Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Optimal Estimation of Genetic Relatedness in High-Dimensional Linear Models

Zijian Guo, Wanjie Wang, T. Tony Cai & Hongzhe Li

To cite this article: Zijian Guo, Wanjie Wang, T. Tony Cai & Hongzhe Li (2019) Optimal Estimation of Genetic Relatedness in High-Dimensional Linear Models, Journal of the American Statistical Association, 114:525, 358-369, DOI: 10.1080/01621459.2017.1407774

To link to this article: https://doi.org/10.1080/01621459.2017.1407774

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 525, 358–369, Theory and Methods
https://doi.org/10.1080/01621459.2017.1407774

Taylor & Francis
Taylor & Francis Group

Optimal Estimation of Genetic Relatedness in High-Dimensional Linear Models

Zijian Guo[a], Wanjie Wang[b], T. Tony Cai[c], and Hongzhe Li[d]

[a]Department of Statistics and Biostatistics, Rutgers University, New Brunswick, NJ; [b]Department of Statistics and Applied Probability, National University of Singapore, Singapore; [c]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA; [d]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

ABSTRACT

Estimating the genetic relatedness between two traits based on the genome-wide association data is an important problem in genetics research. In the framework of high-dimensional linear models, we introduce two measures of genetic relatedness and develop optimal estimators for them. One is genetic covariance, which is defined to be the inner product of the two regression vectors, and another is genetic correlation, which is a normalized inner product by their lengths. We propose functional de-biased estimators (FDEs), which consist of an initial estimation step with the plug-in scaled Lasso estimator, and a further bias correction step. We also develop estimators of the quadratic functionals of the regression vectors, which can be used to estimate the heritability of each trait. The estimators are shown to be minimax rate-optimal and can be efficiently implemented. Simulation results show that FDEs provide better estimates of the genetic relatedness than simple plug-in estimates. FDE is also applied to an analysis of a yeast segregant dataset with multiple traits to estimate the genetic relatedness among these traits. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2016
Revised October 2017

KEYWORDS
Genetic correlations; Genome-wide association studies; Inner product; Minimax rate of convergence; Quadratic functional

1. Introduction

1.1. *Motivation and Background*

Genome-wide association studies (GWAS) have led to identification of thousands of genetic variants or single nucleotide polymorphisms (SNPs) that are associated with various complex phenotypes (Manolio 2010). Results from these GWAS have shown that many complex phenotypes share common genetic variants, including various autoimmune diseases (Zhernakova, van Diemen, and Wijmenga 2009) and psychiatric disorders (Lee et al. 2013). These empirical evidences of shared genetic etiology for various phenotypes provide important insights of common pathophysiologies for related disorders that can be explored for drug repositioning and for studying disease etiology. Such knowledge of genetic sharing can potentially be explored to increase the accuracy of genetic risk prediction (Wray, Goddard, and Visscher 2007; Purcell et al. 2009; Maier et al. 2015). The concept of genetic relatedness or genetic correlations has been proposed to describe the shared genetic associations within pairs of quantitative traits based on GWAS data. This is in contrast to the traditional approaches of estimating co-heritability based on twin or family studies, where measurements of both traits are required on the same set of individuals. Due to the availability of GWAS datasets of many important traits, there has been significant recent interest in methods for quantifying and estimating the genetic relatedness between two traits based on large-scale genetic association data.

Several measures of genetic relatedness have been proposed using GWAS data. Lee et al. (2012) and Yang et al. (2013) extended the mixed-effect model framework to estimate genetic covariance and genetic correlation between two traits. In their models, each individual's trait value is associated with a random genetic effect, which is correlated across individuals by virtue of sharing some of the genetic variants affecting the traits, and an environmental random effect. Co-heritability is then defined as the square root of the ratio of the covariance of the genetic random effects to the product of the total variances. The mixed-effect model approach requires knowledge of the identity of the causal variants, and hence the covariance matrix. This is however not available. Lee et al. (2012) and Yang et al. (2013) approximated the genetic relationship between every pair of individuals across the set of causal variants by the genetic relationship across the set of all genotyped variants. However, the very large number of variants used for estimating the genetic correlations, most of them likely not causative, might mask out the correlations on the set of causal variants, leading to inaccurate and suboptimal estimation of heritability (Golan and Rosset 2011). Bulik-Sullivan et al. (2015) studied the genetic relatedness based on another random effects model for the two traits and developed a cross-trait linkage disequilibrium (LD) score regression to estimate the genetic covariance and genetic correlation. This approach shares similarity with the mixed-effect model approach of Yang et al. (2013) but has the advantages of only using the GWAS summary statistics. Lee and van der Wer (2016) developed an algorithm

CONTACT Hongzhe Li hongzhe@pennmedicine.upenn.edu Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

for multivariate linear mixed model analysis and demonstrated its use in estimating co-heritability.

To alleviate the difficulty of estimating the covariance matrix in the commonly used mixed effect model framework of estimating the heritability or co-heritability, we take a regression approach with fixed genetic effects in high-dimensional settings. High-dimensional linear regression provides a natural framework for GWAS to identify the trait-associated genetic variants, and its advantages over the simple univariate analysis have been demonstrated (Wu et al. 2009). The study of heritability in high-dimensional regression analysis has been studied by Bonnet et al. (2015), Verzelen and Gassiat (2016), and Janson, Barber, and Candes (2016). However, high-dimensional regression analysis has not been explored to study the genetic relatedness between two traits based on genetic association data. The goal of this article is to define two quantities that can be used to measure the genetic relatedness between a pair of traits based on GWAS data in the framework of high-dimensional linear models. Our definitions of the genetic relatedness reflect covariance or correlation of the trait-associated genetic variants. This is different from the mixed-effects model-based approaches where the genetic relatedness is defined through the variance/covariance matrix of the individual-specific random effects and the data from all the genetic variants are used to approximate the true covariance matrix.

1.2. Definition and Problem Formulation

A pair of trait values $(\mathbf{y}, \mathbf{w})$ are modeled as a linear combination of p genetic variants and an error term that includes environmental and unmeasured genetic effects,

$$\begin{aligned} \boldsymbol{y}_{n_1\times 1} &= \boldsymbol{X}_{n_1\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n_1\times 1} \\ \text{and} \quad \boldsymbol{w}_{n_2\times 1} &= \boldsymbol{Z}_{n_2\times p}\boldsymbol{\gamma}_{p\times 1} + \boldsymbol{\delta}_{n_2\times 1}, \end{aligned} \tag{1}$$

where the rows $\boldsymbol{X}_{i\cdot}$ are iid p-dimensional sub-Gaussian random vectors with covariance matrix $\boldsymbol{\Sigma}$, the rows $\boldsymbol{Z}_{i\cdot}$ are iid p-dimensional sub-Gaussian random vectors with covariance matrix $\boldsymbol{\Gamma}$, and the error $(\boldsymbol{\epsilon}, \boldsymbol{\delta})^{\intercal}$ follows the multivariate normal distribution with mean zero and covariance

$$\begin{pmatrix} \sigma_1^2 \mathbf{I}_{n_1\times n_1} & \mathbf{0}_{n_1\times n_2} \\ \mathbf{0}_{n_2\times n_1} & \sigma_2^2 \mathbf{I}_{n_2\times n_2} \end{pmatrix}$$

and is assumed to be independent of $\boldsymbol{X}$ and $\boldsymbol{Z}$.

In the study of genetic relatedness, the pair of traits $\boldsymbol{y}$ and $\boldsymbol{w}$ are assumed to have mean zero, and the jth column of $\boldsymbol{X}$, $\boldsymbol{X}_{\cdot j}$, and the jth column of $\boldsymbol{Z}$, $\boldsymbol{Z}_{\cdot j}$, are the numerically coded genetic markers at the jth genetic variant and are assumed to have mean zero and variance 1. Under this model, if the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, for the ith observation,

$$\begin{aligned} \operatorname{var}(\boldsymbol{y}_i) &= \sum_j \boldsymbol{\beta}_j^2 + \sigma_1^2 = \|\boldsymbol{\beta}\|_2^2 + \sigma_1^2, \\ \text{and} \quad \operatorname{var}(\boldsymbol{w}_i) &= \sum_j \boldsymbol{\gamma}_j^2 + \sigma_2^2 = \|\boldsymbol{\gamma}\|_2^2 + \sigma_2^2, \end{aligned}$$

therefore $||\boldsymbol{\beta}||_2^2/(\|\boldsymbol{\beta}\|_2^2 + \sigma_1^2)$ and $||\boldsymbol{\gamma}||_2^2/(\|\boldsymbol{\gamma}\|_2^2 + \sigma_2^2)$ can then be interpreted as the narrow sense heritability (Bulik-Sullivan et al. 2015).

Based on this model, one measure of genetic relatedness is the inner product of the regression coefficients

$$\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle, \tag{2}$$

which measures the shared genetic effects between these two traits. Bulik-Sullivan et al. (2015) defined this quantity as the genetic covariance due to the p genetic variants. Alternatively, a normalized inner product called genetic correlation, that is, the ratio

$$\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle}{\|\boldsymbol{\beta}\|_2 \|\boldsymbol{\gamma}\|_2} \mathbf{1}(\|\boldsymbol{\beta}\|_2 \|\boldsymbol{\gamma}\|_2 > 0), \tag{3}$$

can also be used. In the case where one of $\|\boldsymbol{\beta}\|_2$ and $\|\boldsymbol{\gamma}\|_2$ is vanishing, the ratio is defined as zero, which indicates no correlation between two traits when one of the regression vector is zero. With this normalization, $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is always between -1 and 1 and can be used to compare the genetic relatedness among multiple pairs. Note that to exhibit genetic correlation, the directions of effect must also be consistently aligned.

Although Bulik-Sullivan et al. (2015) defined (2) and (3) as genetic covariance and genetic correlation, they treated $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as random vectors with a particular covariance form and then proposed to apply LD regression to estimate the expectation of $\langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle$. The focus of this article is to develop estimators for $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ based on two GWAS data with genotype data measured on the same set of genetic markers, denoted by $(\boldsymbol{y}_i, \boldsymbol{X}_{i\cdot}, i = 1, \ldots, n_1)$ and $(\boldsymbol{w}_i, \boldsymbol{Z}_{i\cdot}, i = 1, \ldots, n_2)$.

1.3. Methods and Main Results

A naive estimator is to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ first and then plug-in the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ into the expressions (2) and (3). For the problem of interest, usually there are more genetic markers than the sample size, that is, $p \geq \max\{n_1, n_2\}$. However, for any trait, one expects that only a few of these markers have nonzero effects. One can apply any high-dimensional sparse regression methods such as Lasso (Tibshirani 1996), scaled Lasso (Sun and Zhang 2012), and marginal regression with screening (McCarthy et al. 2008; Fan, Han, and Gu 2012) to estimate these sparse regression coefficients. The aforementioned plug-in estimators, however, have several drawbacks in estimating the genetic relatedness. The Lasso approach shrinks the estimation toward 0, in particular, some weak effects might be shrunken to 0, yet the accumulation of these weak effects may contribute significantly to the trait variability. It is possible that some genetic variants may have strong effects on one trait and weak effects on the other trait. Due to shrinkage, the plug-in of Lasso type estimators fails to capture this part of contribution to genetic relatedness from such genetic variants. Marginal regression calculates the regression score between the trait and each single marker (i.e., $\boldsymbol{y}$ and $\boldsymbol{X}_{\cdot j}$, $1 \leq j \leq p$), and screen for the large scores. This approach also suffers from the existence of weak effects, as the marginal scores must be large enough to survive in the screening step.

