Statistica Sinica **34** (2024), 1023-1043
doi:https://doi.org/10.5705/ss.202021.0386

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

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

[1]*Stanford University,* [2]*Rutgers University, and* [3]*University of Pennsylvania*

Abstract: We examine statistical inference for genetic relatedness between binary traits, based on individual-level genome-wide association data. Specifically, for high-dimensional logistic regression models, we define parameters characterizing the cross-trait genetic correlation, genetic covariance, and trait-specific genetic variance. We develop a novel weighted debiasing method for the logistic Lasso estimator and propose computationally efficient debiased estimators. Further more, we study the rates of convergence for these estimators and establish their asymptotic normality 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, and the size and power of the proposed tests. Numerical studies under both model-generated data and simulated genetic data show the superiority of the proposed methods. By analyzing a real data set on autoimmune diseases, we demonstrate their ability to obtain novel insights about the shared genetic architecture between 10 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) 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 the 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 the genetic

*Corresponding author.

relatedness between complex traits helps to 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)). Compared with methods of traditional approaches from family studies, where measurements of both traits are required for the same individuals, those based on GWAS enjoy the advantages of increased sample sizes and a reduced risk of confounding or ascertainment biases, and thus have greater potential for large-scale analyses 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, Guo and Chow (2012); Lee et al. (2013)). These models decompose the phenotypic variance into genetic and residual variance components, and define the genetic correlation as that 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)). Computationally efficient methods have been developed based on the cross-trait linkage disequilibrium (LD) score regression (Bulik-Sullivan et al. (2015); Ning, Pawitan and Shen (2020)) to estimate a genetic correlation using GWAS summary statistics over a large set of SNPs. This approach relies on the classical asymptotics, which do not consider the high dimensionality of the SNPs relative to the sample sizes, resulting in possibly inaccurate inference results (Zhao and Zhu (2019a)). Other approaches, such as those of Shi et al. (2017), Lu et al. (2017), and Guo et al. (2021a), explore differences in local genetic correlations between traits using genome partitioning based on genomic annotations. Weissbrod, Flint and Rosset (2018) notes that many existing methods are geared primarily toward quantitative traits. Thus, applying them directly to data sets with binary outcomes may suffer from reduced statistical power. They propose a mixed-effects model for estimating the genetic correlation between binary traits.

In this study, we take a high-dimensional regression approach, with fixed genetic effects to identify trait-associated genetic variants and quantify the genetic relatedness between two traits. An important advantage of a multiple regression over the simple univariate regression is its potential to identify more trait-associated variants (Wu et al. (2009)). Existing studies on heritability or co-heritability in a high-dimensional regression framework include, for example, those of Bonnet, Gassiat and Lévy-Leduc (2015), Janson, Barber and Candes (2017), Verzelen and Gassiat (2018), Guo et al. (2019), Zhao and Zhu (2019a), Zhao and Zhu (2019b), and Guo et al. (2021c). Under the linear regression model, Guo et al. (2019) propose bias-corrected estimators for the genetic covariance and correlation parameters, based on individual-level GWAS data, and Zhao and Zhu (2019a) propose consistent estimators for a polygenic risk score and a genetic

correlation, based on GWAS summary statistics. However, these works focus on the genetic relatedness between continuous traits, and do not provide inference procedures such as statistical tests.

