Statistica Sinica **32** (2022), 1027-1048
doi:https://doi.org/10.5705/ss.202019.0445

OPTIMAL ESTIMATION OF SIMULTANEOUS SIGNALS USING ABSOLUTE INNER PRODUCT WITH APPLICATIONS TO INTEGRATIVE GENOMICS

Rong Ma, T. Tony Cai and Hongzhe Li

University of Pennsylvania

Abstract: Integrating the summary statistics from a genome-wide association study and expression quantitative trait loci data provides a powerful way of identifying genes with expression levels that are potentially associated with complex diseases. We introduce a parameter called T-score that quantifies the genetic overlap between a gene and the disease phenotype based on the summary statistics, based on the mean values of two Gaussian sequences. Specifically, given two independent samples $\mathbf{x}_n \sim N(\theta, \mathbf{\Sigma}_1)$ and $\mathbf{y}_n \sim N(\mu, \mathbf{\Sigma}_2)$, the T-score is defined as $\sum_{i=1}^n |\theta_i \mu_i|$, a nonsmooth functional, that characterizes the number of shared signals between two absolute normal mean vectors $|\theta|$ and $|\mu|$. Using approximation theory, estimators are constructed and shown to be minimax rate-optimal and adaptive over various parameter spaces. Simulation studies demonstrate the superiority of the proposed estimators over existing methods. Lastly, the method is applied to an integrative analysis of heart failure genomics data sets and we identify several genes and biological pathways that are potentially causal to human heart failure.

Key words and phrases: Approximation theory, eQTL, GWAS, minimax lower bound, non-smooth functional.

1. Introduction

1.1. Integrating summary data from genome-wide association studies and expression quantitative trait loci studies

Integrative genomics aims to integrate various biological data sets for the systematic discovery of a genetic basis that underlies and modifies a human disease (Giallourakis et al. (2005). To realize its full potential in genomic research, methods are needed that exhibit both computational efficiency and a theoretical guarantee for such integrative analyses. This study proposes a method that combines data sets from genome-wide association studies (GWASs) and expression quantitative trait loci (eQTL) studies in order to identify genetically regulated disease genes. Furthermore, we provide an integrative view of the underlying bi-

Corresponding author: Hongzhe Li, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: hongzhe@upenn.edu.

ological mechanism of complex diseases, such as heart failure. GWAS results have revealed that the majority of single nucleotide polymorphisms (SNPs) associated with a disease lie in noncoding regions of the genome (Hindorff et al. (2009)). These SNPs likely regulate the expression of a set of downstream genes that may have effects on diseases (Nicolae et al. (2010)). On the other hand, eQTL studies measure the association between both cis- and trans- SNPs and the expression levels of genes, which characterizes how genetic variants regulate transcriptions. A key next step in human genetic research is to explore whether these intermediate cellular level eQTL signals are located in the same loci ("colocalize") as GWAS signals and potentially mediate the genetic effects on disease, as well as finding disease genes with eQTL that overlap significantly with the set of loci associated with the disease (He et al. (2013)).

This study focuses on an integrative analysis of the summary statistics of GWAS and eQTL studies performed on possibly different sets of subjects. Owing to the privacy and confidentiality concerns of GWAS/eQTL participants, raw genotype data are often not available. Instead, most published papers provide summary statistics that include single SNP analysis results, such as the estimated effect size, its p-value, and the minor allele frequency. Based on these summary statistics, we propose a method that identifies potential disease genes by measuring their genetic overlaps with the disease. In particular, we propose a gene-specific measure, the T-score, that characterizes the total number of simultaneous SNP signals that share the same loci in both GWAS and eQTL studies of relevant normal tissues. Such a measure enables us to prioritize genes with expression levels that may underlie and modify human disease (Zhao et al. (2017)).

Treating SNP-specific GWAS and eQTL summary z-score statistics (as obtained for linear or logistic regression coefficients) as two independent sequences of Gaussian random variables, we define the T-score as the sum of the product of the absolute values of two normal means over a given set of n SNPs. Specifically, for any individual gene g, we denote $\mathbf{x}_n^g$ as the vector of z-scores from an eQTL study, and $\mathbf{y}_n$ as the vector of z-scores from a GWAS. We assume $\mathbf{x}_n^g \sim N(\theta^g, \mathbf{\Sigma}_1)$ and $\mathbf{y}_n \sim N(\mu, \mathbf{\Sigma}_2)$, for some $\theta^g, \mu \in \mathbb{R}^n$, and covariance matrices $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbb{R}^{n\times n}$ with unit diagonals. The T-score for gene g is then defined as

$$T\text{-score}(g) = \sum_{i=1}^{n} |\theta_i^g \mu_i|, \tag{1.1}$$

where the summation is over a given set of n SNPs. The T-score quantifies the number of simultaneous signals contained in two Gaussian mean vectors, regardless of the directions of the signals. Intuitively, a large T-score would possibly result from a large number of contributing components i with means θ_i^g and μ_i that are simultaneously large in absolute values. The supports (nonzero coordinates) of the mean vectors θ (hereafter, we omit its dependence on g for simplicity) and μ are assumed to have sparse overlaps, because it has been observed that, for a relatively large set of SNPs, only a small subset are associated with both a disease and gene expression (He et al. (2013)). After proper normalizations that account for study sample sizes, the number of SNPs, and effect sizes (see Section 2.5), we estimate the T-scores for all of the genes using summary statistsics. This enables us to identify and prioritize genetically regulated candidate disease genes. Furthermore, the T-scores can be used in a gene set enrichment analysis to identify disease-associated gene sets and pathways, or to quantify the genetic sharing between different complex traits using the GWAS summary statistics (Bulik-Sullivan et al. (2015)).

1.2. Justification of the absolute inner product

The T-score $\sum_{i=1}^n |\theta_i\mu_i|$ measures the overall signal overlap, regardless of the directions of the individual signal components. Although there are other quantities, such as $\sum_{i=1}^n \theta_i^2\mu_i^2$, that achieve a similar purpose, the T-score is closely related to the genetic correlation or genetic relatedness widely used in the genetic literature (Bulik-Sullivan et al. (2015)).

Suppose y and w are two traits, and for a given SNP with genotype score x, the marginal regression functions $y_i = \alpha_x + x_i\beta_x + \epsilon_i$ and $w_i = \eta_x + x_i\gamma_x + \delta_i$ hold for some coefficients (α_x, β_x) and (η_x, γ_x), respectively, where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_{x1}^2)$ and $\delta_i \overset{i.i.d.}{\sim} N(0, \sigma_{x2}^2)$, for $i = 1, 2, \ldots, N$ observations. For GWAS and eQTL data, one can treat y as a phenotype of interest and w as the expression level of a gene. In the above models, $x_i\beta_x$ and $x_i\gamma_x$ are the sample-specific marginal genetic effects due to the SNP x, and one can calculate their sample covariance as

$$\mathrm{Cov}_x = \frac{1}{N}\sum_{i=1}^{N}(x_i\beta_x - \bar{x}\beta_x)(x_i\gamma_x - \bar{x}\gamma_x) = \beta_x\gamma_x \cdot \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \tag{1.2}$$

where $\bar{x} = N^{-1}\sum_{i=1}^N x_i$. On the other hand, suppose for simplicity that the noise variances σ_{x1}^2 and σ_{x2}^2 are known. Then the z-scores based on the least square estimators $\hat{\beta}_x$ and $\hat{\gamma}_x$ satisfy

$$Z_{x1} = \frac{\hat{\beta}_x}{\sigma_{x1}/\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}} \sim N\left(\frac{\beta_x}{\sigma_{x1}/\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}, 1\right)$$

and

$$Z_{x2} = \frac{\hat{\gamma}_x}{\sigma_{x2}/\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}} \sim N\left(\frac{\gamma_x}{\sigma_{x2}/\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}, 1\right).$$