We propose a two-step procedure to estimate the genetic relatedness measure $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ defined in (2), where step 1 is involved with estimating the inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by the plug-in scaled Lasso estimator, and step 2 is involved with correcting the plug-in scaled Lasso estimator. Similar two-step procedures are proposed to estimate the quadratic functionals $\|\boldsymbol{\beta}\|_2^2$ and $\|\boldsymbol{\gamma}\|_2^2$. To estimate the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ defined in (3), we plug-in the estimators of the inner product and quadratic functionals into the definition (3). Due to the correction step, we name our estimators as functional de-biased estimators (FDEs).

FDEs are shown to achieve the minimax optimal convergence rates of estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The optimality of FDEs results from the unique way of balancing the bias and variance for estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. To illustrate this, we focus on estimation of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, take the plug-in estimator of the scaled Lasso estimators (Sun and Zhang 2012), and the plug-in of the de-biased Lasso estimators (Javanmard and Montanari 2014; van de Geer et al. 2014; Zhang and Zhang 2014) as examples and compare them with FDE estimator of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Note that the scaled Lasso estimator achieves the optimal convergence rate of estimating the whole vector β and the de-biased estimator achieves the optimal convergence rate of estimating the single coordinate β_i. However, simply plugging in the scaled Lasso estimators or the de-biased Lasso estimators does not lead to a good estimator of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ since the plug-in estimator of scaled Lasso estimators suffers from a large bias and the plug-in estimator of de-biased Lasso estimators suffers from the inflation of variance.

In contrast, FDE estimator of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ balances the bias and variance in the optimal way. Specifically, in the correction step of FDE estimator, the bias caused by plugging in the scaled Lasso estimator is corrected through adding the minimum amount of variance. As demonstrated in the simulation studies, FDE consistently outperforms the plug-in estimator of the scaled Lasso estimators and the plug-in estimator of the de-biased Lasso estimators. In addition, FDEs do not suffer from dependency among genetic markers. FDEs work for a broad class of dependency structure of genetic markers.

The theoretical analysis given in Section 3 establishes the optimal convergence rates of estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\|\beta\|_2^2$, and $\|\gamma\|_2^2$. To facilitate the discussion, we control the ℓ_2 norm of regression coefficients β and γ as $c\eta_0 \leq \|\boldsymbol{\beta}\|_2 \leq CM_0$ and $c\eta_0 \leq \|\boldsymbol{\gamma}\|_2 \leq CM_0$ where c, C are positive constant independent of n, p. Here, we present the most interesting regime where the signals are strong in the sense of $\eta_0 \geq C\sqrt{k \log p/n}$, where p is the dimension, n is the sample size, k is the maximum sparsity of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and C is a positive constant independent of k, n, p. We have shown that the optimal rate of convergence for estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\|\boldsymbol{\beta}\|_2^2$, and $\|\boldsymbol{\gamma}\|_2^2$ is

$$M_0\left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n}\right) + \frac{k \log p}{n}.$$

The optimal rate depends not only on p, n, and k, but also the upper bound for the signal strength M_0. In addition, we have shown that the optimal convergence rate of estimating $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is

$$\frac{1}{\eta_0}\left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n}\right) + \frac{1}{\eta_0^2}\frac{k \log p}{n}.$$

In contrast to estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\|\boldsymbol{\beta}\|_2^2$, and $\|\boldsymbol{\gamma}\|_2^2$, the optimal rate scales to the inverse of the lower bound for the signal strength, represented by $1/\eta_0$. The estimators $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$, $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$, and $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ proposed in Section 3 are shown to adaptively achieve the optimal rates for estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathrm{Q}(\boldsymbol{\beta})$, $\mathrm{Q}(\boldsymbol{\gamma})$, and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, respectively.

1.4. Notation and Definitions

Basic notation and definitions used in the rest of the article are defined here. For a matrix $\boldsymbol{X} \in \mathbf{R}^{n\times p}$, $\boldsymbol{X}_{i\cdot}$, $\boldsymbol{X}_{\cdot j}$, and $\boldsymbol{X}_{i,j}$ denote, respectively, the ith row, jth column, and (i, j)th entry of the matrix $\boldsymbol{X}$, $\boldsymbol{X}_{i,-j}$ denotes the ith row of $\boldsymbol{X}$ excluding the jth coordinate, and $\boldsymbol{X}_{-j}$ denotes the submatrix of $\boldsymbol{X}$ excluding the jth column. Let $[p] = \{1, 2, \ldots, p\}$. For a subset $J \subset [p]$, $\boldsymbol{X}_J$ denotes the submatrix of $\boldsymbol{X}$ consisting of columns $\boldsymbol{X}_{\cdot j}$ with $j \in J$ and for a vector $\boldsymbol{x} \in \mathbf{R}^p$, $\boldsymbol{x}_J$ is the subvector of $\boldsymbol{x}$ with indices in J and $\boldsymbol{x}_{-J}$ is the sub-vector with indices in J^c. For a vector $\boldsymbol{x} \in \mathbf{R}^p$, the ℓ_q norm of $\boldsymbol{x}$ is defined as $\|\boldsymbol{x}\|_q = (\sum_{i=1}^q |\boldsymbol{x}_i|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|\boldsymbol{x}\|_0$ denoting the cardinality of nonzero elements of $\boldsymbol{x}$ and $\|\boldsymbol{x}\|_\infty = \max_{1\leq j\leq p} |\boldsymbol{x}_j|$. For a matrix A and $1 \leq q \leq \infty$, $\|\boldsymbol{A}\|_q = \sup_{\|\boldsymbol{x}\|_q=1} \|\boldsymbol{A}\boldsymbol{x}\|_q$ is the matrix ℓ_q operator norm. In particular, $\|\boldsymbol{A}\|_2$ is the spectral norm. For a symmetric matrix $\boldsymbol{A}$, $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote, respectively, the smallest and largest eigenvalue of $\boldsymbol{A}$. For a set S, $|S|$ denotes the cardinality of S. For $a \in \mathbf{R}$, $a_+ = \max\{a, 0\}$ and $\mathrm{sign}(a)$ is the sign of a, that is, $\mathrm{sign}(a) = 1$ if $a > 0$, $\mathrm{sign}(a) = -1$ if $a < 0$ and $\mathrm{sign}(0) = 0$. Define the sub-Gaussian norm $\|\boldsymbol{x}\|_{\psi_2}$ of $\boldsymbol{x} \in \mathbf{R}^p$ as $\|\boldsymbol{x}\|_{\psi_2} = \sup_{\boldsymbol{v}\in S^{p-1}} \sup_{q\geq 1} (\mathbf{E}|\boldsymbol{v}^\intercal \boldsymbol{x}|^q)^{\frac{1}{q}}/\sqrt{q}$ where S^{p-1} is the unit sphere in $\mathbf{R}^p$. The random vector $\boldsymbol{x} \in \mathbf{R}^p$ is defined to be sub-Gaussian if its corresponding sub-Gaussian norm is bounded; see Vershynin (2012) for more on sub-Gaussian random variables. For the design matrices $\boldsymbol{X} \in \mathbf{R}^{n_1\times p}$ and $\boldsymbol{Z} \in \mathbf{R}^{n_2\times p}$, we define the corresponding sample covariance matrices as $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\intercal \boldsymbol{X}/n_1$ and $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{Z}^\intercal \boldsymbol{Z}/n_2$. Let $z_{\alpha/2}$ denote the upper $\alpha/2$ quantile of the standard normal distribution. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \leq Cb_n$ for all n, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. c and C are used to denote generic positive constants that may vary from place to place. For any two sequences of numbers a_n and b_n, we will write $b_n \ll a_n$ if $\limsup b_n/a_n = 0$.

1.5. Organization of the Article

The rest of the article is organized as follows. Section 2 presents the procedures for estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathrm{Q}(\boldsymbol{\beta})$, $\mathrm{Q}(\boldsymbol{\gamma})$, and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in details. In Section 3, minimax convergence rates for the estimation problems are established and the proposed estimators are shown to attain the optimal rates. In Section 4, simulation studies are conducted to evaluate the empirical performance of FDEs. A yeast cross data is used to illustrate the estimators in Section 5. Discussion is provided in Section 6. The proofs of main theorems are present in Section 6. The remaining proofs and the extended simulation studies are given in the supplementary materials.

2. Estimation Methods

2.1. Estimation of $\mathbf{I}(\beta, \gamma)$

Since the inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is of significant interest in its own right, we first consider the estimation of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle$.

The scaled Lasso estimators for high-dimensional linear model (1) are defined through the following optimization algorithm (Sun and Zhang 2012),

$$\{\widehat{\boldsymbol{\beta}}, \hat{\sigma}_1\} = \arg \min_{\boldsymbol{\beta} \in \mathbf{R}^p, \sigma_1 \in \mathbf{R}^+} \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2n_1\sigma_1} + \frac{\sigma_1}{2} + \frac{\lambda_0}{\sqrt{n_1}} \sum_{j=1}^{p} \frac{\|\boldsymbol{X}_{\cdot j}\|_2}{\sqrt{n_1}} |\boldsymbol{\beta}_j|, \tag{4}$$

and

$$\{\widehat{\boldsymbol{\gamma}}, \hat{\sigma}_2\} = \arg \min_{\boldsymbol{\gamma} \in \mathbf{R}^p, \sigma_2 \in \mathbf{R}^+} \frac{\|\boldsymbol{w} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2}{2n_2\sigma_2} + \frac{\sigma_2}{2} + \frac{\lambda_0}{\sqrt{n_2}} \sum_{j=1}^{p} \frac{\|\boldsymbol{Z}_{\cdot j}\|_2}{\sqrt{n_2}} |\boldsymbol{\gamma}_j|, \tag{5}$$

where $\lambda_0 = \sqrt{2.01 \log p}$. To construct an optimal estimator of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, it is helpful to analyze the error of the plug-in estimator $\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} \rangle$,

$$\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} \rangle - \langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle = \langle \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + \langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \rangle - \langle \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \rangle. \tag{6}$$

The last term on the right-hand side, $\langle \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \rangle$ is "small," but the first two terms $\langle \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$ and $\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \rangle$ can be large. This provides the motivation for the proposed estimator, where we first estimate these two terms and then subtract them from $\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} \rangle$ to obtain the final estimator of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

