Statistica Sinica **31** (2021), 1985-2004
doi:https://doi.org/10.5705/ss.202019.0182

REMI: REGRESSION WITH MARGINAL INFORMATION AND ITS APPLICATION IN GENOME-WIDE ASSOCIATION STUDIES

Jian Huang, Yuling Jiao, Jin Liu and Can Yang

University of Iowa, Zhongnan University of Economics and Law, Duke-NUS Medical School and Hong Kong University of Science and Technology

Abstract: We consider the problem of variable selection and estimation in high-dimensional linear regression models when complete data are not accessible, but we do have certain marginal information or summary statistics. This problem is motivated by genome-wide association studies (GWASs) with millions of genotyped single nucleotide polymorphisms (SNPs), which have been widely used to identify risk variants among complex human traits/diseases. With the large number of completed GWASs, statistical methods using summary statistics have become increasingly important because of the inaccessibility of individual-level data. In this study, we propose the regression with marginal information (REMI) method, an ℓ_1 penalized approach with estimated marginal effects and an estimated covariance matrix of the predictors with external reference samples. The proposed method is highly scalable and capable of analyzing multiple GWAS data sets from hundreds of thousands individuals and a large number of SNPs. We also establish an upper bound on the error of the REMI estimator, which has the same order as that of the minimax error bound of the Lasso with complete individual-level data. We conduct simulation studies to evaluate the performance of the proposed method. An interesting finding is that when there is a large number of marginal estimates available with a small number of reference samples, as in a GWAS, the proposed method yields good estimation and prediction results, outperforming the Lasso with complete data, but with a relatively small sample size. We apply the proposed method to the 10 traits GWAS data of the Northern Finland Birth Cohorts program. In particular, the real-data analysis results indicate that a summary-level-based analysis using the REMI outperforms an individual-level-based analysis when the sample size of the summary-level data is larger than that of the individual-level data. In summary, our theoretical and real-data results provide solid support for a summary-level-based analysis. As a result, polygenic risk scores of a wide variety of complex diseases can be obtained using summary statistics with theoretically guaranteed performance. The developed R package and the code to reproduce the results are available at `https://github.com/gordonliu810822/REMI`.

Corresponding author: Jin Liu, Centre for Quantitative Medicine, Program in Health Services and Systems Research, Duke-NUS Medical School, Singapore 169857. E-mail: jin.liu@duke-nus.edu.sg.

Key words and phrases: Genome-wide association studies, high dimensional regression, marginal information, polygenic risk score.

1. Introduction

High-dimensional regressions are widely applied in fields such as medicine, biology, finance, and marketing (Hastie, Tibshirani and Friedman (2009)). Consider the linear regression model that relates a response variable Y to a vector of p predictors $X = (X_1, \ldots, X_p)^T$:

$$Y = \sum_{j=1}^{p} X_j \beta_j^* + \epsilon, \tag{1.1}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ is a vector of regression coefficients, and ϵ is a random error term with mean zero and noise level σ_ϵ^2. In most applications, the data set comprises an $n \times p$ matrix $\mathbf{X}$ of variables in X and a vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ of responses Y collected from n individuals. Given the individual-level data $\{\mathbf{X}, \mathbf{y}\}$, there exist convex (Tibshirani (1996); Candes and Tao (2007)) and nonconvex (Fan and Li (2001); Zhang (2010)) penalized methods for estimating $\boldsymbol{\beta}^*$ with a theoretical guarantee (Zhao and Yu (2006); Meinshausen and Bühlmann (2006); Zhang and Huang (2008); Bickel, Ritov and Tsybakov (2009); Zhang and Zhang (2012)). See also the monographs (Bühlmann and van de Geer (2011); Hastie, Tibshirani and Wainwright (2015)), and the references therein.

Motivated by applications in human genetics, we consider the problem of estimating $\boldsymbol{\beta}^*$ when individual-level data $\{\mathbf{X}, \mathbf{y}\}$ are not accessible, but marginal information is available, such as $\mathbf{X}_j^T \mathbf{y}$ and $\mathbf{X}_j^T \mathbf{X}_j$, for $j = 1, \ldots, p$, where $\mathbf{X}_j$ is the jth column of $\mathbf{X}$. Therefore, we refer to our problem formulation as a "regression with marginal information" (REMI). To make our formulation feasible, we assume the covariance structure of the variables in X can be estimated using a reference panel data set in the form of an $n_{\mathrm{r}} \times p$ data matrix $\mathbf{X}_{\mathrm{r}}$, where n_{r} is the number of samples from the reference panel and $n_{\mathrm{r}} \ll p$. A natural question arises: Without accessing individual-level data, can we use the marginal information and the reference data $\mathbf{X}_{\mathrm{r}}$ to estimate $\boldsymbol{\beta}^*$, assuming observations in $\mathbf{X}_{\mathrm{r}}$ and $\mathbf{X}$ are from the same distribution?

In particular, our problem arises in genome-wide association studies (GWASs), which have been conducted over the past decade to study the genetic basis of human complex phenotypes, including both quantitative traits and complex diseases (Hindorff et al. (2009); Welter et al. (2014); Visscher et al. (2017)). As of April 2018, more than 59,000 unique phenotype-variant (typically single-nucleotide polymorphism, or SNP) associations have been reported

in about 3,300 GWAS publications (see the GWAS Catalog database (`https://www.ebi.ac.uk/gwas/`). An important lesson from the GWASs (Yang et al. (2010); Visscher et al. (2012, 2017)) is that complex phenotypes are highly polygenic; that is, they are often affected by many genetic variants with small effects. Well-known examples include human height (Wood et al. (2014)), psychiatric disorders (Gratten et al. (2014)), and diabetes (Fuchsberger et al. (2016)). Due to the polygenicity, variants with small effects remain largely undiscovered, and large sample sizes are required to explore the genetic architectures of complex phenotypes. As a results, researchers are forming large genomic consortia, such as the genetic investigation of anthropometric traits (GIANT) consortium and the psychiatric genomic consortium (PGC), to maximize sample sizes, aiming at a deeper understanding of these architectures.

Despite the prevalence of data sharing, it is still difficult for a research group to fully access the individual-level genotype data available in a consortium. For example, a core research group from the GIANT consortium reported that they can only access genotype data from about 44,000 individuals (Yang et al. (2015)), even though the total sample size is more than 250,000 for the consortium (Wood et al. (2014)). There are several reasons for the restricted access to individual-level data. First, privacy protection is always a concern when sharing individual-level genotype data. Second, it is often time-consuming to achieve agreement on data sharing among different research groups. Third, many practical issues arise in data transportation and storage. In contrast, GWAS summary statistics are widely available through many public gateways (Genetics (2012)), for example, the download session at the GWAS catalog `https://www.ebi.ac.uk/gwas/downloads/summary-statistics`. Because these summary statistics (e.g., estimated effect sizes, standard errors, and z-values) are often generated by a simple linear regression analysis, summary statistics are essentially marginal information.