We address the following two questions concerning binary traits. How can we define and study the genetic relatedness between two binary traits in a high-dimensional regression framework? How can we perform a valid statistical inference, such as testing hypotheses or constructing confidence intervals (CIs) 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 two traits is defined as the covariance between the log-odds ratios associated with the two traits, that is, genetic covariance$(y, w) = \text{Cov}\left(\log\{\pi_y(X)/(1-\pi_y(X))\}, \log\{\pi_w(X)/(1-\pi_w(X))\}\right)$, which, by definition, admits the following expressions: $\text{Cov}(\log\{\pi_y(X)/(1-\pi_y(X))\}, \log\{\pi_w(X)/(1-\pi_w(X))\}) = \text{Cov}(X^\top\beta, X^\top\gamma) = \beta^\top\Sigma\gamma$. Similarly, we define the genetic variance of the binary trait y as the variance of its associated log-odds ratio, that is, genetic variance$(y) = \text{Var}\left(\log\{\pi_y(X)/(1-\pi_y(X))\}\right)$, which satisfies $\text{Var}\left(\log\{\pi_y(X)/(1-\pi_y(X))\}\right) = \text{Var}(X^\top\beta) = \beta^\top\Sigma\beta$. We define the genetic variance of the trait w as $\text{Var}\left(\log\{\pi_w(X)(1-\pi_w(X))\}\right) = \text{Var}(X^\top\gamma) = \gamma^\top\Sigma\gamma$. Whenever the genetic variances of y and w are both 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, $\text{Corr}(\log\{\pi_y(X)/(1-\pi_y(X))\}, \log\{\pi_w(X)(1-\pi_w(X))\}) = \beta^\top\Sigma\gamma/\sqrt{\beta^\top\Sigma\beta\gamma^\top\Sigma\gamma}$, 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. It is used by Wei and Higgins (2013) to account for correlated outcomes in meta-analysis, and by Bagos (2012) when the data take 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 considering the true covariance structure of the

variants.

We examine 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 CIs, and 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 $A \in \mathbb{R}^{p\times p}$, $\lambda_i(A)$ denotes its ith largest eigenvalue and $\lambda_{\max}(A) = \lambda_1(A)$ and $\lambda_{\min}(A) = \lambda_p(A)$. For a smooth function $f(x)$ defined on $\mathbb{R}$, we denote $\dot{f}(x) = df(x)/dx$ and $\ddot{f}(x) = d^2f(x)/dx^2$. 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 $a_n \le Cb_n$ 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 commonly used to study the 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 is common in cross-trait analyses 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 the 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 $i \in \{1, 2, \ldots, m\}$, $1 \le m \le$

$\min\{n_1, n_2\}$. 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. A discussion of Scenario II is provided in Section S5 of the Supplementary Material (Ma et al. (2021)), because our methods and results in this case are very similar.

2.2. Weighted bias correction and the proposed estimators

Estimating 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 $\gamma^\top\Sigma\gamma$. The novel bias-correction method proposed here yields nearly unbiased estimators of these functionals of interest. We construct the estimators using the following two-step procedure. In the first step, we obtain an initial plug-in estimator of the functional based on the pooled sample covariance matrix $\widehat{\Sigma} = 1/(n_1+n_2)\big[\sum_{i=1}^{n_1} X_{i\cdot}X_{i\cdot}^\top + \sum_{i=1}^{n_2} Z_{i\cdot}Z_{i\cdot}^\top\big]$, and the logistic Lasso estimators

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

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

We begin with the 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 $\widehat{\beta}^\top\widehat{\Sigma}\widehat{\gamma}$, the error of which 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 $\widehat{\beta}^\top(\widehat{\Sigma}-\Sigma)\widehat{\gamma}$ contributes only 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,and 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 $h(u) = e^u/(1+e^u)$. Then by Taylor's expansion, $h(\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}) - h(\alpha+X_{i\cdot}^\top\beta) = e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}X_{i\cdot}^\top(\widehat{\beta}-\beta)/(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2 + e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}}(\widehat{\alpha}-\alpha)/(1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2 + \Delta_i$, 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 $\epsilon_i = y_i - h(\alpha+X_{i\cdot}^\top\beta)$, then

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

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 $(1+e^{\widehat{\alpha}+X_{i\cdot}^{\top}\widehat{\beta}})^2/e^{\widehat{\alpha}+X_{i\cdot}^{\top}\widehat{\beta}}$ 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 relative to the leading term $\big(\sum_{i=1}^{n_1}X_{i\cdot}X_{i\cdot}^{\top}\big)(\widehat{\beta}-\beta)$, we can construct an estimator of $\widehat{\gamma}^{\top}\Sigma(\widehat{\beta}-\beta)$ 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 $\widehat{\beta}^{\top}\Sigma(\widehat{\gamma}-\gamma)$ using