The intuition for estimating $\langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \rangle$ is given first. Since

$$\frac{1}{n_1} \boldsymbol{X}^\intercal (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \frac{1}{n_1} \boldsymbol{X}^\intercal \epsilon, \tag{7}$$

multiplying both sides of (7) by a vector $\boldsymbol{u} \in \mathbf{R}^p$ yields

$$\frac{1}{n_1} \boldsymbol{u}^\intercal \boldsymbol{X}^\intercal (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{u}^\intercal \widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \frac{1}{n_1} \boldsymbol{u}^\intercal \boldsymbol{X}^\intercal \boldsymbol{\epsilon}, \tag{8}$$

which can be written as

$$\begin{aligned} &\frac{1}{n_1} \boldsymbol{u}^\intercal \boldsymbol{X}^\intercal (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) - \langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \rangle \\ &\quad = (\widehat{\boldsymbol{\Sigma}}\boldsymbol{u} - \widehat{\boldsymbol{\gamma}})^\intercal (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \frac{1}{n_1} \boldsymbol{u}^\intercal \boldsymbol{X}^\intercal \boldsymbol{\epsilon}. \end{aligned} \tag{9}$$

If the vector $\boldsymbol{u}$ can be chosen such that the right-hand side of (9) is "small," then $\boldsymbol{u}^\intercal \boldsymbol{X}^\intercal (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n_1$ is a good estimator of $\langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \rangle$. Since the first term on the right-hand side of (9) is upper bounded as $|(\widehat{\boldsymbol{\Sigma}}\boldsymbol{u} - \widehat{\boldsymbol{\gamma}})^\intercal (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})| \leq \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{u} - \widehat{\boldsymbol{\gamma}}\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$, we control the right-hand side of (9) through constructing a projection vector $\boldsymbol{u}$ such that $\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{u} - \widehat{\boldsymbol{\gamma}}\|_\infty$ is constrained and the second term of (9) $\boldsymbol{u}^\intercal \boldsymbol{X}^\intercal \boldsymbol{\epsilon}/n_1$ is controlled through minimizing its variance $\sigma_1^2 \boldsymbol{u}^\intercal \widehat{\boldsymbol{\Sigma}} \boldsymbol{u}/n_1$. This leads to the following convex optimization algorithm for identifying the projection vector $\boldsymbol{u}$ for estimating $\langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \rangle$,

$$\widehat{\boldsymbol{u}}_1 = \arg\min_{\boldsymbol{u} \in \mathbf{R}^p} \left\{ \boldsymbol{u}^\intercal \widehat{\boldsymbol{\Sigma}} \boldsymbol{u} : \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{u} - \widehat{\boldsymbol{\gamma}}\|_\infty \leq \|\widehat{\boldsymbol{\gamma}}\|_2 \frac{\lambda_1}{\sqrt{n_1}} \right\}, \tag{10}$$

where $\lambda_1 = 12\lambda_{\max}^2(\boldsymbol{\Sigma})\sqrt{\log p}$.

Remark 1. The solution of the above optimization problem might not be unique and $\widehat{\boldsymbol{u}}_1$ is defined as any minimizer of the optimization problem. The theory established in Section 3 still holds for any minimizer of (10). The optimization problem (10) is solved through its equivalent Lagrange dual problem, which is computationally efficient and scales well to the high-dimensional problem. See Step 2 in Table 1 for more details.

Once the projection vector $\widehat{\boldsymbol{u}}_1$ is obtained, $\langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \rangle$ is then estimated by $\widehat{\boldsymbol{u}}_1^\intercal \boldsymbol{X}^\intercal (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n_1$. Similarly, the projection vector for estimating $\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}} \rangle$ can be obtained via the convex algorithm

$$\widehat{\boldsymbol{u}}_2 = \arg\min_{\boldsymbol{u} \in \mathbf{R}^p} \left\{ \boldsymbol{u}^\intercal \widehat{\boldsymbol{\Gamma}} \boldsymbol{u} : \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{u} - \widehat{\boldsymbol{\beta}}\|_\infty \leq \|\widehat{\boldsymbol{\beta}}\|_2 \frac{\lambda_2}{\sqrt{n_2}} \right\}, \tag{11}$$

where $\lambda_2 = 12\lambda_{\max}^2(\boldsymbol{\Gamma})\sqrt{\log p}$. Then $\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}} \rangle$ is estimated by $\widehat{\boldsymbol{u}}_2^\intercal \boldsymbol{Z}^\intercal (\boldsymbol{w} - \boldsymbol{Z}\widehat{\boldsymbol{\gamma}})/n_2$.

Table 1. FDE algorithm without sample splitting for estimating the inner product, quadratic functionals, and the normalized inner product.

Input: design matrices: $\boldsymbol{X}, \boldsymbol{Z}$; response vectors: $\boldsymbol{y}, \boldsymbol{w}$; tuning parameters λ_0, λ.
Output: $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$, $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$, and $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

1. Scaled Lasso: Calculate $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ from (4) and (5) with the tuning parameter λ_0.

Inner product calculation:

2. Projection vector $\widehat{\boldsymbol{u}}_1$: Calculate $\widehat{\boldsymbol{u}}_1 = \arg\min_{\boldsymbol{u}} \boldsymbol{u}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{u}/4n_1 + \boldsymbol{u}^\intercal \widehat{\boldsymbol{\gamma}} + \lambda^t \|\boldsymbol{u}\|_1$, where $\lambda^t = \lambda^{t-1}/1.5$, and $\lambda^0 = \lambda/\sqrt{n_1}$. Repeat until $\widehat{\boldsymbol{u}}_1$ cannot be solved with λ^t replaced by λ^{t+1}, or $t \geq 10$.
3. Projection vector $\widehat{\boldsymbol{u}}_2$: Calculate $\widehat{\boldsymbol{u}}_2 = \arg\min_{\boldsymbol{u}} \boldsymbol{u}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{Z} \boldsymbol{u}/4n_2 + \boldsymbol{u}^\intercal \widehat{\boldsymbol{\beta}} + \lambda^t \|\boldsymbol{u}\|_1$, where $\lambda^t = \lambda^{t-1}/1.5$, and $\lambda^0 = \lambda/\sqrt{n_1}$. Repeat until $\widehat{\boldsymbol{u}}_2$ cannot be solved with λ^t replaced by λ^{t+1}, or $t \geq 10$.
4. Correction: $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} \rangle + \widehat{\boldsymbol{u}}_1^\intercal \boldsymbol{X}^\intercal (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n_1 + \widehat{\boldsymbol{u}}_2^\intercal \boldsymbol{Z}^\intercal (\boldsymbol{w} - \boldsymbol{Z}\widehat{\boldsymbol{\gamma}})/n_2$.

Quadratic functional calculation:

5. Projection vector $\widehat{\boldsymbol{u}}_3$: Calculate $\widehat{\boldsymbol{u}}_3 = \arg\min_{\boldsymbol{u}} \boldsymbol{u}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{u}/4n_1 + \boldsymbol{u}^\intercal \widehat{\boldsymbol{\beta}} + \lambda^t \|\boldsymbol{u}\|_1$, where $\lambda^t = \lambda^{t-1}/1.5$, and $\lambda^0 = \lambda/\sqrt{n_1}$. Repeat until $\widehat{\boldsymbol{u}}_3$ cannot be solved with λ^t replaced by λ^{t+1}, or $t \geq 10$.
6. Projection vector $\widehat{\boldsymbol{u}}_4$: Calculate $\widehat{\boldsymbol{u}}_4 = \arg\min_{\boldsymbol{u}} \boldsymbol{u}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{Z} \boldsymbol{u}/4n_2 + \boldsymbol{u}^\intercal \widehat{\boldsymbol{\gamma}} + \lambda^t \|\boldsymbol{u}\|_1$, where $\lambda^t = \lambda^{t-1}/1.5$, and $\lambda^0 = \lambda/\sqrt{n_2}$. Repeat until $\widehat{\boldsymbol{u}}_1$ cannot be solved with λ^t replaced by λ^{t+1}, or $t \geq 10$.
7. Correction: $\widehat{\mathrm{Q}}(\boldsymbol{\beta}) = \left(\|\widehat{\boldsymbol{\beta}}\|^2 + 2\widehat{\boldsymbol{u}}_3^\intercal \boldsymbol{X}^\intercal (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n_1\right)_+$, $\widehat{\mathrm{Q}}(\boldsymbol{\gamma}) = \left(\|\widehat{\boldsymbol{\gamma}}\|^2 + 2\widehat{\boldsymbol{u}}_4^\intercal \boldsymbol{Z}^\intercal (\boldsymbol{w} - \boldsymbol{Z}\widehat{\boldsymbol{\gamma}})/n_2\right)_+$.

Ratio calculation:

8. $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathrm{sign}(\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})) \cdot \min\{\left(|\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})|/\sqrt{\widehat{\mathrm{Q}}(\boldsymbol{\beta})\widehat{\mathrm{Q}}(\boldsymbol{\gamma})}\right)\mathbf{1}\{\widehat{\mathrm{Q}}(\boldsymbol{\beta})\widehat{\mathrm{Q}}(\boldsymbol{\gamma}) > 0\}, 1\}$.

The final estimator $\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ of $\mathrm{I}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is given by

$$\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\rangle + \widehat{\boldsymbol{u}}_1^{\top} \frac{1}{n_1} \boldsymbol{X}^{\top} \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) + \widehat{\boldsymbol{u}}_2^{\top} \frac{1}{n_2}\boldsymbol{Z}^{\top}\left(\boldsymbol{w}-\boldsymbol{Z}\widehat{\boldsymbol{\gamma}}\right). \tag{12}$$

It is clear from the above discussion that the key idea for the construction of the final estimator $\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is to identify the projection vectors $\widehat{\boldsymbol{u}}_1$ and $\widehat{\boldsymbol{u}}_2$ such that $\langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\rangle$ and $\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\rangle$ are well approximated. It will be shown in Section 3 that the estimator $\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is adaptively minimax rate-optimal.

Remark 2. As mentioned, simply plugging in the Lasso, scaled Lasso, or de-biased estimator does not lead to a good estimator of $\mathrm{I}(\boldsymbol{\beta},\boldsymbol{\gamma})$. Another natural approach is to first threshold the de-biased estimator to obtain a sparse estimator of the coefficient vectors (see details in Zhang and Zhang (2014, sec 3.3), Guo et al. (2016, eq. (10))) and then plug-in this thresholded estimator. This estimator is referred to as the thresholded estimator. Simulations in Section 4 demonstrate that the proposed estimator defined in (12) outperforms the three plug-in estimators using the scaled Lasso, de-biased, and thresholded estimators.