To meet the demand for data analysis in GWASs, various statistical methods have been proposed that use marginal information. Using a few hundred human-genome data values from the 1000 Genome Project as a reference panel, we find information on the correlation structure of genetic variants (typically using "linkage disequilibrium" in genetics, or LD). This allows these methods to bypass the individual-level data, using only marginal information. Here, we roughly divide these methods into three categories. The first is methods for heritability estimation. The heritability of a phenotype quantifies the relative importance of genetics and the environment to the phenotype (Visscher, Hill and Wray (2008)). When individual-level data are accessible, linear mixed-model

(LMM)-based approaches (e.g., GCTA, Yang et al. (2010, 2011)) are widely used for heritability estimation (Lee et al. (2011)). In the absence of individual-level data, Bulik-Sullivan et al. (2015) first introduced the LD score regression, called LDScore, for heritability estimation using only summary statistics and reference data from the 1000 Genome Project. Based on the minimal-norm quadratic unbiased estimation criteria, Zhou (2016) proposed a novel method of moments, MQS, for variance component estimation using summary statistics. The second category is methods for association mapping. Heritability estimation provides a global measure that quantifies the overall contribution from genetic factors. Association mapping localizes genetic variants associated with a given phenotype. Recently, several statistical methods have been proposed for association mapping based on summary statistics, including the functional GWAS (FGWAS) (Pickrell (2014)), probabilistic annotation integrator (PAINTOR) (Kichaev et al. (2014)), causal variants identification in associated regions (CAVIAR) (Hormozdiari et al. (2014)), and CAVIAR Bayes factor (CAVIARBF) (Chen et al. (2015)). Although these methods are very useful for performing association mapping on summary statistics, they still have limitations. First, they adopt ad-hoc methods of reducing the computational cost. For example, to avoid a combinatorial search, FGWAS assumes there is only one causal signal in an LD block, and PAINTOR searches no more than two causal variants in its default setting. Second, the statistical analysis is oversimplified in order to overcome estimation difficulties. For example, the noncentrality parameter in PAINTOR and the variance components in CAVIAR and CAVIARBF are pre-fixed, rather than adaptively estimated from the data. The third category contains methods for effect size estimation and risk prediction. Recently, Vilhjálmsson et al. (2015) proposed a Bayesian method, LDpred, for effect size estimation and risk prediction by accounting for LD. Along this line, Hu et al. (2017) proposed AnnoPred to improve LDpred by incorporating functional information on the human genome. However, neither LDpred nor AnnonPred should be considered as a marginal-information-based method because they both require individual-level data as validation data for their parameter tuning.

Most recently, Ning et al. (2017) proposed the selection operator for jointly analyzing multiple variants (SOJO), which uses GWAS summary statistics to identify association signals in complex traits. However, they do not consider the theoretical properties of this approach.

Although existing statistical methods have shown good empirical performance in GWAS data analyses, there are a number of open questions related to regressions with marginal information. First, the sample size of the reference

panel is often very small. For example, there are only about 370 samples from the 1000 Genome Project that can be used as references when analyzing GWAS data on European ancestry. It remains unclear why such a small sample size is good enough for exploring the correlation structure of a large number of variables (i.e., SNPs). Second, the theoretical properties of existing methods for effect size estimation and prediction errors have not been studied. Third, the algorithms are often time-consuming, because they need to run thousands of Markov chain Monte Carlo (MCMC) iterations. Positive answers to these questions will benefit GWAS data analysis. For example, the polygenic risk score of a disease can be established without accessing individual-level data, instead using summary statistics and a few reference samples. This saves on computational effort without sacrificing statistical accuracy.

We propose a statistically guaranteed method, the regression with marginal information (REMI) method, to address the above open questions. The rest of this paper is organized as follows. In Section 2, we introduce our REMI model and discuss its use with GWAS data. In Section 3, we present an efficient coordinate descent algorithm, and discuss practical issues in implementing this algorithm. In Section 4, we establish the error bound and prediction error of the proposed method. In particular, our theoretical results explain why a small number of samples (i.e., n_{r}) from the reference panel can be good enough for effect size estimation and risk prediction. In Section 5, we present the results of simulation studies and a real-data analysis. In particular, our results indicate that a summary-level-based analysis using the REMI method outperforms an individual-level-based analysis when the sample size in summary-level data is larger than that of the individual-level data.

2. The REMI Model

2.1. The REMI model

For the linear regression model in (1.1), if the individual-level data $(\mathbf{y}, \mathbf{X})$ are available, a basic approach for estimating $\boldsymbol{\beta}^*$ in high-dimensional settings is the Lasso (Tibshirani (1996)). The Lasso estimator is given by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.1}$$

where $\|\cdot\|_1$ is the ℓ_1 norm, and $\lambda \geq 0$ is a regularization parameter. In our problem, however, the individual-level data $\{\mathbf{X}, \mathbf{y}\}$ are not accessible. Hence, a direct application of the Lasso is not feasible. Note that several other im-

portant penalized methods have been proposed, including the smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)) and minimax concave penalty (MCP) (Zhang (2010)). We focus on the Lasso penalty below, although our proposed approach can also be based on the other penalties.

We now describe our proposed REMI model with the Lasso penalty. Rewrite (2.1) as

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) + \lambda\|\boldsymbol{\beta}\|_1 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \frac{2}{n}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \lambda\|\boldsymbol{\beta}\|_1, \end{aligned} \tag{2.2}$$

where the second term only involves the inner product of the optimization variable $\boldsymbol{\beta}$ and marginal information, say, $\widetilde{\mathbf{y}} = \mathbf{X}^T\mathbf{y}/n$, which we assume is available. The difficulty comes from the first term, where $\mathbf{X}^T\mathbf{X}/n$ is unknown because $\mathbf{X}$ is not observed. Motivated by applications in GWASs, we assume there exists a reference $n_{\mathrm{r}} \times p$ data matrix $\mathbf{X}_{\mathrm{r}}$, where the rows of $\mathbf{X}_{\mathrm{r}}$ are independent and identically distributed (i.i.d.) and have the same distribution, with the covariance matrix $\boldsymbol{\Sigma}$ as the rows of $\mathbf{X}$. Therefore, both $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^T\mathbf{X}/n$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}} = \mathbf{X}_{\mathrm{r}}^T\mathbf{X}_{\mathrm{r}}/n_{\mathrm{r}}$ can be viewed as estimators of $\boldsymbol{\Sigma}$ and we propose solving the following optimization problem to estimate $\boldsymbol{\beta}^*$:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{c}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n_{\mathrm{r}}}\boldsymbol{\beta}^T\mathbf{X}_{\mathrm{r}}^T\mathbf{X}_{\mathrm{r}}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\widetilde{\mathbf{y}} + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.3}$$

where $\widehat{\boldsymbol{\beta}}^{\mathrm{c}}$ denotes the estimator using the reference covariance matrix. Clearly, the above model (2.3) uses only the marginal correlation between $\mathbf{X}$ and $\mathbf{y}$, with the covariance matrix estimated using an external reference panel $\mathbf{X}_{\mathrm{r}}$.

