mathbb{R}$, we denote $\dot{f}(x) = df(x)/dx$ and $\ddot{f}(x) =$

$d^2 f(x)/dx^2$. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$, $a_n \ll b_n$ or $b_n \gg a_n$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n)$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \leq C b_n$ for all n. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2. Estimation of Genetic Relatedness

2.1 Genetic Relatedness under Various Settings of Data Availability

We consider two types of data collection scenarios that are commonly adopted for studying genetic relatedness between two traits based on individual-level GWAS data. Data sets obtained from these two scenarios are widely available in current genetic research. In the first scenario, measurements of two traits along with the subject genotypes are obtained from different groups of unrelated individuals. In other words, there are two independent data sets, each containing measurements of a single trait and genotypes for a group of unrelated individuals. This scenario arise commonly when researchers attempt to conduct a cross-trait analysis based on multiple independent GWAS data. In the second scenario, measurements of multiple traits of interest along with the subject genotypes may be obtained from a same group of unrelated individuals. This type of data set is also widely

available by virtue of many large-scale studies such as UK Biobank (Sudlow et al., 2015). The above two scenarios are formally defined as follows.

Scenario (I): Data from independent samples. The observations are $\{(y_i, X_{i\cdot})\}_{i=1}^{n_1}$ and $\{(w_i, Z_{i\cdot})\}_{i=1}^{n_2}$, where $X_{i\cdot}$ and $Z_{i\cdot}$ are drawn independently from some probability measure P_θ on $\mathbb{R}^p$ with covariance matrix Σ, and y_i and w_i are generated based on (1.1) and (1.2), respectively.

Scenario (II): Data from overlapped samples. The observations are $\{(y_i, X_{i\cdot})\}_{i=1}^{n_1}$ and $\{(w_i, Z_{i\cdot})\}_{i=1}^{n_2}$, where $Z_{i\cdot} = X_{i\cdot}$ for $i \in \{1, 2, ..., m\}$, $1 \leq m \leq \min\{n_1, n_2\}$. The samples in $\{Z_{i\cdot}\}_{i=1}^{m}$, $\{X_{i\cdot}\}_{i=m+1}^{n_1}$ and $\{Z_{i\cdot}\}_{i=m+1}^{n_2}$ are drawn independently from some probability measure P_θ on $\mathbb{R}^p$ with covariance matrix Σ, and y_i and w_i are generated from (1.1) and (1.2), respectively.

Note that Scenario (I) corresponds to Scenario (II) with $m = 0$. In what follows, we introduce our main results by focusing on Scenario (I) to avoid unnecessary complications in the notation. The discussions under Scenario II are delayed to Section S5 of the Supplement (Ma et al., 2021) as our methods and results in this case are very similar.

2.2 Weighted Bias Correction and the Proposed Estimators

Estimation of the genetic correlation R can be reduced to estimating the genetic covariance functional $\beta^\top\Sigma\gamma$ and the genetic variance functionals $\beta^\top\Sigma\beta$ and $\gamma^\top\Sigma\gamma$, respectively. The novel bias correction method developed here will lead to nearly unbiased estimators of these functionals of interest, and the construction of which can be summarized as the following two-step procedure. In the first step, an initial plug-in estimator of the functional is obtained based on the pooled sample covariance matrix $\widehat{\Sigma} = \frac{1}{n_1+n_2}\big[\sum_{i=1}^{n_1} X_{i\cdot}X_{i\cdot}^\top + \sum_{i=1}^{n_2} Z_{i\cdot}Z_{i\cdot}^\top\big]$, and the logistic Lasso estimators

$$(\widehat{\alpha},\widehat{\beta}) = \arg\min_{\alpha,\beta}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\Big\{-y_i(\alpha+\beta^\top X_{i\cdot}) + \log(1+e^{\alpha+\beta^\top X_{i\cdot}})\Big\} + \lambda(\|\beta\|_1 + |\alpha|)\right],$$
$$(\widehat{\zeta},\widehat{\gamma}) = \arg\min_{\zeta,\gamma}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\Big\{-w_i(\zeta+\gamma^\top Z_{i\cdot}) + \log(1+e^{\zeta+\gamma^\top Z_{i\cdot}})\Big\} + \lambda(\|\gamma\|_1 + |\zeta|)\right], \tag{2.1}$$

with $\lambda = C\sqrt{\log p/n}$ for some constant $C > 0$. In the second step, the final estimator is obtained by modifying the initial estimator with a carefully designed bias correction term.

We begin with genetic covariance functional $\beta^\top\Sigma\gamma$. With the logistic Lasso estimators (2.1) and $\widehat{\Sigma}$, the corresponding plug-in estimator is defined as $\widehat{\beta}^\top\widehat{\Sigma}\widehat{\gamma}$, whose error can be decomposed as $\widehat{\beta}^\top\widehat{\Sigma}\widehat{\gamma} - \beta^\top\Sigma\gamma = \widehat{\gamma}^\top\Sigma(\widehat{\beta}-\beta) + \widehat{\beta}^\top\Sigma(\widehat{\gamma}-\gamma) - (\widehat{\beta}-\beta)^\top\Sigma(\widehat{\gamma}-\gamma) + \widehat{\beta}^\top(\widehat{\Sigma}-\Sigma)\widehat{\gamma}$. It turns out that the term

$\widehat{\beta}^\top(\widehat{\Sigma}-\Sigma)\widehat{\gamma}$ only contributes to the variance of the plug-in estimator, the terms $\widehat{\gamma}^\top\Sigma(\widehat{\beta}-\beta)$ and $\widehat{\beta}^\top\Sigma(\widehat{\gamma}-\gamma)$ contribute to the leading order bias of the plug-in estimator, whereas the contribution from $(\widehat{\beta}-\beta)^\top\Sigma(\widehat{\gamma}-\gamma)$ is negligible. Therefore, the bias of the plug-in estimator can be further reduced by estimating $\widehat{\gamma}^\top\Sigma(\widehat{\beta}-\beta)$ and $\widehat{\beta}^\top\Sigma(\widehat{\gamma}-\gamma)$ directly. To accomplish this, set $h(u)=\frac{e^u}{1+e^u}$, then by Taylor's expansion, $h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta})-h(\alpha+X_{i\cdot}^\top\beta)=\frac{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}X_{i\cdot}^\top(\widehat{\beta}-\beta)}{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}+\frac{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}(\widehat{\alpha}-\alpha)}{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}+\Delta_i$, where $\Delta_i=\ddot{h}[X_{i\cdot}'^\top\{t\beta'+(1-t)\widehat{\beta}'\}]\{X_{i\cdot}'^\top(\widehat{\beta}'-\beta')\}^2$ for some $t\in(0,1)$, $\beta'=(\alpha,\beta^\top)^\top$, $\widehat{\beta}'=(\widehat{\alpha},\widehat{\beta}^\top)^\top$ and $X_{i\cdot}'=(1,X_{i\cdot}^\top)^\top$. Furthermore, if we define $\epsilon_i=y_i-h(\alpha+X_{i\cdot}^\top\beta)$,

$$
\begin{aligned}
&\{h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta})-y_i\}X_{i\cdot}\\
&=\left\{\frac{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}X_{i\cdot}^\top(\widehat{\beta}-\beta)+\frac{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}(\widehat{\alpha}-\alpha)+\Delta_i-\epsilon_i\right\}X_{i\cdot}\\
&=\frac{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}X_{i\cdot}X_{i\cdot}^\top(\widehat{\beta}-\beta)+(\Delta_i-\epsilon_i)X_{i\cdot}+\frac{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}(\widehat{\alpha}-\alpha)X_{i\cdot}.
\end{aligned}
$$

