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: In this paper, we consider the problem of variable selection and estimation in high-dimensional linear regression models when the complete data are not accessible, but only certain marginal information or summary statistics are available. This problem is motivated from the genome-wide association studies (GWAS) with millions of genotyped single nucleotide polymorphisms (SNPs) that have been widely used to identify risk variants among complex human traits/diseases. With a large number of completed GWAS, statistical methods using summary statistics become more and more important because of inaccessibility to individual-level data. In this study, we propose REMI, an ℓ_1 penalized approach with estimated marginal effects and an estimated covariance matrix of the predictors with an 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 up-

per bound on the error of the REMI estimator, which has the same order as that of the minimax error bound of 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 is in GWAS, the proposed method yields good estimation and prediction results, and outperforms the Lasso with complete data but with a relatively small sample size. We also demonstrated the application of the proposed method on the ten traits GWAS data from the Northern Finland Birth Cohorts program. In particular, the real data analysis results indicate that summary-level-based analysis by REMI can outperform individual-level-based analysis when the sample size in summary-level data is larger than that of individual-level data. In summary, both of our theoretical and real data results provide a solid support of summary-level-based analysis. As a result, polygenic risk score of a wide variety of complex diseases can be obtained using summary statistics with theoretically guaranteed performance. The developed R package and the codes to reproduce all the results are available at `https://github.com/gordonliu810822/REMI`.

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

1. Introduction

High dimensional regression has been widely applied in various fields, such as medicine, biology, finance and marketing (Hastie et al., 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 the vector of regression coefficients and ϵ is the random error term with mean zero and noise level σ_ϵ^2. In most applications, the data set is comprised of an $n \times p$ matrix $\mathbf{X}$ for variables in X and a vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ for response Y collected from n individuals. Given the individual-level data $\{\mathbf{X}, \mathbf{y}\}$, there exist convex (Tibshirani, 1996; Candes et al., 2007) and nonconvex (Fan and Li, 2001; Zhang, 2010) penalized methods for estimating $\boldsymbol{\beta}^*$ with theoretical guarantee (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Zhang and Huang, 2008; Bickel et al., 2009a; Zhang and Zhang, 2012), just name a few. Also see the monographs (Bühlmann and Van De Geer, 2011; Hastie et al., 2015) and the references therein.

Motivated from the applications in human genetics, we consider the problem of estimating $\boldsymbol{\beta}^*$ when the individual-level data $\{\mathbf{X}, \mathbf{y}\}$ is not accessible but the marginal information is available, such as $\mathbf{X}_j^T \mathbf{y}$ and $\mathbf{X}_j^T \mathbf{X}_j$, $j = 1, \ldots, p$, where $\mathbf{X}_j$ is the j-th column of $\mathbf{X}$. For this reason, we refer to our problem formulation as "regression with marginal information" (REMI). To make our formulation feasible, we also assume that information of the covariance structure of variables in X can be estimated via a

reference panel data set in the form of an $n_{\rm r} \times p$ data matrix $\mathbf{X}_{\rm r}$, where $n_{\rm r}$ is the number of samples from the reference panel and $n_{\rm r} \ll p$. A natural question arises: Without accessing the individual-level data, can we use marginal information together with the reference data $\mathbf{X}_{\rm r}$ to estimate $\boldsymbol{\beta}^*$, assuming observations in $\mathbf{X}_{\rm r}$ and $\mathbf{X}$ are from the same distribution?

In particular, our problem arises in genome-wide association studies (GWAS), 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 in short) associations have been reported in about 3,300 publications of GWAS (see the GWAS Catalog database (https://www.ebi.ac.uk/gwas/). An important lesson from GWAS (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), as well as diabetes (Fuchsberger et al., 2016). Due to the polygenicity, variants with small effects largely remain undiscovered yet and large sample sizes are required in exploring genetic architectures of complex phe-

notypes. Researchers world-wide are forming large genomic consortia, such as the genetic investigation of anthropometric traits (GIANT) consortium and the psychiatric genomic consortium (PGC), to maximize sample size, aiming at a deeper understanding of the genetic architecture of complex phenotypes.

Although much efforts have been made for data sharing, it is still very 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 GAINT consortium reported that they can only access genotype data from about 44,000 individuals (Yang et al., 2015) while 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 the individual-level data. First, privacy protection is always a big concern in sharing individual-level genotype data. Second, it is often time-consuming to achieve an agreement on data-sharing among different research groups. Third, many practical issues arise in data transportation and storage. In contrast, summary statistics from GWAS are widely available through many public gateways (Genetics, 2012), e.g., 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

simple linear regression analysis, summary statistics are essentially marginal information.

To meet the great demand of data analysis in GWAS, various statistical methods have been proposed to utilize marginal information. Using a few hundreds of human genome data from the 1000 Genome Project as a reference panel, information on the correlation structure of genetic variants (typically using "linkage disequilibrium" in genetics, or LD for short) becomes available. This allows these methods to bypass the individual-level data but only use marginal information. Here we roughly divide these methods into three categories: (a) Methods for heritability estimation. Heritability of a phenotype quantifies the relative importance of genetics and environment to the phenotype (Visscher et al., 2008). When individual-level data is 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, named LDScore, for heritability estimation only using summary statistics and the 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 with summary statistics.

