HYPOTHESIS TESTING FOR HIGH-DIMENSIONAL REGRESSION UNDER EXTREME PHENOTYPE SAMPLING OF CONTINUOUS TRAITS

A DISSERTATION

SUBMITTED ON THE THIRTIETH DAY OF MARCH 2018

TO THE DEPARTMENT OF GLOBAL BIOSTATISTICS AND DATA SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF PUBLIC HEALTH AND TROPICAL MEDICINE

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY

Chao Xu

APPROVED:

Hong-Wen Deng, PhD date

Yu-Ping Wang, PhD date

Hui Shen, PhD date

Lan-Juan Zhao, PhD date
Abstract

Extreme phenotype sampling (EPS) is a broadly-used design to identify candidate genetic factors contributing to the variation of quantitative traits. By enriching the signals in the extreme phenotypic samples within the top and bottom percentiles, EPS can boost the study power compared with the random sampling with the same sample size. The existing statistical methods for EPS data test the variants/regions individually. However, many disorders are caused by multiple genetic factors. Therefore, it is critical to simultaneously model the effects of genetic factors, which may increase the power of current genetic studies and identify novel disease-associated genetic factors in EPS. The challenge of the simultaneous analysis of genetic data is that the number (p ~10,000) of genetic factors is typically greater than the sample size (n ~1,000) in a single study. The standard linear model would be inappropriate for this p>n problem due to the rank deficiency of the design matrix. An alternative solution is to apply a penalized regression method – the least absolute shrinkage and selection operator (LASSO).

LASSO can deal with this high-dimensional (p>n) problem by forcing certain regression coefficients to be zero. Although the application of LASSO in genetic studies under random sampling has been widely studied, its statistical inference and testing under EPS remain unknown. We propose a novel sparse model (EPS-LASSO) with hypothesis test for high-dimensional regression under EPS based on a decorrelated score function to investigate the genetic associations, including the gene expression and rare variant analyses. The comprehensive simulation shows EPS-LASSO outperforms existing methods with superior power when the effects are large and stable type I error and FDR control. Together with the real data analysis of genetic study for obesity, our results
indicate that EPS-LASSO is an effective method for EPS data analysis, which can account for correlated predictors.
Acknowledgments

I would like to express my deepest appreciation to my advisor, Dr. Hong-Wen Deng, for his broad foresight, profound knowledge, insightful mentoring, and all his support for my study and life. It is my fortune to learn from and work with him for all these years.

I am sincerely grateful for the superb committee and the Department of Global Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University. I came from a different country with a different academic background. I received the finest and rigid education in biostatistics and data science from the faculty and staff at the Department and the University. There are too many charitable people to be listed here.

I would also like to thank my parents, family, and friends. I could not reach this point without their greatest kindness, patience, and encouragement. I dedicate this dissertation to my parents, family, mentors, and friends in all my life.
# Table of Contents

I. Background and Significance ........................................................................................................ 1

II. Literature Review ...................................................................................................................... 7
   - Linear model and hypothesis testing .................................................................................... 7
   - LASSO regression ................................................................................................................ 9
   - LASSO in genetic data analyses ....................................................................................... 12

III. Research Question .................................................................................................................. 15

IV. Methods and Data .................................................................................................................. 17
   - Sparse regression for EPS genetic data analysis ............................................................... 17
   - Hypothesis testing for EPS-LASSO(-RV) ......................................................................... 20
   - Simulation design ............................................................................................................. 23
   - Real data analyses ........................................................................................................... 26
   - Source code .................................................................................................................... 28

V. Simulation Result .................................................................................................................... 29
   - Performance of EPS-LASSO .......................................................................................... 29
   - Performance of EPS-LASSO-RV .................................................................................. 40

VI. Application to Gene Expression Analysis ............................................................................. 46

VII. Application to Rare Variant Analysis .................................................................................. 49

VIII. Discussion ............................................................................................................................ 51

IX. Conclusion ............................................................................................................................. 57
X. Appendix........................................................................................................................................................................... 59
  X.I The selection of criterion for tuning parameter ................................................................. 59
  X.II Refitted estimator vs non-refitted estimator ................................................................. 61
  X.III Proof of the asymptotic distribution of the Score and Wald statistics ................. 62
  X.IV Using same and using different data ....................................................................... 65
  X.V Screening aided EPS-LASSO .................................................................................. 67

References............................................................................................................................................................................. 69.
I. Background and Significance

Genetic data analysis is a way to identify candidate variants, genes or genomic regions which contribute to specific diseases by testing the correlations between those genetic factors and disease status\(^1\). A genetic data analysis or a genetic study can be conducted at different genetic levels, such as genome-wide single nucleotide polymorphism (SNP), DNA methylation and transcriptomic gene expression.