In order to construct a good estimator of $\Sigma(\widehat{\beta}-\beta)$, we rescale each item $\{h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta})-y_i\}X_{i\cdot}$ by a sample-specific weight $\frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}$ so that

$$
\begin{aligned}
&\sum_{i=1}^{n_1}\frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}\{h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta})-y_i\}X_{i\cdot}\\
&=\left(\sum_{i=1}^{n_1}X_{i\cdot}X_{i\cdot}^\top\right)(\widehat{\beta}-\beta)+\sum_{i=1}^{n_1}\frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}(\Delta_i-\epsilon_i)X_{i\cdot}+(\widehat{\alpha}-\alpha)\sum_{i=1}^{n_1}X_{i\cdot}.
\end{aligned}
$$

Consequently, as long as the last two terms in the above equation are negligible comparing to the leading term $\left(\sum_{i=1}^{n_1}X_{i\cdot}X_{i\cdot}^\top\right)(\widehat{\beta}-\beta)$, an estimator

of $\widehat{\gamma}^\top\Sigma(\widehat{\beta}-\beta)$ can be constructed as

$$\widehat{\gamma}^\top\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}\{h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta})-y_i\}X_{i\cdot}. \tag{2.2}$$

Similarly, we can estimate the error term $\widehat{\beta}^\top\Sigma(\widehat{\gamma}-\gamma)$ by

$$\widehat{\beta}^\top\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{(1+e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}})^2}{e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}}}\{h(\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma})-w_i\}Z_{i\cdot}. \tag{2.3}$$

As a result, in light of the error decomposition, a bias-corrected estimator for $\beta^\top\Sigma\gamma$ is defined as

$$\begin{aligned}\widehat{\beta^\top\Sigma\gamma} = \widehat{\beta}^\top\widehat{\Sigma}\widehat{\gamma} &- \widehat{\gamma}^\top\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}\{h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta})-y_i\}X_{i\cdot}\\ &- \widehat{\beta}^\top\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{(1+e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}})^2}{e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}}}\{h(\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma})-w_i\}Z_{i\cdot}.\end{aligned} \tag{2.4}$$

The above estimator modifies the simple plug-in estimator by adding a carefully constructed bias-correction term accounting for the leading order bias of the plug-in estimator. The bias-correction terms (2.2) and (2.3) are weighted averages, where the weights, originated from the nonlinearity of the link function, reflect each sample's contribution to the overall bias.

In the same vein of our construction of the estimator $\widehat{\beta^\top\Sigma\gamma}$, bias-corrected estimators for the genetic variances can be defined similarly as

$$\widehat{\beta^\top\Sigma\beta} = \widehat{\beta}^\top\widehat{\Sigma}\widehat{\beta} - 2\widehat{\beta}^\top\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}\{h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta})-y_i\}X_{i\cdot}, \tag{2.5}$$

$$\widehat{\gamma^\top\Sigma\gamma} = \widehat{\gamma}^\top\widehat{\Sigma}\widehat{\gamma} - 2\widehat{\gamma}^\top\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{(1+e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}})^2}{e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}}}\{h(\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma})-w_i\}Z_{i\cdot} \tag{2.6}$$

Based on the above genetic variance and covariance estimators, a natural estimator of the genetic correlation is $\bar{R} = \frac{\widehat{\beta^\top \Sigma \gamma}}{\sqrt{\widehat{\beta^\top \Sigma \beta}\widehat{\gamma^\top \Sigma \gamma}}}$. Taking into account the actual range of R, we propose its final estimator as

$$\widehat{R} = \begin{cases} \bar{R}, & \text{if } (\widehat{\beta^\top \Sigma \gamma})^2 < \widehat{\beta^\top \Sigma \beta}\widehat{\gamma^\top \Sigma \gamma} \\ 0, & \text{if } \widehat{\beta^\top \Sigma \beta}\widehat{\gamma^\top \Sigma \gamma} = 0 \\ \text{sign}(\bar{R}), & \text{otherwise} \end{cases} . \tag{2.7}$$

Compared to the existing methods for constructing debiased estimators in high-dimensional regression (Zhang and Zhang, 2014; Javanmard and Montanari, 2014a,b; van de Geer et al., 2014; Cai and Guo, 2017; Guo et al., 2019; Ma et al., 2020; Cai and Guo, 2020; Cai et al., 2021; Guo et al., 2021), our proposed method has two distinct advantages. Firstly, the proposed estimators can be directly obtained from their explicit expressions as in (2.4) to (2.7), which only rely on the initial logistic Lasso estimator, and simple plug-in procedures. Its main computational task is to solve for the initial Lasso estimator, which can be efficiently done with a standard tuning process (Section 5), and therefore is more scalable to the large data sets in genetic studies. This is very different from the existing methods, which, in addition to the initial estimator, involve solving other high-dimensional optimization problems for bias correction, which are computationally challenging, time-consuming, and subject to difficult

tuning processes. Secondly, with our carefully constructed weighted bias-correction method, many commonly used but stringent technical conditions can be avoided. This significantly expands the range of applicability of our proposed methods; see also the discussions after Theorems 1 and 5.

3. Confidence Intervals and Statistical Tests

As an important consequence, it can be shown that each of the above proposed estimators is asymptotically normally distributed. This can be used to construct confidence intervals and statistical tests for the functionals.

Specifically, it can be shown that the estimator $\widehat{\beta^\top\Sigma\gamma}$ has variance $v^2 = \frac{n_1+n_2}{n_1}E\{\eta_i^{(X)}(\widehat{\gamma}^\top X_{i\cdot})^2\}+\frac{n_1+n_2}{n_2}E\{\eta_i^{(Z)}(\widehat{\beta}^\top Z_{i\cdot})^2\}+E\{\widehat{\beta}^\top(X_{i\cdot}X_{i\cdot}^\top-\Sigma)\widehat{\gamma}\}^2$, where $\eta_i^{(X)} = \frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^4e^{\alpha+X_{i\cdot}^\top\beta}}{(1+e^{\alpha+X_{i\cdot}^\top\beta})^2e^{2\widehat{\alpha}+2X_{i\cdot}^\top\widehat{\beta}}}$ and $\eta_i^{(Z)} = \frac{(1+e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}})^4e^{\zeta+Z_{i\cdot}^\top\gamma}}{(1+e^{\zeta+Z_{i\cdot}^\top\gamma})^2e^{2\widehat{\zeta}+2Z_{i\cdot}^\top\widehat{\gamma}}}$. Intuitively, the parameters β and γ in the above expressions can be estimated by their initial Lasso estimators, so that a moment estimator of the asymptotic variance can be defined as $\widehat{v}^2 = \frac{n_1+n_2}{n_1^2}\sum_{i=1}^{n_1}\frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}(\widehat{\gamma}^\top X_{i\cdot})^2 + \frac{n_1+n_2}{n_2^2}\sum_{i=1}^{n_2}\frac{(1+e^{\widehat{\eta}+Z_{i\cdot}^\top\widehat{\gamma}})^2}{e^{\widehat{\eta}+Z_{i\cdot}^\top\widehat{\gamma}}}(\widehat{\beta}^\top Z_{i\cdot})^2+\frac{1}{n_1+n_2}\big\{\sum_{i=1}^{n_1}(\widehat{\beta}X_{i\cdot}X_{i\cdot}^\top\widehat{\gamma}-\widehat{\beta}\widehat{\Sigma}\widehat{\gamma})^2+\sum_{i=1}^{n_2}(\widehat{\beta}Z_{i\cdot}Z_{i\cdot}^\top\widehat{\gamma}-\widehat{\beta}\widehat{\Sigma}\widehat{\gamma})^2\big\}$. Hence, an $(1-\alpha)$-level CI for the genetic covariance is $\mathrm{CI}_\alpha(\beta^\top\Sigma\gamma,\mathcal{D}) = \big[\widehat{\beta^\top\Sigma\gamma}-\widehat{\rho},\widehat{\beta^\top\Sigma\gamma}+\widehat{\rho}\big]$, where $\widehat{\rho} = \frac{z_{\alpha/2}\widehat{v}}{\sqrt{n_1+n_2}}$ and $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the upper $\alpha/2$-quantile of the standard normal distribution. Similarly, the asymptotic variance of the genetic variance estimator $\widehat{\beta^\top\Sigma\beta}$ can be derived