(b) Methods for association mapping. Heritability estimation provides a global measure which quantifies the overall contribution from genetic factors while association mapping is to localize genetic variants associated with a given phenotype. Recently, a few statistical methods have been proposed for association mapping based on summary statistics, including FGWAS (Pickrell, 2014), PAINTOR (Kichaev et al., 2014), CAVIAR (Hormozdiari et al., 2014), and CAVIARBF (Chen et al., 2015). Although these methods are very useful for performing association mapping on summary statistics, they still have their limitations. On one hand, they adopted some ad-hoc ways to reduce computational cost. For example, to avoid a combinatorial search, FGWAS assumes that there is only one causal signal in an LD block and PAINTOR searches no more than two causal variants in its default setting. On the other hand, statistical analysis is oversimplified to overcome estimation difficulties. For example, the non-centrality parameter in PAINTOR and the variance components in CAVIAR and CAVIARBF are pre-fixed rather than adaptively estimated from data. (c) 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) further introduced AnnoPred to improve LDpred by incorporating functional

information in human genome. However, neither LDpred or AnnonPred should be considered as a marginal-information-based method because it requires individual-level data as validation data for its parameter tuning. Most recently, Ning et al. (2017) proposed SOJO to use GWAS summary statistics in identifying association signals in complex traits, but they did not consider the theoretical properties of this approach.

Although the existing statistical methods have shown a good empirical performance in GWAS data analysis, there are a number of open questions on REMI. 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 reference for analyzing GWAS data in 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 the existing methods for effect size estimation and prediction error have not been studied. Third, the algorithms are often time-consuming as they need to run thousands of Markov Chain Monte Carlo (MCMC) iterations. Positive answers to these questions can bring a great benefit to GWAS data analysis. For example, polygenic risk score of a variety of diseases can be established without accessing individual-level data but by using summary statistics and a few

number of reference samples. In such a way, a huge amount of computational efforts can be saved without sacrificing statistical accuracy.

In this paper, we propose a statistically guaranteed method, Regression with marginal information (REMI), 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 REMI in GWAS. In Section 3, we present an efficient coordinate descent algorithm and discuss some 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 show the results from both simulation studies and real data analysis. In particular, our results indicate that summary-level-based analysis by REMI can outperform individual-level-based analysis when the sample size in summary-level data is larger than that of individual-level data.

2. The REMI model

2.1 The REMI model

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

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

where $\|\cdot\|_1$ is the ℓ_1 norm and $\lambda \geq 0$ is a regularization parameter. In our problem, however, the individual-level data $\{\mathbf{X}, \mathbf{y}\}$ is not accessible. Hence, direct application of the Lasso is not feasible here. We note that several other important penalized methods have been proposed, including SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). We will 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

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) + \lambda\|\boldsymbol{\beta}\|_1$$

$$= \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}/n - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}/n + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.2}$$

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 since $\mathbf{X}$ is not observed. Motivated by the application in GWAS, we assume that there exists a reference $n_\mathrm{r} \times p$ data matrix $\mathbf{X}_\mathrm{r}$, where the rows of $\mathbf{X}_\mathrm{r}$ are i.i.d. and have the same distribution with 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 $\boldsymbol{\Sigma}$. So we propose to solve the following optimization problem to estimate $\boldsymbol{\beta}^*$:

$$\widehat{\boldsymbol{\beta}}^\mathrm{c} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\mathbf{X}_\mathrm{r}^T\mathbf{X}_\mathrm{r}\boldsymbol{\beta}/n_\mathrm{r} - 2\boldsymbol{\beta}^T\widetilde{\mathbf{y}} + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.3}$$

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

2.2 REMI in GWAS

In the context of GWAS, the available marginal information may not be $\widetilde{\mathbf{y}} = \mathbf{X}^T\mathbf{y}/n$ but summary statistics $\{\widehat{\beta}_j^\mathrm{m}, \hat{s}_j^2\}_{j=1,\ldots,p}$ from 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 superscript $^{\mathrm{m}}$ is used to denote marginal information. $\widehat{\beta}_j^{\mathrm{m}}$ and $\hat{s}_j^2$ are

the estimated effect size and its variance for SNP j, respectively. Due 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

$\widehat{\boldsymbol{\beta}}^{\mathrm{m}} = [\widehat{\beta}_1^{\mathrm{m}}, \ldots, \widehat{\beta}_p^{\mathrm{m}}]^T$, $\hat{\mathbf{s}}^2 = [\hat{s}_1^2, ..., \hat{s}_p^2]^T$ be the vectors collecting estimated

effect sizes and estimated variance, respectively, and $\widehat{\mathbf{S}}$ be a $p \times p$ diagonal

matrix with $\hat{s}_j$ being its j-th diagonal element. Further, we introduce a

$p \times p$ diagonal matrix $\widehat{\mathbf{D}} = \mathrm{diag}(\hat{d}_j)$ with its j-th diagonal element being

the sample standard deviation of $\mathbf{X}_j$, i.e., $\hat{d}_j = \sqrt{\frac{\mathbf{X}_j^T\mathbf{X}_j}{n}}$, and correlation

matrix $\widehat{\mathbf{R}} = [\hat{r}_{jk}] \in \mathbb{R}^{p \times p}$ with $\hat{r}_{jk} = \frac{\mathbf{X}_j^T\mathbf{X}_k}{(\mathbf{X}_j^T\mathbf{X}_j)^{1/2}(\mathbf{X}_k^T\mathbf{X}_k)^{1/2}}$. Noticing 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