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

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 that accounts 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, 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 $\bar{R} = \widehat{\beta^\top\Sigma\gamma}/\sqrt{\widehat{\beta^\top\Sigma\beta}\widehat{\gamma^\top\Sigma\gamma}}$. Taking into account the actual range of R, we propose its final estimator as

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

Compared with 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, Cai and Li (2020); Cai and Guo (2020); Cai, Guo and Ma (2023); Guo et al. (2021b)), our proposed method has two distinct advantages. First, the proposed estimators can be obtained directly from their explicit expressions, as in (2.4) to (2.7), which rely only 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 achieved efficiently using a standard tuning process (Section 5), and therefore is more scalable to the large data sets in genetic studies. In contrast, existing methods involve solving other high-dimensional optimization problems, in addition to the initial estimator, for bias correction. These additional problems are computationally challenging, time-consuming, and subject to difficult tuning processes. Second, by using our carefully constructed weighted bias-correction method, we can avoid many commonly used, but stringent technical conditions. This significantly expands the range of applicability of our proposed methods; see also the discussions after Theorems 1 and 5.

3. CIs 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 CIs and statistical tests for the functionals.

Specifically, it can be shown that the estimator $\widehat{\beta^\top\Sigma\gamma}$ has variance $v^2 = (n_1+n_2)/n_1 E\{\eta_i^{(X)}(\widehat{\gamma}^\top X_{i\cdot})^2\} + ((n_1+n_2)/n_2)E\{\eta_i^{(Z)}(\widehat{\beta}^\top Z_{i\cdot})^2\} + E\{\widehat{\beta}^\top(X_{i\cdot}X_{i\cdot}^\top - \Sigma)\widehat{\gamma}\}^2$, where $\eta_i^{(X)} = (1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^4 e^{\alpha+X_{i\cdot}^\top\beta}/((1+e^{\alpha+X_{i\cdot}^\top\beta})^2 e^{2\widehat{\alpha}+2X_{i\cdot}^\top\widehat{\beta}})$ and

$\eta_i^{(Z)} = (1+e^{\widehat{\zeta}+Z_{i\cdot}^\top\widehat{\gamma}})^4 e^{\zeta+Z_{i\cdot}^\top\gamma}/((1+e^{\zeta+Z_{i\cdot}^\top\gamma})^2 e^{2\widehat{\zeta}+2Z_{i\cdot}^\top\widehat{\gamma}})$. Intuitively, the parameters β and γ in the above expressions can be estimated using their initial Lasso estimators. Thus, we can define a moment estimator of the asymptotic variance as $\widehat{v}^2 = ((n_1+n_2)/n_1^2)\sum_{i=1}^{n_1}((1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2/e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})(\widehat{\gamma}^\top X_{i\cdot})^2 + ((n_1+n_2)/n_2^2)\sum_{i=1}^{n_2}((1+e^{\widehat{\eta}+Z_{i\cdot}^\top\widehat{\gamma}})^2/e^{\widehat{\eta}+Z_{i\cdot}^\top\widehat{\gamma}})(\widehat{\beta}^\top Z_{i\cdot})^2 + (1/(n_1+n_2))\{\sum_{i=1}^{n_1}(\widehat{\beta}X_{i\cdot}X_{i\cdot}^\top\widehat{\gamma} - \widehat{\beta}\widehat{\Sigma}\widehat{\gamma})^2 + \sum_{i=1}^{n_2}(\widehat{\beta}Z_{i\cdot}Z_{i\cdot}^\top\widehat{\gamma} - \widehat{\beta}\widehat{\Sigma}\widehat{\gamma})^2\}$. Hence, a $(1-\alpha)$-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 $\widehat{\rho} = z_{\alpha/2}\widehat{v}/\sqrt{n_1+n_2}$, 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 $v_\beta^2 = (4(n_1+n_2)/n_1)E\{\eta_i^{(X)}(\beta^\top X_{i\cdot})^2\} + E\{\beta^\top(X_{i\cdot}X_{i\cdot}^\top - \Sigma)\beta\}^2$, which can be estimated using $\widehat{v}_\beta^2 = (4(n_1+n_2)/n_1^2)\sum_{i=1}^{n_1}((1+e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})^2/e^{\widehat{\alpha}+X_{i\cdot}^\top\widehat{\beta}})(\widehat{\beta}^\top X_{i\cdot})^2 + (1/(n_1+n_2))\{\sum_{i=1}^{n_1}(\widehat{\beta}X_{i\cdot}X_{i\cdot}^\top\widehat{\beta} - \widehat{\beta}\widehat{\Sigma}\widehat{\beta})^2 + \sum_{i=1}^{n_2}(\widehat{\beta}Z_{i\cdot}Z_{i\cdot}^\top\widehat{\beta} - \widehat{\beta}\widehat{\Sigma}\widehat{\beta})^2\}$. Then, a $(1-\alpha)$-level CI 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 $\widehat{\rho}_\beta = z_{\alpha/2}\widehat{v}_\beta/\sqrt{n_1+n_2}$; $\mathrm{CI}_\alpha(\gamma^\top\Sigma\gamma,\mathcal{D})$ can be obtained by symmetry.