as $v_\beta^2 = \frac{4(n_1+n_2)}{n_1} E\{\eta_i^{(X)}(\widehat{\beta}^\top X_{i\cdot})^2\} + E\{\widehat{\beta}^\top (X_{i\cdot}X_{i\cdot}^\top - \Sigma)\widehat{\beta}\}^2$, which can be estimated by $\widehat{v}_\beta^2 = \frac{4(n_1+n_2)}{n_1^2}\sum_{i=1}^{n_1} \frac{(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2}{e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}}(\widehat{\beta}^\top X_{i\cdot})^2 + \frac{1}{n_1+n_2}\Big\{\sum_{i=1}^{n_1}(\widehat{\beta}X_{i\cdot}X_{i\cdot}^\top\widehat{\beta} - \widehat{\beta}\widehat{\Sigma}\widehat{\beta})^2 + \sum_{i=1}^{n_2}(\widehat{\beta}Z_{i\cdot}Z_{i\cdot}^\top\widehat{\beta} - \widehat{\beta}\widehat{\Sigma}\widehat{\beta})^2\Big\}$. Then, an $(1-\alpha)$-level confidence interval for $\beta^\top\Sigma\beta$ is $\mathrm{CI}_\alpha(\beta^\top\Sigma\beta, \mathcal{D}) = \big[\widehat{\beta^\top\Sigma\beta} - \widehat{\rho}_\beta, \widehat{\beta^\top\Sigma\beta} + \widehat{\rho}_\beta\big]$, where $\widehat{\rho}_\beta = \frac{z_{\alpha/2}\widehat{v}_\beta}{\sqrt{n_1+n_2}}$. The confidence interval $\mathrm{CI}_\alpha(\gamma^\top\Sigma\gamma, \mathcal{D})$ can be obtained by symmetry.

The confidence interval for the genetic correlation R is a direct consequence of the Slutsky's theorem. Specifically, for the estimator $\widehat{R}$ defined in (2.7), whenever $\widehat{\beta^\top\Sigma\beta}\,\widehat{\gamma^\top\Sigma\gamma} \neq 0$, we can estimate its asymptotic variance by $\widehat{v}_R^2 = \frac{\widehat{v}^2}{\widehat{\beta^\top\Sigma\beta}\,\widehat{\gamma^\top\Sigma\gamma}}$, and define the corresponding $(1-\alpha)$-level confidence interval as $\mathrm{CI}_\alpha(R, \mathcal{D}) = \big[\widehat{R} - \widehat{\rho}_R, \widehat{R} + \widehat{\rho}_R\big] \cap [-1, 1]$, where $\widehat{\rho}_R = \frac{z_{\alpha/2}\widehat{v}_R}{\sqrt{n_1+n_2}}$.

Converting the above CIs, we obtain statistical tests for each of the null hypotheses $H_{0,1} : \beta^\top\Sigma\gamma = B_0$,$H_{0,2} : \beta^\top\Sigma\beta = Q_0$, and $H_{0,3} : R = R_0$, for some $B_0 \in \mathbb{R}$, $Q_0 \geq 0$ and $R_0 \in [-1, 1]$. Specifically, we define test statistics $T_1 = \frac{\sqrt{n_1+n_2}(\widehat{\beta^\top\Sigma\gamma} - B_0)}{\widehat{v}}$, $T_2 = \frac{\sqrt{n_1+n_2}(\widehat{\beta^\top\Sigma\beta} - Q_0)}{\widehat{v}_\beta}$ and $T_3 = \frac{\sqrt{n_1+n_2}(\widehat{R} - R_0)}{\widehat{v}_R}$, so that for each $\ell \in \{1, 2, 3\}$, to obtain an α-level test, we reject the null hypothesis $H_{0,\ell}$ whenever $|T_\ell| > z_{\alpha/2}$.

4. Theoretical Properties

4.1 Rates of Convergence and Optimality

The random covariates are characterized by the following conditions.

(A1) For each $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$, $X_{i\cdot}$ and $Z_{j\cdot}$ are centred *i.i.d.* sub-Gaussian random vectors where $\Sigma = E(X_{i\cdot}X_{i\cdot}^\top) \in \mathbb{R}^{p\times p}$ satisfies $M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M$ for some constant $M > 1$.

(A2) There exists a positive constant c_0 such that $E\left(\frac{\beta^\top X_{i\cdot}X_{i\cdot}^\top\gamma}{\beta^\top\Sigma\gamma} - 1\right)^2 > c_0$.

About the regression coefficients, we denote $k = \max\{\|\beta\|_0, \|\gamma\|_0\}$, $U(\beta,\gamma) = \max\{\|\beta\|_2, \|\gamma\|_2\}$ and $L(\beta,\gamma) = \min\{\|\beta\|_2, \|\gamma\|_2\}$. We assume

(A3) $\max\{|\alpha|, |\zeta|\} \leq C$ and $U(\beta,\gamma) \leq C$ for some constant $C > 0$.

Intuitively, assumptions (A1) and (A3) imply that the marginal case probabilities $P(y_i = 1)$ and $P(w_i = 1)$ are balanced, or bounded away from 0 and 1, whereas (A2) ensures the asymptotic variances does not diminish.

For technical reasons, for each trait we split the corresponding samples into halves so that the initial Lasso estimation step and the rest of the steps such as covariance estimation and bias-correction are conducted on independent data sets. Without loss of generality, we assume under Scenario I there are $2(n_1+n_2)$ samples in $\mathcal{D}$, divided into two disjoint subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, each containing n_1+n_2 independent samples, with n_1 samples corresponding to trait y_i and n_2 samples corresponding to trait w_i. The initial Lasso estimators are obtained from $\mathcal{D}_1$, whereas the sample covariance, the bias-correction terms and the asymptotic variance estimators are based on $\mathcal{D}_2$ and the initial Lasso estimators. We emphasize that the sample split-

ting procedure is only used to facilitate the theoretical analysis, and is not needed in practice. We demonstrate this point numerically in Section 5; see also Section 7 for more discussions.

The following theorem concerns the rate of convergence of the bias-corrected estimators $\widehat{\beta^\top\Sigma\gamma}$ and $\widehat{\beta^\top\Sigma\beta}$; the results for $\widehat{\gamma^\top\Sigma\gamma}$ are similar.

Theorem 1 (Rates of Convergence)**.** *Suppose (A1) and (A3) hold,* $n_1 \asymp n_2 \asymp n$ *and* $k \lesssim \frac{n}{\log p \log n}$*. Then, for sufficiently large* (n, p) *and any* $t > 0$*,*

$$|\widehat{\beta^\top\Sigma\gamma} - \beta^\top\Sigma\gamma| \lesssim \frac{tU(\beta,\gamma)}{\sqrt{n}} + \{1 + U(\beta,\gamma)\sqrt{\log n}\}\frac{k\log p}{n}, \tag{4.1}$$

$$|\widehat{\beta^\top\Sigma\beta} - \beta^\top\Sigma\beta| \lesssim \frac{t\|\beta\|_2}{\sqrt{n}} + (1 + \|\beta\|_2\sqrt{\log n})\frac{k\log p}{n}, \tag{4.2}$$

with probability at least $1 - p^{-c} - n^{-c} - t^{-2}$ *for some constant* $c > 0$*.*

In Theorem 1, in addition to the mild sparsity condition, the consistency of the proposed estimators only require the balanced marginal case probabilities through (A1) and (A3), and the general sub-Gaussian design with a regular covariance matrix, which includes many important cases such as Gaussian, bounded, and binary designs, or any combinations of them. It makes the proposed methods widely applicable to various practical settings.