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}/n - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}/n + \lambda\|\boldsymbol{\beta}\|_1,$$

$$= \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\widehat{\mathbf{D}}\widehat{\mathbf{R}}\widehat{\mathbf{D}}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\widehat{\mathbf{D}}^2\widehat{\boldsymbol{\beta}}^{\mathrm{m}} + \lambda\|\boldsymbol{\beta}\|_1,$$

$$\approx \arg\min_{\boldsymbol{\beta}} \frac{\mathbf{y}^T\mathbf{y}}{n^2}\boldsymbol{\beta}^T\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{R}}\widehat{\mathbf{S}}^{-1}\boldsymbol{\beta} - 2\frac{\mathbf{y}^T\mathbf{y}}{n^2}\boldsymbol{\beta}^T\widehat{\mathbf{S}}^{-2}\widehat{\boldsymbol{\beta}}^{\mathrm{m}} + \lambda\|\boldsymbol{\beta}\|_1,$$

$$= \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{R}}\widehat{\mathbf{S}}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\widehat{\mathbf{S}}^{-2}\widehat{\boldsymbol{\beta}}^{\mathrm{m}} + \widetilde{\lambda}\|\boldsymbol{\beta}\|_1,$$

where $\widetilde{\lambda} = \frac{n^2}{\mathbf{y}^T\mathbf{y}}\lambda$ and the approximation holds in the case of polygenicity.

Because $\widetilde{\lambda}$ is a tuning parameter that scales λ with a constant factor $(\frac{n^2}{\mathbf{y}^T\mathbf{y}})$,

we slightly abuse λ for $\widetilde{\lambda}$ and propose to solve the following optimization

problem

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

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 using correlation information. Similar to REMI (2.3) in which covariance matrix $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^T\mathbf{X}/n$ needs to be estimated, here correlation matrix $\widehat{\mathbf{R}}$ needs to be estimated by samples from the reference panel $\mathbf{X}_{\mathrm{r}}$. We refer (2.3) as REMI-C and (2.4) as the REMI-R, respectively.

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

$$\widehat{\beta}_j^{\mathrm{c}}(\lambda) = \arg\min_{\beta_j} \widehat{\sigma}_{jj}\beta_j^2 - 2\left(\widetilde{y}_j - \sum_{k\neq j}\widehat{\beta}_k^{\mathrm{c}}\widehat{\sigma}_{jk}\right)\beta_j + \lambda|\beta_j|, \tag{3.5}$$

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 the warm start and some other tricks as described in Friedman et al. (2010). In particular, we generate a sequence of $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_D)$ equally spaced in logarithm 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 warm start. The path algorithm is described in Algorithm 1.

Output: Solution path for $\widehat{\boldsymbol{\beta}}^{c}(\boldsymbol{\lambda})$.

for $l = 1, 2, \ldots, D$ **do**

    Initialize $\widehat{\boldsymbol{\beta}}^{c}(\lambda_l) = \widehat{\boldsymbol{\beta}}^{c}(\lambda_{l-1})$, if $l > 1$; $\widehat{\boldsymbol{\beta}}(\lambda_l) = \mathbf{0}$, if $l = 1$

    repeat

        for $j = 1, 2, \ldots, p$ **do**

$$\eta_j = \widetilde{y}_j - \sum_{k \neq j} \widehat{\beta}_k^{c}(\lambda)\widehat{\sigma}_{jk}$$

$$\widehat{\beta}_j^{c}(\lambda) \leftarrow \mathrm{S}(\eta_j, \lambda/2)/\widehat{\sigma}_{jj}$$

        end

    until *Convergence*;

end

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

Similar to REMI-C, an efficient coordinate descent algorithm can developed for solving REMI-R (2.4). The efficient path algorithm is given in

Algorithm 2.

Output: Solution path $\widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\boldsymbol{\lambda})$.

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**

$$\eta_j = \frac{\widehat{\beta}_j^{\mathrm{m}}}{\hat{s}_j^2} - \frac{1}{\hat{s}_j} \sum_{k \neq j} \frac{\widehat{\beta}_k^{\mathrm{r}} \hat{r}_{jk}}{\hat{s}_k}$$

$$\widehat{\beta}_j^{\mathrm{r}}(\lambda) \leftarrow \mathrm{S}(\eta_j, \lambda/2) \times \hat{s}_j^2$$

    end

  until *Convergence*;

end

Algorithm 2: Path algorithm to solve REMI-R (2.4) with a sequence

of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)$

3.2 Reference panel

In REMI-R model (2.4), it involves the cohort-based estimated correlation

matrix. Based on the nature of the correlation patterns of the SNPs, $\mathbf{R}$ can

be approximated by 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, respectively (Berisa and Pickrell, 2016).

Then we calculate 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 et al., 2005), where we used $\kappa = 0.9$ as default (the estimate of $\boldsymbol{\beta}^*$ is insensitive to κ, Pasaniuc et al. (2014)). Thus, similar to Zhu and Stephens (2017), REMIs and its individual-level-data counterpart will produce approximately the same inferential results within a region. After plugging $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{R}}^{\mathrm{r}}$ in (2.4), we can use the coordinate-descent algorithm to obtain $\widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda)$ (Algorithm 2).