The product of the mean values of the above z-scores satisfies

$$\mathbb{E}Z_{x1}\mathbb{E}Z_{x2} = \frac{\beta_x \gamma_x}{\sigma_{x1}\sigma_{x2}/\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{\text{Cov}_x}{\sigma_{x1}\sigma_{x2}}. \tag{1.3}$$

Therefore, in terms of the Gausssian sequence model considered in this paper, the T-score is a parameter measuring the sum of absolute normalized sample covariances between the marginal genetic effects across a set of n SNPs; that is, for a set S of SNPs, the corresponding T-score satisfies

$$T\text{-score} = \sum_{x \in S} |\mathbb{E}Z_{x1}\mathbb{E}Z_{x2}| = \sum_{x \in S} \frac{|\text{Cov}_x|}{\sigma_{x1}\sigma_{x2}}, \tag{1.4}$$

which measures the overall simultaneous genetic effect of the SNPs in S.

1.3. Related works

Statistically, an estimation of the T-score involves estimating a nonsmooth functional, the absolute value function, of Gaussian random variables. Unlike estimating smooth functionals such as linear or quadratic functionals (Ibragimov and Khas'minskii (1985); Donoho and Nussbaum (1990); Fan (1991); Efromovich and Low (1994); Cai and Low (2005, 2006)), where some natural unbiased estimators are available, much less is known about estimating nonsmooth functionals. Using approximation theory, Cai and Low (2011) established the minimax risk and constructed a minimax optimal procedure for estimating a nonsmooth functional. More recently, this idea has been adapted to statistical information theory to estimate nonsmooth functionals, such as the Rényi entropy, support size, and L_1-norm (Jiao et al. (2015); Jiao, Han and Weissman (2016); Wu and Yang (2016, 2019); Acharya et al. (2016)). In particular, Collier, Comminges and Tsybakov (2020) obtained sharp minimax rates for estimating the L_γ-norm for $\gamma \le 1$ under a single sparse Gaussian sequence model, where the optimal rates are achieved by estimators that depend on knowledge of the underlying sparsity. Nonetheless, how to estimate the absolute inner product of two Gaussian mean vectors (T-score) with a sparse overlap as adaptively as possible remains unknown.

In the statistical genetics and genomics literature, several approaches have been proposed for integrating GWAS and eQTL data sets. Under the colocalization framework, methods such as those of Nica et al. (2010) and Giambartolomei et al.

(2014) were developed to detect colocalized SNPs. However, these methods do not directly identify the potential causal genes. Under the transcriptome-wide association study (TWAS) framework, Zhu et al. (2016) proposed a summary data-based Mendelian randomization method for causal gene identification, by posing several structural causality assumptions. Gamazon et al. (2015) developed a gene-based association method called PrediXcan that directly tests the molecular mechanisms through which a genetic variation affects a phenotype. Nevertheless, there is still a need for a quantitative measure of the genetic sharing between the genes and the disease that can be estimated from GWAS/eQTL summary statistics.

As a related, but different quantity, the genetic covariance ρ, proposed by Bulik-Sullivan et al. (2015), as a measure of the genetic sharing between two traits, can be expressed using our notation as $\rho = \sum_{i=1}^{n} \theta_i \mu_i$. In addition to the difference due to the absolute value function, in the original definition of genetic covariance ρ, the mean vectors θ and μ represent the conditional effect sizes (i.e., conditional on all other SNPs in the genome). In contrast, the mean vectors in our T-score correspond to the marginal effect sizes, making them directly applicable to the standard GWAS/eQTL summary statistics. In addition, unlike the linkage disequilibrium (LD) score regression approach considered in Bulik-Sullivan et al. (2015), our proposed method takes advantage of the fact that the support overlap between θ and μ is expected to be very sparse.

1.4. Main contributions

We propose an estimator of the T-score, based on the idea of thresholding and truncating the best polynomial approximation estimator. To the best of our knowledge, this is the first result related to the estimation of the absolute inner product of two Gaussian mean vectors. Under the framework of statistical decision theory, the minimax lower bounds are obtained, and we show that our proposed estimators are minimax rate-optimal over various parameter spaces. In addition, our results indicate that the proposed estimators are locally adaptive to the unknown sparsity level and the signal strength (Section 2). Our simulation study shows the strong empirical performance and robustness of the proposed estimators in various settings, and provides guidelines for using our proposed estimators in practice (Section 3). An analysis of GWAS and eQTL data sets of heart failure using the proposed method identifies several important genes that are functionally relevant to the etiology of human heart failure (Section 4).

2. Minimax Optimal Estimation of T-score

2.1. Minimax lower bounds

We start by establishing the minimax lower bounds for estimating the T-score over various parameter spaces. Throughout, we denote $T(\theta, \mu) = \sum_{i=1}^{n} |\theta_i \mu_i|$. For a vector $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$, we define $\|a\|_\infty = \max_{1 \le j \le n} |a_i|$. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists an absolute constant C such that $a_n \le C b_n$, for all n, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

For both practical and theoretical interest, we focus on the class of mean vector pairs (θ, μ) with only a small fraction of support overlaps. Specifically, for any $s < n$, we define the parameter space for (θ, μ) as $D(s) = \{(\theta, \mu) \in \mathbb{R}^n \times \mathbb{R}^n : |\text{supp}(\theta) \cap \text{supp}(\mu)| \le s\}$. Intuitively, in addition to the sparsity s, the difficulty of estimating $T(\theta, \mu)$ should also rely on the magnitudes of the mean vectors θ and μ and the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. To this end, we define the parameter space for $(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ as $D^\infty(s, L_n) = \big\{(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) : (\theta, \mu) \in D(s), \max(\|\theta\|_\infty, \|\mu\|_\infty) \le L_n, \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_n\big\}$, where both s and L_n can growth with n. In particular, to construct estimators that are as adaptive as possible, and to avoid unnecessary complexities of extra logarithmic terms, we calibrate the sparsity $s \asymp n^\beta$, for some $0 < \beta < 1$. Throughout, we consider the normalized loss function as the squared distance scaled by n^{-2}, and define the estimation risk for some estimator $\hat{T}$ as $\mathcal{R}(\hat{T}) = (1/n^2)\mathbb{E}(\hat{T} - T(\theta, \mu))^2$. To simplify our statement, we define the rate function $\psi(s, n) = \min\big\{\log\big(1 + (n/s^2)\big), L_n^2\big\} + \min\{\log s, L_n^2\}/\log^2 s$. The following theorem establishes the minimax lower bound over $D^\infty(s, L_n)$.

Theorem 1. *Let $\mathbf{x_n} \sim N(\theta, \mathbf{\Sigma}_1)$ and $\mathbf{y_n} \sim N(\mu, \mathbf{\Sigma}_2)$ be multivariate Gaussian random vectors, where $(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in D^\infty(s, L_n)$. Then,*

$$\inf_{\hat{T}} \sup_{(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in D^\infty(s, L_n)} \mathcal{R}(\hat{T}) \gtrsim \frac{L_n^2 s^2 \psi(s, n)}{n^2}, \tag{2.1}$$

where $\hat{T}$ is any estimator based on $(\mathbf{x_n}, \mathbf{y_n})$.