The transcriptomic gene expression analyses examine the RNA transcripts expressed in the specific tissues or cells via high-throughput techniques. The microarray and RNA sequencing (RNA-Seq) are two most widely used technologies for gene expression analyses. The data collected from both microarray and RNA-Seq are continuous. All the genes in the genome are screened to identify novel candidate genes which may contribute to the disease development. Given the \(\sim 25,000\) genes in the whole genome, the number of genes in the gene expression analysis is typically in \(\sim 10,000\). The gene expression data can also be used in sample classification and phenotype prediction\(^2\)\(^,\)\(^3\).

SNP is a variation in a single nucleotide that occurs at a specific position in the genome. For a specific SNP with major allele denoted by A and minor allele denoted by a, there are three genotypes A/A, A/a, and a/a. The genotype can also be coded as 0, 1, and 2, which is the number of minor alleles in the genotype. Accordingly, the genotype follows a Binomial distribution \(B(2, \text{MAF})\), where MAF stands for the minor allele frequency. In the SNP analyses, three genetic models are usually considered according to the different ways that the alleles affect the disease risk. They are dominant model: any genotypes with minor allele A/a or a/a carry the same effect; recessive model: only a/a
carries effect; additive model: the effect of a/a is twice of A/a. Among the three genetic models, the additive model is mostly used. The genome-wide association study (GWAS) is a typical example of large-scale SNP analyses. In GWAS, ~1 million SNPs are scanned to find the disease-associated markers. Although GWAS have successfully identified a large number of common SNPs (MAF – minor allele frequency > 0.05) involved in complex diseases, most of them only explained a limited proportion of heritability. The low frequency and rare variants with MAF <0.05 are expected to play an important role in uncovering gene-phenotype relationships as well as the missing heritability unexplained by the traditional GWAS. With the capability to sequence the whole genome, the SNP analyses using the next-generation sequencing is expected to detect candidate variants across the full MAF spectrum, especially the rare variants which may explain the missing heritability.

The disease phenotype or traits involved in a genetic data analysis commonly appears in the type of continuous, binary, and survivorship. For instance, in an obesity study looking into the body mass index (BMI), the BMI value is a continuous trait. While the infected status for an infectious disease is an example of the binary trait. The survivorship is a kind of phenotype often used in cancer studies. Depending on different types of traits considered, the generalized linear model is employed to infer and test the associations between genetic factors and the continuous or binary phenotype. The analysis of survivorship requires specific survival model, such as the Cox proportional hazards model. In this study, we are working on the study sampling from continuous traits.

Extreme phenotype sampling (EPS) is a study design in genetic research including gene expression and SNP analyses. In EPS, the genetic associations between genetic
factors and disease are studied using the subjects with extreme phenotypes instead of random samples\textsuperscript{14}. The extreme phenotypes are usually from the two ends of the distribution of a quantitative trait (Fig. I-1). For example, the BMI study in the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) involved 267 subjects with BMI values within top \textasciitilde 2.3\% and 178 subjects with BMI values within bottom \textasciitilde 1.5\% out of a total of 11,468 subjects\textsuperscript{15, 16}. By pooling the extreme phenotypic samples, it can boost the association testing power by enriching the signals, such as the rare variants\textsuperscript{17}. Rare variants are usually defined as the SNP with MAF < 0.01. The rare variants are expected to play an important role in uncovering gene-phenotype relationships. However, the detection and association power for rare variants are low in random sampling studies. Several recent studies have shown the EPS is beneficial for the rare variant studies. Because the enrichment of rare causal variants in extreme samples leads to an increase of power in detection and association, and a decrease in study effort and cost\textsuperscript{17, 18}.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{EPS_distribution.png}
\caption{Truncated distribution of EPS. Dashed lines indicate the truncated points and the two tails are kept for EPS.}
\end{figure}