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 and it is straightforward to develop the same strategy for REMI-C. Similarly to the Lasso solver (Friedman et al., 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 pre-specified with the default value at 0.05. Note that $\lambda_{\max} = \max\left\{2\hat{\beta}_j^{\mathrm{m}}/\hat{s}_j^2\right\}_{j=1,\ldots,p}$. We search for optimal value λ value using BIC,

$$\mathrm{BIC}(\lambda_l) = L(\widehat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda_l)) + \log(n)\mathrm{df}(\lambda_l). \tag{3.6}$$

Zou et al. (2007) showed that the number of nonzero coefficients is an unbiased estimate for the degrees of freedom of Lasso. We choose $\mathrm{df}(\lambda_l)$ to be the number of nonzero coefficients given λ_l. To make fair comparison of

REMIs with Lasso, we also use BIC to select the regularization parameter when individual-level-data is accessible.

4. Theoretical properties

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

Since in real applications of genetic data the underlying signal is not exactly sparse but with many small components. Here, we assume that the target $\boldsymbol{\beta}^*$ is weak sparse, i.e., 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

$$s = |\mathcal{A}| \le n, \quad \|\boldsymbol{\beta}_{\mathcal{I}}^*\|_\infty \le 2\sigma_{\boldsymbol{\epsilon}}\sqrt{\frac{\log p}{n}}, \tag{4.7}$$

since the signal whose magnitude smaller than this order is undetectable. Then

$$\widetilde{\mathbf{y}} = \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\epsilon}} = \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{\mathcal{A}}^* + \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{\mathcal{I}}^* + \widetilde{\boldsymbol{\epsilon}}, \tag{4.8}$$

where, $\widetilde{\boldsymbol{\epsilon}} = \mathbf{X}^T\boldsymbol{\epsilon}/n$. Let $C_1 \ge \sqrt{\|\text{diag}(\boldsymbol{\Sigma})\|_\infty}$, $C_2 \ge \max_{j=1,\ldots p}\{\|\boldsymbol{\Sigma}_j\|_1\}$, $C_3 \ge \|\boldsymbol{\beta}_{\mathcal{A}}^*\|_1$, $C_4 \ge \|\boldsymbol{\beta}_{\mathcal{I}}^*\|_1$. Recall the restrict eigenvalue (Bickel et al., 2009b) of

$\widehat{\boldsymbol{\Sigma}}_r$ is defined as

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

where

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

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

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

$$\|\widehat{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_{\mathcal{A}}^*\| \leq \frac{6}{\phi_0}(\frac{C_1(C_3+C_4)}{\sqrt{C}}\sqrt{\frac{s\log p}{n}} + \frac{C_1+\sqrt{C}C_2}{\sqrt{C}}\sigma_\epsilon\sqrt{\frac{s\log p}{n}} + \frac{C_1 C_3}{\sqrt{C}}\sqrt{\frac{s\log p}{n_r}}).$$

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

$$\|\mathbf{X}_{\text{new}}(\widehat{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_{\mathcal{A}}^*)\|_2^2/n_{\text{new}} \leq \mathcal{O}((\sigma_\epsilon\sqrt{\frac{s\log p}{n}} + \sqrt{\frac{s\log p}{n_r}})^2)(1 + s^2(\sqrt{\frac{\log p}{n_r}} + \sqrt{\frac{\log p}{n_{\text{new}}}})).$$

Remark 2. The assumption that $\mathbf{X}_r$ are i.i.d sub-Gaussian samples drawn from population with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ implies the restricted eigenvalue condition $\phi_{\widehat{\boldsymbol{\Sigma}}_r} \geq \phi_0$ holds for some positive ϕ_0 with high probability as long as $n_r \geq \mathcal{O}(s\log p)$ (Van De Geer et al., 2009; Vershynin, 2010;

Huang et al., 2018). As shown in Theorem 1, with the help of reference panel, we can also get an accurate estimator by (2.3) even if we only have marginal information in 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 REMI model (2.3) achieve the minimax optimal rate as that of the Lasso (Raskutti et al., 2011) if the number of samples of reference penal n_{r} is at the same order of the number of individual-level samples n. Moreover if the magnitude of the significant entries larger than $\mathcal{O}(\sigma_\epsilon \sqrt{\frac{s \log p}{n}} + \sqrt{\frac{s \log p}{n_{\mathrm{r}}}})$, the estimated support $\mathrm{supp}(\widehat{\boldsymbol{\beta}}^{\mathrm{c}})$ coincide with the true significant set $\mathcal{A}$.

5. Numerical Studies

5.1 Simulation studies

In simulation studies, we compare REMI-C (2.3), REMI-R (2.4) and Lasso using individual-level data. To avoid unrealistic LD pattern in simulation, we used the genotype data $\mathbf{X}$ from RPGEH data set (Hoffmann et al., 2011). The RPGEH data set provided 657,184 genotyped SNPs for 62,313 European individuals. We performed strict quality control on data using PLINK (Purcell et al., 2007). We excluded SNPs with a minor allele frequency less than 1%, having missing values in more than 1% of the individuals or with a Hardy-Weinberg equilibrium p-value below 0.0001. Moreover,

we removed one member of pairs with genetic relatedness larger than 0.05.
Finally, there remained 53,940 samples for 550,482 SNPs.