2.2. Estimation of $\mathrm{Q}(\beta)$ and $\mathrm{Q}(\gamma)$

To estimate the normalized inner product $\mathrm{R}(\boldsymbol{\beta},\boldsymbol{\gamma})$, it is necessary to estimate the quadratic functionals $\mathrm{Q}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ and $\mathrm{Q}(\boldsymbol{\gamma}) = \|\boldsymbol{\gamma}\|_2^2$. To this end, we randomly split the data $(\boldsymbol{y},\boldsymbol{X})$ into two subsamples $(\boldsymbol{y}^{(1)},\boldsymbol{X}^{(1)})$ with sample size $n_1/2$ and $(\boldsymbol{y}^{(2)},\boldsymbol{X}^{(2)})$ with sample size $n_1/2$ and the data $(\boldsymbol{w},\boldsymbol{Z})$ into two subsamples $(\boldsymbol{w}^{(1)},\boldsymbol{Z}^{(1)})$ with sample size $n_2/2$ and $(\boldsymbol{w}^{(2)},\boldsymbol{Z}^{(2)})$ with sample size $n_2/2$.

With a slight abuse of notation, let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ denote the optimizers of the scaled Lasso algorithm (4) applied to $(\boldsymbol{y}^{(1)},\boldsymbol{X}^{(1)})$ and (5) applied to $(\boldsymbol{w}^{(1)},\boldsymbol{Z}^{(1)})$, respectively. For the scaled Lasso algorithms, the sample sizes n_1 and n_2 are replaced by $n_1/2$ and $n_2/2$, respectively. Again, the simple plug-in estimator $\|\widehat{\boldsymbol{\beta}}\|_2^2$ of $\mathrm{Q}(\boldsymbol{\beta})$ is not a good estimator of $\|\boldsymbol{\beta}\|_2^2$ because of the following error decomposition,

$$\|\widehat{\boldsymbol{\beta}}\|_2^2 - \|\boldsymbol{\beta}\|_2^2 = 2\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle - \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2, \tag{13}$$

where the second term on the right-hand side of (13) is "small," but the first can be large. Specifically, the term $2\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\rangle$ is estimated first and then is added to $\|\widehat{\boldsymbol{\beta}}\|_2^2$ to obtain the final estimator of $\|\boldsymbol{\beta}\|_2^2$. To estimate $\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\rangle$, a projection vector $\boldsymbol{u}$ is identified such that the following difference is controlled,

$$\begin{aligned} &\frac{1}{n_1/2}\boldsymbol{u}^{\top}(\boldsymbol{X}^{(2)})^{\top}(\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}) - \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\rangle \\ &\quad = (\boldsymbol{u}^{\top}\widehat{\boldsymbol{\Sigma}}^{(2)} - \widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \frac{1}{n_1/2}\boldsymbol{u}^{\top}(\boldsymbol{X}^{(2)})^{\top}\boldsymbol{\epsilon}, \end{aligned} \tag{14}$$

with $\widehat{\boldsymbol{\Sigma}}^{(2)} = (\boldsymbol{X}^{(2)})^{\top}\boldsymbol{X}^{(2)}/(n_1/2)$. Define the projection vector $\widehat{\boldsymbol{u}}_3$ as the solution to the following optimization algorithm:

$$\widehat{\boldsymbol{u}}_3 = \arg\min_{\boldsymbol{u}\in\mathbf{R}^p}\left\{\boldsymbol{u}^{\top}\widehat{\boldsymbol{\Sigma}}^{(2)}\boldsymbol{u} : \|\widehat{\boldsymbol{\Sigma}}^{(2)}\boldsymbol{u} - \widehat{\boldsymbol{\beta}}\|_\infty \le \|\widehat{\boldsymbol{\beta}}\|_2\frac{\lambda_1}{\sqrt{n_1/2}}\right\}, \tag{15}$$

where $\lambda_1 = 12\lambda_{\max}^2(\boldsymbol{\Sigma})\sqrt{\log p}$. We then estimate $\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\rangle$ by $\widehat{\boldsymbol{u}}_3^{\top}(\boldsymbol{X}^{(2)})^{\top}(\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}})/(n_1/2)$ and propose the final estimator of $\|\boldsymbol{\beta}\|_2^2$ as

$$\widehat{\mathrm{Q}}(\boldsymbol{\beta}) = \left(\|\widehat{\boldsymbol{\beta}}\|_2^2 + 2\widehat{\boldsymbol{u}}_3^{\top}\frac{1}{n_1/2}\left(\boldsymbol{X}^{(2)}\right)^{\top}\left(\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}\right)\right)_+ . \tag{16}$$

Similarly, the estimator of $\|\boldsymbol{\gamma}\|_2^2$ is given by

$$\widehat{\mathrm{Q}}(\boldsymbol{\gamma}) = \left(\|\widehat{\boldsymbol{\gamma}}\|_2^2 + 2\widehat{\boldsymbol{u}}_4^{\top}\frac{1}{n_2/2}\left(\boldsymbol{Z}^{(2)}\right)^{\top}\left(\boldsymbol{w}^{(2)} - \boldsymbol{Z}^{(2)}\widehat{\boldsymbol{\gamma}}\right)\right)_+ , \tag{17}$$

where

$$\widehat{\boldsymbol{u}}_4 = \arg\min_{\boldsymbol{u}}\left\{\boldsymbol{u}^{\top}\widehat{\boldsymbol{\Gamma}}^{(2)}\boldsymbol{u} : \|\widehat{\boldsymbol{\Gamma}}^{(2)}\boldsymbol{u} - \widehat{\boldsymbol{\gamma}}\|_\infty \le \|\widehat{\boldsymbol{\gamma}}\|_2\frac{\lambda_2}{\sqrt{n_2/2}}\right\}, \tag{18}$$

with $\widehat{\boldsymbol{\Gamma}}^{(2)} = (\boldsymbol{Z}^{(2)})^{\top}(\boldsymbol{Z}^{(2)})/(n_2/2)$ and $\lambda_2 = 12\lambda_{\max}^2(\boldsymbol{\Gamma})\sqrt{\log p}$.

Remark 3. Sample splitting is used here for the purpose of the theoretical analysis. In the simulation study (Section 4), the performance of the proposed estimator without sample splitting is investigated; see Steps 5–7 in Table 1. The proposed estimator without sample splitting performs even better numerically than with sample splitting since more observations are used in constructing the initial estimators $\|\widehat{\boldsymbol{\beta}}\|_2^2$ and $\|\widehat{\boldsymbol{\gamma}}\|_2^2$ and the projection vectors $\widehat{\boldsymbol{u}}_3$ and $\widehat{\boldsymbol{u}}_4$.

2.3. Estimation of $\mathrm{R}(\beta, \gamma)$

Given the estimators $\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})$, $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$, and $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$ constructed in Sections 2.1 and 2.2, a natural estimator for the normalized inner product $\mathrm{R}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is given by

$$\begin{aligned} \widehat{\mathrm{R}}(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \mathrm{sign}\left(\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})\right) \\ &\quad \cdot \min\left\{\frac{|\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})|}{\sqrt{\widehat{\mathrm{Q}}(\boldsymbol{\beta})\widehat{\mathrm{Q}}(\boldsymbol{\gamma})}}\mathbf{1}\left\{\widehat{\mathrm{Q}}(\boldsymbol{\beta})\widehat{\mathrm{Q}}(\boldsymbol{\gamma}) > 0\right\}, 1\right\}, \end{aligned} \tag{19}$$

where $\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})$, $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$, and $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$ are estimators of $\langle\boldsymbol{\beta},\boldsymbol{\gamma}\rangle$, $\|\boldsymbol{\beta}\|_2^2$, and $\|\boldsymbol{\gamma}\|_2^2$ defined in (12), (16), and (17), respectively. It is possible that one of $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$ and $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$ is 0 if $\|\boldsymbol{\beta}\|_2^2$ and $\|\boldsymbol{\gamma}\|_2^2$ are close to zero. In this case, the normalized inner product $\mathrm{R}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is estimated as 0. Since $\mathrm{R}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is always between -1 and 1, the estimator $\widehat{\mathrm{R}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is truncated to ensure that it is within the range. The FDE algorithm without sample splitting for calculating the estimators $\widehat{\mathrm{I}}(\boldsymbol{\beta},\boldsymbol{\gamma})$, $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$, $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$, and $\widehat{\mathrm{R}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is detailed in Table 1.

3. Theoretical Analysis

3.1. Upper Bound Analysis

The samples sizes n_1 and n_2 are assumed to be of the same order, that is, $n_1 \asymp n_2$. Let $n = \min\{n_1, n_2\}$ be the smallest of two sample sizes. The following assumptions are introduced to facilitate the theoretical analysis.

(A1) The population covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ satisfy $1/M_1 \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le M_1$, and $1/M_1 \le \lambda_{\min}(\boldsymbol{\Gamma}) \le \lambda_{\max}(\boldsymbol{\Gamma}) \le M_1$, where $M_1 \ge 1$ is a positive constant. The random design matrix $\boldsymbol{X}$ is assumed to be independent of the other random design matrix $\boldsymbol{Z}$. The noise levels σ_1 and σ_2 satisfy $\max\{\sigma_1, \sigma_2\} \le M_2$, where $M_2 > 0$ is a positive constant.

(A2) The ℓ_2 norms of the coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are bounded away from zero in the sense that

$$\min \ \{\|\boldsymbol{\beta}\|_2, \|\boldsymbol{\gamma}\|_2\} \ge \eta_0 \ge C\sqrt{k\log p/n}, \quad \text{where} \quad k = \max\{\|\boldsymbol{\beta}\|_0, \|\boldsymbol{\gamma}\|_0\}. \tag{20}$$

Assumption (A1) places a condition on the spectrum of the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ and an upper bound on the noise levels σ_1 and σ_2. Assumption (A2) requires that the total strengths of the signals have to be bounded away from zero by η_0, which is only used in the upper bound analysis of the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

The following theorem establishes the convergence rates of the estimators $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$, and $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$, proposed in (12), (16), and (17), respectively.

Theorem 1. Suppose the assumption (A1) holds and $k = \max\{\|\boldsymbol{\beta}\|_0, \|\boldsymbol{\gamma}\|_0\} \le cn/\log p$ for some $c > 0$. Then for any fixed constant $0 < \alpha < 1/4$, with probability at least $1 - 4\alpha - p^{-c_0}$, we have

$$\left|\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})\right| \lesssim (\|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\gamma}\|_2) \times \left(\frac{z_{\alpha/2}}{\sqrt{n}} + \frac{k\log p}{n}\right) + \frac{k\log p}{n}, \tag{21}$$

$$\left|\widehat{\mathrm{Q}}(\boldsymbol{\beta}) - \mathrm{Q}(\boldsymbol{\beta})\right| \lesssim \|\boldsymbol{\beta}\|_2 \left(\frac{z_{\alpha/2}}{\sqrt{n}} + \frac{k\log p}{n}\right) + \frac{k\log p}{n}, \tag{22}$$

$$\left|\widehat{\mathrm{Q}}(\boldsymbol{\gamma}) - \mathrm{Q}(\boldsymbol{\gamma})\right| \lesssim \|\boldsymbol{\gamma}\|_2 \left(\frac{z_{\alpha/2}}{\sqrt{n}} + \frac{k\log p}{n}\right) + \frac{k\log p}{n}, \tag{23}$$

where c_0 is a positive constant.