To establish the optimality of the proposed genetic covariance estimator, our next result concerns the minimax lower bound for estimating $\beta^\top\Sigma\gamma$.

To this end, we define the parameter space for $\theta = (\beta, \gamma, \Sigma)$ as

$$\Theta(k, L_n) = \left\{ (\beta, \gamma, \Sigma) : \begin{array}{c} \max\{\|\beta\|_0, \|\gamma\|_0\} \le k, U(\beta, \gamma) \le L_n \\ M^{-1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le M \end{array} \right\}$$

for some constant $M > 1$, and denote $\xi = \beta^\top \Sigma \gamma$.

Theorem 2 (Minimax Lower Bound)**.** *Suppose* X_i *and* $Z_i \overset{i.i.d.}{\sim} N(0, \Sigma)$ *for* $i = 1, ..., n$, *and* $k \lesssim \min\left\{p^\nu, \frac{n}{\log p}\right\}$ *for some* $0 < \nu < 1/2$. *Then*

$$\inf_{\widehat{\xi}} \sup_{\theta \in \Theta(k, L_n)} P_\theta\left(|\widehat{\xi} - \xi| \gtrsim \frac{L_n^2}{\sqrt{n}} + \min\left\{ \frac{L_n}{\sqrt{n}} + k\frac{\log p}{n}, L_n^2 \right\} \right) \ge c \qquad (4.3)$$

for some constant $c > 0$.

By Theorem 1, a uniform upper bound over the parameter space $\Theta(k, L_n)$ can be obtained as $\sup_{\theta \in \Theta(k, L_n)} P_\theta\big(|\widehat{\beta^\top \Sigma \gamma} - \beta^\top \Sigma \gamma| \lesssim \frac{tL_n}{\sqrt{n}} + (1 + L_n\sqrt{\log n})\frac{k \log p}{n}\big) \ge 1 - p^{-c} - n^{-c} - t^{-2}$. Combining this with the lower bound from Theorem 2, we conclude that, for all $k \lesssim \min\{\frac{n}{\log p \log n}, p^\nu\}$ with any $\nu \in (0, 1/2)$, and $\sqrt{\frac{k \log p}{n}} \lesssim L_n \lesssim 1$, our genetic covariance estimator $\widehat{\beta^\top \Sigma \gamma}$ is minimax rate-optimal over $\Theta(k, L_n)$, up to a $\sqrt{\log n}$ factor. In particular, in this case, the exact rate optimality of $\widehat{\beta^\top \Sigma \gamma}$ is guaranteed over the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p \sqrt{\log n}}$, or the weak signal regime $L_n \lesssim (\log n)^{-1/2}$, over which the minimax rate is $\frac{L_n}{\sqrt{n}} + \frac{k \log p}{n}$. Moreover, this suggests that the uncertainty due to the covariance estimation $\widehat{\beta}^\top(\widehat{\Sigma} - \Sigma)\widehat{\gamma}$ in the plug-in estimator is fundamental and may not be removed as for the leading order biases.

Theorem 3 (Rate of Convergence)**.** *Suppose (A1) (A2) and (A3) hold,* $n_1 \asymp n_2 \asymp n$, $k \ll \frac{n}{\log p \log n}$ *and* $L(\beta,\gamma) \gg \sqrt{k \log p/n}$. *Then* $|\widehat{R} - R| \to 0$ *in probability. In particular, for sufficiently large* (n,p) *and any constant* $t > \sqrt{2}$, *with probability at least* $1 - 2t^{-2}$, *it holds that*

$$|\widehat{R} - R| \lesssim \frac{t\{U(\beta,\gamma) + U^2(\beta,\gamma)\}}{L^2(\beta,\gamma)\sqrt{n}} + \frac{1 + U(\beta,\gamma)\sqrt{\log n}}{L^2(\beta,\gamma)} \cdot \frac{k \log p}{n}. \quad (4.4)$$

Comparing to Theorem 1, the consistency of $\widehat{R}$ requires an additional condition (A2) and a lower bound on the minimal effect size. These conditions are necessary to ensure the true genetic variances are bounded away from zero and the genetic correlation is well-defined.

4.2 Theoretical Properties of the Inference Procedures

We establish the asymptotic normality of the proposed bias-corrected estimators and provide theoretical justifications of the CIs and the statistical tests. We start with a theorem that provides a refined analysis of the estimation errors and consequently the asymptotic normality of the estimators.

Theorem 4 (Asymptotic Normality)**.** *Suppose (A1) (A2) (A3) hold,* $n_1 \asymp n_2 \asymp n$, $k \lesssim \frac{n}{\log p \log n}$ *and* $L(\beta,\gamma) \gg \sqrt{k \log p/n}$. *Then*

1. *It holds that* $\widehat{\beta^\top \Sigma \gamma} - \beta^\top \Sigma \gamma = A_n + B_n$, *where* $P\big\{A_n \lesssim \{U(\beta,\gamma)\sqrt{\log n} + 1\}\frac{k \log p}{n}\big\} \geq 1 - p^{-c} - n^{-c}$, *and* $\frac{\sqrt{n_1+n_2}B_n}{v}\big|\mathcal{D}_1 \to_d N(0,1)$ *as* $(n,p) \to \infty$.

Additionally, if $k \ll \frac{U(\beta,\gamma)\sqrt{n}}{\{1+U(\beta,\gamma)\sqrt{\log n}\}\log p}$, *we establish the asymptotic normality* $\frac{\sqrt{n_1+n_2}(\widehat{\beta^\top\Sigma\gamma}-\beta^\top\Sigma\gamma)}{v}\big|\mathcal{D}_1 \to_d N(0,1)$.

2. *It holds that* $\widehat{\beta^\top\Sigma\beta}-\beta^\top\Sigma\beta = A'_n+B'_n$, *where* $P\big\{A'_n \lesssim (\|\beta\|_2\sqrt{\log n}+1)\frac{k\log p}{n}\big\} \geq 1-p^{-c}-n^{-c}$, *and* $\frac{\sqrt{n_1+n_2}B'_n}{v_\beta}\big|\mathcal{D}_1 \to_d N(0,1)$ *as* $(n,p)\to\infty$. *Additionally, if* $k \ll \frac{\|\beta\|_2\sqrt{n}}{[1+\|\beta\|_2\sqrt{\log n}]\log p}$, *we establish the asymptotic normality* $\frac{\sqrt{n_1+n_2}(\widehat{\beta^\top\Sigma\beta}-\beta^\top\Sigma\beta)}{v_\beta}\big|\mathcal{D}_1 \to_d N(0,1)$.

The second part of the theorem applies to the estimator $\widehat{\gamma^\top\Sigma\gamma}$ by symmetry. A direct consequence of Theorems 1 and 4, in combination with Slutsky's theorem, is the following theorem concerning the asymptotic normality of the genetic correlation estimator $\bar{R}$ in Section 2.2.

Theorem 5 (Asymptotic Normality)**.** *Under the conditions of Theorem 4, if* $k \ll \min\{\frac{n}{\log p\log n}, \frac{U(\beta,\gamma)\sqrt{n}}{\{1+U(\beta,\gamma)\sqrt{\log n}\}\log p}\}$, *we have* $\frac{\sqrt{n_1+n_2}(\bar{R}-R)}{v_R}\big|\mathcal{D}_1 \to_d N(0,1)$ *as* $(n,p)\to\infty$.