2.2. REMI in GWAS

In the context of a GWAS, rather than have $\widetilde{\mathbf{y}} = \mathbf{X}^T\mathbf{y}/n$ as marginal information, we may only have the summary statistics $\{\widehat{\beta}_j^{\mathrm{m}}, \hat{s}_j^2\}_{j=1,\ldots,p}$ from the univariate linear regression:

$$\widehat{\beta}_j^{\mathrm{m}} = (\mathbf{X}_j^T\mathbf{X}_j)^{-1}\mathbf{X}_j^T\mathbf{y}, \quad \hat{s}_j^2 = (n\mathbf{X}_j^T\mathbf{X}_j)^{-1}(\mathbf{y} - \mathbf{X}_j\widehat{\beta}_j^{\mathrm{m}})^T(\mathbf{y} - \mathbf{X}_j\widehat{\beta}_j^{\mathrm{m}}),$$

where the superscript m is used to denote marginal information, and $\widehat{\beta}_j^{\mathrm{m}}$ and $\hat{s}_j^2$ are the estimated effect size and its variance, respectively, for SNP j. Owing to the polygenicity of many complex phenotypes, the standard errors can be well approximated by $\hat{s}_j \approx \sqrt{(n\mathbf{X}_j^T\mathbf{X}_j)^{-1}\mathbf{y}^T\mathbf{y}}$ (Zhu and Stephens (2017)). Let $\widehat{\boldsymbol{\beta}}^{\mathrm{m}} =$

$[\widehat{\beta}_1^{\mathrm{m}}, \ldots, \widehat{\beta}_p^{\mathrm{m}}]^T$, $\hat{\mathbf{s}}^2 = [\hat{s}_1^2, \ldots, \hat{s}_p^2]^T$ be the vectors collecting estimated effect sizes and estimated variances, respectively, and let $\widehat{\mathbf{S}}$ be a $p \times p$ diagonal matrix, with $\hat{s}_j$ as its jth diagonal element. Furthermore, we introduce a $p \times p$ diagonal matrix $\widehat{\mathbf{D}} = \mathrm{diag}(\hat{d}_j)$, with its jth diagonal element being the sample standard deviation of $\mathbf{X}_j$, that is, $\hat{d}_j = \sqrt{\mathbf{X}_j^T \mathbf{X}_j / n}$, and a correlation matrix $\widehat{\mathbf{R}} = [\hat{r}_{jk}] \in \mathbb{R}^{p \times p}$, with $\hat{r}_{jk} = \mathbf{X}_j^T \mathbf{X}_k / ((\mathbf{X}_j^T \mathbf{X}_j)^{1/2} (\mathbf{X}_k^T \mathbf{X}_k)^{1/2})$. Noting that $\hat{d}_j^2 \widehat{\beta}_j^{\mathrm{m}} = \mathbf{X}_j^T \mathbf{y}/n$ and $n^2 \hat{d}_j^2 \hat{s}_j^2 \approx \mathbf{y}^T \mathbf{y}$, the REMI formulation (2.2) becomes

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \ \frac{1}{n} \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} - \frac{2}{n} \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \lambda \|\boldsymbol{\beta}\|_1, \\
&= \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \ \boldsymbol{\beta}^T \widehat{\mathbf{D}} \widehat{\mathbf{R}} \widehat{\mathbf{D}} \boldsymbol{\beta} - 2 \boldsymbol{\beta}^T \widehat{\mathbf{D}}^2 \widehat{\boldsymbol{\beta}}^{\mathrm{m}} + \lambda \|\boldsymbol{\beta}\|_1, \\
&\approx \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \ \frac{\mathbf{y}^T \mathbf{y}}{n^2} \boldsymbol{\beta}^T \widehat{\mathbf{S}}^{-1} \widehat{\mathbf{R}} \widehat{\mathbf{S}}^{-1} \boldsymbol{\beta} - 2 \frac{\mathbf{y}^T \mathbf{y}}{n^2} \boldsymbol{\beta}^T \widehat{\mathbf{S}}^{-2} \widehat{\boldsymbol{\beta}}^{\mathrm{m}} + \lambda \|\boldsymbol{\beta}\|_1, \\
&= \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \ \boldsymbol{\beta}^T \widehat{\mathbf{S}}^{-1} \widehat{\mathbf{R}} \widehat{\mathbf{S}}^{-1} \boldsymbol{\beta} - 2 \boldsymbol{\beta}^T \widehat{\mathbf{S}}^{-2} \widehat{\boldsymbol{\beta}}^{\mathrm{m}} + \widetilde{\lambda} \|\boldsymbol{\beta}\|_1,
\end{aligned}$$

where $\widetilde{\lambda} = (n^2/(\mathbf{y}^T \mathbf{y}))\lambda$, and the approximation holds in the case of polygenicity. Because $\widetilde{\lambda}$ is a tuning parameter that scales λ with a constant factor $(n^2/(\mathbf{y}^T \mathbf{y}))$, we slightly abuse λ for $\widetilde{\lambda}$, and propose solving the following optimization problem:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{r}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \ L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1, \tag{2.4}$$

where $L(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \widehat{\mathbf{S}}^{-1} \widehat{\mathbf{R}} \widehat{\mathbf{S}}^{-1} \boldsymbol{\beta} - 2 \boldsymbol{\beta}^T \widehat{\mathbf{S}}^{-2} \widehat{\boldsymbol{\beta}}^{\mathrm{m}}$, and $\widehat{\boldsymbol{\beta}}^{\mathrm{r}}$ denotes the estimates based on the correlation information. Similarly to the REMI (2.3), in which the covariance matrix $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^T \mathbf{X}/n$ needs to be estimated, the correlation matrix $\widehat{\mathbf{R}}$ needs to be estimated using samples from the reference panel $\mathbf{X}_{\mathrm{r}}$. We refer (2.3) as REMI-C, and to (2.4) as REMI-R.

3. Algorithm and Practical Issues

3.1. Algorithm

Here, we adopt the widely used coordinate descent algorithm, which updates one parameter at a time, say $\widehat{\beta}_j^{\mathrm{c}}$, keeping all other parameters fixed at their current values. Thus, the sub-problem for parameter $\widehat{\beta}_j^{\mathrm{c}}$ can be written as

$$\widehat{\beta}_j^{\mathrm{c}}(\lambda) = \underset{\beta_j}{\mathrm{argmin}} \, \widehat{\sigma}_{jj} \beta_j^2 - 2 \left(\widetilde{y}_j - \sum_{k \neq j} \widehat{\beta}_k^{\mathrm{c}} \widehat{\sigma}_{jk} \right) \beta_j + \lambda |\beta_j|, \tag{3.1}$$

Algorithm 1 Path algorithm to solve REMI-C (2.3) with a sequence of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)$.