The CI for the genetic correlation R is a direct consequence of 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 $\widehat{v}_R^2 = \widehat{v}^2/(\widehat{\beta^\top\Sigma\beta}\widehat{\gamma^\top\Sigma\gamma})$, and define the corresponding $(1-\alpha)$-level CI as $\mathrm{CI}_\alpha(R,\mathcal{D}) = \big[\widehat{R} - \widehat{\rho}_R, \widehat{R} + \widehat{\rho}_R\big] \cap [-1,1]$, where $\widehat{\rho}_R = z_{\alpha/2}\widehat{v}_R/\sqrt{n_1+n_2}$.

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 $T_1 = \sqrt{n_1+n_2}(\widehat{\beta^\top\Sigma\gamma} - B_0)/\widehat{v}$, $T_2 = \sqrt{n_1+n_2}(\widehat{\beta^\top\Sigma\beta} - Q_0)/\widehat{v}_\beta$, and $T_3 = \sqrt{n_1+n_2}(\widehat{R} - R_0)/\widehat{v}_R$, 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 centered independent and identically distributed *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 $E(\beta^\top X_{i\cdot}X_{i\cdot}^\top\gamma/(\beta^\top\Sigma\gamma) - 1)^2 > c_0$.

For 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 zero and one, whereas (A2) ensures that the asymptotic variances do not diminish.

For technical reasons, for each trait, we split the corresponding samples into halves, so that the initial Lasso estimation step and the other steps, such as the covariance estimation and bias-correction, are conducted on independent data sets. Without loss of generality, we assume under Scenario I that 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 $\mathcal{D}_1$; the sample covariance, bias-correction terms, and asymptotic variance estimators are based on $\mathcal{D}_2$ and the initial Lasso estimators. Note that the sample-splitting 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.

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* $k \lesssim n/(\log p \log n)$. *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 requires only the balanced marginal case probabilities from (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. Thus the proposed methods are 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* $i=1,\ldots,n$*, and* $k \lesssim \min\{p^\nu, n/\log p\}$*, for some* $0<\nu<1/2$*. Then,*

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

for some constant $c>0$*.*

By Theorem 1, a uniform upper bound over the parameter space $\Theta(k,L_n)$ can be obtained as $\sup_{\theta\in\Theta(k,L_n)} P_\theta\big(|\widehat{\beta^\top\Sigma\gamma} - \beta^\top\Sigma\gamma| \lesssim tL_n/\sqrt{n} + (1+L_n\sqrt{\log n})\, k\log p/n\big) \geq 1 - p^{-c} - n^{-c} - t^{-2}$. Combining this with the lower bound from Theorem 2, we conclude that, for all $k \lesssim \min\{n/(\log p\log n), p^\nu\}$, with any $\nu\in(0,1/2)$, and $\sqrt{k\log p/n} \lesssim L_n \lesssim 1$, 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 $k \lesssim \sqrt{n}/\log p\sqrt{\log n}$, or the weak signal regime $L_n \lesssim (\log n)^{-1/2}$, over which the minimax rate is $L_n/\sqrt{n} + k\log p/n$. 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,* $n_1 \asymp n_2 \asymp n$*,* $k \ll n/(\log p\log n)$*, 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*