Some remarks about the technical innovations leading to the above theorems are in order. Firstly, distinct from the existing works on the statistical inference in high-dimensional logistic regression, the proposed methods do not require the commonly assumed but stringent theoretical conditions such as the bounded individual probability condition (van de Geer, 2008; van de Geer et al., 2014; Ning and Liu, 2017; Ma et al., 2020; Guo et al., 2021)

where $P(y_i = 1|X_{i\cdot}) \in (\delta, 1-\delta)$ for all $1 \leq i \leq n$ and some $\delta \in (0, 1/2)$, the sparse inverse population Hessian condition (van de Geer et al., 2014; Belloni et al., 2016; Ning and Liu, 2017; Janková and van de Geer, 2018) or the sparse precision condition (Ma et al., 2020). Secondly, from a practical point of view, the removal of these technical assumptions significantly expands the range of applicability of the proposed methods. For example, as was argued by Cai et al. (2021) and Xia et al. (2020), in practice, the bounded individual probability and the sparse inverse population Hessian conditions are seldom satisfied or can be verified from the data. In contrast, the balanced marginal case probability condition holds easily and can be checked based on the observed outcomes.

Built upon Theorems 4 and 5, theoretical justifications such as the asymptotic coverage probability and the expected length of the proposed CIs $\mathrm{CI}_\alpha(\beta^\top\Sigma\gamma, \mathcal{D})$, $\mathrm{CI}_\alpha(\beta^\top\Sigma\beta, \mathcal{D})$ and $\mathrm{CI}_\alpha(R, \mathcal{D})$ can be obtained.

Theorem 6 (Confidence Intervals)**.** *Under the conditions of Theorem 4, for any constant* $0 < \alpha < 1$*, if* $k \ll \min\{\frac{n}{\log p \log n}, \frac{U(\beta,\gamma)\sqrt{n}}{\{1+U(\beta,\gamma)\sqrt{\log n}\}\log p}\}$*, then*

1. *(Coverage)* $\underline{\lim}_{n,p\to\infty} P_\theta\{\beta^\top\Sigma\gamma \in \mathrm{CI}_\alpha(\beta^\top\Sigma\gamma, \mathcal{D})\} \geq 1-\alpha$, $\underline{\lim}_{n,p\to\infty} P_\theta\{\beta^\top\Sigma\beta \in \mathrm{CI}_\alpha(\beta^\top\Sigma\beta, \mathcal{D})\} \geq 1-\alpha$, *and* $\underline{\lim}_{n,p\to\infty} P_\theta\{R \in \mathrm{CI}_\alpha(R, \mathcal{D})\} \geq 1-\alpha$;

2. *(Length) if we denote* $L\{\mathrm{CI}_\alpha(\cdot, \mathcal{D})\}$ *as the length of* $\mathrm{CI}_\alpha(\cdot, \mathcal{D})$*, then*

with probability at least $1 - p^{-c}$, *we have* $L\{\mathrm{CI}_\alpha(\beta^\top \Sigma \gamma, \mathcal{D})\} \asymp \frac{U(\beta,\gamma)}{\sqrt{n}}$, $L\{\mathrm{CI}_\alpha(\beta^\top \Sigma \beta, \mathcal{D})\} \asymp \frac{\|\beta\|_2}{\sqrt{n}}$. *and* $L\{\mathrm{CI}_\alpha(R, \mathcal{D})\} \asymp \frac{1}{L(\beta,\gamma)\sqrt{n}}$.

This theorem implies that the statistical tests proposed in Section 3 have the following theoretical properties concerning their sizes and powers under certain local alternatives.

Corollary 1 (Hypotheses Testing)**.** *Under the conditions of Theorem 6,*

1. *(Size) for each* $\ell \in \{1, 2, 3\}$, *for any constant* $0 < \alpha < 1$, *under the null hypothesis* $H_{0,\ell}$, *we have* $\overline{\lim}_{n,p\to\infty} P_\theta(|T_\ell| > z_{\alpha/2}) \leq \alpha$;

2. *(Power) for any* $0 < \delta < 1$, *there exists some* $c > 0$ *such that, for any* $|\beta^\top \Sigma \gamma - B_0| \geq cU(\beta, \gamma)n^{-1/2}$, $\underline{\lim}_{n,p\to\infty} P_\theta(|T_1| > z_{\alpha/2}) \geq 1 - \delta$; *for any* $|\beta^\top \Sigma \beta - Q_0| \geq c\|\beta\|_2 n^{-1/2}$, $\underline{\lim}_{n,p\to\infty} P_\theta(|T_2| > z_{\alpha/2}) \geq 1 - \delta$; *and for any* $|R - R_0| \geq cL^{-1}(\beta, \gamma)n^{-1/2}$, $\underline{\lim}_{n,p\to\infty} P_\theta(|T_3| > z_{\alpha/2}) \geq 1 - \delta$.

5. Simulations

5.1 Evaluations with Simulated Genetic Data

To justify our proposed methods for analyzing real genetic data sets, we carried out numerical experiments under the settings where the covariates were simulated genotypes with possible LD structures that resembled those of the human genome, and the inferences were made at a chromosomal basis.

Specifically, focusing on the Scenario I with $n_1 = n_2 = n$, for given choices of p and n, using the `R` package `sim1000G` (Dimitromanolakis et al., 2019), we generated genotypes of $2n$ unrelated individuals containing p SNPs based on the sequencing data over a region (GrCH37: bp 40,900 to bp 2,000,000) on chromosome 9 of 503 European samples from the 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium, 2015), and a comprehensive haplotype map integrated over 1,184 reference individuals (International HapMap 3 Consortium, 2010); see Section S4 of the Supplement for the resulting correlation matrix among the generated SNPs. The true effect sizes for the two binary traits were generated such that for each trait there were 25 associated SNPs with 12 of them shared by both traits. The effect sizes of the associated SNPs were uniformly drawn from $[-1, 1]$. For reasons of practical interest, we mainly focused on the estimation, confidence intervals and hypothesis testing about the genetic correlation parameter. The results about the genetic covariance and variance can be found in Section 5.2 below and Section S4 of the Supplement.

For parameter estimation, in addition to our proposed estimators ("pro"), we also considered (i) the simple plug-in ("plg") estimators $\hat{\beta}^\top\hat{\Sigma}\hat{\gamma}$, $\hat{\beta}^\top\hat{\Sigma}\hat{\beta}$ and $\hat{R}_{plg} = \frac{\hat{\beta}^\top\hat{\Sigma}\hat{\gamma}}{\sqrt{\hat{\beta}^\top\hat{\Sigma}\hat{\beta}\hat{\gamma}^\top\hat{\Sigma}\hat{\gamma}}}$; (ii) the component-wise projected Lasso ("lpj") estimators $\breve{\beta}^\top\hat{\Sigma}\breve{\gamma}$, $\breve{\beta}^\top\hat{\Sigma}\breve{\beta}$ and $\hat{R}_{lpj} = \frac{\breve{\beta}^\top\hat{\Sigma}\breve{\gamma}}{\sqrt{\breve{\beta}^\top\hat{\Sigma}\breve{\beta}\breve{\gamma}^\top\hat{\Sigma}\breve{\gamma}}}$ where each component

of $\breve{\beta}$ and $\breve{\gamma}$ was the debiased Lasso estimator implemented by the function `lasso.proj` in the R package `hdi` under default setting; and (iii) the component-wise projected Ridge ("rpj") estimators $\breve{\beta}^\top\hat{\Sigma}\breve{\gamma}$, $\breve{\beta}^\top\hat{\Sigma}\breve{\beta}$ and $\hat{R}_{rpj} = \frac{\breve{\beta}^\top\hat{\Sigma}\breve{\gamma}}{\sqrt{\breve{\beta}^\top\hat{\Sigma}\breve{\beta}\breve{\gamma}^\top\hat{\Sigma}\breve{\gamma}}}$ where each component of $\breve{\beta}$ and $\breve{\gamma}$ were obtained from the function `ridge.proj` in the R package `hdi` under the default setting. For the proposed method, we used cross-validation to determine the tuning parameter (see Section S4.1 for details). Table 1 contains the empirical estimation errors (square-roots of the empirical mean-squared errors) for the genetic correlation estimators, which demonstrates the superior performance of the proposed method.