From the above theorem and the definition of the rate function $\psi(s, n)$, when $\beta \in (0, 1/2)$, (2.1) becomes

$$\inf_{\hat{T}} \sup_{(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in D^\infty(s, L_n)} \mathcal{R}(\hat{T}) \gtrsim \frac{L_n^2 s^2}{n^2} \min\{\log n, L_n^2\}, \tag{2.2}$$

when $\beta \in (1/2, 1)$, we have

$$\inf_{\hat{T}} \sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^{\infty}(s,L_n)} \mathcal{R}(\hat{T}) \gtrsim \frac{L_n^2 s^2}{n^2 \log^2 n} \min\{\log n, L_n^2\}, \tag{2.3}$$

and when $\beta = 1/2$, we have $\inf_{\hat{T}} \sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^{\infty}(s,L_n)} \mathcal{R}(\hat{T}) \gtrsim L_n^2 s^2/n^2$.

2.2. Optimal estimators of the T-score using a polynomial approximation

In general, the proposed estimators are based on the idea of an optimal estimation of the absolute value of normal means, as studied by Cai and Low (2011). They applied the best polynomial approximation of the absolute value function to obtain the optimal estimator and the minimax lower bound. Specifically, they defined the $2K$-degree polynomial $G_K(x) = (2/\pi)T_0(x) + (4/\pi)\sum_{k=1}^{K}(-1)^{k+1}(T_{2k}(x)/(4k^2-1)) \equiv \sum_{k=0}^{K} g_{2k}x^{2k}$, where $T_k(x) = \sum_{j=0}^{[k/2]}(-1)^j(k/(k-j))\binom{k-j}{j} 2^{k-2j-1}x^{k-2j}$ are Chebyshev polynomials. Then, for any $X \sim N(\theta,1)$, if H_k are Hermite polynomials with respect to the standard normal density ϕ such that $d^k\phi(y)/dy^k = (-1)^k H_k(y)\phi(y)$, the estimator based on $\tilde{S}_K(X) \equiv \sum_{k=0}^{K} g_{2k}M_n^{-2k+1}H_{2k}(X)$ for some properly chosen K and M_n has some optimality properties for estimating $|\theta|$. This important result motivates our construction of the optimal estimators of the T-score.

We begin by considering the setting where $\mathbf{x}_n = (x_1,\ldots,x_n)^\top \sim N(\theta,\mathbf{I}_n)$ and $\mathbf{y}_n = (y_1,\ldots,y_n)^\top \sim N(\mu,\mathbf{I}_n)$. To estimate $T(\theta,\mu)$, we first split each sample into two copies, one for testing, and the other for the estimation. Specifically, for $x_i \sim N(\theta_i,1)$, we generate x_{i1} and x_{i2} from x_i by letting $z_i \sim N(0,1)$ and setting $x'_{i1} = x_i + z_i$ and $x'_{i2} = x_i - z_i$. Let $x_{il} = x'_{il}/\sqrt{2}$, for $l = 1,2$. Then, $x_{il} \sim N(\theta'_i,1)$, for $l=1,2$ and $i=1,\ldots,n$, with $\theta'_i = \theta_i/\sqrt{2}$. Similarly, we construct $y_{il} \sim N(\mu'_i,1)$, for $l=1,2$ and $i=1,\ldots,n$, with $\mu'_i = \mu_i/\sqrt{2}$. Because $T(\theta,\mu) = 2T(\theta',\mu')$, estimating $T(\theta,\mu)$ using $\{x_i,y_i\}_{i=1}^n$ is equivalent to estimating $T(\theta',\mu')$ using $\{x_{il},y_{il}\}_{i=1}^n, l=1,2$.

In light of the estimator $\tilde{S}_K(X)$, we consider a slightly adjusted statistic $S_K(X) = \sum_{k=1}^{K} g_{2k}M_n^{-2k+1}H_{2k}(X)$, and define its truncated version $\delta_K(X) = \min\{S_K(X), n^2\}$, with $M_n = 8\sqrt{\log n}$ and $K \geq 1$ to be specified later. The above truncation is important in reducing the variance of $\delta_K(X)$. By the sample splitting procedure, we construct an estimator of $|\theta'_i|$ as

$$\hat{V}_{i,K}(x_i) = \delta_K(x_{i1})I(|x_{i2}| \leq 2\sqrt{2\log n}) + |x_{i1}|I(|x_{i2}| > 2\sqrt{2\log n}),$$

and a similar estimator of $|\mu'_i|$ as $\hat{V}_{i,K}(y_i)$. To further exploit the sparse structure, we also consider their thresholded version,

$$\hat{V}^S_{i,K}(x_i) = \delta_K(x_{i1})I(\sqrt{2\log n} < |x_{i2}| \le 2\sqrt{2\log n}) + |x_{i1}|I(|x_{i2}| > 2\sqrt{2\log n}),$$

as an estimator of $|\theta'_i|$ and, similarly, $\hat{V}^S_{i,K}(y_i)$ for $|\mu'|$. Intuitively, both $\hat{V}_{i,K}(x_i)$ and $\hat{V}^S_{i,K}(x_i)$ are hybrid estimators: $\hat{V}_{i,K}(x_i)$ is a composition of an estimator based on a polynomial approximation designed for small to moderate observations (less than $2\sqrt{2\log n}$ in absolute value) and the simple absolute value estimator applied to large observations (larger than $2\sqrt{2\log n}$ in absolute value). In contrast, $\hat{V}^S_{i,K}(x_i)$ has an additional thresholding procedure for small observations (less than $\sqrt{2\log n}$ in absolute value). Consequently, we propose two estimators of $T(\theta,\mu)$, namely,

$$\widehat{T}_K = 2\sum_{i=1}^n \hat{V}_{i,K}(x_i)\hat{V}_{i,K}(y_i), \tag{2.4}$$

as the hybrid nonthresholding estimator, and

$$\widehat{T}^S_K = 2\sum_{i=1}^n \hat{V}^S_{i,K}(x_i)\hat{V}^S_{i,K}(y_i) \tag{2.5}$$

as the hybrid thresholding estimator. Both estimators rely on K, a tuning parameter to be specified later.

2.3. Theoretical properties and minimax optimality

The following theorem provides the risk upper bounds of $\widehat{T}_K$ and $\widehat{T}^S_K$ over $D^\infty(s, L_n)$.