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

Compared with 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 that 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 for the CIs and 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), *and* (A3) *hold,* $n_1 \asymp n_2 \asymp n$*,* $k \lesssim n/(\log p\log n)$*, and* $L(\beta,\gamma) \gg \sqrt{k\log p/n}$*. Then, we have the following:*

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

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

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

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* $k \ll \min\{n/(\log p\log n), U(\beta,\gamma)\sqrt{n}/(\{1+U(\beta,\gamma)\sqrt{\log n}\}\log p)\}$, *we have* $(\sqrt{n_1+n_2}(\bar{R}-R)/v_R)\big|\mathcal{D}_1 \to_d N(0,1)$ *as* $(n,p)\to\infty$.

Some remarks about the technical innovations leading to the above theorems are in order. First, in contrast to existing works on statistical inference in high-dimensional logistic regression, the proposed methods do not require several 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, Cai and Li (2020); Guo et al. (2021b)), 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, Chernozhukov and Wei (2016); Ning and Liu (2017); Janková and van de Geer (2018)), and the sparse precision condition (Ma, Cai and Li (2020)). Second, from a practical viewpoint, removing these technical assumptions significantly expands the range of applicability of the proposed methods. For example, as argued by Cai, Guo and Ma (2023) and Xia, Nan and Li (2020), in practice, the bounded individual probability and the sparse inverse population Hessian conditions are seldom satisfied or verifiable from the data. In contrast, the balanced marginal case probability condition holds easily, and can be checked based on the observed outcomes.

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

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

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* $L\{\mathrm{CI}_\alpha(\beta^\top\Sigma\gamma,\mathcal{D})\} \asymp U(\beta,\gamma)/\sqrt{n}$, $L\{\mathrm{CI}_\alpha(\beta^\top\Sigma\beta,\mathcal{D})\} \asymp \|\beta\|_2/\sqrt{n}.$*, and* $L\{\mathrm{CI}_\alpha(R,\mathcal{D})\} \asymp 1/(L(\beta,\gamma)\sqrt{n})$.

This theorem implies that the statistical tests proposed in Section 3 have the following theoretical properties related to their size and power under certain local alternatives.

Corollary 1 (Hypothesis Testing). *Under the conditions of Theorem* 6*, we have the following:*

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 based on simulated genetic data

To justify our proposed methods for analyzing real genetic data sets, we carried out numerical experiments under settings in which the covariates are simulated genotypes with possible LD structures that resemble those of the human genome, and inferences are made at a chromosomal basis. Specifically, focusing on 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 generate 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)). We also generate a comprehensive haplotype map integrated over 1,184 reference individuals (International HapMap 3 Consortium (2010)); see Section S4 of the Supplementary Material 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 are 25 associated SNPs, with 12 of them shared by both traits. The effect sizes of the associated SNPs are drawn uniformly from $[-1,1]$. For reasons of practical interest, we focus mainly on the estimation, CIs and hypothesis testing about the genetic correlation parameter. The results for the genetic covariance and variance can be found in Section 5.2 below and Section S4 of the Supplementary Material.

For the 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 $\hat{R}_{plg} = \hat{\beta}^\top\hat{\Sigma}\hat{\gamma}/\sqrt{\hat{\beta}^\top\hat{\Sigma}\hat{\beta}\hat{\gamma}^\top\hat{\Sigma}\hat{\gamma}}$; (ii) the component-wise projected Lasso ("lpj") estimators $\check{\beta}^\top\hat{\Sigma}\check{\gamma}$, $\check{\beta}^\top\hat{\Sigma}\check{\beta}$, and $\hat{R}_{lpj} = \check{\beta}^\top\hat{\Sigma}\check{\gamma}/\sqrt{\check{\beta}^\top\hat{\Sigma}\check{\beta}\check{\gamma}^\top\hat{\Sigma}\check{\gamma}}$, where each component of $\check{\beta}$ and $\check{\gamma}$ is the debiased Lasso estimator implemented by the function `lasso.proj` in the R package `hdi` using the default settings; and (iii) the component-wise projected Ridge ("rpj") estimators $\check{\beta}^\top\hat{\Sigma}\check{\gamma}$, $\check{\beta}^\top\hat{\Sigma}\check{\beta}$, and $\hat{R}_{rpj} = \check{\beta}^\top\hat{\Sigma}\check{\gamma}/\sqrt{\check{\beta}^\top\hat{\Sigma}\check{\beta}\check{\gamma}^\top\hat{\Sigma}\check{\gamma}}$, where each component of $\check{\beta}$ and $\check{\gamma}$ is obtained from the function `ridge.proj` in the R package `hdi` using the default settings. For the proposed method, we use 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 demonstrate the superior performance of the proposed method.