The upper bound of estimating $\langle\boldsymbol{\beta}, \boldsymbol{\gamma}\rangle$ not only depends on k, n, and p, but also scales to the signal strengths $\|\boldsymbol{\beta}\|_2$ and $\|\boldsymbol{\gamma}\|_2$. For the estimation of the quadratic functional $\mathrm{Q}(\boldsymbol{\beta})$ (or $\mathrm{Q}(\boldsymbol{\gamma})$), the convergence rate depends on $\|\boldsymbol{\beta}\|_2$ (or $\|\boldsymbol{\gamma}\|_2$). The following theorem establishes the convergence rate of the estimator $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ proposed in (19).

Theorem 2. Suppose the assumptions (A1) and (A2) hold and $k = \max\{\|\boldsymbol{\beta}\|_0, \|\boldsymbol{\gamma}\|_0\} \le cn/\log p$ for some $c > 0$. Then for any fixed constant $0 < \alpha < 1/4$, with probability at least $1 - 4\alpha - p^{-c_0}$, we have

$$\left|\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})\right| \lesssim \frac{1}{\eta_0}\left(\frac{z_{\alpha/2}}{\sqrt{n}} + \frac{k\log p}{n}\right) + \frac{1}{\eta_0^2}\frac{k\log p}{n}, \tag{24}$$

where c_0 is a positive constant.

In contrast to Theorem 1, Theorem 2 requires the extra assumption (A2) on the signal strengths $\|\boldsymbol{\beta}\|_2$ and $\|\boldsymbol{\gamma}\|_2$. The convergence rate of estimating $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is scaled to the inverse of the signal strength, $1/\eta_0$. This is different from the error bound in Theorem 1, where the estimation accuracy is scaled to the signal strength. The lower bound results established in Theorem 3 will demonstrate the necessity of Assumption (A2) for estimation of $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

3.2. Minimax Lower Bounds

This section establishes the minimax lower bounds of estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathrm{Q}(\boldsymbol{\beta})$, $\mathrm{Q}(\boldsymbol{\gamma})$, and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. We first introduce parameter spaces for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1, \boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2)$, which is defined as the product of parameter spaces for $\boldsymbol{\theta}_{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1)$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2)$. We define the following parameter space for both $\boldsymbol{\theta}_{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1)$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2)$,

$$\mathcal{G}(k, M_0) = \left\{(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1) : \|\boldsymbol{\beta}\|_0 \le k, \ \|\boldsymbol{\beta}\|_2 \le M_0, \ \frac{1}{M_1} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le M_1, \ \sigma_1 \le M_2\right\}, \tag{25}$$

where $M_1 \ge 1$ and $M_2 > 0$ are positive constants. The parameter space defined in (25) requires that the signal β contains less than k nonzero coefficients and the ℓ_2 norm $\|\beta\|_2$ is upper bounded by M_0, where M_0 is allowed to grow with n and p. The lower bound results in Theorem 3 show that the estimation difficulties of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathrm{Q}(\boldsymbol{\beta})$, and $\mathrm{Q}(\boldsymbol{\gamma})$ depend on M_0. The other conditions $1/M_1 \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le M_1$ and $\sigma_1 \le M_2$ are regularity conditions. Based on the definition (25), the parameter space for $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1, \boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2)$ is defined as a product of two parameter spaces,

$$\Theta(k, M_0) = \{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1, \boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2) \ : \ (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1) \in \mathcal{G}(k, M_0), \ (\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2) \in \mathcal{G}(k, M_0)\}. \tag{26}$$

For establishing optimal bounds of $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, we define the following parameter space

$$\Theta(k, \eta_0) = \{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1, \boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2) \ : \ (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1) \in \mathcal{G}(k, \eta_0), \ (\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \sigma_2) \in \mathcal{G}(k, \eta_0)\}, \tag{27}$$

where

$$\mathcal{G}(k, \eta_0) = \left\{(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_1) : \|\boldsymbol{\beta}\|_0 \le k, \ \|\boldsymbol{\beta}\|_2 \ge \eta_0, \ \frac{1}{M_1} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le M_1, \ \sigma_1 \le M_2\right\},$$

with $\eta_0 \ge 0$. In contrast to the parameter space $\mathcal{G}(k, M_0)$ where $\|\beta\|_2$ is upper bounded by M_0, the parameter space $\mathcal{G}(k, \eta_0)$ requires the signal strength $\|\beta\|_2$ to be lower bounded by η_0, where η_0 is allowed to grow with n and p. The lower bound in Theorem 3 shows that the estimation difficulty of $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ depends on η_0.

The following theorem establishes the minimax lower bounds for the convergence rates of estimating the inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, the quadratic functionals $\mathrm{Q}(\boldsymbol{\beta})$ and $\mathrm{Q}(\boldsymbol{\gamma})$ and the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

Theorem 3. Suppose $k \le c\min\{n/\log p, p^{\nu}\}$ for some constants $c > 0$ and $0 \le \nu < \frac{1}{2}$. Then

$$\inf_{\widetilde{\mathrm{I}}} \sup_{\boldsymbol{\theta}\in\Theta(k,M_0)} \mathbb{P}_{\boldsymbol{\theta}}\left(\left|\widetilde{\mathrm{I}} - \mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})\right| \gtrsim \min\left\{M_0\left(\frac{1}{\sqrt{n}} + \frac{k\log p}{n}\right) + \frac{k\log p}{n}, M_0^2\right\}\right) \ge \frac{1}{4}, \tag{28}$$

$$\inf_{\widetilde{\mathrm{Q}}} \sup_{\boldsymbol{\theta}_{\boldsymbol{\beta}} \in \mathcal{G}(k,M_0)} \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{\beta}}} \Big(\big|\widetilde{\mathrm{Q}} - \mathrm{Q}(\boldsymbol{\beta})\big| \gtrsim \min\left\{ M_0 \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) + \frac{k \log p}{n}, M_0^2 \right\} \Big) \geq \frac{1}{4}, \quad (29)$$

$$\inf_{\widetilde{\mathrm{Q}}} \sup_{\boldsymbol{\theta}_{\boldsymbol{\gamma}} \in \mathcal{G}(k,M_0)} \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}} \left(\big|\widetilde{\mathrm{Q}} - \mathrm{Q}(\boldsymbol{\gamma})\big| \gtrsim \min\left\{ M_0 \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) + \frac{k \log p}{n}, M_0^2 \right\} \right) \geq \frac{1}{4}, \quad (30)$$

$$\inf_{\widetilde{\mathrm{R}}} \sup_{\boldsymbol{\theta} \in \Theta(k,\eta_0)} \mathbb{P}_{\boldsymbol{\theta}} \Big(\big|\widetilde{\mathrm{R}} - \mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})\big| \gtrsim \min\left\{ \frac{1}{\eta_0} \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) + \frac{1}{\eta_0^2} \frac{k \log p}{n}, 1 \right\} \Big) \geq \frac{1}{4}. \quad (31)$$

Remark 4. Estimation of quadratic functionals has been extensively studied in the classical Gaussian sequence model. See, for example, Donoho and Nussbaum (1990); Efromovich and Low (1996); Laurent and Massart (2000); Cai and Low (2005, 2006); Collier, Comminges, and Tsybakov (2015) for details. In the regime $k \leq c \min\{n/\log p, p^{\nu}\}$ for some constants $c > 0$ and $0 < \nu < \frac{1}{2}$, Theorem 2 in Collier, Comminges, and Tsybakov (2015) gives a lower bound, $\min\{M_0/\sqrt{n} + k \log p/n, M_0^2\}$, for estimating $\|\beta\|_2^2$ in the sequence model. In contrast, an extra term $M_0 k \log p/n$ appears in the lower bound given in (29) for estimating $\|\beta\|_2^2$ in high-dimensional linear regression. One intuitive reason for this extra term is that high-dimensional linear regression is involved with an extra inverse process than Gaussian sequence model. Estimation of the quadratic functional $\|\beta\|_2^2$ in high-dimensional linear regression is fundamentally harder than that in the Gaussian sequence model. For the high-dimensional linear regression, the estimation lower bound $k \log p/n$ in (29) can also be established by the general lower bounds developed in Cai and Guo (2017). See Section 8 in Cai and Guo (2017) for details.

3.3. Optimality of FDEs

In this section, we establish the optimality of FDEs by combining Theorems 1 and 2 over the parameter spaces $\Theta(k, M_0)$ and $\Theta(k, \eta_0)$ defined in (26) and (27), respectively.

Corollary 1. Suppose $k \leq c \min\{n/\log p, p^{\nu}\}$ and $M_0 \geq \eta_0 \geq C\sqrt{k \log p/n}$ for some constants $C, c > 0$ and $0 \leq \nu < \frac{1}{2}$. Then

$$\sup_{\boldsymbol{\theta} \in \Theta(k,M_0)} \mathbb{P}_{\boldsymbol{\theta}} \left(\big|\widehat{\mathrm{I}} - \mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})\big| \gtrsim M_0 \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) + \frac{k \log p}{n} \right) \geq 1 - 4\alpha - p^{-c_0}, \quad (32)$$

$$\sup_{\boldsymbol{\theta}_{\boldsymbol{\beta}} \in \mathcal{G}(k,M_0)} \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{\beta}}} \left(\big|\widehat{\mathrm{Q}} - \mathrm{Q}(\boldsymbol{\beta})\big| \gtrsim M_0 \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) + \frac{k \log p}{n} \right) \geq 1 - 4\alpha - p^{-c_0}, \quad (33)$$

$$\sup_{\boldsymbol{\theta}_{\boldsymbol{\gamma}} \in \mathcal{G}(k,M_0)} \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}} \left(\big|\widehat{\mathrm{Q}} - \mathrm{Q}(\boldsymbol{\gamma})\big| \gtrsim M_0 \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) + \frac{k \log p}{n} \right) \geq 1 - 4\alpha - p^{-c_0}, \quad (34)$$

$$\sup_{\boldsymbol{\theta} \in \Theta(k,\eta_0)} \mathbb{P}_{\boldsymbol{\theta}} \left(\big|\widehat{\mathrm{R}} - \mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})\big| \gtrsim \frac{1}{\eta_0} \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) + \frac{1}{\eta_0^2} \frac{k \log p}{n} \right) \geq 1 - 4\alpha - p^{-c_0}, \quad (35)$$

where $0 < \alpha < 1/4$ and c_0 is a positive constant.

Combined with Theorem 3, Corollary 1 implies that, for $M_0 \geq C\sqrt{k \log p/n}$, the estimators $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\widehat{\mathrm{Q}}(\boldsymbol{\beta})$, $\widehat{\mathrm{Q}}(\boldsymbol{\gamma})$ proposed in (12), (16), and (17) achieve the minimax lower bounds (28), (29), and (30) within a constant factor, that is, FDEs are minimax rate-optimal. On the other hand, if $M_0 \ll \sqrt{k \log p/n}$, estimation of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathrm{Q}(\boldsymbol{\beta})$, and $\mathrm{Q}(\boldsymbol{\gamma})$ is uninteresting as the trivial estimator 0 achieves the minimax lower bound in this case. For estimation of $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, under the assumption $\eta_0 \geq C\sqrt{k \log p/n}$, Corollary 1 shows that the estimator $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ given in (19) achieves the minimax lower bound $1/\eta_0 \times (1/\sqrt{n} + k \log p/n) + 1/\eta_0^2 \times k \log p/n$ in (31). Hence, $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is the rate-optimal estimator of $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ under the assumption (A2). When $\eta_0 \ll \sqrt{k \log p/n}$, estimation of $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ becomes trivial as the simple estimator 0 attains the minimax lower bound. This demonstrates the necessity of the assumption (A2) in Theorem 2.