Theorem 2. *Let $\mathbf{x_n} \sim N(\theta, \mathbf{\Sigma}_1)$ and $\mathbf{y_n} \sim N(\mu, \mathbf{\Sigma}_2)$ be multivariate Gaussian random vectors with $(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2) \in D^\infty(s,L_n)$ and $s \asymp n^\beta$. Then,*

1. *for any $\beta \in (0,1)$ and K being any finite positive integer, we have*

$$\sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^\infty(s,L_n)} \mathcal{R}(\widehat{T}^S_K) \lesssim \frac{(L_n^2 + \log n)s^2\log n}{n^2}; \tag{2.6}$$

   if, in addition, $L_n \le (\sqrt{2}-1)\sqrt{\log n}$, then

$$\sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^\infty(s,L_n)} \mathcal{R}(\widehat{T}^S_K) \lesssim \frac{s^2L_n^4}{n^2} + \frac{\log^2 n}{n^{5/2}} + \frac{L_n^2\log n}{n^2}; \tag{2.7}$$

2. *for any $\beta \in (1/2,1)$ and $K = r\log n$, for some $0 < r < (2\beta-1)/12$, we*

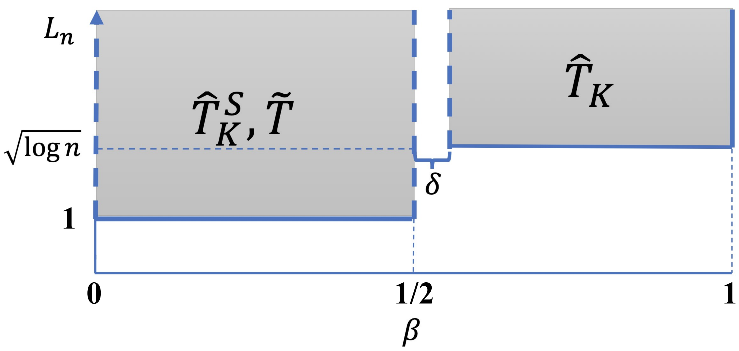

Figure 1. A graphical illustration of the regions where the proposed estimators are minimax optimal and adaptive. Here, $\widehat{T}_K^S$ has K being any finite positive integer, and $\widehat{T}_K$ has $K = r\log n$, for some $0 < r < \delta/6$.

have

$$\sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^{\infty}(s,L_n)} \mathcal{R}(\widehat{T}_K) \lesssim \frac{(L_n^2 + 1/\log n)s^2}{n^2\log n}. \tag{2.8}$$

Over the sparse region $\beta \in (0, 1/2)$, the risk upper bounds (2.6) and (2.7) along with the minimax lower bound (2.2) implies that $\widehat{T}_K^S$, with K being any finite positive integer, is minimax rate-optimal over $D^{\infty}(s, L_n)$ when $L_n \gtrsim 1$, where the minimax rate is

$$\inf_{\hat{T}} \sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^{\infty}(s,L_n)} \mathcal{R}(\hat{T}) \asymp \frac{L_n^2 s^2}{n^2}\min\{\log n, L_n^2\}. \tag{2.9}$$

When $L_n \lesssim 1$, the problem is less interesting, because in this case, the trivial estimator zero attains the minimax rate $L_n^4 s^2/n^2$. Over the dense region $\beta \in (1/2, 1)$, the nonthresholding estimator $\widehat{T}_K$, with $K = r\log n$ for some small r, is minimax rate-optimal over $D^{\infty}(s, L_n)$, for $L_n \gtrsim \sqrt{\log n}$, where the minimax rate is

$$\inf_{\hat{T}} \sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^{\infty}(s,L_n)} \mathcal{R}(\hat{T}) \asymp \frac{L_n^2 s^2}{n^2\log n}. \tag{2.10}$$

In both cases, the tuning parameter K plays an important role in controlling the bias–variance trade-off. An important consequence of our results concerns the local adaptivity of $\widehat{T}_K$ and $\widehat{T}_K^S$ with respect to s and L_n. Specifically, for any $\delta > 0$, the estimator $\widehat{T}_K$ with $K = r\log n$, for some $0 < r < \delta/6$, is simultaneously rate-optimal for any $L_n \gtrsim \sqrt{\log n}$ and any $\beta \in (1/2+\delta, 1)$, whereas the estimator $\widehat{T}_K$, with K being any finite positive integer, is simultaneously rate-optimal for any $L_n \gtrsim 1$ and $\beta \in (0, 1/2)$; see Figure 1.

Unfortunately, even with appropriate choices of K, neither $\widehat{T}_K^S$ nor $\widehat{T}_K$ is simultaneously optimal across all $\beta \in (0,1)$. However, because the difference between the optimal rates of convergence between (2.9) and (2.10) is only a factor of $\log n$, in practice, even when $\beta \in (1/2, 1)$, the thresholding estimator $\widehat{T}_K^S$ performs just as well as the nonthresholding estimator $\widehat{T}_K$. See Section 3 for detailed numerical studies.

2.4. Sparse estimation using simple thresholding

According to our previous analysis, if the parameter space is very sparse, that is, $\beta \in (0, 1/2)$, the proposed estimator $\widehat{T}_K^S$ is minimax optimal if we choose K as any constant positive integer. In other words, any constant degree polynomial approximation suffices for the optimal estimation of $T(\theta,\mu)$, including the constant function. Thus in this case, the polynomial approximation is essentially redundant for our purpose.

In light of the above observation, we consider the simple thresholding estimator $\widetilde{T} = 2\sum_{i=1}^n \hat{U}_i(x_i)\hat{U}_i(y_i)$, where $\hat{U}_i(x_i) = |x_{i1}|I(|x_{i2}| > 2\sqrt{2\log n})$. Our next theorem obtains the risk upper bound of $\widetilde{T}$ over $D^\infty(s, L_n)$, which along with (2.2) from Theorem 1, shows that $\widetilde{T}$ is also minimax optimal and adaptive over any sparsity level $\beta \in (0, 1/2)$ and $L_n \gtrsim 1$.

Theorem 3. *Let $\mathbf{x_n} \sim N(\theta, \mathbf{\Sigma}_1)$ and $\mathbf{y_n} \sim N(\mu, \mathbf{\Sigma}_2)$ be multivariate Gaussian random vectors with $(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in D^\infty(s, L_n)$. Then,*

$$\sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^\infty(s,L_n)} \mathcal{R}(\widetilde{T}) \lesssim \frac{(L_n^2 + \log n)s^2 \log n}{n^2}. \tag{2.11}$$

If, in addition, $L_n \leq \sqrt{2\log n}$, then

$$\sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D^\infty(s,L_n)} \mathcal{R}(\widetilde{T}) \lesssim \frac{s^2 L_n^4}{n^2} + \frac{\log^2 n}{n^3} + \frac{L_n^2 \log n}{n^2}. \tag{2.12}$$

Because our simple thresholding estimator $\widetilde{T}$ completely drops the polynomial components in $\widehat{T}_K^S$, its variance is significantly reduced. As a result, we find that as long as $\max(\|\theta\|_\infty, \|\mu\|_\infty) \leq \sqrt{n}$, the condition $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_n$ can be removed without changing the rate of convergence. To this end, we define the enlarged parameter space

$$D_0^\infty(s, L_n) = \left\{ (\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) : \begin{array}{l} (\theta,\mu) \in D(s), \max(\|\theta\|_\infty, \|\mu\|_\infty) \leq L_n, \\ \mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \succeq 0, \mathbf{\Sigma}_1 \text{ and } \mathbf{\Sigma}_2 \text{ have unit diagonals.} \end{array} \right\}.$$

In particular, because $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ have unit diagonals, the sample splitting procedure (Section 2.1) still applies, which leads only to a $1/2$-scaling of the off-diagonal entries of the covariance matrices.