As individual-level-based analyses often suffer from limited sample sizes
due to the restricted access of individual-level data, summary-level-based
analyses may have advantages because the sample sizes are often much
larger. To simulate this situation, we prefixed the sample size for individual-
level-based analyses at $n_{\text{ind}} = 3,000$. Specifically we randomly selected
n_{ind} samples from 53,940 individuals in the RPGEH data set to form the
genotype matrix $\mathbf{X} \in \mathbb{R}^{n_{\text{ind}} \times p}$, where $p = 19,865$ was the total number of
the genotyped SNP 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 $(h^2 = \frac{\text{Var}(\mathbf{X}\boldsymbol{\beta})}{\text{Var}(\mathbf{X}\boldsymbol{\beta}^*) + \sigma_\epsilon^2})$ was controlled at 0.2, 0.3, 0.4 and 0.5. Here
$\boldsymbol{\beta}^*$ was the vector of true effect size with sparsity α, i.e., $\alpha \times p$ entries in $\boldsymbol{\beta}^*$
were nonzero and they were 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 of individual-level
data analysis.

To generate summary-level data, we varied sample size n from 3,000
to 50,000. We generated individual-level data as we described above and
then we ran simple linear regression on $\{\mathbf{X}_j, \mathbf{y}\}$, $j = 1, \ldots, p$ to obtain

$\{\widehat{\boldsymbol{\beta}}^{(m)}, \hat{\mathbf{s}}^2\}$. After that, we pretended that we did not have individual-level data $\{\mathbf{X}, \mathbf{y}\}$ and then only used $\mathbf{X}^T\mathbf{y}$ and $\{\widehat{\boldsymbol{\beta}}^{(m)}, \hat{\mathbf{s}}^2\}$ as the input for REMI-C and REMI-R, respectively. We used 379 European samples from the 1000 Genome Project data as the reference panel to estimate covariance matrix (REMI-C) or correlation matrix (REMI-R), as detailed in Section 3.2. For each replication, we held out 200 independent samples to evaluate prediction accuracy. In total, we summarized our results based on 50 replications for each setting.

We compared the performance of REMI-C, REMI-R and the Lasso using individual-level data in terms of variable selection and prediction. Specifically, we used partial area under the receiver operating characteristic (ROC) curve (partial AUC) for variable selection performance and the Pearson's correlation coefficient between predicted and observed phenotypes for prediction performance. The results of this simulation study are shown in Figure 1 and Figure 2. First, we can observe that the difference between REMI-C and REMI-R is nearly invisible. This justifies the approximation made in REMI-R. Second, when the sample size ($n = $ 3,000 or 5,000) in summary-level data is similar to that of individual-level data ($n_{\text{ind}} = 3,000$), the performance of variable selection and prediction for REMIs (both REMI-C and REMI-R) is similar to that of Lasso. Third,

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 REMIs can have big advantages over the Lasso when the sample sizes of summary-level data become much larger.

5.2 Real data analysis

To demonstrate the utility of REMIs, we first compared Lasso and REMIs based on the GWAS data set from the Northern Finland Birth Cohorts program (NFBC1966) (Sabatti et al., 2009). The NFBC1966 data set contains information for 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 have been genotyped. We performed strict quality control on data using PLINK (Purcell et al., 2007). We first excluded individuals having discrepancies between reported sex and sex determined from the X chromosome. We also excluded SNPs with a minor allele frequency less than 1%, having missing values in more than 1% of the individuals or with a Hardy-Weinberg equilibrium p-value below 0.0001. In particular, we se-
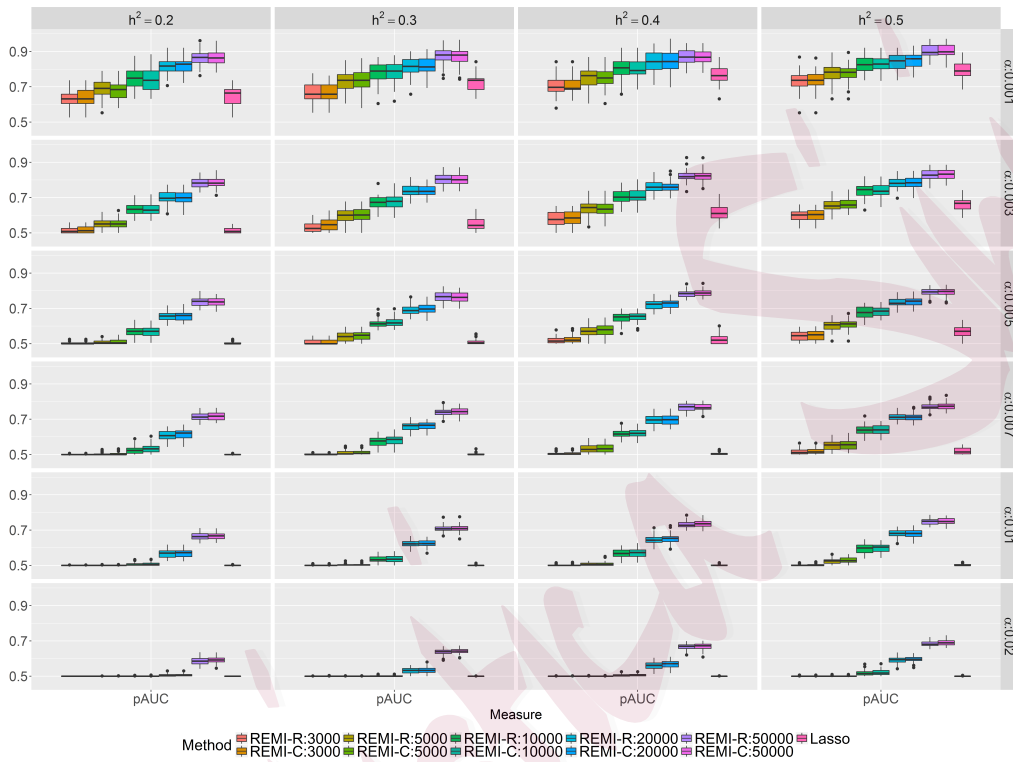
Figure 1: The comparison of variable selection performance of REMIs (REMI-R and REMI-C) for summary-statistics data with Lasso for individual-level data with sample size 3000. The sample size used to produce summary statistics was varied and denoted as $n \in \{3,000, 5,000, 10,000, 20,000, 50,000\}$. We used partial AUC to measure the variable selection performance.