4. Simulation Evaluations and Comparisons

We compare the finite-sample performance of several estimators of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ using simulations. These estimators included plug-in scaled Lasso estimator (Sun and Zhang 2012), plug-in de-biased estimator (Javanmard and Montanari 2014; van de Geer et al. 2014; Zhang and Zhang 2014), plug-in thresholded estimator (Zhang and Zhang 2014, sec. 3.3), and the proposed estimator FDE. Specifically, they are defined as

- FDE: The inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in (12) and the ratio $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in (19). We consider FDE with sample splitting (FDE-S) and without sample splitting (FDE-NS) for $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.
- Plug-in scaled Lasso estimator (Lasso): The inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by $\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} \rangle$ and the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by
$$[\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} \rangle / (\|\widehat{\boldsymbol{\beta}}\|_2 \|\widehat{\boldsymbol{\gamma}}\|_2)] \mathbf{1}\{\|\widehat{\boldsymbol{\beta}}\|_2 \|\widehat{\boldsymbol{\gamma}}\|_2 > 0\}.$$
- Plug-in de-biased estimator (De-biased): Denote the de-biased Lasso estimators as $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$. The inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by $\langle \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}} \rangle$ and the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by $[\langle \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}} \rangle / (\|\widetilde{\boldsymbol{\beta}}\|_2 \|\widetilde{\boldsymbol{\gamma}}\|_2)] \mathbf{1}\{\|\widetilde{\boldsymbol{\beta}}\|_2 \|\widetilde{\boldsymbol{\gamma}}\|_2 > 0\}$.
- Plug-in thresholded estimator (Thresholded): Denote the thresholded estimators as $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$. The inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by $\langle \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}} \rangle$ and the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is estimated by $[\langle \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}} \rangle / (\|\bar{\boldsymbol{\beta}}\|_2 \|\bar{\boldsymbol{\gamma}}\|_2)] \mathbf{1}\{\|\bar{\boldsymbol{\beta}}\|_2 \|\bar{\boldsymbol{\gamma}}\|_2 > 0\}$.

Implementation of the de-biased, thresholded and FDE estimators requires the scaled Lasso estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ in the initial step. The scaled Lasso estimator is implemented by the equivalent square root Lasso algorithm (Belloni, Chernozhukov, and Wang 2011). The theoretical tuning parameter is $\lambda_0 = \sqrt{2.01 \log p/n}$, which may be conservative in the numerical studies. Instead, the tuning parameter is chosen as $\lambda_0 = b\sqrt{2.01 \log p}$. However, the performances of all estimators are evaluated across a grid of tuning parameter values $b \in \{0.25, 0.5, 0.75, 1\}$ (see supplementary material, Section A.1). The results showed that $b = 0.5$ was a good choice for all the estimators. Hence, $\lambda_0 = 0.5\sqrt{2.01 \log p/n}$ was used for the

Table 2. Mean square errors (MSE) of the estimates of the inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ for various signal strength parameters. Lasso: plug-in estimator with the scaled Lasso estimator; De-biased: plug-in estimator with the de-biased estimator; Thresholded: plug-in estimator with the thresholded estimator; FDE: the proposed estimator $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$; FDE-S: the proposed estimator $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with sample splitting; FDE-NS: the proposed estimator $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ without sample splitting.

		Strength parameters, (τ_1, τ_2)							
		(1.8, 0.4)	(2.2, 0.3)	(2.6, 0.2)	(3, 0.1)	(0.1, 1.6)	(0.2, 1.4)	(0.3, 1.2)	(0.4, 1)
$\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$	Truth	8.088	7.414	5.841	3.370	1.797	3.145	4.044	4.493
		MSE							
	Lasso	9.295	11.564	12.560	7.279	2.377	4.889	5.409	4.800
	De-biased	1.733	2.191	2.324	1.386	0.449	0.838	0.985	0.886
	Thresholded	2.029	3.377	6.463	5.789	1.877	3.024	2.432	1.546
	FDE	1.847	2.471	2.662	2.118	0.734	0.995	1.028	0.986
$\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$	Truth	0.5314	0.5314	0.5314	0.5314	0.5314	0.5314	0.5314	0.5314
		MSE							
	Lasso	0.0023	0.0075	0.0332	0.1260	0.1574	0.0624	0.0227	0.0087
	De-biased	0.0208	0.0415	0.0864	0.1590	0.1736	0.1068	0.0627	0.0373
	Thresholded	0.0045	0.0139	0.0585	0.1753	0.0964	0.0981	0.0389	0.0153
	FDE-S	0.0337	0.0303	0.0621	0.0678	0.2130	0.1199	0.0694	0.0616
	FDE-NS	0.0036	0.0064	0.0163	0.0580	0.0892	0.0237	0.0116	0.0061

numerical studies in this section and Section 5. To implement the FDE algorithm, the other tuning parameter λ is chosen as $\sqrt{2.01 \log p/n}$ for the correction Steps 2, 3, 5, and 6 in Table 1.

Comparisons of estimates of $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are presented below. Results on estimating the quadratic functionals are presented in the supplementary material, Section A.2. For each setting, with the parameters $(p, n_1, n_2, s, s_1, s_2)$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$, $F_{\boldsymbol{\beta}}$, $F_{\boldsymbol{\gamma}}$ specified, we generate the data and compare different methods as follows:

1. Generate sets $S_1 \subset [p]$ and $S_2 \subset [p]$, with $|S_1| = s_1$, $|S_2| = s_2$ and $|S_1 \cap S_2| = s$. For $\boldsymbol{\beta} \in \mathbf{R}^p$ and $\boldsymbol{\gamma} \in \mathbf{R}^p$, generate $\boldsymbol{\beta}_j \sim F_{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma}_l \sim F_{\boldsymbol{\gamma}}$, for $j \in S_1, l \in S_2$, and set $\boldsymbol{\beta}_j = 0$ and $\boldsymbol{\gamma}_l = 0$, for $j \notin S_1, l \notin S_2$.
2. Generate $\boldsymbol{X}_{i\cdot} \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$, $1 \le i \le n_1$, and $\boldsymbol{Z}_{i\cdot} \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Gamma})$, $1 \le i \le n_2$.
3. Generate the noise $\boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} N(0, 1)$, $1 \le i \le n_1$, and $\boldsymbol{\delta}_i \overset{\text{iid}}{\sim} N(0, 1)$, $1 \le i \le n_2$. Generate the outcome as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $\boldsymbol{w} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\delta}$.
4. With $\boldsymbol{X}$, $\boldsymbol{y}$, $\boldsymbol{Z}$, and $\boldsymbol{w}$, estimate $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ through different estimators.
5. Repeat Steps 2–4 for L times.

We evaluate the performance of an estimator by the mean squared error (MSE), which is defined as

$$\mathrm{MSE}(\widehat{\mathrm{T}}) = \frac{1}{L} \sum_{l=1}^{L} (\widehat{\mathrm{T}}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{w}; l) - \mathrm{T})^2, \qquad (36)$$

for a given quantity T and its estimate $\widehat{\mathrm{T}}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{w}; l)$ from lth replication. We consider two different settings with two sets of parameters $(p, n_1, n_2, s, s_1, s_2)$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$, $F_{\boldsymbol{\beta}}$, and $F_{\boldsymbol{\gamma}}$ and the simulation for each setting is repeated $L = 300$ times.

Experiment 1. The parameters are set as follows, $(p, n_1, n_2) = (600, 400, 400)$, the sparsity parameters $(s, s_1, s_2) = (15, 30, 25)$, and the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ satisfy $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Gamma}_{ij} = (0.8)^{|i-j|}$. For given positive values τ_1 and τ_2, the signals of $\boldsymbol{\beta}$ satisfy that $\boldsymbol{\beta}_{j_i} = (1 + i/s_1)\tau_1/2$, for $j_i \in S_1, i = 1, 2, \ldots, s_1$, and the signals in $\boldsymbol{\gamma}$ satisfy that $\boldsymbol{\gamma}_j = \tau_2$ for $j \in S_2$. This simulation aims to investigate the case where the coefficients for one regression are much larger than the other by varying the signal strength parameters as $(\tau_1, \tau_2) \in \{(3.0, 0.1), (2.6, 0.2), (2.2, 0.3), (1.8, 0.4), (0.1, 1.6), (0.2, 1.4), (0.3, 1.2), (0.4, 1.0)\}$.

The results are summarized in Table 2. For all combinations of the signal strength parameters, in terms of estimating the inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, FDE consistently outperformed the plug-in estimates with Lasso and thresholded Lasso. Moreover, with increasing difference between τ_1 and τ_2, the advantage of FDE over the plug-in estimate using Lasso or thresholded Lasso became larger. The same results were observed for estimation of the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, where FDE-NS had consistent better performance than other methods. Although de-biased performed well in terms of estimating $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, it performed much worse than FDE-NS for estimating $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

As discussed in Section 2, the sample splitting of estimating the normalized inner product is simply proposed to facilitate the theoretical analysis and might not be necessary for the algorithm. Our simulation results indicated that the proposed estimator without sample splitting (FDE-NS) performed quite well in all settings, even better than FDE-S, because more samples were used for estimation and correction steps. Such observations led us to use the proposed estimator without sample splitting (FDE-NS) in the real data analysis in Section 5.

Experiment 2. The parameters are set as follows, $(p, n_1, n_2) = (800, 400, 400)$, signal strength parameters $(\tau_1, \tau_2) = (0.2, 0.1)$ and the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ satisfy $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Gamma}_{ij} = (0.8)^{|i-j|}$. The signals of $\boldsymbol{\beta}$ follow that $\boldsymbol{\beta}_{j_i} = (1 + i/s_1)\tau_1/2$, for $j_i \in S_1$, $i = 1, 2, \ldots, s_1$, and the signals in $\boldsymbol{\gamma}$ satisfy that $\boldsymbol{\gamma}_j = \tau_2$ for $j \in S_2$. This simulation setting is set to investigate the relationship between the performance of estimators and the signal sparsity level and vary the sparsity levels of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as $(s_1, s_2) \in \{(40, 40), (50, 50), (60, 60), (70, 70), (80, 80), (90, 90), (100, 100), (110, 110)\}$, and fix the number of common signals at $s = 20$. Since the number of the associated variants is very large for both coefficient vectors, large values of τ_1 and τ_2 would induce strong signals such that all the methods perform well. Instead, we consider a more challenging setting where the signal magnitude is small, that is, $\tau_1 = 0.2$ and $\tau_2 = 0.1$.