Theorem 4. *Let* $\mathbf{x_n} \sim N(\theta, \mathbf{\Sigma}_1)$ *and* $\mathbf{y_n} \sim N(\mu, \mathbf{\Sigma}_2)$*, where* $(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in D_0^\infty(s, L_n)$ *and* $L_n \lesssim \sqrt{n}$*. Then, we have*

$$\sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D_0^\infty(s,L_n)} \mathcal{R}(\widetilde{T}) \lesssim \frac{(L_n^2 + \log n)s^2 \log n}{n^2}. \tag{2.13}$$

By definition, we have $D^\infty(s, L_n) \subset D_0^\infty(s, L_n)$. It then follows from Theorems 1 and 4 that for any $\beta \in (0, 1/2)$ and $L_n \lesssim \sqrt{n}$,

$$\inf_{\hat{T}} \sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in D_0^\infty(s,L_n)} \mathcal{R}(\hat{T}) \asymp \frac{s^2 L_n^2}{n^2} \cdot \min\{\log n, L_n^2\}, \tag{2.14}$$

where the minimax optimal rate can be attained by $\widetilde{T}$ when $L_n \geq \sqrt{\log n}$, and by the trivial estimator zero when $L_n < \sqrt{\log n}$. This establishes the minimax optimality and adaptivity of $\widetilde{T}$ over $D_0^\infty(n^\beta, L_n)$, for any $\beta \in (0, 1/2)$ and $L_n \gtrsim \sqrt{\log n}$. This result confirms an important advantage of $\widetilde{T}$ over $\widehat{T}_K^S$, namely, its guaranteed theoretical performance over arbitrary correlation structures, which complies with the fact that in many applications the observations are not necessarily independent. For further discussions on estimations with nonidentity covariances or unknown covariances, see Section S5.2 of the Supplementary Material.

2.5. Normalization, LD and the use of the T-score

Dealing with LD among the SNPs (Reich et al. (2001); Daly et al. (2001); Pritchard and Przeworski (2001)) is essential in any genetic studies. In this study, we follow Bulik-Sullivan et al. (2015) and propose using the normalized T-score

$$\text{Normalized } T\text{-score}(g) = \frac{\sum_{i=1}^n |\theta_i^g \mu_i|}{\|\theta^g\|_2 \|\mu\|_2}$$

as a measure of the genetic overlap between gene g and the outcome disease. In particular, the estimation of the ℓ_2 norms $\|\theta^g\|_2$ and $\|\mu\|_2$, or in our context, the SNP-heritability of the traits (Yang et al. (2010)), can be easily accomplished using summary statistics. As a result, every normalized T-score lies between zero and one, which is scale invariant (e.g., invariance to sample sizes and SNP effect sizes) and comparable across many different genes or studies. In addition, as argued by Bulik-Sullivan et al. (2015), the normalized T-score is less sensitive to the choice of the n-SNP sets.

Moreover, in Theorem 4, we show that the simple thresholding estimator $\widetilde{T}$ does not require the independence of the z-scores, which theoretically guarantees its applicability in the presence of an arbitrary LD structure among the SNPs. However, our theoretical results concerning $\widehat{T}_K$ and $\widehat{T}_K^S$ rely on such an independence assumption. In our simulation studies, we found that the empirical performance (including optimality) of $\widehat{T}_K$ and $\widehat{T}_K^S$ is not likely affected by the dependence due to the LD structure. As a result, our proposed estimation method, although partially analyzed under the independence assumption, can be directly applied to the summary statistics, without specifying the underlying LD or covariance structure.

The T-score can be used to identify disease genes and pathways using GWAS and eQTL data. For each gene, we estimate the T-score using our proposed estimators and the vectors of z-scores from the GWAS and eQTL studies. After obtaining the estimated T-scores for all genes and the corresponding SNP-heritability, we rank the genes by the order of their normalized T-scores. As a result, genes with the highest ranks are considered important in providing insights into the biological mechanisms of a disease. For a gene set or pathway analysis, we obtain the normalized T-scores T_j, for $1 \leq j \leq J$, for a given gene set S, and then calculate the Kolmogorov–Smirnov test statistic, defined as $\sup_t |(1/k)\sum_{j\in S} I(T_j \leq t) - (1/k')\sum_{j\in S^c} I(T_j \leq t)|$, where k and k' are the numbers of genes in S and S^c, respectively. For a given gene set, the significance of this test implies that the gene set S is enriched by genes that share similar genetic variants to those for the disease of interest, suggesting their relevance to the etiology of the disease. See Section 4 for detailed applications.

3. Simulation Studies

This section demonstrates and compares the empirical performance of our proposed estimators and some alternative estimators under various settings.

Simulation under multivariate Gaussian models. We generate a pair of n-dimensional vectors, denoted as $\mathbf{x}_n$ and $\mathbf{y}_n$, with $n = 1.5 \times 10^5, 3 \times 10^5$ and 5×10^5, from the multivariate normal distributions $N(\theta, \mathbf{\Sigma})$ and $N(\mu, \mathbf{\Sigma})$, respectively. We choose $s \in \{50, 100, 200, 400, 800\}$, which cover the regions $s \leq \sqrt{n}$ and $s > \sqrt{n}$, and generate (θ, μ) as follows: 1) the supports of θ and μ are randomly sampled from the coordinates, with the nonzero components generated from Unif(1,10); and 2) we partition the coordinates of θ and μ into blocks of size 10 and randomly pick $s/10$ blocks as the support, to which we assign symmetric triangle-shaped values, the maximal value of which is generated from Unif(5,10). The above signal structures are referred to as Sparse Pattern I and II, respectively. For the covariance matrix $\mathbf{\Sigma}$, we consider a global covariance $\mathbf{\Sigma} = \mathbf{I}$ and two block-wise covariances $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ (see Section S6 of the Supplementary Material for their explicit forms). We evaluate our proposed estimators $\widehat{T}_K^S$, $\widehat{T}_K$, and $\widetilde{T}$, as well as (1) the hybrid thresholding estimator without sample splitting, denoted as $\widehat{T}_K^{S*}$, and (2) the naive estimator $\overline{T}$, which simply calculates the absolute inner product of the observed vectors. For $\widehat{T}_K^S$ and $\widehat{T}_K^{S*}$, we fix $K = 8$, whereas for $\widehat{T}_K$, we set $K = \lfloor (1/12)\log n \rfloor$. Each setting was repeated 100 times, and the performance was evaluated using the empirical version of the rescaled mean square error $\text{RMSE}(\hat{T}) = (1/s)\sqrt{\mathbb{E}(\hat{T} - T)^2}$. Table 1 reports the empirical RMSE of the five estimators under the settings with independent observations. For brevity, the results under correlated observations are given in Tables S6.1 and S6.2 of the Supplementary Material. In general, $\widehat{T}_K^S$, $\widetilde{T}$, and $\widehat{T}_K^S$ perform similarly, with $\widehat{T}_K^S$ performing slightly better, although all are superior to the naive estimator $\overline{T}$. Here, $\widehat{T}_K^{S*}$ outperforms the other estimators in all settings, possibly because of the reduced variability as a result of not using sample splitting. Because the sample splitting is needed only to facilitate our theoretical analysis, in applications, we suggest using $\widehat{T}_K^{S*}$ for better performance. Moreover, Tables S6.1 and S6.2 in the Supplementary Material show that the proposed estimators are robust to the underlying sparsity patterns and the covariance structures.