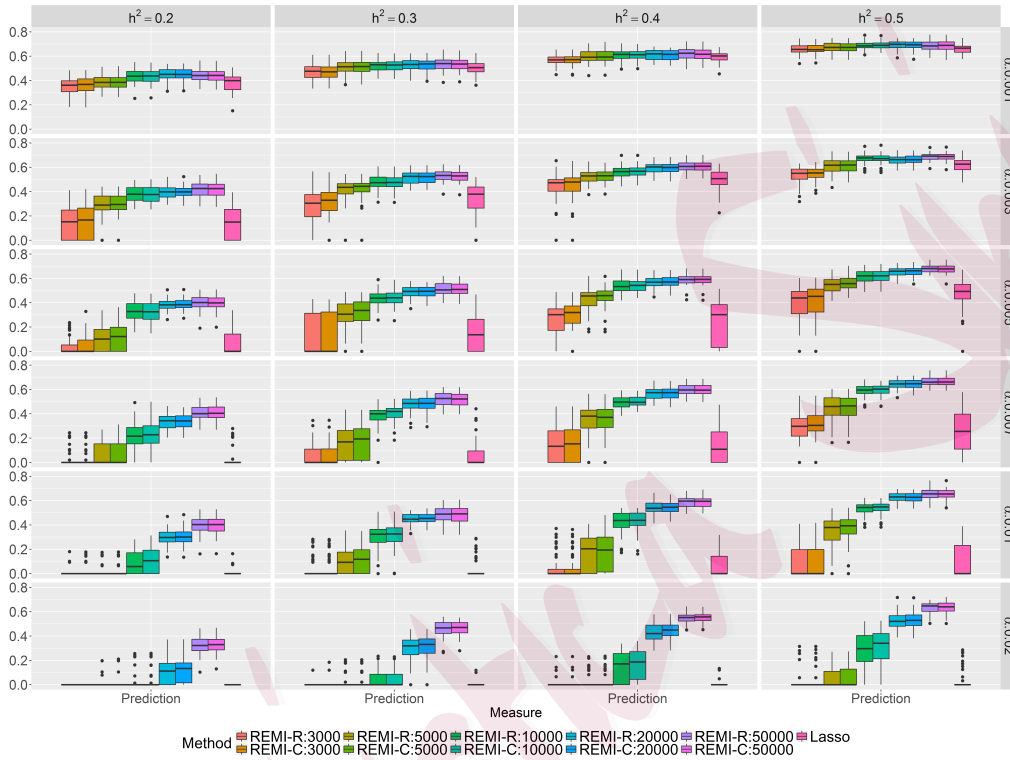Figure 2: Prediction accuracy of REMIs with summary-level data and the Lasso for individual-level data of sample size 3,000. The sample size used to simulate summary statistics was varied $n \in \{3,000, 5,000, 10,000, 20,000, 50,000\}$. Prediction accuracy is measured by the Pearson's correlation coefficient between predicted and observed phenotypes.

lected well-imputed variants from HapMap 3 reference panel (Consortium et al., 2010). After the strict quality control, 5,123 individuals with 310,975 SNPs in NFBC1966 were remaining for the further analysis. As we have the individual-level data, it is possible to run REMI-C, REMI-R and Lasso for all these traits. The solution paths using Lasso, REMI-R, and REMI-C for the ten metabolic traits in NFBC1966 data set are presented in Figures S1. The dotted vertical bars in these two figures indicate the corresponding selected tuning parameters based on BIC. One can see that the differences among solution paths for the Lasso using individual-level data, REMI-C and REMI-R using summary statistics are very minor, which is consistent with results from our simulation studies in Section 5.1.

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 made available. Therefore, we applied REMI-R to analyze summary statistics for ten GWASs of complex phenotypes. The source of the GWASs is given in Table S1. Because the individuals of summary-level data sets were all from European ancestry, we used 379 European-ancestry samples in 1000 Genome Project (Consortium et al., 2012) as a reference panel to estimate correlation matrix. Due to the quality of SNPs in the summary statistics, we restricted our analysis to a set of common and well-imputed variants from the HapMap 3 refer-

ence panel (Consortium et al., 2010), which included 1,197,724 SNPs in total. Figure 3 shows the Manhattan plots of 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 REMI-R for all other nine traits are shown in Figure S2.