For confidence intervals, we compared our proposed CIs ("pro") with an alternative bootstrap CIs. Specifically, the bootstrap CIs are based on the plg estimators calculated from 100 observations sampled from the original data set, so that the final CIs are constructed based on the empirical distributions of 500 bootstrap estimators. Table 2 contains the averaged coverage probabilities and lengths of the proposed and the plg-based bootstrap CIs, denoted as "boot," with 500 rounds of simulation for each setting. Our results suggest the desirable coverage and shorter length of the proposed CIs. Finally, for hypotheses testing, we evaluated the empirical type I errors and the statistical powers of both our proposed tests and the plg-based boot-

strap tests under the setup where the effect sizes were generated with an additional constraint $|\beta^\top \Sigma \gamma| > 3$. Table 3 contains the empirical type I errors and statistical powers of the proposed tests over different settings, each based on 500 rounds of simulations. Our results suggest empirical validity of the proposed test, and its advantage over the bootstrap tests. Although in Tables 2 and 3, the proposed method became a little conservative when n increased from 200 to 400, which was likely due to the limitation of our empirically determined tuning parameter, we still observed greater power for the test and shorter lengths for the CIs with larger n, and in both cases the advantage over the alternative methods. For more simulations under a slightly different setting of association structure, see Section S4.5 of the Supplement (Table S8).

5.2 Evaluation with Model-Generated Data

We consider the high-dimensional setting $p > n$, and set the sparsity level as $k = 25$. For the true regression coefficients, given the support $\mathcal{S}$ such that $|\mathcal{S}| = k$, we generated β_j and γ_j uniformly from $[-1, 1]$ for all $j \in \mathcal{S}$. For the design covariates, we focused on Scenario I, where $n_1 = n_2 = n$ and the covariates were generated from a multivariate Gaussian distribution with covariance matrix as either $\Sigma = \Sigma_B$, where Σ_B is a $p \times p$ blockwise diagonal

Table 1: Estimation errors for the genetic correlation under simulated genetic data with $k = 25$. pro: proposed estimators; plg: simple plug-in estimators; lpj: component-wise projected Lasso estimators; rpj: the component-wise projected Ridge estimators.

p	$n = 200$				300				400			
	pro	plg	lpj	rpj	pro	plg	lpj	rpj	pro	plg	lpj	rpj
700	0.09	0.12	0.15	0.16	0.09	0.11	0.14	0.13	0.08	0.11	0.13	0.12
800	0.08	0.10	0.15	0.14	0.08	0.11	0.15	0.11	0.09	0.11	0.15	0.12
900	0.09	0.13	0.16	0.15	0.11	0.12	0.15	0.13	0.07	0.11	0.14	0.11
1000	0.10	0.12	0.14	0.15	0.09	0.11	0.14	0.12	0.08	0.09	0.14	0.09

Table 2: Coverage and length of the CIs for the genetic correlation under simulated genetic data with $\alpha = 0.05$.

p	$n = 200$				300				400			
	coverage		length		coverage		length		coverage		length	
	pro	boot	pro	boot	pro	boot	pro	boot	pro	boot	pro	boot
700	96.4	82.4	0.30	0.37	97.6	85.8	0.26	0.39	97.0	82.6	0.27	0.41
800	97.0	85.4	0.29	0.37	98.0	82.5	0.27	0.39	98.2	85.2	0.26	0.39
900	96.6	84.2	0.31	0.36	96.8	86.2	0.26	0.38	97.6	84.0	0.25	0.39
1000	97.5	86.0	0.30	0.34	97.6	80.0	0.26	0.36	97.8	84.9	0.26	0.41

Table 3: Type I errors and powers for testing the genetic correlation under simulated genetic data with $\alpha = 0.05$.

p	$n = 200$				300				400			
	type I error		power		type I error		power		type I error		power	
	pro	boot	pro	boot	pro	boot	pro	boot	pro	boot	pro	boot
700	0.04	0.41	0.47	0.72	0.04	0.35	0.63	0.68	0.02	0.34	0.69	0.65
800	0.04	0.42	0.46	0.74	0.03	0.37	0.59	0.71	0.03	0.34	0.70	0.66
900	0.04	0.42	0.45	0.70	0.03	0.35	0.64	0.66	0.02	0.32	0.69	0.73
1000	0.06	0.41	0.42	0.71	0.02	0.36	0.63	0.70	0.02	0.36	0.68	0.70

matrix of 10 identical unit diagonal Toeplitz matrices whose off-diagonal entries descend from 0.3 to 0 (see Section S4.1 of the Supplement for its explicit form), or $\Sigma = \Sigma_E$ where Σ_E is an exchangeable covariance matrix with unit diagonals and off-diagonals being 0.2. The numerical result on each setting was based on 500 rounds of simulations.

For parameter estimation, we evaluated the proposed method and the three alternative methods defined in the previous section. The results, due to space limit, were put in Section S4.2 of the Supplement (Tables S1, S2), which demonstrated the superiority of each of the proposed estimators over the alternatives. Under the same simulation setups, we evaluated and

compared different method for constructing 95% CIs for the parameters. Specifically, we compared our proposed CIs ("pro") with two alternative bootstrap CIs, based on 500 plg estimators or rpj estimators calculated from 100 observations sampled from the original data set. Table 4 contains the averaged coverage probabilities and lengths of the proposed and the plg-based bootstrap CIs ("boot") under the blockwise diagonal covariant matrix. For reason of space, the results under the exchangeable covariance, and the results for the rpj-based bootstrap CIs, whose coverage were in general poorer than the plg-based CIs for $\beta^\top\Sigma\gamma$ and $\beta^\top\Sigma\beta$, and only slightly better than the plg-based CIs for R, are delayed to Section S4.3 of the Supplement (Tables S3-S5). In general our proposed CIs achieve the 95% nominal confidence levels whereas the bootstrap CIs are off-target or biased. In particular, for the genetic correlation R, the proposed CI has better coverage and smaller length. In addition, our proposed methods were computationally more efficient than the bootstrap CIs as the averaged running time (MacBook Pro with 2.2 GHz 6-Core Intel Core i7) for the proposed CIs is only about 1 second whereas the bootstrap CIs takes more than 1.6 mins for the plg-based CIs and 1 hour for the rpj-based CIs on average. When the sample size increased from 300 to 500, the empirical coverage of the proposed CIs for $\beta^\top\Sigma\gamma$ and R seemed to inflate slightly,

which again was likely due to our empirically determined tuning parameter. Nevertheless, the proposed CIs had shorter length for larger n, and its advantage over the alternative method was notable.

For hypothesis testing, we also compared the empirical type I errors and statistical powers of our proposed tests and the plg-based bootstrap tests, demonstrating the empirical superiority of the proposed method. For reason of space, we relegate our simulation results to Section S4.4 (Tables S6 and S7) of the Supplement.