Simulation under model-generated GWAS and eQTL data allowing for population stratification. In order to justify our proposed methods for an integrative analysis of GWAS and eQTL data, we carried out additional numerical experiments under more realistic settings. Here, the GWAS-based genotypes are simulated allowing for population stratification, and the corresponding z-scores are calculated from a case-control study that adjusts for population structure using principal component (PC) scores. Specifically, for the GWAS data, we adopted the simulation settings from Astle and Balding (2009), where 1,000 cases and 1,000 controls are drawn from a population of 6,000 individuals, partitioned into three equal-sized subpopulations. Ancestral minor allele fractions are generated from Unif(0.05,0.5) for all 10,000 unlinked SNPs. For each SNP, subpopulation allele fractions are generated from the beta-binomial model $\text{Beta}\big(((1-F)/F)p, ((1-F)/F)(1-p)\big)$ with a population divergence parameter $F = 0.1$. We simulate the disease phenotype under a logistic regression model with 20 SNP markers, each with effect size 0.4. The population disease prevalence is 0.05. To obtain z-scores, we fit a marginal logistic regression for each

Table 1. Empirical RMSE under the covariance $\mathbf{\Sigma} = \mathbf{I}_n$. $\widehat{T}_K^{S*}$: the hybrid thresholding estimator without sample splitting; $\widehat{T}_K^S$: the hybrid thresholding estimator; $\widetilde{T}$: the simple thresholding estimator; $\widehat{T}_K$: the hybrid nonthresholding estimator; $\overline{T}$: the naive estimator that calculates the absolute inner product of observed vectors.

$n/10^4$	s	$\widehat{T}_K^{S*}$	$\widehat{T}_K^S$	$\widetilde{T}$	$\widehat{T}_K$	$\overline{T}$	$\widehat{T}_K^{S*}$	$\widehat{T}_K^S$	$\widetilde{T}$	$\widehat{T}_K$	$\overline{T}$
		Sparse Pattern I					Sparse Pattern II				
15	50	10.54	20.85	27.47	25.14	1,910.3	8.69	26.79	32.9	28.84	1,909.2
	100	11.41	21.00	27.92	25.63	954.3	8.08	26.33	32.64	28.75	954.3
	200	10.30	21.19	30.83	28.01	476.9	8.42	25.83	32.33	28.54	476.9
	400	10.01	20.57	29.24	26.78	238.0	8.64	25.88	31.67	27.84	238.0
	800	10.58	22.36	29.99	27.05	118.8	9.20	25.48	31.16	27.61	118.7
30	50	9.50	20.51	30.13	27.7	3,819.4	10.72	28.11	33.67	29.73	3,819.8
	100	11.07	25.85	33.66	29.98	1,909.3	9.20	27.90	34.36	30.04	1,908.6
	200	10.60	22.19	30.3	27.09	954.4	9.71	25.89	31.88	28.27	954.1
	400	10.54	22.22	30.08	26.85	476.9	10.73	27.79	32.3	28.61	476.7
	800	10.86	23.52	30.62	27.24	238.2	8.62	26.67	34.2	30.11	238.0
50	50	12.27	27.30	32.18	28.67	6,363.4	12.02	25.78	27.07	24.37	6,365.3
	100	11.25	24.86	30.69	27.29	3,182.4	8.54	29.67	35.99	31.4	3,182.5
	200	11.02	22.48	29.39	25.88	1,591.3	9.98	29.13	34.21	29.94	1,591.3
	400	11.40	23.42	29.86	26.45	795.4	12.51	25.28	28.06	25.09	795.2
	800	10.85	22.85	29.40	26.11	397.2	10.23	27.05	32.69	28.84	397.2

Table 2. Empirical RMSE for simulated GWAS and eQTL data. $\widehat{T}_K^{S*}$: the hybrid thresholding estimator without sample splitting; $\widehat{T}_K^S$: the hybrid thresholding estimator; $\widetilde{T}$: the simple thresholding estimator; $\widehat{T}_K$: the hybrid nonthresholding estimator; $\overline{T}$: the naive estimator that calculates the absolute inner product of observed vectors.

s	$\widehat{T}_K^{S*}$	$\widehat{T}_K^S$	$\widetilde{T}$	$\widehat{T}_K$	$\overline{T}$
5	19.61	32.26	40.45	34.25	1,318.1
10	17.42	35.27	39.87	36.80	638.9
15	13.92	31.78	36.50	34.50	425.6
20	12.77	29.18	32.72	30.52	317.7

SNP, accounting for the first two PCs of the genotypes. For the eQTL data, 10,000 unlinked SNPs are generated independently with minor allele fractions from Unif(0.05,0.5). The gene expression levels of 2,000 samples are simulated under a linear regression model with covariates being s SNP markers that overlap with the GWAS SNPs. Each has an effect size of 0.5, and the errors are drawn independently from the standard normal distribution. The eQTL z-scores are obtained from a marginal linear regression. The above simulations were repeated

500 times. The population mean of the z-scores corresponding to the truly associated SNP markers are approximated using the sample mean of the z-scores. Table 2 shows the empirical RMSEs for the five estimators with $s \in \{5, 10, 15, 20\}$. Again, our proposed estimators $\widehat{T}_K$, $\widehat{T}_K^S$, and $\widetilde{T}$ outperform the naive estimator $\overline{T}$ across all settings, and $\widehat{T}_K^{S*}$ performs even better. The numerical results agree with our simulations under the multivariate Gaussian settings, suggesting the applicability of our proposed methods for integrating GWAS and eQTL data.

4. Integrative Data Analysis of Human Heart Failure

Finally, we apply our proposed estimation procedure to identify genes with expressions that are possibly causally linked to heart failure by integrating GWAS and eQTL data. The GWAS results were obtained from a heart failure genetic association study at the University of Pennsylvania, a prospective study of patients recruited from the University of Pennsylvania, Case Western Reserve University, and the University of Wisconsin, where genotype data were collected from 4,523 controls and 2,206 cases using the Illumina OmniExpress Plus array. The GWAS summary statistics were calculated controlling for age, gender, and the first two principal components of the genotypes.

The heart failure eQTL data were obtained from the MAGNet eQTL study (`https://www.med.upenn.edu/magnet/index.shtml`), where left ventricular free-wall tissue were collected from 136 donor hearts without heart failure. Genotype data were collected using Affymetrix genome-wide SNP array 6.0, and only markers in a Hardy–Weinberg equilibrium with minor allele frequencies above 5% were considered. Gene expression data were collected using Affymetrix GeneChip ST1.1 arrays, normalized using RMA (Irizarry et al. (2003)), and batch-corrected using ComBat (Johnson, Li and Rabinovic (2007)). To obtain a common set of SNPs, the SNPs were imputed using 1,000 Genomes Project data. Summary statistics for the MAGNet eQTL data were obtained using the fast marginal regression algorithm of Sikorska et al. (2013), controlling for age and gender.

4.1. Ranking of potential heart failure causal genes

After matching the SNPs of the eQTL and GWAS data, we had a total of 347,019 SNPs and 19,081 genes with expression data available. Given the results of the simulation studies, throughout, we use $\widehat{T}_K^{S*}$ with $K = 8$ to estimate the T-scores. The analysis then follows from Section 2.5 so that the genes are ordered by their normalized T-scores. To assess that the top scored genes indeed represent true biological signals, we calculated the T-scores for two "null data sets" created using permutations. For the first data set, we randomly permuted the labels of the SNPs of the GWAS z-scores by sampling without replacement, before estimating the normalized T-scores using the eQTL z-scores. For the second data set, we permuted the labels of the SNPs of the GWAS z-scores in a circular manner, similarly to Cabrera et al. (2012). Specifically, for each chromosome, we randomly chose one SNP as the start of the chromosome, and moved the SNPs on the fragment before this SNP to the end. Such a cyclic permutation preserves the local dependence of the z-scores. By permuting the data from one phenotype, we break the original matching of the z-scores between the two phenotypes. The permutation was performed 50 times, and we obtained the null distribution of T-scores based on the permuted data. Figure 2 shows the ranked normalized T-scores based on the original data and box plots of the ranked z-scores based on 50 permutations of the z-scores. We find that all of the top-ranked genes have larger T-scores than those based on permutations. In addition, about 30 top-ranked genes in the top plot and about 10 top-ranked genes in the bottom plot have true T-scores larger than all T-scores from the permuted data sets. This confirms that the top-ranked genes based on their estimated normalized T-scores are not due to random noise, and indeed represent a sharing of genetic variants between heart failure and gene expression levels.