Output: Solution path for $\widehat{\boldsymbol{\beta}}^{\mathrm{c}}(\boldsymbol{\lambda})$.
for $l = 1, 2, \ldots, D$ **do**
  Initialize $\widehat{\boldsymbol{\beta}}^{\mathrm{c}}(\lambda_l) = \widehat{\boldsymbol{\beta}}^{\mathrm{c}}(\lambda_{l-1})$, if $l > 1$; $\widehat{\boldsymbol{\beta}}(\lambda_l) = \mathbf{0}$, if $l = 1$
  repeat
    for $j = 1, 2, \ldots, p$ **do**
      $\eta_j = \widetilde{y}_j - \sum_{k \neq j} \widehat{\beta}^{\mathrm{c}}_k(\lambda)\widehat{\sigma}_{jk}$
      $\widehat{\beta}^{\mathrm{c}}_j(\lambda) \leftarrow \mathrm{S}(\eta_j, \lambda/2)/\widehat{\sigma}_{jj}$
    end
  until *Convergence*;
end

where $\widehat{\sigma}_{jk}$ is an element in $\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}} = [\widehat{\sigma}_{jk}] \in \mathbb{R}^{p \times p}$. An efficient path algorithm can be developed based on a warm start and some other tricks, as described in Friedman, Hastie and Tibshirani (2010). In particular, we generate a sequence of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)$, equally spaced in logarithm form, with $\lambda_1 = \lambda_{\max}$ and $\lambda_D = \tau\lambda_{\max}$, where $\lambda_{\max}$ is the minimum λ that shrinks all parameters to zero, and τ is usually set to 0.05. For each λ, we use the solution of (2.3) from the last λ value as the warm start. The path algorithm is described in Algorithm 1.

Similarly to REMI-C, an efficient coordinate descent algorithm can be developed to solve REMI-R (2.4). The efficient path algorithm is given in Algorithm 2.

3.2. Reference panel

The REMI-R model in (2.4) involves a cohort-based estimated correlation matrix. Based on the nature of the correlation patterns of the SNPs, $\mathbf{R}$ can be approximated using a block diagonal matrix. Specifically, we first partition the whole genome into L blocks ($L = 1{,}703$ for European ancestry, and $L = 1{,}445$ for Asian ancestry) (Berisa and Pickrell (2016)). Then, we calculate the empirical correlation matrix $\widehat{\mathbf{R}}_{\mathrm{emp}}$ for each LD-block. To ensure a stable numerical result, we apply a simple shrinkage estimator to obtain $\widehat{\mathbf{R}}^{\mathrm{r}} = \kappa\widehat{\mathbf{R}}_{\mathrm{emp}} + (1 - \kappa)\mathbf{I}$ within each block (Schäfer and Strimmer (2005)), where we use $\kappa = 0.9$ as the default (the estimate of $\boldsymbol{\beta}^*$ is insensitive to κ, Pasaniuc et al. (2014)). Thus, similarly to Zhu and Stephens (2017), the REMI methods and their individual-level-data counterparts will produce approximately the same inferential results within a region. After substituting $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{R}}^{\mathrm{r}}$ into (2.4), we use the coordinate descent algorithm to obtain $\widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda)$ (Algorithm 2).

Algorithm 2 Path algorithm to solve REMI-R (2.4) with a sequence of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)$.

Output: Solution path for $\widehat{\boldsymbol{\beta}}^{\mathrm{c}}(\boldsymbol{\lambda})$.
for $l = 1, 2, \ldots, D$ **do**
  Initialize $\widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda_l) = \widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda_{l-1})$, if $l > 1$; $\widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda_l) = \mathbf{0}$, if $l = 1$
  repeat
    for $j = 1, 2, \ldots, p$ **do**
      $\eta_j = \widehat{\beta}_j^{\,\mathrm{m}}/\hat{s}_j^2 - (1/\hat{s}_j)\sum_{k \neq j} \widehat{\beta}_k^{\,\mathrm{r}} \hat{r}_{jk}/\hat{s}_k$
      $\widehat{\beta}_j^{\,\mathrm{r}}(\lambda) \leftarrow \mathrm{S}(\eta_j, \lambda/2) \times \hat{s}_j^2$
    end
  until *Convergence*;
end

3.3. Choice of regularization parameter λ

The REMIs have one regularization parameter, λ. Here, we briefly show how to choose this parameter for REMI-R; it is straightforward to develop the same strategy for REMI-C. Similarly to the Lasso solver (Friedman, Hastie and Tibshirani (2010)), we generate a sequence of $\boldsymbol{\lambda}$ from $\lambda_{\max}$ to $\tau\lambda_{\max}$, where $\lambda_{\max}$ is the minimum value of λ that shrinks all parameters to zero, and τ is prespecified with a default value of 0.05. Note that $\lambda_{\max} = \max\{2\hat{\beta}_j^{\mathrm{m}}/\hat{s}_j^2\}_{j=1,\ldots,p}$. We search for the optimal value of λ using they Bayesian information criterion (BIC),

$$\mathrm{BIC}(\lambda_l) = L(\widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda_l)) + \log(n)\mathrm{df}(\lambda_l). \tag{3.2}$$

Zou, Hastieand and Tibshirani (2007) showed that the number of nonzero coefficients is an unbiased estimate for the degrees of freedom of the Lasso. We choose $\mathrm{df}(\lambda_l)$ to be the number of nonzero coefficients, given λ_l. To fairly compare the REMIs with the Lasso, we also use the BIC to select the regularization parameter when individual-level data are accessible.

4. Theoretical Properties

In this section, we give nonasymptotic bounds on the estimation error $\|\boldsymbol{\beta}^*_{\mathcal{A}} - \widehat{\boldsymbol{\beta}}^{\mathrm{c}}\|$ and the prediction error $\|\mathbf{X}_{\mathrm{new}}\boldsymbol{\beta}^*_{\mathcal{A}} - \mathbf{X}_{\mathrm{new}}\widehat{\boldsymbol{\beta}}^{\mathrm{c}}\|^2/n_{\mathrm{new}}$, where $\mathcal{A}$ denotes an index of significant entries of $\boldsymbol{\beta}^*$, and $\boldsymbol{\beta}^*_{\mathcal{A}}$ denotes the vector supported on $\mathcal{A}$.