The results are summarized in Table 3. Clearly, FDE outperformed the other methods. When the signals became denser, the improvement of FDE over other methods was more

Table 3. Mean square errors (MSE) of the estimates of the inner product $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the normalized inner product $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ for various sparsity parameters. Lasso: plug-in estimator with the scaled Lasso estimator; De-biased: plug-in estimator with the de-biased estimator; Thresholded: plug-in estimator with the thresholded estimator; FDE: the proposed estimator $\widehat{\mathrm{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$; FDE-S: the proposed estimator $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with sample splitting; FDE-NS: the proposed estimator $\widehat{\mathrm{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ without sample splitting.

		Sparsity parameter, s_1 ($s_2 = s_1$)							
		40	50	60	70	80	90	100	110
$\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$	Truth	0.190	0.170	0.219	0.212	0.179	0.221	0.183	0.221
		MSE							
	Lasso	0.032	0.025	0.039	0.036	0.024	0.035	0.023	0.028
	De-biased	0.015	0.015	0.018	0.017	0.024	0.027	0.040	0.066
	Thresholded	0.027	0.021	0.031	0.029	0.020	0.025	0.018	0.018
	FDE	0.020	0.014	0.021	0.022	0.011	0.013	0.008	0.008
$\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$	Truth	0.4027	0.2908	0.3122	0.2592	0.1914	0.2110	0.1573	0.1725
		MSE							
	Lasso	0.1157	0.0517	0.0539	0.0370	0.0166	0.0180	0.0097	0.0079
	De-biased	0.1267	0.0601	0.0659	0.0411	0.0160	0.0173	0.0063	0.0059
	Thresholded	0.1392	0.0687	0.0732	0.0504	0.0262	0.0277	0.0155	0.0142
	FDE-S	0.1154	0.1225	0.0779	0.0574	0.0456	0.0499	0.0450	0.0493
	FDE-NS	0.0847	0.0340	0.0368	0.0294	0.0115	0.0091	0.0055	0.0047

pronounced. For estimation of $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, the results showed that FDE-NS consistently outperformed other estimators. As the number of signals increased, the MSE corresponding to FDE-NS decreased quickly.

5. Genetic Relatedness Yeast Colony Growths Based on Genome-Wide Association Data

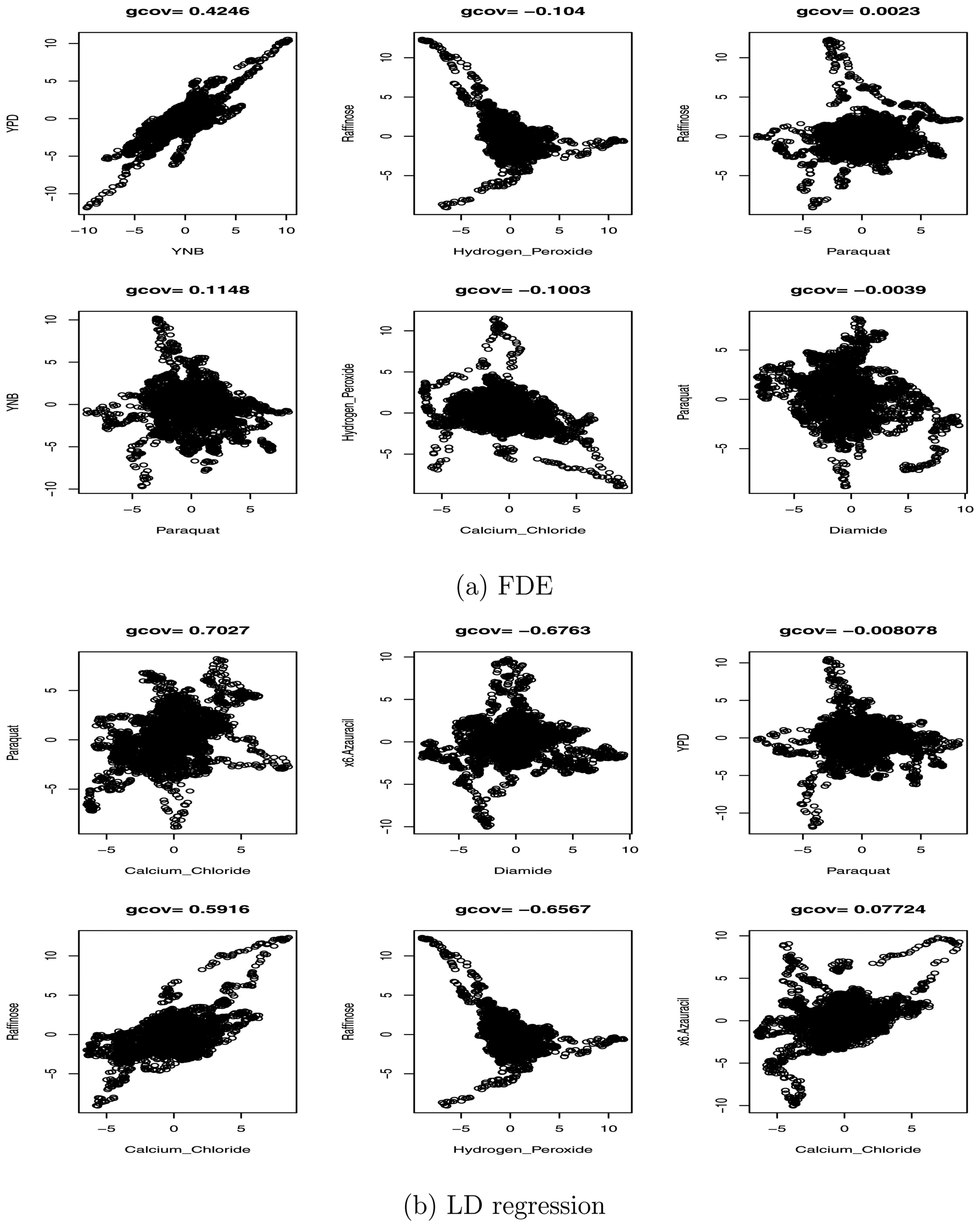
Bloom et al. (2013) reported a large-scale genome-wide association study of 46 quantitative traits based on 1008 *Saccharomyces cerevisiae* segregants crossbred from a laboratory strain and a wine strain. The dataset included 11,623 unique genotype markers. Since many of these markers are highly correlated and differ only in a few samples, Bloom et al. (2013) further selected a set of 4410 markers that are weakly dependent based on the linkage disequilibrium information. Specifically, these markers were selected by picking one marker closest to each centimorgan position on the genetic map. The maker genotypes are coded as 1 or −1, according to which strain it came from and satisfy the sub-Gaussian conditions. The traits of interest were the end-point colony size normalized by the control growth under 46 different growth media, including hydrogen peroxide, diamide, calcium, yeast nitrogen base (YNB), yeast extract peptone dextrose (YPD), etc. Bloom et al. (2013) showed that the genetic variants are associated with many of such trait values. It is therefore important to genetic relatedness among these related traits.

To demonstrate the genetic relatedness among these traits, eight traits were considered, including the normalized colony sizes under calcium chloride (calcium), diamide, hydrogen peroxide (hydrogen), paraquat, raffinose, 6 azauracil (azauracil), YNB, and YPD. Each trait was normalized to have variance 1, so the quadratic norm represents the total genetic effects for each trait and an estimate of the heritability. FDE was applied to every pair of these 8 traits without sample splitting, for a total of 28 pairs. The results are summarized in Table 4, including estimates of the heritability, genetic covariance, and genetic correlation for each of the 28 pairs. The genetic heritability of these traits ranged from 0.22 for raffinose to 0.67 for YPD. About two-thirds of these pairs had an estimated genetic correlation smaller than 0.1, indicating relatively weak genetic correlations among these traits.

To further demonstrate the genetic relatedness among these pairs, for each trait, a Z-score was calculated based on regressing the trait value $\boldsymbol{y}$ on genetic marker $\boldsymbol{X}_{\cdot j}$, for $i \leq j \leq p$. A larger absolute value of the Z-score statistic implies a stronger effect of the marker on the trait. For any pair of traits, the scatterplot of the Z-statistics provides a way of revealing the shared genetic relationship between them. The scatterplots of the Z-scores for all 28 pairs of traits are included in Section D of the supplemental materials. Figure 1(a) shows the plots of several pairs of the traits, including the pairs with a large positive $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, YPD versus YNB and Paraquat versus YNB, pairs with a large negative $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, raffinose versus hydrogen and calcium versus hydrogen, and pairs with $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ near 0, including paraquat versus diamide and paraquat versus raffinose. The plot clearly indicates a strong positive genetic covariance between YPD and YNB. The genetic covariance between paraquat and YNB/YPD

Table 4. FDE estimation for the heritability (bold diagonals), genetic covariance (upper diagonals), and genetic correlation (lower diagonals) among for each pair of eight colony growth traits of the yeast segregants.

Traits	Calcium	Diamide	Hydrogen	Paraquat	Raffinose	Azauracil	YNB	YPD
Calcium	**0.3314**	−0.0189	−0.1003	0.0084	0.0927	0.0095	0.0656	−0.0134
Diamide	−0.0286	**0.4390**	0.0598	−0.0039	0.0500	0.0446	−0.0159	0.0803
Hydrogen	−0.1579	0.0942	**0.4033**	0.0576	−0.1040	0.0601	0.0672	0.0637
Paraquat	0.0117	−0.0053	0.0799	**0.5199**	0.0023	0.0365	0.1148	0.1029
Raffinose	0.1972	0.1065	−0.2213	0.0049	**0.2208**	0.0137	0.0830	0.0331
Azauracil	0.0172	0.0809	0.1089	0.0661	0.0248	**0.3045**	−0.0259	0.0703
YNB	0.0968	−0.0235	0.0991	0.1693	0.1224	−0.0383	**0.4594**	0.4246
YPD	−0.0164	0.0983	0.0779	0.1259	0.0405	0.0860	0.5195	**0.6680**



(a) FDE

(b) LD regression

Figure 1. Scatterplots of marginal regression Z-score statistics for six pairs of traits ranked by the estimated genetic covariance ($gcov$) based on FDE (a) or LD regression (b), including the pairs with large positive genetic covariance (left panel), negative genetic covariance (middle panel), and small genetic covariance (right panel).

is smaller. raffinose/hydrogen and calcium/hydrogen pair clearly show negative genetic correlation. There are several genetic variants with very large effects on hydrogen, but they are not associated with the other traits such as raffinose and calcium. The shared genetic variants are relatively weak, leading to smaller genetic covariances. The plots on the bottom show the pairs of traits with weak genetic covariances. These plots indicate that the proposed genetic correlation measures can indeed capture the genetic sharing among different related traits.

Figure 2 shows the six pairs of the phenotypes ranked by the estimated genetic correlations FDE, including two with the largest positive genetic correlations, two with the largest negative genetic correlations, and two with the small genetic correlations. The pairs identified agree with the marginal Z-scores very well.