4.2. Gene set enrichment analysis

To complete our analysis, we finish this section with a gene set enrichment analysis (GSEA) (Subramanian et al. (2005)), using the normalized T-scores to identify the biological processes associated with heart failure. In the following analysis, we removed genes with low expression and small variability across the samples, which resulted in a total of 6,355 genes. The method described in Section 2.5 was applied to the gene sets from Gene Ontology (GO) (Botstein et al. 2000), which contain at least 10 genes, and 5,023 biological processes were tested. Figure S6.1 in the Supplementary Material presents directed acyclic graphs of the GO biological processes linked to the most significant GO terms from the simultaneous signal GSEA analysis. Table 4 shows the top six GO biological processes identified from the GSEA analysis. Among them, regulation of skeletal muscle contraction, the linoleic acid metabolic process, and calcium ion regulation are strongly implicated in human heart failure. Murphy et al. (2011) showed that skeletal muscle reflexes are essential to the initiation and regulation of the cardiovascular response to exercise, and an alteration of this reflex mechanism can happen in disease states such as hypertension and heart failure. In Farvid et al. (2014), a thorough meta-analysis supported a significant inverse association be-

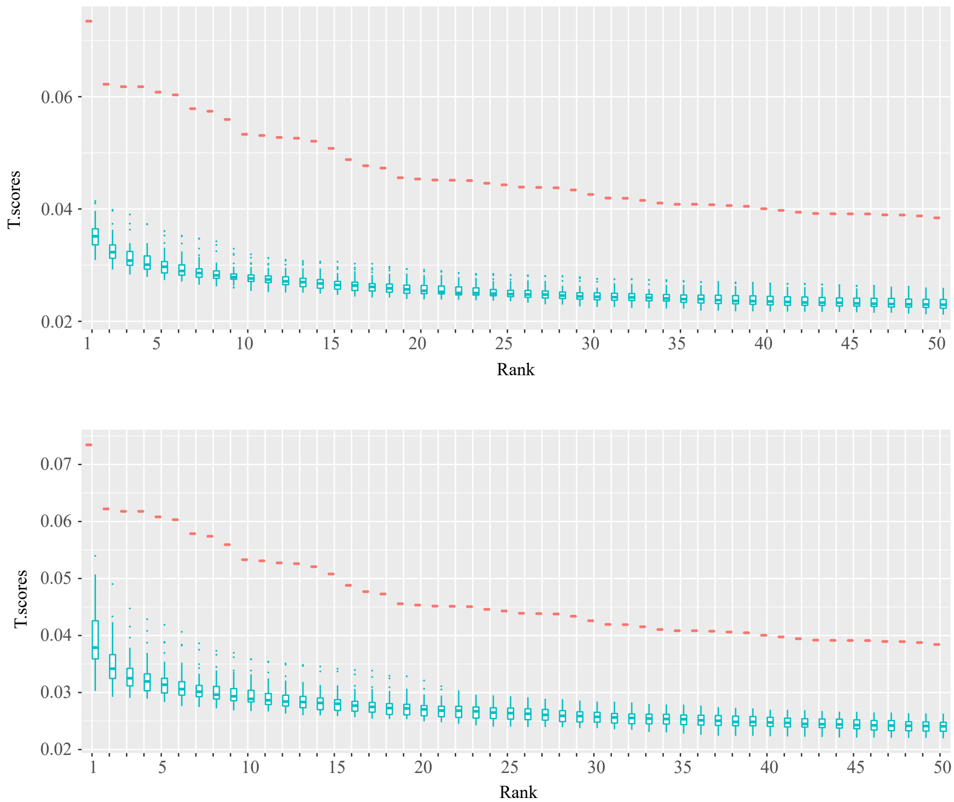

Figure 2. Estimated score (in short bars) for top 50 genes and the boxplots of the top scores based on 50 permutations. Top: random permutation of the GWAS scores; bottom: cyclic permutations of the GWAS scores.

tween dietary linoleic acid intake, when replacing either carbohydrates or saturated fat, and the risk of coronary heart disease. Moreover, the importance of calcium-dependent signaling in heart failure was reported in Marks (2003), who suggested that impaired calcium release causes decreased muscle contraction (systolic dysfunction), and defective calcium removal hampers relaxation (diastolic dysfunction).

Table 3 lists the top eight highest ranked genes, along with their biological annotations. All of the genes are either directly or indirectly associated with human heart failure, including those related to fibrotic myocardial degeneration, Wnt signalling activity, and heart-valve development. It is interesting that our proposed methods can identify these relevant genes using only the gene expression data measured on normal heart tissue.

Table 3. Top eight genes associated with heart failure based on the estimated normalized T-scores and their functional annotations.

Gene Name	Annotations
TMEM37	voltage-gated ion channel activity (Chen et al. (2007))
ADCY7	adenylate cyclase activity; fibrotic myocardial degeneration (Nojiri et al. (2006))
C1QC	Wnt signaling activity; associated with heart failure (Naito et al. (2012))
FAM98A	associated with ventricular septal defect (Liu et al. (2018)
BMP2	associated with heart-valve development (Rivera-Feliciano and Tabin (2006))
SLCO2B1	organic anion transporter; associated with cardiac glycoside (Mikkaichi et al. (2004))
C1QA	Wnt signaling activity; associated with heart failure (Naito et al. (2012))
FCGR2B	intracellular signaling activity; associated with vascular disease pathogenesis (Tanigaki et al. (2015))

Table 4. Top six GO biological processes associated with heart failure based on the gene set enrichment analysis.

GO term	p-value
Biological Process	
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	7.9×10^{-7}
linoleic acid metabolic process	1.0×10^{-6}
regulation of skeletal muscle contraction by calcium ion signaling	3.4×10^{-6}
positive regulation of sequestering of calcium ion	3.4×10^{-6}
cellular response to caffeine	1.0×10^{-5}
cellular response to purine-containing compound	1.0×10^{-5}

5. Discussion

This study considers the optimal estimation over sparse parameter spaces. In Section 2, the minimax rates of convergence were established for the parameter spaces $D^{\infty}(n^{\beta}, L_n)$ with $\beta \in (0, 1/2) \cup (1/2, 1)$, leaving a gap at $\beta = 1/2$. Our theoretical analysis suggests a lower bound (2.1) with the rate function $\psi(s, n) \asymp 1$, which cannot be attained by any of our proposed estimators. Nevertheless, in Section S5.1 of the Supplementary Material, we confirm that $L_n^2 s^2/n^2$ is the minimax rate of convergence for $\beta = 1/2$ by proposing an estimator achieving such a rate.