For the CIs, we compare our proposed CIs ("pro") with 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 hypothesis testing, we evaluate the empirical type-I errors and statistical power of the proposed tests and the plg-based bootstrap tests in a setup in which the effect sizes are generated using an additional constraint $|\beta^\top\Sigma\gamma| > 3$. Table 3 shows the empirical type-I errors and statistical power of the proposed tests over different settings, each based on 500 rounds of simulations. Our results suggest that the proposed test is empirically valid and has advantages over the bootstrap tests. In Tables 2 and 3, the proposed method becomes a little conservative when n increases from 200 to 400, likely because of the limitation of our empirically determined tuning parameter. Nevertheless, we still observe 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 additional simulations under a slightly different setting of the association structure, see Section S4.5 of the Supplementary Material (Table S8).

5.2. Evaluation based on 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 generate β_j and γ_j uniformly from $[-1, 1]$, for all $j \in \mathcal{S}$. For the design covariates, we focus on Scenario I, where $n_1 = n_2 = n$. The covariates

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
1,000	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
1,000	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 power when 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
1,000	0.06	0.41	0.42	0.71	0.02	0.36	0.63	0.70	0.02	0.36	0.68	0.70

are generated from a multivariate Gaussian distribution with covariance matrix as either $\Sigma = \Sigma_B$, where Σ_B is a $p \times p$ blockwise-diagonal matrix of 10 identical unit-diagonal Toeplitz matrices, with off-diagonal entries that descend from 0.3 to 0 (see Section S4.1 of the Supplementary Materia for its explicit form), or $\Sigma = \Sigma_E$, where Σ_E is an exchangeable covariance matrix with unit diagonals and off-diagonals equal to 0.2. The numerical result for each setting is based on 500 rounds of simulations.

For the parameter estimation, we evaluate the proposed method and the three alternative methods defined in the previous section. The results are provided in Section S4.2 of the Supplementary Material (Tables S1, S2), which show the superiority of each of the proposed estimators over the alternatives. For the same simulation setups, we evaluate and compare various methods for constructing 95% CIs for the parameters. Specifically, we compare 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. In general, the coverage of the rpj-based bootstrap CIs is poorer than that of the plg-based CIs for $\beta^\top\Sigma\gamma$ and $\beta^\top\Sigma\beta$, and only slightly better than that of the plg-based CIs for R; these results and those under the exchangeable covariance are provided in Supplementary Material Section S4.3 (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 a smaller length. In addition, our proposed methods are computationally more efficient than the bootstrap CIs, because 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 take more than 1.6 minutes for the plg-based CIs, and 1 hour for the rpj-based CIs. When the sample size increases from 300 to 500, the empirical coverage of the proposed CIs for $\beta^\top\Sigma\gamma$ and R seems to inflate slightly, which is again likely due to our empirically determined tuning parameter. Nevertheless, the proposed CIs have a shorter length for larger n, and the advantage of the proposed method over the alternative methods is evident.

For the hypothesis testing, we also compare the empirical Type-I errors and statistical power of our proposed tests and the plg-based bootstrap tests, demonstrating the empirical superiority of the proposed method; these results are provided in Section S4.4 (Tables S6 and S7) of the Supplementary Material.

6. Analysis of 10 Pediatric Autoimmune Diseases

We investigate the genetic correlations between each pair of 10 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). We identified the subjects with a disease and controls 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 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 10 diseases, we evaluate their chromosome-specific genetic relatedness by estimating and performing hypothesis testing on the genetic correlation parameter for each of the 22 autosomes. By focusing on the chromosome-specific genetic correlations, we can make a better inference

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 CIs.