Besides the effect size estimation, we evaluated prediction performance using 5,123 samples from the NFBC1966 (Sabatti et al., 2009). To make a fair comparison with Lasso, we first split all 5,123 samples into ten folds. On the one hand, we applied REMI-R on the summary statistics for these 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 NFBC1966 data set, we calculated the predicted phenotypic values and evaluated the Pearson's correlation coefficients between the predicted phenotypic values and the observed ones. On the other hand, we fitted the Lasso on the individual-level NFBC1966 data using the same ten-fold data for cross-validation. Specifically, we randomly selected nine folds of individual-level data as the training set to fit the Lasso, and evaluated prediction accuracy of the fitted model using the remaining one fold. Note that we used the same remaining fold to evaluate the prediction accuracy of the fitted REMI-R model. The prediction performance for REMI-R and Lasso

is shown in Figure 4. Clearly, the prediction performance of REMI-R out-performs the standard Lasso as the sample size in the summary statistics for these lipid traits are around 100,000 but the individual-level data contains only 5,123×9/10 samples. These real data results indicate the great advantage of REMI over the Lasso for risk prediction.
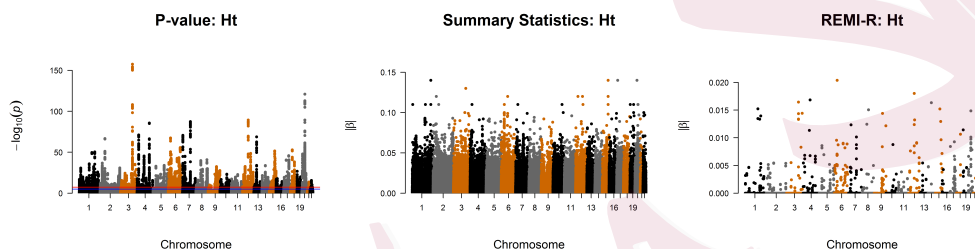

Figure 3: Manhattan plots of analysis result of human height for: $-\log_{10}p$-value, $|\widehat{\boldsymbol{\beta}}^{\mathrm{m}}|$ from marginal analysis and $|\widehat{\boldsymbol{\beta}}^{\mathrm{r}}|$ from REMI-R.

6. Discussion

In this study, we proposed a novel approach for high-dimensional regression analysis when only marginal regression information and an external reference panel data set are available. Our work is motivated from combining information from multiple GWAS. To date, a large number of GWAS have been conducted to find genetic factors associated with complex traits. Due to the need for privacy protection and issues in data-sharing of individual-
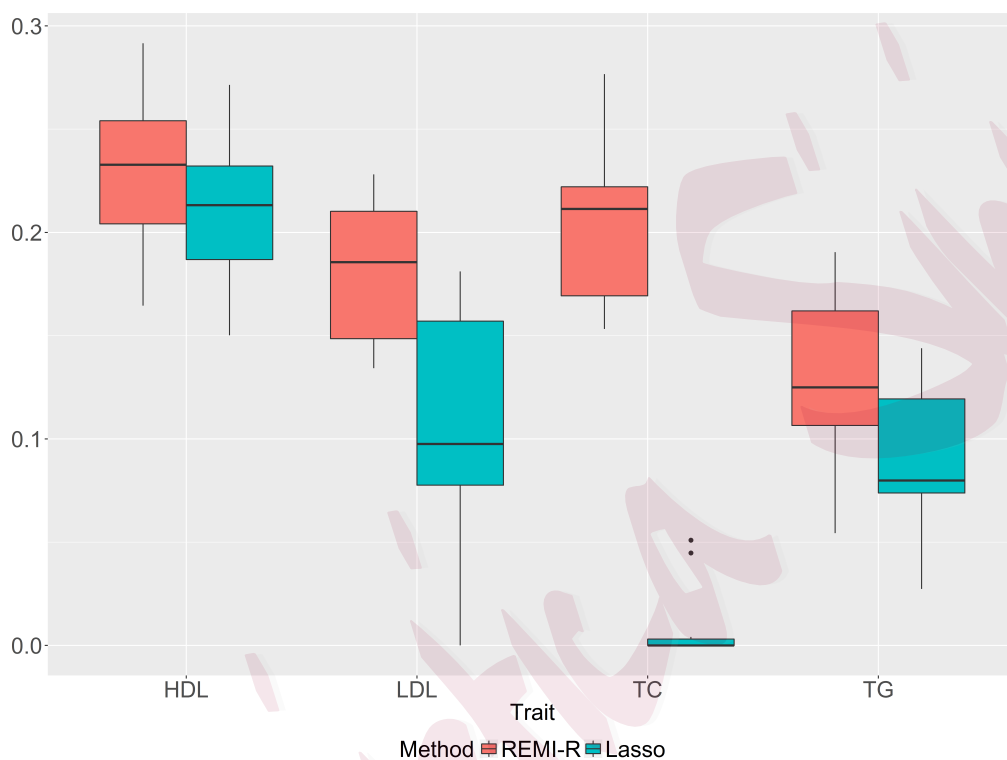
Figure 4: Prediction accuracy (measured by the Pearson's correlation coefficients) of REMI-R and the Lasso for HDL, LDL, TC and TG in the NFBC1966 data sets, where REMI-R was fitted using independent summary-level data and the Lasso was fitted using the individual-level data from NFBC1966. Their prediction accuracies were evaluated on 1/10 of the NFBC1966 data set holded out for testing.