In real applications using genetic data, the underlying signal is not exactly sparse, but contains many small components. Here, we assume the target $\boldsymbol{\beta}^*$ is weakly sparse; that is, in addition to some significant components indexed by $\mathcal{A}$, there may be many nonzero entries in $\boldsymbol{\beta}^*$ with very small magnitude, as indexed by $\mathcal{I} = \mathcal{A}^c$. Let $\boldsymbol{\beta}^*_{\mathcal{I}}$ be the vector supported on $\mathcal{I}$. It is reasonable to assume that

$$s = |\mathcal{A}| \le n, \quad \|\boldsymbol{\beta}^*_{\mathcal{I}}\|_\infty \le 2\sigma_{\boldsymbol{\epsilon}}\sqrt{\frac{\log p}{n}}, \tag{4.1}$$

because a signal with a magnitude smaller than this order is undetectable. Then,

$$\widetilde{\mathbf{y}} = \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\epsilon}} = \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{A}} + \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{I}} + \widetilde{\boldsymbol{\epsilon}}, \tag{4.2}$$

where, $\widetilde{\boldsymbol{\epsilon}} = \mathbf{X}^T\boldsymbol{\epsilon}/n$. Let $C_1 \ge \sqrt{\|\mathrm{diag}(\boldsymbol{\Sigma})\|_\infty}$, $C_2 \ge \max_{j=1,\ldots,p}\{\|\boldsymbol{\Sigma}_j\|_1\}$, $C_3 \ge \|\boldsymbol{\beta}^*_{\mathcal{A}}\|_1$, and $C_4 \ge \|\boldsymbol{\beta}^*_{\mathcal{I}}\|_1$. The restricted eigenvalue (Bickel, Ritov and Tsybakov (2009)) of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}$ is defined as

$$\phi_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}} = \min_{0\neq\mathbf{v}\in\mathcal{C}_{\mathcal{A},3}} \frac{\mathbf{v}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\mathbf{v}}{\|\mathbf{v}\|_2^2},$$

where

$$\mathcal{C}_{\mathcal{A},3} = \{\mathbf{v} \in \mathcal{R}^p : \|\mathbf{v}_{\mathcal{I}}\|_1 \le 3\|\mathbf{v}_{\mathcal{A}}\|_1\}.$$

Theorem 1. *Assume the rows of* $\mathbf{X}$ *and* $\mathbf{X}_{\mathrm{r}}$ *are i.i.d sub-Gaussian samples drawn from a population with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}$, $\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}$ *satisfies the restricted eigenvalue condition with* $\phi_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}} \ge \phi_0 > 0$, *and the noise vector* $\boldsymbol{\epsilon}$ *is mean-zero sub-Gaussian with noise level* σ_ϵ, *and* $n \ge 4/C\log p$, $n_{\mathrm{r}} \ge 4/C\log p$. *Take* $\lambda \ge 2\lambda_0 = 4(((C_1(C_3+C_4))/\sqrt{C})\sqrt{\log p/n} + ((C_1+\sqrt{C}C_2)/\sqrt{C})\sigma_\epsilon\sqrt{\log p/n} + (C_1C_3/\sqrt{C})\sqrt{\log p/n_{\mathrm{r}}})$.

(i) With probability at least $1 - 3/p^2 - 1/p^3$, *we have*

$$\|\widehat{\boldsymbol{\beta}}^{\mathrm{c}} - \boldsymbol{\beta}^*_{\mathcal{A}}\| \le \frac{6}{\phi_0}\left(\frac{C_1(C_3+C_4)}{\sqrt{C}}\sqrt{\frac{s\log p}{n}} + \frac{C_1+\sqrt{C}C_2}{\sqrt{C}}\sigma_\epsilon\sqrt{\frac{s\log p}{n}} + \frac{C_1C_3}{\sqrt{C}}\sqrt{\frac{s\log p}{n_{\mathrm{r}}}}\right).$$

(ii) Suppose we observe $\mathbf{X}_{\mathrm{new}} \in \mathcal{R}^{n_{\mathrm{new}}\times p}$, *the rows of which are sampled from the same distribution as that of* $\mathbf{X}$. *Then, with probability at least* $1-3/p^2-1/p^3$, *the prediction error satisfies*

$$\frac{\|\mathbf{X}_{\mathrm{new}}(\widehat{\boldsymbol{\beta}}^{\mathrm{c}} - \boldsymbol{\beta}^*_{\mathcal{A}})\|_2^2}{n_{\mathrm{new}}} \le \mathcal{O}\left(\left(\sigma_\epsilon\sqrt{\frac{s\log p}{n}} + \sqrt{\frac{s\log p}{n_{\mathrm{r}}}}\right)^2\right)\left(1 + s^2\left(\sqrt{\frac{\log p}{n_{\mathrm{r}}}} + \sqrt{\frac{\log p}{n_{\mathrm{new}}}}\right)\right).$$

Remark 1. The assumption that $\mathbf{X}_{\mathrm{r}}$ are i.i.d sub-Gaussian samples drawn from a population with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ implies that the restricted

eigenvalue condition $\phi_{\widehat{\mathbf{\Sigma}}_{\mathrm{r}}} \geq \phi_0$ holds for some positive ϕ_0 with high probability, as long as $n_{\mathrm{r}} \geq \mathcal{O}(s \log p)$ (Van De Geer and Bühlmann (2009); Vershynin (2010); Huang et al. (2018)). As shown in Theorem 1, with the help of a reference panel, we can obtain an accurate estimator using (2.3), even if we have only marginal information in a high-dimension setting, as long as $n \geq \mathcal{O}(s \log p)$ and $n_{\mathrm{r}} \geq \mathcal{O}(s \log p)$. Furthermore, the estimation error of the REMI model in (2.3) achieves the minimax optimal rate of the Lasso (Raskutti, Wainwright and Yu (2011)) if the number of samples in the reference panel n_{r} is of the same order as the number of individual-level samples n. Moreover, if the magnitude of the significant entries is larger than $\mathcal{O}(\sigma_\epsilon \sqrt{s \log p/n} + \sqrt{s \log p/n_{\mathrm{r}}})$, then the estimated support $\mathrm{supp}(\widehat{\boldsymbol{\beta}}^{\mathrm{c}})$ coincides with the true significant set $\mathcal{A}$.

5. Numerical Studies

5.1. Simulation studies

In our simulation studies, we compare the REMI-C (2.3), REMI-R (2.4), and Lasso using individual-level data. To avoid unrealistic LD patterns in the simulations, we used the genotype data $\mathbf{X}$ from the RPGEH data set (Hoffmann et al. (2011)). The RPGEH data set provides 657,184 genotyped SNPs for 62,313 European individuals. We perform strict quality control on the data using PLINK (Purcell et al. (2007)). We exclude SNPs with a minor allele frequency of less than 1%, those with missing values in more than 1% of individuals, and those with a Hardy–Weinberg equilibrium p-value below 0.0001. Moreover, we remove one member of pairs with genetic relatedness larger than 0.05. Finally, there remained 53,940 samples for 550,482 SNPs.