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
1,000	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
1,000	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
1,000	98.6	5.44	57.8	2.18	90.4	6.70	30.0	2.73	98.2	0.29	76.6	0.36

from a limited sample size for many diseases, and obtain insights on the genomic regions that relate the two diseases of interest.

For each subject, after removing the SNPs with a minor allele frequency less than 0.05, we have a total of 475,324 SNPs were obtained across 22 autosomes (see Supplementary Material for details). To apply our proposed methods, for each pair of diseases, we randomly split the controls into two groups of equal size, combine them with each of the cases, and fit two high-dimensional logistic regressions between the disease outcomes and the SNPs to obtain the initial logistic Lasso estimators for each disease. Then we obtain the bias-corrected estimators, where the sample covariance matrix is calculated based on all samples. Moreover, using our proposed method, we test the individual null hypothesis that the chromosome-specific genetic correlation is zero between each pair of diseases in order to identify i) diseases that are genetically associated; and ii) specific chromosomes in which diseases have a 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 with larger absolute values are 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, between CVID and JIA on chromosome 8, and between CD and PSOR on chromosome 13. Many pairs of these diseases showed genetic relatedness at the

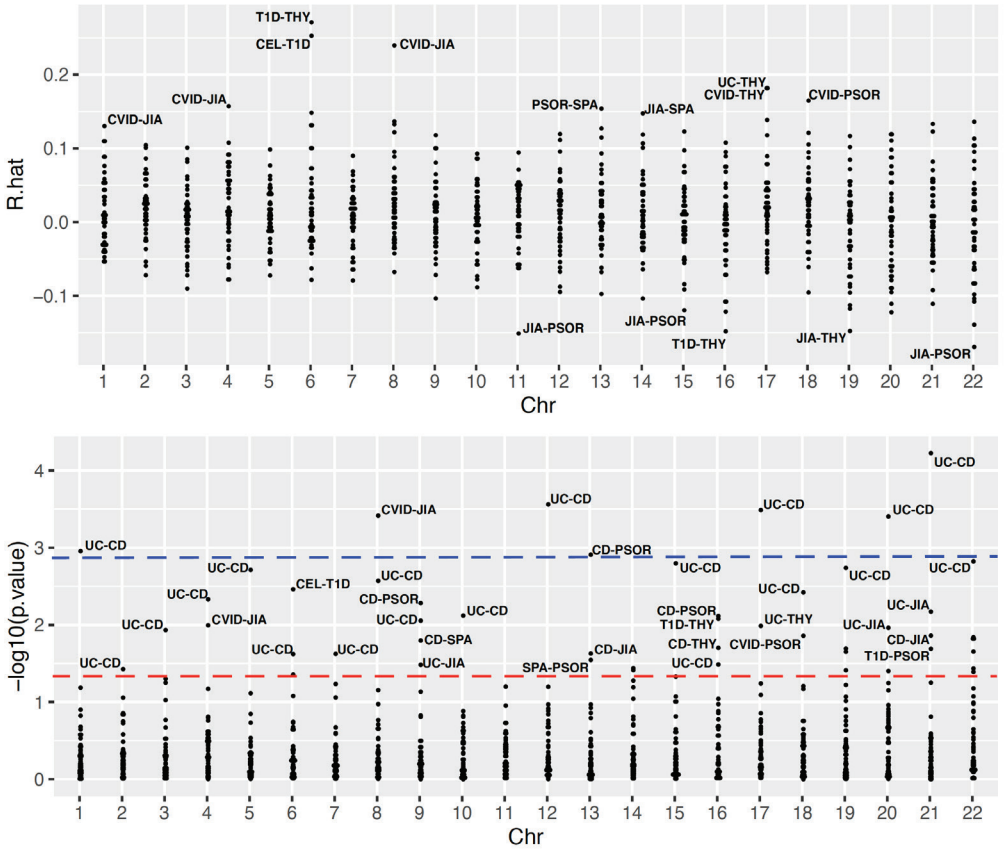

Figure 1. Analysis of genetic sharing between 10 autoimmune diseases. Top panel: estimated genetic correlations between each pair of diseases on each autosome. Bottom panel: negative log-transformed p-values for each pair of diseases, based on the proposed method. The red and blue dashed lines represent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

nominal p-value of 0.05. However, however, because of the small sample sizes, they do 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. However, our proposed methods also locate genetic sharing with specific chromosomes and provide theoretically valid uncertainty quantifications.