In some applications, we may need to consider nonsparse parameter spaces. In this case, our theoretical analysis shows that the estimator $\widehat{T}_K$ with $K = r\log n$, for some small constant $r > 0$, can still be applied. Specifically, from our proof of Theorem 1 and Theorem 2, it follows that if we define the nonsparse parameter space as $\mathcal{D}_U^{\infty}(L_n) = \big\{(\theta, \mu, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) : (\theta, \mu) \in \mathbb{R}^n \times \mathbb{R}^n, \max(\|\theta\|_\infty, \|\mu\|_\infty) \leq L_n, \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_n\big\}$, with $L_n \gtrsim \sqrt{\log n}$, then for $\mathbf{x_n} \sim N(\theta, \mathbf{\Sigma}_1)$ and $\mathbf{y_n} \sim N(\mu, \mathbf{\Sigma}_2)$, the minimax rate $\inf_{\hat{T}} \sup_{(\theta,\mu,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in\mathcal{D}_U^{\infty}(L_n)} \mathcal{R}(\hat{T}) \asymp L_n/\log n$ can be attained by the above $\widehat{T}_K$.

In light of our genetic applications, it is also natural and interesting to consider parameter spaces where θ and μ are both sparse in themselves. Specifically, assuming the triple sparsity of θ, μ, and $\{\theta_i\mu_i\}_{i=1}^n$, interesting phase transitions might exist, where the minimax rates and the optimal estimators could be different from those reported here. In addition to the estimation problems, it is also of interest to conduct hypothesis testing and to construct confidence intervals for the T-score. These problems are technically challenging owing to the nonsmooth functional. We leave these important problems for future research.

Supplementary Material

The online Supplementary Material includes the proofs of the main theorems. Supplementary notes, figures, and tables are also included.

Acknowledgments

The authors are grateful to the Editor, the Associate Editor, and two anonymous referees for their constructive comments. R.M. would like to thank Mark G. Low for his helpful discussions. The research reported in this publication was supported by NIH grants R01GM123056 and R01GM129781 and NSF grant DMS-1712735.

References

Acharya, J., Das, H., Orlitsky, A. and Suresh, A. T. (2016). A unified maximum likelihood approach for optimal distribution property estimation. *arXiv preprint arXiv:1611.02960.*

Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science* **24**, 451–471.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. et al. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**, 25–29.

Bulik-Sullivan, B., 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.

Cabrera, C. P., Navarro, P., Huffman, J. E., Wright, A. F., Hayward, C., Campbell, H. et al. (2012). Uncovering networks from genome-wide association studies via circular genomic permutation. *G3-Genes. Genom. Genet.* **2**, 1067–1075.

Cai, T. T. and Low, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Stat.* **33**, 2930–2956.

Cai, T. T. and Low, M. G. (2006). Optimal adaptive estimation of a quadratic functional. *Ann. Stat.* **34**, 2298–2325.

Cai, T. T. and Low, M. G. (2011). Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Stat.* **39**, 1012–1041.

Chen, R.-S., Deng, T.-C., Garcia, T., Sellers, Z. M. and Best, P. M. (2007). Calcium channel γ subunits: A functionally diverse protein family. *Cell Biochem. and Biophys.* **47**, 178–186.

Collier, O., Comminges, L. and Tsybakov, A. B. (2020). On estimation of nonsmooth functionals of sparse normal means. *Bernoulli* **26**, 1989–2020.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics* **29**, 229–232.

Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *Journal of Complexity* **6**, 290–323.

Efromovich, S. and Low, M. G. (1994). Adaptive estimates of linear functionals. *Probability Theory and Related Fields* **98**, 261–275.

Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Stat.* **19**, 1273–1294.

Farvid, M. S., Ding, M., Pan, A., Sun, Q., Chiuve, S. E., Steffen, L. M., Willett, W. C. and Hu, F. B. (2014). Dietary linoleic acid and risk of coronary heart disease: A systematic review and meta-analysis of prospective cohort studies. *Circulation* **130**, 1568–1578.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J. et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098.

Giallourakis, C., Henson, C., Reich, M., Xie, X. and Mootha, V. K. (2005). Disease gene discovery through integrative genomics. *Annu. Rev. Genomics Hum. Genet.* **6**, 381–406.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C. and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10**, e1004383.

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X. and Li, H. (2013). Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* **92**, 667–680.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* **106**, 9362–9367.

Ibragimov, I. A. and Khas'minskii, R. Z. (1985). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theor. Probab. Appl+* **29**, 18–32.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Jiao, J., Han, Y. and Weissman, T. (2016). Minimax estimation of the L_1 distance. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, 750–754. IEEE.

Jiao, J., Venkat, K., Han, Y. and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory* **61**, 2835–2885.

Johnson, W. E., Li, C. and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127.

Liu, H., Hu, Y., Zhuang, B., Yin, J., Chen, X. H., Wang, J., Li, M.-M., Xu, J., Wang, X. Y., Yu, Z. B. and Han, S. P. (2018). Differential expression of circRNAs in embryonic heart tissue associated with ventricular septal defect. *Int. J. Med. Sci.* **15**, 703–712.

Marks, A. R. (2003). Calcium and the heart: A question of life and death. *The Journal of Clinical Investigation* **111**, 597–600.

Mikkaichi, T., Suzuki, T., Tanemoto, M., Ito, S. and Abe, T. (2004). The organic anion transporter (OATP) family. *Drug Metabolism and Pharmacokinetics* **19**, 171–179.

Murphy, M. N., Mizuno, M., Mitchell, J. H. and Smith, S. A. (2011). Cardiovascular regulation by skeletal muscle reflexes in health and disease. *Am. J. Physiol. Heart. Circ. Physiol.*

Naito, A. T., Sumida, T., Nomura, S., Liu, M. L., Higo, T., Nakagawa, A. et al. (2012). Complement C1q activates canonical wnt signaling and promotes aging-related phenotypes. *Cell* **149**, 1298–1313.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I. and Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics* **6**, e1000895.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics* **6**, e1000888.

Nojiri, H., Shimizu, T., Funakoshi, M., Yamaguchi, O., Zhou, H., Kawakami, S. et al. (2006). Oxidative stress causes heart failure with impaired mitochondrial respiration. *Journal of Biological Chemistry* **281**, 33789–33801.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* **69**, 1–14.

Reich, D. E., Cargill, M., Bolk, S. and Ireland, J. (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

Rivera-Feliciano, J. and Tabin, C. J. (2006). Bmp2 instructs cardiac progenitors to form the heart-valve-inducing field. *Developmental Biology* **295**, 580–588.

Sikorska, K., Lesaffre, E., Groenen, P. F. and Eilers, P. H. (2013). GWAS on your notebook: Fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics* **14**, 166.

Subramanian, A., Tamayo, P., Mootha, V. K. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550.

Tanigaki, K., Sundgren, N., Khera, A., Vongpatanasin, W., Mineo, C. and Shaul, P. W. (2015). Fcγ receptors and ligands and cardiovascular disease. *Circulation Research* **116**, 368–384.

Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory* **62**, 3702–3720.

Wu, Y. and Yang, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Stat.* **47**, 857–883.

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.

Zhao, S. D., Cai, T. T., Cappola, T. P., Margulies, K. B. and Li, H. (2017). Sparse simultaneous signal detection for identifying genetically controlled disease genes. *J. Am. Stat.*

Assoc. **112**, 1032–1046.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E. et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**, 481–487.

Rong Ma

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

E-mail: rongm@upenn.edu

T. Tony Cai

Department of Statistics, The Wharton School, 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 2019; accepted October 2020)