Because individual-level-based analyses often suffer from limited sample sizes, owing to restricted access to individual-level data, summary-level-based analyses may have advantages because their sample sizes are often much larger. To simulate this situation, we pre-fixed the sample size for the individual-level-based analyses at $n_{\mathrm{ind}} = 3{,}000$. Specifically, we randomly selected n_{ind} of the 53,940 individuals in the RPGEH data set to form the genotype matrix $\mathbf{X} \in \mathbb{R}^{n_{\mathrm{ind}} \times p}$, where $p = 19{,}865$ is the total number of genotyped SNPs on chromosomes 16, 17, and 18. Then, the phenotype vector $\mathbf{y}$ was generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and the heritability ($h^2 = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta})/(\mathrm{Var}(\mathbf{X}\boldsymbol{\beta}^*) + \sigma_\epsilon^2)$) was controlled at 0.2, 0.3, 0.4, and 0.5. Here, $\boldsymbol{\beta}^*$ is the vector of the true effect size, with sparsity α; that is, $\alpha \times p$ entries in $\boldsymbol{\beta}^*$ are nonzero and sampled from $\mathcal{N}(0, 1)$. In our simulation study, we varied α in $\{0.001, 0.003, 0.005, 0.007, 0.01, 0.02\}$. With $\{\mathbf{X}, \mathbf{y}\}$ at hand, the standard Lasso can be applied, serving as a reference for the

individual-level data analysis.

To generate summary-level data, we varied the sample size n from 3,000 to 50,000. We generated individual-level data as described above, and then ran a simple linear regression on $\{\mathbf{X}_j, \mathbf{y}\}$, for $j = 1, \ldots, p$, to obtain $\{\widehat{\boldsymbol{\beta}}^{(m)}, \hat{\mathbf{s}}^2\}$. Then, we pretended that we did not have individual-level data $\{\mathbf{X}, \mathbf{y}\}$, using only $\mathbf{X}^T\mathbf{y}$ and $\{\widehat{\boldsymbol{\beta}}^{(m)}, \hat{\mathbf{s}}^2\}$ as the inputs for REMI-C and REMI-R, respectively. We used 379 European samples from the 1000 Genome Project data as the reference panel to estimate the covariance matrix (REMI-C) and correlation matrix (REMI-R), as discussed in Section 3.2. For each replication, we used 200 independent samples to evaluate the prediction accuracy. Finally, we summarized our results based on 50 replications for each setting.

We compare the performance of the REMI-C, REMI-R, and Lasso using individual-level data in terms of their variable selection and prediction. Specifically, we use the partial area under the receiver operating characteristic (ROC) curve (partial AUC) for the variable selection performance, and use the Pearson correlation coefficient between the predicted and the observed phenotypes for the prediction performance. The results of this simulation study are shown in Figure 1 and Figure 2. First, we observe very little difference between REMI-C and REMI-R. This justifies the approximation made in REMI-R. Second, when the sample size (n = 3,000 or 5,000) of the summary-level data is similar to that of the individual-level data (n_{ind} = 3,000), REMI-C and REMI-R perform similarly to the Lasso in terms of variable selection and prediction. Third, the REMIs gradually outperform the Lasso as the sample size increases from 5,000 to 50,000 for both variable selection and prediction. This clearly indicates that the REMIs have an advantage over the Lasso when the sample size of the summary-level data becomes much larger.

5.2. Real-data analysis

To demonstrate the utility of the REMIs, we first compare the Lasso and the REMIs using the GWAS data set from the Northern Finland Birth Cohorts program (NFBC1966) (Sabatti et al. (2009)). The NFBC1966 data set contains information on 5,402 individuals, with a selected list of phenotypic data related to cardiovascular disease, including high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), triglycerides (TG), C-reactive protein (CRP), glucose, insulin, body mass index (BMI), systolic (SysBP), and diastolic (DiaBP) blood pressure. For each individual, 364,590 SNPs are genotyped. We perform strict quality control on the data using PLINK (Purcell et al. (2007)). We first exclude those individuals with discrepancies between their reported sex

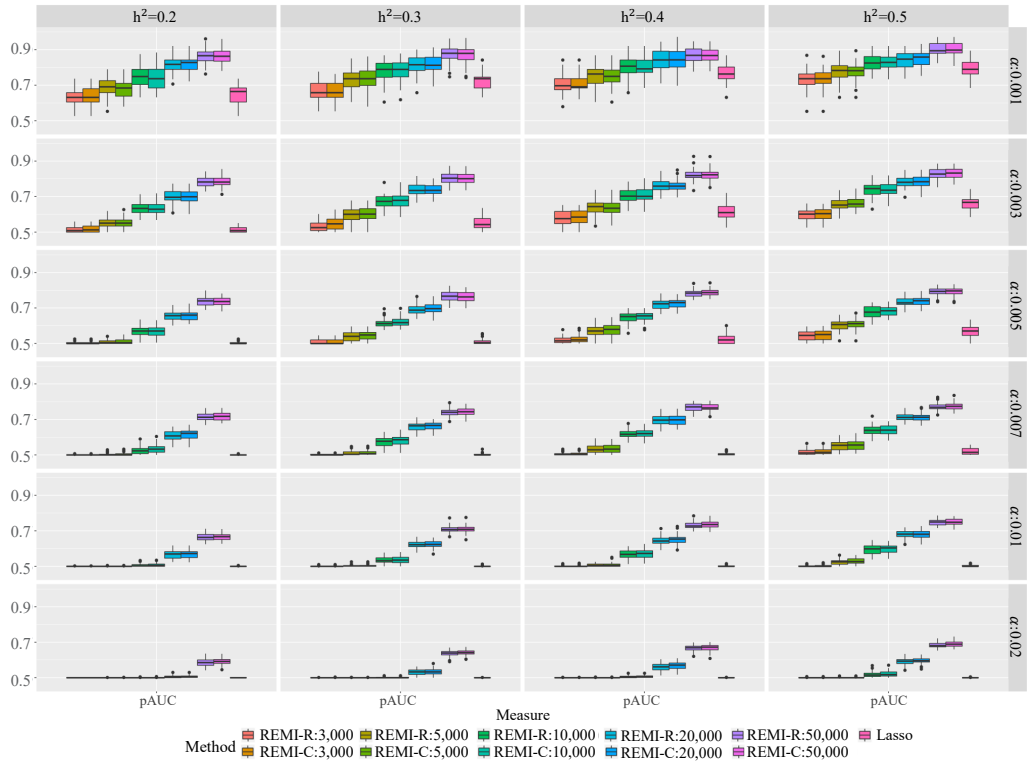

Figure 1. A comparison of the variable selection performance of the REMIs (REMI-R and REMI-C) using summary-statistics data with the Lasso using individual-level data, with sample size 3,000. The sample size used to produce the summary statistics was varied and denoted as $n \in \{3000, 5000, 10000, 20000, 50000\}$. We use the partial AUC to measure the variable selection performance.

and the sex determined from the X chromosome. We also exclude SNPs with a minor allele frequency of less than 1%, those with missing values in more than 1% of individuals, and those with a Hardy–Weinberg equilibrium p-value below 0.0001. In particular, we select well-imputed variants from the HapMap 3 reference panel (The International HapMap 3 Consortium (2010)). After the strict quality control, 5,123 individuals with 310,975 SNPs in NFBC1966 remained for further analysis. Because we have the individual-level data, it is possible to run the REMI-C, REMI-R, and Lasso for all these traits. The solution paths using the three methods for the 10 metabolic traits in the NFBC1966 data set are presented in Figure S1. The dotted vertical bars indicate the corresponding selected tuning parameters based on the BIC. As shown, there are only minor differences between the solution paths for the three methods, which is consistent with the results of our simulation studies discussed in Section 5.1.