7. Discussion

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

sample size while producing efficient and valid statistical inferences under mild conditions on the sparsity andthe effect size of the true associations, and on the covariance structure or linkage disequilibrium of the variants. Many works have tried to improve the speed of optimization and operation for genome-scale and ultrahigh-dimensional data sets. For example, Qian et al. (2019) propose a new computational framework in which scalable Lasso solutions can be obtained for a 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 a whole-genome sequencing scale.

Supplementary Material

The Supplementary Material includes proofs of the main theorems and the technical lemmas, additional simulation results, supplementary notes, figures and tables.

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.

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

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

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

Bulik-Sullivan, B.-S., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh. P.-R. et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236–1241.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics* **45**, 615–646.

Cai, T. T. and Guo, Z. (2020). Semi-supervised inference for explained variance in high-dimensional regression and its applications. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* **82**, 391–419.

Cai, T. T., Guo, Z. and Ma, R. (2023). Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association* **118**, 1319–1332.

Dimitromanolakis, A., Xu, J., Krol, A. and Briollais, L. (2019). sim1000g: A user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics* **20**, Article 26.

Guo, H., Li, J. J., Lu, Q. and Hou, L. (2021a). Detecting local genetic correlations with scan statistics. *Nature Communications* **12**, 2033.

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

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

Guo, Z., Wang, W., Cai, T. T. and Li, H. (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association* **114**, 358–369.

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58.

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

Janson, L., Barber, R. F. and Candes, E. (2017). Eigenprism: Inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* **79**, 1037–1065.

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

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

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H. et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* **45**, 984–994.

Lee, S. H., Wray, N. R., Goddard, M. E. and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88**, 294–305.

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. and Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542.

Li, Y. R., Li, J., Zhao, S. D., Bradfield, J. P., Mentch, F. D., Maggadottir, S. M. et al. (2015). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine* **21**, 1018–1027.

Lu, Q., Li, B., Ou, D., Erlendsdottir, M., Powles, R. L., Jiang, T., et al. (2017). A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *American Journal of Human Genetics* **101**, 939–964.

Ma, R., Cai, T. T. and Li, H. (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association* **116**, 984–998.

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

Maier, R., Moser, G., Chen, G.-B., Ripke, S., Cross-Disorder Working Group of the Psychiatric Genomics Consortium, Coryell, W. et al. (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *American Journal of Human Genetics* **96**, 283–294.

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

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

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

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

Shi, H., Mancuso, N., Spendlove, S. and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *American Journal of Human Genetics* **101**, 737–751.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J. 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**, e1001779.

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A. et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics* **50**, 229–237.

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36**, 614–645.

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

van Rheenen, W., Peyrot, W. J., Schork, A. J. and Wray, N. R. (2019). Genetic correlations of polygenic disease traits: From theory to practice. *Nature Reviews Genetics* **20**, 567–581.

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

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

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

Weissbrod, O., Flint, J. and Rosset, S. (2018). Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. *American Journal of Human Genetics* **103**, 89–99.

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

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

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* **76**, 217–242.

Zhang, Y., Cheng, Y., Ye, Y., Jiang, W., Lu, Q. and Zhao, H. (2020). Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. *Briefings in Bioinformatics* **22**, bbaa442.

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

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

Rong Ma

Department of Statistics, Stanford University, Stanford, CA 02135, USA.

E-mail: rongm@stanford.edu

Zijian Guo

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.

E-mail: zijguo@stat.rutgers.edu

T. Tony Cai

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

E-mail: tcai@wharton.upenn.edu

Hongzhe Li

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

E-mail: hongzhe@upenn.edu

(Received November 2021; accepted September 2022)