6. Analysis of Ten Pediatric Autoimmune Diseases

We investigate the genetic correlations between each pair of ten pediatric autoimmune diseases, including autoimmune thyroiditis (THY), psoriasis (PSOR), juvenile idiopathic arthritis (JIA), ankylosing spondylitis (AS), common variable immunodeficiency (CVID), celiac disease (CEL), Crohn's disease (CD), ulcerative colitis (UC), type 1 diabetes (T1D) and systemic lupus erythematosus (SLE). The diseased subjects and controls were identified either directly from previous studies or from de-identified samples and associated electronic medical records in the genomics biorepository at The Children's Hospital of Philadelphia (Li et al., 2015). The data set includes 10,718 normal controls, 97 THY cases, 107 AS cases, 100 PSOR

Table 4: Coverage and length of the CIs with $\Sigma = \Sigma_B$, $\alpha = 0.05$ and sparsity $k = 25$.

pro: proposed estimators; boot: the plg-based bootstrap confidence intervals.

p	$\beta^\top\Sigma\gamma$				$\beta^\top\Sigma\beta$				R			
	pro		boot		pro		boot		pro		boot	
	cov	len	cov	len	cov	len	cov	len	cov	len	cov	len
	$n = 300$											
700	94.8	6.24	46.4	2.05	94.4	7.61	13.5	2.42	96.6	0.35	76.0	0.37
800	97.4	7.72	47.8	1.91	92.4	7.89	13.2	2.30	95.0	0.37	76.4	0.36
900	93.6	5.59	50.2	1.85	93.8	6.71	14.6	2.27	96.4	0.34	73.6	0.35
1000	93.2	5.85	42.6	1.93	92.6	7.88	7.2	2.39	93.0	0.32	76.4	0.36
	$n = 400$											
700	96.0	6.11	56.6	2.30	92.0	7.85	30.0	2.96	96.6	0.32	76.6	0.37
800	97.4	5.91	55.4	2.20	92.4	7.47	22.8	2.63	96.2	0.32	74.4	0.37
900	96.6	5.81	51.0	2.19	90.6	7.32	21.6	2.69	96.6	0.31	73.0	0.37
1000	93.8	5.65	47.8	2.07	90.4	7.11	19.8	2.58	93.4	0.31	72.6	0.36
	$n = 500$											
700	99.0	5.71	61.0	2.40	95.2	6.93	43.2	2.92	98.6	0.30	73.4	0.37
800	98.6	5.70	60.6	2.38	93.4	7.07	41.2	2.83	97.2	0.29	78.0	0.37
900	99.2	5.92	58.0	2.32	92.6	7.36	31.2	2.88	98.4	0.30	76.6	0.36
1000	98.6	5.44	57.8	2.18	90.4	6.70	30.0	2.73	98.2	0.29	76.6	0.36

cases, 173 CEL cases, 254 SLE cases, 308 CVID cases, 865 UC cases, 1086 T1D cases, 1123 JIA cases, and 1922 CD cases. Specifically, for each pair of the ten diseases, we evaluated their chromosome-specific genetic relatedness by estimating and performing hypotheses testing about the genetic correlation parameter on each of the 22 autosomes. By focusing on the chromosome-specific genetic correlations, we are able to make better inference with limited sample sizes for many diseases, and to obtain insights on the genomic regions that relate the two diseases of interest.

For each subject, after removing the SNPs with minor allele frequency less than 0.05, a total of 475,324 SNPs were obtained across 22 autosomes (see Supplement for details). To apply our proposed methods, for each pair of diseases, we randomly split the controls into two groups of equal size, combined them with each of the cases and fitted two high-dimensional logistic regressions between the disease outcomes and the SNPs to obtain the initial logistic Lasso estimators for each disease. Then the bias-corrected estimators were obtained, where the sample covariance matrix were calculated based on all the samples. Moreover, using our proposed method, we tested the individual null hypothesis that the chromosome-specific genetic correlation is zero between each pair of diseases in order to identify i) the diseases that are genetically associated and ii) the specific chromosome

where the diseases have shared genetic architecture.

The results are summarized in Figure 1. The top panel shows the estimated chromosome-specific genetic correlations between each pair of diseases, where the disease pairs having larger absolute values were annotated. The bottom panel shows the negative log-transformed p-values for each pair of diseases. Our tests suggest strong genetic sharing between UC and CD on chromosomes 1, 12, 17, 20 and 21, CVID and JIA on chromosome 8, and CD and PSOR on chromosome 13. Many pairs of these diseases showed genetic relatedness at the nominal p-value of 0.05, however, due to small sample sizes, they did not reach the statistical significance after the Bonferroni adjustment of multiple comparisons. Note that the pairs UC and CD, and CVID and JIA were also found to be statistically significant by Li et al. (2015) using different measures of genetic sharing, yet our proposed methods were able to additionally locate the genetic sharing to specific chromosomes and provide theoretically valid uncertainty quantifications.

7. Discussion

In this paper, a statistical inference framework for studying the genetic relatedness between two binary traits was introduced under the high-dimensional logistic regression models. Our model allows the number of SNPs to far ex-

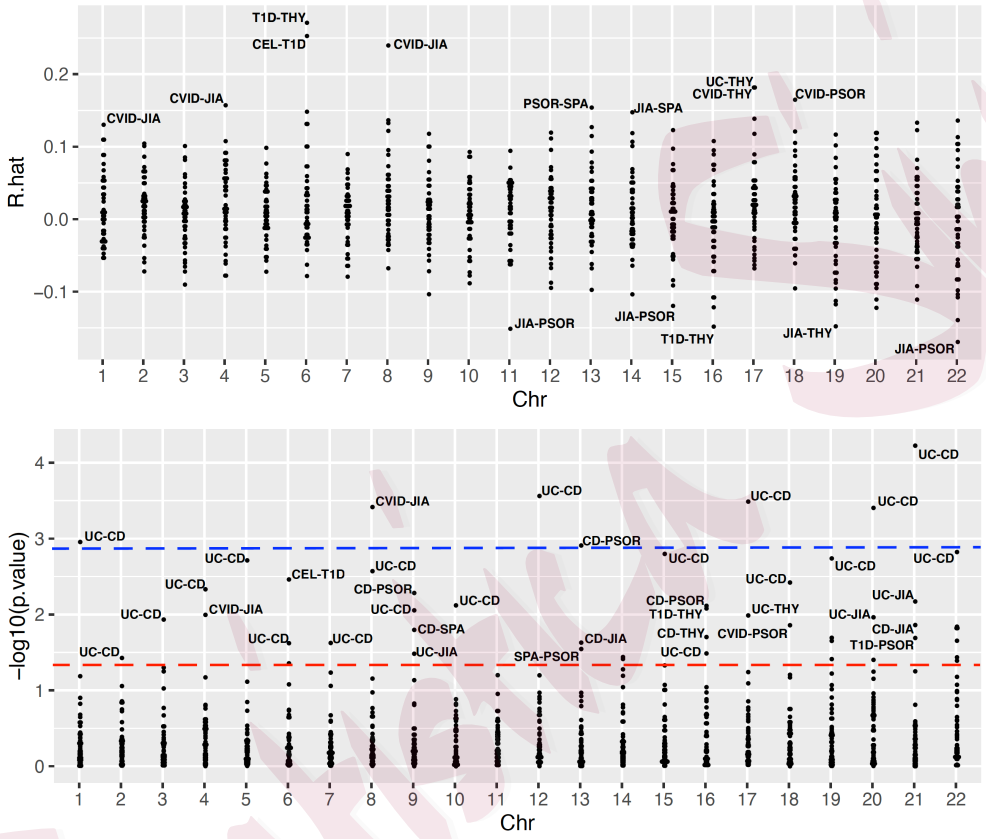

Figure 1: Analysis of genetic sharing between 10 autoimmune diseases. Top panel: estimated genetic correlations between each pair of diseases on each autosome. Bottom panel: negative log-transformed p-values for each pair of diseases, based on proposed method. The red and blue dashed lines represent the original and Bonferroni-adjusted significance level at 0.05.

ceed the sample sizes while producing efficient and valid statistical inference under mild conditions on sparsity and effect size of the true associations, and the covariance structure or linkage disequilibrium of the variants. Many efforts have been made to improve the speed of optimization and operation for genome-scale and ultrahigh-dimensional data sets. For example, in Qian et al. (2019), a new computational framework was proposed so that scalable Lasso solutions can be obtained for very large Biobank data set involving about 300,000 individuals and 800,000 genetic variants. We expect that these new computational methods will increase the utility of the proposed methods in genetic correlation analysis at whole genome sequencing scale.