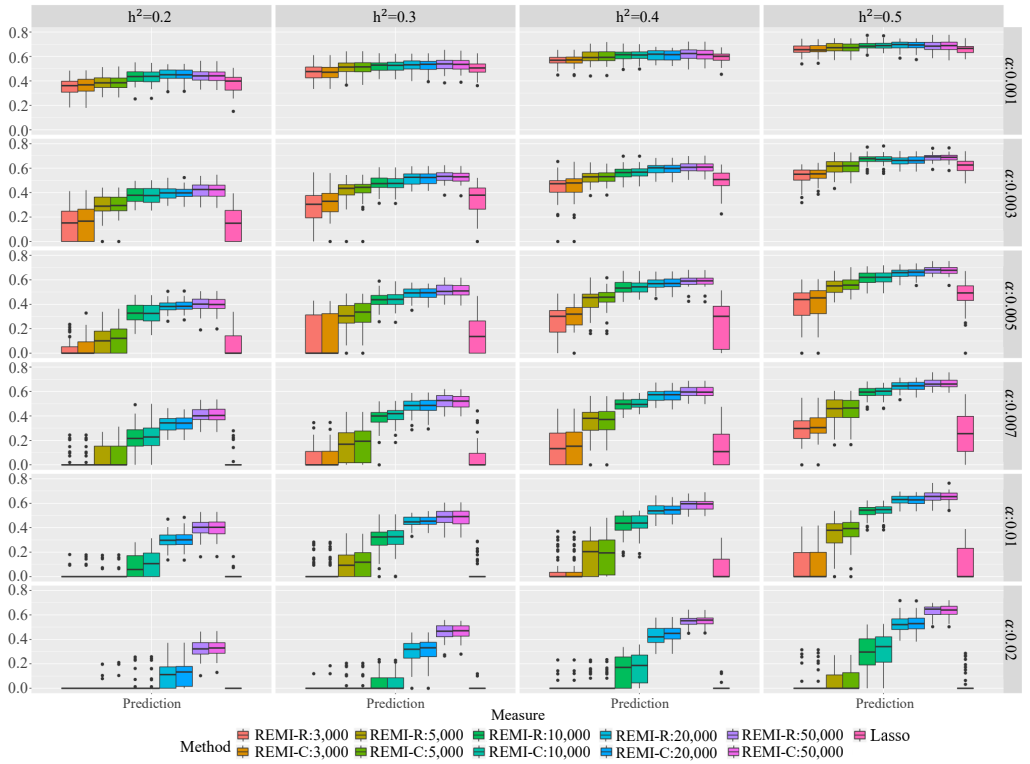

Figure 2. The prediction accuracy of the REMIs using summary-level data with the Lasso using individual-level data, with a sample size 3,000. The sample size used to simulate the summary statistics varies as $n \in \{3000, 5000, 10000, 20000, 50000\}$. The prediction accuracy is measured using the Pearson correlation coefficient between the predicted and the observed phenotypes.

In the released GWAS summary-level data sets, it is often the case that $\{\widehat{\boldsymbol{\beta}}^{(m)}, \hat{\mathbf{s}}^2\}$ rather than the inner product $\mathbf{X}^T\mathbf{y}$ is available. Therefore, we apply the REMI-R to analyze summary statistics for 10 GWASs of complex phenotypes. The source of each GWAS is given in Table S1. Because the individuals that make up the summary-level data sets are all from European ancestry, we use 379 European-ancestry samples from the 1000 Genome Project (The 1000 Genomes Project Consortium. (2012)) as a reference panel to estimate the correlation matrix. Owing to the quality of the SNPs in the summary statistics, we restrict our analysis to a set of common and well-imputed variants from the HapMap 3 reference panel (The International HapMap 3 Consortium (2010)), which includes 1,197,724 SNPs in total. Figure 3 shows the Manhattan plots of the summary statistics for height (Ht), including $-\log_{10}(p\text{-value})$, $|\widehat{\boldsymbol{\beta}}^{\mathrm{m}}|$, and $|\widehat{\boldsymbol{\beta}}^{\mathrm{r}}|$. The Manhattan plots of the absolute effect sizes from the REMI-R for the other

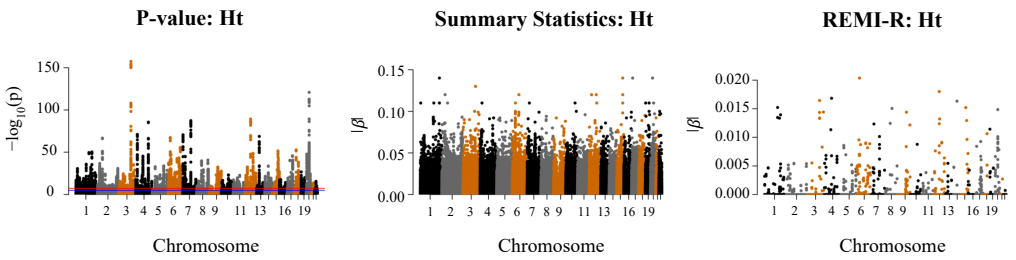

Figure 3. Manhattan plots of analysis results for human height: -$\log_{10}p$-value, $|\widehat{\boldsymbol{\beta}}^{\mathrm{m}}|$ from the marginal analysis, and $|\widehat{\boldsymbol{\beta}}^{\mathrm{r}}|$ from REMI-R.

nine traits are shown in Figure S2.