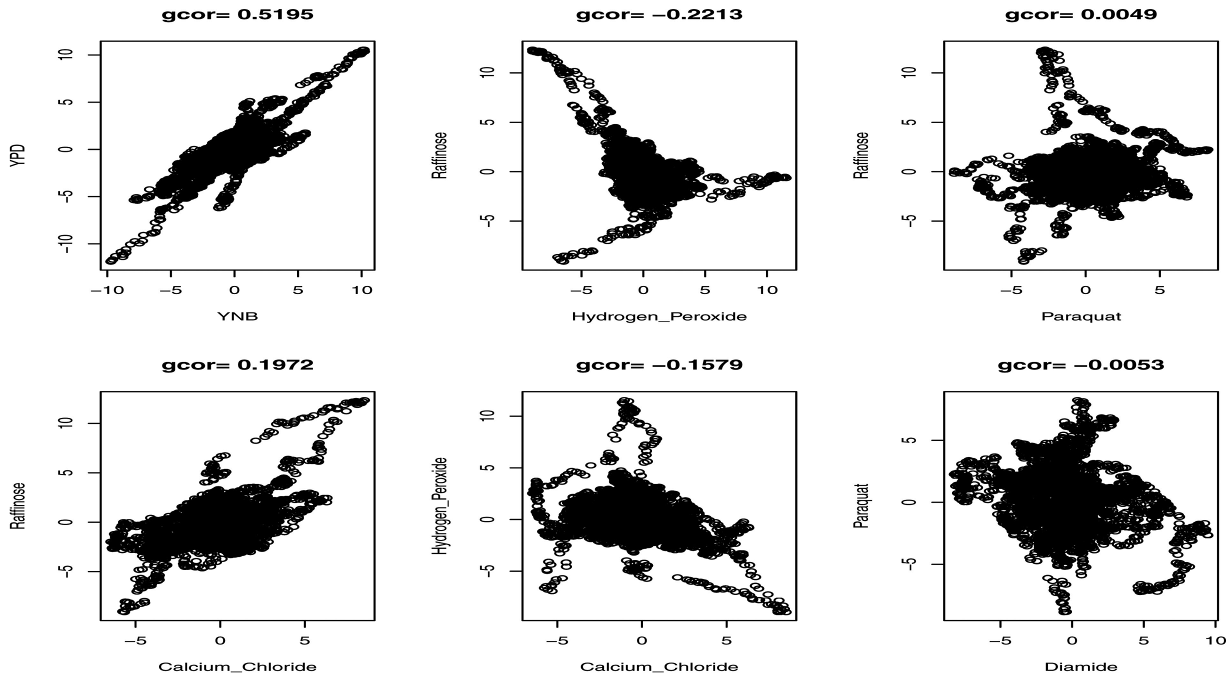


Figure 2. Scatterplots of marginal regression Z-score statistics for six pairs of traits ranked by the estimated genetic correlation (*gcor*) based on FDE, including the pairs with large positive genetic correlation (left panel), negative genetic correlation (middle panel), and small genetic correlation (right panel).

As a comparison, we also obtained the estimated genetic covariance for each pair of the traits using the LD regression methods proposed by Bulik-Sullivan et al. (2015). The pairs of traits with large positive, negative, or weak estimated covariance are presented in Figure 1(b). The pairs with the largest positive and negative estimated covariance are different from those two pairs identified by FDE. Comparison of the scatterplots of the Z-score statistics in Figure 1 indicates the pairs identified by FDE seem to agree with the marginal Z-statistics better.

6. Discussion

Motivated by the problems of estimating the genetic relatedness between two traits using the GWAS data, we have considered the problem of estimating the different functionals of the regression coefficients of two linear models, including the inner product $\langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle$, the quadratic functionals $\mathrm{Q}(\boldsymbol{\beta})$ and $\mathrm{Q}(\boldsymbol{\gamma})$, and the ratio $\mathrm{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The proposed method is different from plugging in the de-biased estimators proposed in Javanmard and Montanari (2014); van de Geer et al. (2014); Zhang and Zhang (2014). The correction procedures are implemented on the inner product and quadratic functionals directly, which balance the bias and variance uniquely for these functionals and hence result in minimax rate optimal estimators. The proposed estimators were shown in simulations to result in smaller estimation errors than directly plugging in these de-biased estimators across different settings. Results from analysis of the yeast segregants data suggested that the yeast colony growth sizes were under similar genetic controls under certain growth medias such as YPD and YNB, but this was not true for all pairs of growth media considered.

The algorithm for obtaining the these estimates only involves applying the Lasso several times, which can be implemented efficiently using the coordinate descent algorithms. The Matlab codes to implement the proposed estimation methods are available at *http://statgene.med.upenn.edu/software.html*. An important future research is to quantify uncertainty of these proposed estimators and the upper bound analysis of (21)–(23) and (24) indicates the possibility of constructing confidence intervals, centering at the proposed estimators and of parametric length $1/\sqrt{n}$, under additional sparsity and other regularity conditions.

Supplementary Material

The supplementary materials present extended simulations in Section A. In Section B, we prove Theorem 2. In Section C, we prove (28) and (31) in Theorem 3. We also provide detailed proofs of extra lemmas in Section D. In Section E, we present detailed results of real data analysis.

Acknowledgment

The authors thank Alexandre Tsybakov for helpful discussion on Section 3.2, and the reviewer and AE for helpful comments.

Funding

The research of Wanjie Wang and Hongzhe Li was supported in part by NIH grants CA127334 and GM097505.

References

Belloni, A., Chernozhukov, V., and Wang, L. (2011), “Square-Root Lasso: Pivotal Recovery of Sparse Signals Via Conic Programming,” *Biometrika*, 98, 791–806. [364]

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., and Kruglyak, L. (2013), “Finding the Sources of Missing Heritability in a Yeast Cross,” *Nature*, 494, 234–237. [366]

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. [359]

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., Neale, B. M., ReproGen Consortium, Psychiatric Genomics Consortium, and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium (2015), "An Atlas of Genetic Correlations Across Human Diseases and Traits," *Nature Genetics*, 47, 1236–1241. [358,359,368]

Cai, T. T., and Guo, Z. (2017), "Accuracy Assessment for High-dimensional Linear Regression," *The Annals of Statistics*, 46, 1807–1836. [364]

Cai, T. T., and Low, M. G. (2005), "Nonquadratic Estimators of a Quadratic Functional," *The Annals of Statistics*, 33, 2930–2956. [364]

——— (2006), "Optimal Adaptive Estimation of a Quadratic Functional," *The Annals of Statistics*, 34, 2298–2325. [364]

Collier, O., Comminges, L., and Tsybakov, A. B. (2015), "Minimax Estimation of Linear and Quadratic Functionals on Sparsity Classes," *The Annals of Statistics*, 45, 923–958. [364]

Donoho, D. L., and Nussbaum, M. (1990), "Minimax Quadratic Estimation of a Quadratic Functional," *Journal of Complexity*, 6, 290–323. [364]

Efromovich, S., and Low, M. (1996), "On Optimal Adaptive Estimation of a Quadratic Functional," *The Annals of Statistics*, 24, 1106–1125. [364]

Fan, J., Han, X., and Gu, W. (2012), "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence," *Journal of the American Statistical Association*, 107, 1019–1035. [359]

Golan, D., and Rosset, S. (2011), "Accurate Estimation of Heritability in Genome Wide Studies Using Random Effects Models," *Bioinformatics*, 27, i317–i323. [358]

Guo, Z., Kang, H., Cai, T. T., and Small, D. S. (2016), "Confidence Intervals for Causal Effects with Invalid Instruments Using Two-stage Hard Thresholding with Voting," arXiv:1603.05224. [362]

Janson, L., Barber, R. F., and Candes, E. (2016), "Eigenprism: Inference for High Dimensional Signal-to-Noise Ratios," *Journal of the Royal Statistical Society*, Series B, 79, 1037–1065. [359]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-dimensional Regression," *The Journal of Machine Learning Research*, 15, 2869–2909. [360,364,368]

Laurent, B., and Massart, P. (2000), "Adaptive Estimation of a Quadratic Functional by Model Selection," *The Annals of Statistics*, 28, 1302–1338. [364]

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., Mowry, B. J., Thapar, A., Goddard, M. E., and Witte, J. S. (2013), "Genetic Relationship Between Five Psychiatric Disorders Estimated from Genome-Wide SNPS," *Nature Genetics*, 45, 984–994. [358]

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. [358]

Lee, S. H., and van der Wer, J. H. (2016), "Mtg2: An Efficient Algorithm for Multivariate Linear Mixed Model Analysis Based on Genomic Information," *Bioinformatics*, 32, 1420–1422. [358]

Maier, R., Moser, G., Chen, G.-B., Ripke, S., Coryell, W., Potash, J. B., Scheftner, W. A., Shi, J., Weissman, M. M., Hultman, C. M., Landen, M., Levinosn, D. F., Kendler, K. S., Smoller, J. W., Wray, N. R., Lee. S. H., and Cross-Disorder Working Group of the Psychiatric Genomics Consortium (2015), "Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder," *The American Journal of Human Genetics*, 96, 283–294. [358]

Manolio, T. A. (2010), "Genomewide Association Studies and Assessment of the Risk of Disease," *New England Journal of Medicine*, 363, 166–176. [358]

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008), "Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges," *Nature Reviews Genetics*, 9, 356–369. [359]

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Sklar, P., Ruderfer, D. M., McQuillin, A., Morris, D. W., and International Schizophrenia Consortium (2009), "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder," *Nature*, 460, 748–752. [358]

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 101, 269–284. [359,360,364]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [359]

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. [360,364,368]

Vershynin, R. (2012), "Introduction to the Non-asymptotic Analysis of Random Matrices," in *Compressed Sensing: Theory and Applications*, eds. Y. Eldar, and G. Kutyniok, Cambridge, UK: Cambridge University Press, pp. 210–268. [360]

Verzelen, N., and Gassiat, E. (2016), "Adaptive Estimation of High-Dimensional Signal-to-Noise Ratios," *Bernoulli*, 24, 3683–3710. [359]

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007), "Prediction of Individual Genetic Risk to Disease from Genome-wide Association Studies," *Genome Research*, 17, 1520–1528. [358]

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. [359]

Yang, L., Neale, B. M., Liu, L., Lee, S. H., Wray, N. R., Ji, N., Li, H., Qian, Q., Wang, D., Li, J., Faraone, S. V., Wang, Y., and Psychiatric GWAS Consortium: ADHD Subgroup (2013), "Polygenic Transmission and Complex Neuro Developmental Network for Attention Deficit Hyperactivity Disorder: Genome-Wide Association Study of Both Common and Rare Variants," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 162, 419–430. [358]

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, 76, 217–242. [360,362,364,368]

Zhernakova, A., van Diemen, C. C., and Wijmenga, C. (2009), "Detecting Shared Pathogenesis from the Shared Genetics of Immune-related Diseases," *Nature Reviews Genetics*, 10, 43–45. [358]