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature 526*(7571), 68.

Bagos, P. G. (2012). On the covariance of two correlated log-odds ratios. *Statistics in Medicine 31*(14), 1418–1431.

Belloni, A., V. Chernozhukov, and Y. Wei (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics 34*(4), 606–619.

Bonnet, A., E. Gassiat, and C. Lévy-Leduc (2015). Heritability estimation in high dimensional sparse linear mixed models. *Electronic Journal of Statistics 9*(2), 2099–2129.

REFERENCES

Bulik-Sullivan, B., H. K. Finucane, V. Anttila, et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics 47*(11), 1236.

Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Stat. 45*(2), 615–646.

Cai, T. T. and Z. Guo (2020). Semi-supervised inference for explained variance in high-dimensional regression and its applications. *J. R. Stat. Soc. B 82*, 391–419.

Cai, T. T., Z. Guo, and R. Ma (2021). Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, 1–14.

Dimitromanolakis, A., J. Xu, A. Krol, and L. Briollais (2019). sim1000g: a user-friendly genetic variant simulator in r for unrelated individuals and family-based designs. *BMC Bioinformatics 20*(1), 1–9.

Guo, H., J. J. Li, Q. Lu, and L. Hou (2021). Detecting local genetic correlations with scan statistics. *Nature Communications 12*(1), 1–13.

Guo, Z., P. Rakshit, D. S. Herman, and J. Chen (2021). Inference for the case probability in high-dimensional logistic regression. *The Journal of Machine Learning Research 22*(1), 11480–11533.

Guo, Z., C. Renaux, P. Bühlmann, and T. Cai (2021). Group inference in high dimensions with applications to hierarchical testing. *Electronic Journal of Statistics 15*(2), 6633–6676.

Guo, Z., W. Wang, T. T. Cai, and H. Li (2019). Optimal estimation of genetic relatedness in

high-dimensional linear models. *J. Am. Stat. Assoc. 114*(525), 358–369.

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature 467*(7311), 52.

Janková, J. and S. van de Geer (2018). De-biased sparse PCA: Inference and testing for eigenstructure of large covariance matrices. *arXiv preprint arXiv:1801.10567*.

Janson, L., R. F. Barber, and E. Candes (2017). Eigenprism: inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. B 79*(4), 1037.

Javanmard, A. and A. Montanari (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*(1), 2869–2909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Lee, S. H., S. Ripke, B. M. Neale, S. V. Faraone, S. M. Purcell, R. H. Perlis, B. J. Mowry, A. Thapar, M. E. Goddard, J. S. Witte, et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature genetics 45*(9), 984.

Lee, S. H., N. R. Wray, M. E. Goddard, and P. M. Visscher (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet. 88*(3), 294–305.

Lee, S. H., J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray (2012). Estimation

of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics 28*(19), 2540–2542.

Li, Y. R., J. Li, S. D. Zhao, J. P. Bradfield, F. D. Mentch, S. M. Maggadottir, C. Hou, D. J. Abrams, D. Chang, F. Gao, et al. (2015). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine 21*(9), 1018.

Lu, Q., B. Li, D. Ou, M. Erlendsdottir, R. L. Powles, T. Jiang, Y. Hu, D. Chang, C. Jin, W. Dai, et al. (2017). A powerful approach to estimating annotation-stratified genetic covariance via gwas summary statistics. *Am. J. Hum. Genet. 101*(6), 939–964.

Ma, R., T. T. Cai, and H. Li (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *J. Am. Stat. Assoc.*, 1–15.

Ma, R., Z. Guo, T. T. Cai, and H. Li (2021). Supplement to "statistical inference for genetic relatedness based on high-dimensional logistic regression".

Maier, R., G. Moser, G.-B. Chen, et al. (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet. 96*(2), 283–294.

Mukherjee, R., N. S. Pillai, and X. Lin (2015). Hypothesis testing for high-dimensional sparse binary regression. *Ann. Stat. 43*(1), 352–381.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Stat. 45*(1), 158–195.

Ning, Z., Y. Pawitan, and X. Shen (2020). High-definition likelihood inference of genetic correlations across human complex traits. *Nature Genetics 52*(8), 859–864.

O'Connor, L. J. and A. L. Price (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics 50*(12), 1728–1734.

Qian, J., W. Du, Y. Tanigawa, M. Aguirre, R. Tibshirani, M. A. Rivas, and T. Hastie (2019). A fast and flexible algorithm for solving the lasso in large-scale and ultrahigh-dimensional problems. *BioRxiv*, 630079.

Shi, H., N. Mancuso, S. Spendlove, and B. Pasaniuc (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet. 101*(5), 737–751.

Sudlow, C., J. Gallacher, N. Allen, et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos Med 12*(3), e1001779.

Turley, P., R. K. Walters, O. Maghzian, et al. (2018). Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature Genetics 50*(2), 229–237.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Stat. 36*(2), 614–645.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat. 42*(3), 1166–1202.

Van Rheenen, W., W. J. Peyrot, A. J. Schork, et al. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet. 20*(10), 567–581.

Vattikuti, S., J. Guo, and C. C. Chow (2012). Heritability and genetic correlations explained by common snps for metabolic syndrome traits. *PLoS Genetics 8*(3), e1002637.

Verzelen, N. and E. Gassiat (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli 24*(4B), 3683–3710.

Wei, Y. and J. P. Higgins (2013). Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine 32*(7), 1191–1205.

Weissbrod, O., J. Flint, and S. Rosset (2018). Estimating snp-based heritability and genetic correlation in case-control studies directly and with summary statistics. *Am. J. Hum. Genet. 103*(1), 89–99.

Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics 25*(6), 714–721.

Xia, L., B. Nan, and Y. Li (2020). A revisit to de-biased lasso for generalized linear models. *arXiv preprint arXiv:2006.12778*.

Yang, J., B. Benyamin, B. P. McEvoy, et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature Genetics 42*(7), 565–569.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. B 76*(1), 217–242.

REFERENCES

Zhang, Y., Y. Cheng, Y. Ye, W. Jiang, Q. Lu, and H. Zhao (2020). Comparison of methods for estimating genetic correlation between complex traits using gwas summary statistics. *bioRxiv*.

Zhao, B. and H. Zhu (2019a). Cross-trait prediction accuracy of high-dimensional ridge-type estimators in genome-wide association studies. *arXiv preprint arXiv:1911.10142*.

Zhao, B. and H. Zhu (2019b). On genetic correlation estimation with summary statistics from genome-wide association studies. *arXiv preprint arXiv:1903.01301*.

Department of Statistics, Stanford University, Stanford, CA 02135

E-mail: (rongm@stanford.edu)

Department of Statistics, Rutgers University, Piscataway, NJ 08854

E-mail: (zijguo@stat.rutgers.edu)

Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104

E-mail: (tcai@wharton.upenn.edu)

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

E-mail: (hongzhe@upenn.edu)