level data, it is important to be able to effectively make full use of the summary statistics from separate studies. In contrast to the limited sample size in individual-level data based GWAS analysis, a prominent feature of summary-level data analysis is that it can effectively make use of multiple data sets, 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, it is 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 have conducted comprehensive simulations and real data analysis to demonstrate the utility of REMI. The experimental results show that the performance of REMI can be very similar to the Lasso when the sample sizes of summary-level data and individual-level data are the same. In genetic analysis, summary-level data sets are much easier to access and their sample sizes are often orders of magnitude larger than that of individual-level data sets. Consequently, REMI can be superior to the existing methods requiring complete data by taking advantages of the larger sample sizes, as

demonstrated in our real data example. In summary, both of our theoretical and real data results provide a solid support of summary-level-based analysis. As a result, polygenic risk score of a wide variety of complex diseases can be obtained using summary statistics with theoretically guaranteed performance.

Acknowledgements

References

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations., *Bioinformatics (Oxford, England)* **32**(2): 283.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009a). Simultaneous analysis of Lasso and

REFERENCES

Dantzig selector, *Ann. Statist.* **37**(4): 1705–1732.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009b). Simultaneous analysis of lasso and dantzig selector, *The Annals of Statistics* pp. 1705–1732.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*, Springer.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., of the Psychiatric Genomics Consortium, S. W. G. et al. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies, *Nature genetics* **47**(3): 291–295.

Candes, E., Tao, T. et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n, *The Annals of Statistics* **35**(6): 2313–2351.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A. and Schaid, D. J. (2015). Fine mapping causal variants with an approximate bayesian method using marginal test statistics, *Genetics* **200**(3): 719–736.

Consortium, . G. P. et al. (2012). An integrated map of genetic variation from 1,092 human genomes, *Nature* **491**(7422): 56.

Consortium, I. H. . et al. (2010). Integrating common and rare genetic variation in diverse human populations, *Nature* **467**(7311): 52.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle

REFERENCES

properties, *J. Amer. Statist. Assoc.* **96**(456): 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): 1.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J. et al. (2016). The genetic architecture of type 2 diabetes, *Nature* .

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**(6): 782–790.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning (2nd Edition)*, Springer.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*, CRC press.

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, *Proceedings of the National Academy of Sciences* **106**(23): 9362–9367.

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., Cawley, S., Chung, E., Connell, S., Eshragh, J. et al. (2011). Next generation genome-wide association

REFERENCES

tool: design and coverage of a high-throughput european-optimized snp array, *Genomics* **98**(2): 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**(2): 497–508.

Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X. and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases, *PLOS Computational Biology* **13**(6): 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**(4): A2062–A2086.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P. and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies, *PLoS Genet* **10**(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**(3): 294–305.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso, *Ann. Statist.* **34**(3): 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,

REFERENCES

The American Journal of Human Genetics **101**(6): 903–912.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N. and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment, *Bioinformatics* **30**(20): 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**(4): 559–573.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses, *The American Journal of Human Genetics* **81**(3): 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**(10): 6976–6994.

Sabatti, C., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U. et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population, *Nature genetics* **41**(1): 35.

Schäfer, J., Strimmer, K. et al. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical applications in genetics and molecular biology* **4**(1): 32.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc.*

REFERENCES

Ser. B **58**(1): 267–288.

Van De Geer, S. A., Bühlmann, P. et al. (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., Genovese, G., Loh, P.-R., Bhatia, G., Do, R. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores, *The American Journal of Human Genetics* **97**(4): 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**(1): 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**(4): 255–266.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation, *The American Journal of Human Genetics* **101**(1): 5–22.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, *Nucleic Acids Research* **42**(D1): D1001–D1006.

REFERENCES

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z. et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height, *Nature Genetics* **46**(11): 1173–1186.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., Esko, T., Milani, L. et al. (2015). Genome-wide genetic homogeneity between sexes and populations for human height and body mass index, *Human molecular genetics* p. ddv443.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010). Common snps explain a large proportion of the heritability for human height, *Nature genetics* **42**(7): 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**(1): 76–82.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.* **38**(2): 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**(4): 1567–1594.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems, *Statist. Sci.* **27**(4): 576–593.

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

REFERENCES

Zhou, X. (2016). A unified framework for variance component estimation with summary statis-
tics in genome-wide association studies, *bioRxiv* p. 042846.

Zhu, X. and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statis-
tics from genome-wide association studies, *The annals of applied statistics* **11**(3): 1561.

Zou, H., Hastie, T., Tibshirani, R. et al. (2007). On the degrees of freedom of the lasso, *The
Annals of Statistics* **35**(5): 2173–2192.

Department of Statistics and Actuarial Science

University of Iowa

Iowa City, IA 52242

USA

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

Department of Statistics

Zhongnan University of Economics and Law

China

E-mail: (yulingjiaomath@whu.edu.cn)

Centre for Quantitative Medicine

Program in Health Services & Systems Research

Duke-NUS Medical School

Singapore 169857

REFERENCES

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

Department of Mathematics

Hong Kong University of Science and Technology

Hong Kong

E-mail: (macyang@ust.hk)