In addition to the effect size estimation, we evaluate the prediction performance using 5,123 samples from the NFBC1966 (Sabatti et al. (2009)). To make a fair comparison with the Lasso, we first split all 5,123 samples into 10 folds. First, we apply REMI-R to the summary statistics for the lipid traits listed in Table S1. Again, we used 379 European-ancestry samples from the 1000 Genome Project as a reference panel. For each of 10 folds in the NFBC1966 data set, we calculate the predicted phenotypic values and evaluate the Pearson correlation coefficients between the predicted and the observed phenotypic values. Then, we fit the Lasso on the individual-level NFBC1966 data using the same 10-fold data for cross-validation. Specifically, we randomly select nine folds of the individual-level data as the training set to fit the Lasso, and evaluate the prediction accuracy of the fitted model using the remaining one fold. Note that we use the same remaining fold to evaluate the prediction accuracy of the fitted REMI-R model. The prediction performance of the REMI-R and Lasso methods is shown in Figure 4. Clearly, the REMI-R outperforms the standard Lasso in terms of prediction performance. This is because the sample size of the summary statistics for these lipid traits is around 100,000, whereas the individual-level data contain only $5{,}123\times 9/10$ samples. These real-data results indicate the advantage of the REMI over the Lasso for risk prediction.

6. Conclusion

We proposed a novel approach for high-dimensional regression analyses when only marginal regression information and an external reference panel data set are available. Our work is motivated by combining information from multiple GWASs. To date, a large number of GWASs have been conducted to find genetic factors associated with complex traits. Owing to the need for privacy protection

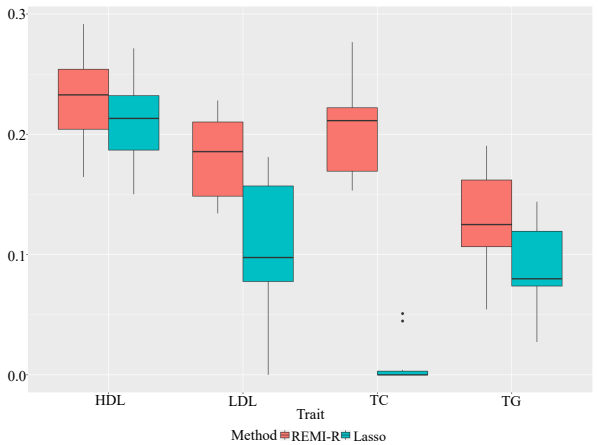

Figure 4. The prediction accuracy (measured by the Pearson correlation coefficients) of the REMI-R and the Lasso for HDL, LDL, TC and TG in the NFBC1966 data sets, where the REMI-R was fitted using independent summary-level data, and the Lasso was fitted using individual-level data from NFBC1966. The prediction accuracies are evaluated on 1/10 of the NFBC1966 data set retained for testing.

and issues related to the sharing of individual-level data, it is important to make full use of the summary statistics from separate studies. In contrast to the limited sample sizes in GWAS analyses based on individual-level data, a prominent feature of a summary-level data analysis is that it uses multiple data sets effectively, which leads to a much larger combined sample size.

Under mild conditions, we prove that the REMI estimator (2.3) based on the marginal information and the reference panel achieves the minimax optimal rate estimation error under reasonable conditions. In particular, the requirement on the size of the reference panel data is quite mild, only in the order of the logarithm of the model dimension. Our theoretical result successfully explains why a relatively small reference sample can be good enough for accurate estimation and prediction in real applications. We conducted comprehensive simulations and a real-data analysis to demonstrate the utility of the REMI method. The experiment results show that the REMI performs similarly to the Lasso when the sample sizes of the summary-level and individual-level data are the same. In genetic analyses, summary-level data sets are much easier to access and their sample sizes are often orders of magnitude larger than those of individual-level data sets. Consequently, the REMI method can be superior to existing methods that require complete data by taking advantage of the larger sample sizes, as demonstrated in our real-data example. In summary, our theoretical and real-data results provide solid support for summary-level-based analyses. As a result,

the polygenic risk score of complex disease can be obtained using summary statistics, with theoretically guaranteed performance.

Supplementary Material

The online Supplementary Material contains technical details and supplementary figures and tables.

Acknowledgments

This work was supported in part by the National Science Funding of China (61501389, 11871474), the Hong Kong Research Grant Council (22302815, 12316 116, 12301417, and 16307818), The Hong Kong University of Science and Technology (startup grant R9405), Duke-NUS Medical School WBS (R-913-200-098-263), and the Ministry of Education, Singapore (MOE2016-T2-2-029, MOE2018-T2-1-046, MOE2018-T2-2-006). All computational work for this study was performed using the resources of the National Supercomputing Centre, Singapore (`https://www.nscc.sg`).

References

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N. et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A. et al. (2015). Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719–736.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J. et

al. (2016). The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47.

Genetics, N. (2012). Asking for more. *Nature Genetics* **44**, 733.

Gratten, J., Wray, N. R., Keller, M. C. and Visscher, P. M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature Neuroscience* **17**, 782–790.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. Springer, New York.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367.

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y. et al. (2011). Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79–89.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508.

Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F. et al. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLOS Computational Biology* **13**, e1005589.

Huang, J., Jiao, Y., Lu, X. and Zhu, L. (2018). Robust decoding from 1-bit compressive sampling with ordinary and regularized least squares. *SIAM Journal on Scientific Computing* **40**, A2062–A2086.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L. et al. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722.

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

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436–1462.

Ning, Z., Lee, Y., Joshi, P. K., Wilson, J. F., Pawitan, Y. and Shen, X. (2017). A selection operator for summary association statistics reveals allelic heterogeneity of complex traits. *The American Journal of Human Genetics* **101**, 903–912.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J. et al. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics* **94**, 559–573.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57**, 6976–6994.

Sabatti, C., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G. et al. (2009).

Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* **41**, 35–46.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**, 32.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

Van De Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3**, 1360–1392.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics* **97**, 576–592.

Visscher, P. M., Brown, M. A., McCarthy, M. I. and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics* **90**, 7–24.

Visscher, P. M., Hill, W. G. and Wray, N. R. (2008). Heritability in the genomics era–concepts and misconceptions. *Nature Reviews Genetics* **9**, 255–266.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. et al. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics* **101**, 5–22.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H. et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–D1006.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S. et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Nolte, I. M. et al. (2015). Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Human Molecular Genetics* **24**, 7445–7449.

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.

Yang, J., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76–82.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–1594.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27**, 576–593.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

Zhou, X. (2016). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The Annals of Applied Statistics* **11**, 2027–2051.

Zhu, X. and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statis-

tics from genome-wide association studies. *The Annals of Applied Atatistics* **11**, 1561–1592.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the Lasso. *The Annals of Statistics* **35**, 2173–2192.

Jian Huang

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA.

E-mail: j.huang@polyu.edu.hk

Yuling Jiao

Department of Statistics, Zhongnan University of Economics and Law, Hubei, China.

E-mail: yulingjiaomath@whu.edu.cn

Jin Liu

Centre for Quantitative Medicine, Program in Health Services and Systems Research, Duke-NUS Medical School, Singapore 169857.

E-mail: jin.liu@duke-nus.edu.sg

Can Yang

Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong.

E-mail: macyang@ust.hk

(Received May 2019; accepted February 2020)