The distribution of extreme trait values in EPS breaks the normality assumption of many statistical methods. The methods applicable to random sampling data may not be
appropriate for the EPS data analysis. With a broad application in genetic studies, numerous methods have been proposed to analyze data from EPS. Case-control methods treat the samples with extremely high and low trait values as cases and controls. Then the standard statistical methods for group comparison can be used, such as t-test to find genes with differential expression levels or frequency/proportion test to identify candidate variants with different allele frequencies\textsuperscript{19, 20}. However, the case-control methods disregard the quantitative trait values, wherein much of the genetic information may reside. Considering the inefficient of the case-control methods, Huang et al. developed a likelihood-based method that makes full use of the available extreme phenotype data to identify causal variants and showed an improved power performance\textsuperscript{14}. Further, Barnett et al. proposed a novel statistical method specialized for testing rare variant effects based on the optimal Sequence Kernel Association Test (SKAT) using continuous phenotypes in the EPS analyses\textsuperscript{21, 22}. Given the truncated Gaussian distribution of the phenotypical data in EPS, they derived the asymptotical distribution of the statistic proposed by the optimal SKAT. Then a continuous extreme phenotypes SKAT (CEP-SKATO) method was proposed to do the region based rare variant analyses.

Some diseases are caused by mutations in a single gene, such as sickle cell disease and Huntington’s disease. Many other disorders, however, are caused by multiple contributing genetic factors. Therefore, it is critical to simultaneously model the effects of genes or variants. The joint modeling may increase the power of current genetic studies under EPS and identify novel disease-associated genetic factors.

On the other hand, most of the existing methods are single marker/region test, where the genetic markers or regions are tested individually, and a strict significance level is used to
address the multiple testing problem. Although multiple rare variants are usually examined by pooling into regions, the pooled regions are still tested individually. The difficulty of the simultaneous analysis of genetic data is that the number (p) of genetic factors is often much greater than the study sample size (n) in a single genetic study. For example, the number of genes involved is often around 20,000 for transcriptomic gene expression analysis. The number of variants for GWAS is even greater. Meanwhile, the sample size for an individual genetic study is commonly less than or close to 10,000. The standard linear model would not be appropriate for this high-dimensional (p >> n) joint modeling analysis due to the rank deficiency of the design matrix. An alternative solution is to apply the penalized regression methods. For example, the least absolute shrinkage and selection operator (LASSO)\textsuperscript{23}.

LASSO is a regression model that can deal with the high dimensional (p > n) problem by forcing certain regression coefficients to be zero. Since it was proposed, several extensions have been increasingly studied, including the Group LASSO that considers the predefined group structures of predictor variables when applying LASSO\textsuperscript{24}. Meanwhile, these LASSO based approaches have been studied for the application in genetic studies, such as differential gene expression analysis\textsuperscript{25}, the GWAS\textsuperscript{26}, network analysis\textsuperscript{27}, methylation analysis\textsuperscript{28} and so on. In a recent study, Larson et al. proposed a gene-based LASSO and exon-based sparse group LASSO for the rare variant enrichment analysis\textsuperscript{29}.

Further, statistical inference of LASSO has been broadly studied. Various algorithms were proposed to solve LASSO, including LARS\textsuperscript{30}, GLMNET\textsuperscript{31}, SLEP\textsuperscript{32}. Procedures considering noise level were also investigated, such as the Scaled\textsuperscript{33} and Square-root\textsuperscript{34} LASSO. Despite that the limiting distribution of the LASSO estimator has been studied
since 2000\textsuperscript{35}, it was only until recently that the hypothesis testing and/or confidence intervals for LASSO were well shaped under different conditions\textsuperscript{36, 37}. Several studies proposed a debiased method for the hypothesis testing for the sparse linear or generalized linear model with Gaussian or non-Gaussian noise\textsuperscript{38-41}. Ning and Liu proposed a general testing framework based on a decorrelated score function approach and applied it to the linear regression, Gaussian graphical model and additive hazards model\textsuperscript{42, 43}. The SLOPE LASSO was proposed to control the false discovery rate (FDR)\textsuperscript{44}. In spite of these advancements, the statistical inference and testing for the sparse regression model under EPS remain unknown.

In view of the challenges in EPS data analyses and advantages of LASSO, we propose to employ the LASSO based approach for genetic study under EPS. We fully use the non-normal distributed phenotypic data to simultaneously detect the disease-associated genes and/or variants.
II. Literature Review

Linear model and hypothesis testing

Given a data set of \( n \) independent and identically distributed (i.i.d.) pairs \((x_i, y_i), i = 1, 2, ..., n\), where \( x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T \) is the set of the \( p \) predictor variables and \( y_i \) is the response variable for the subject \( i \), we have the linear model (LM):

\[
Y = X\beta + \epsilon, \quad \text{(Equation 1)}
\]

where the design matrix \( X_{n \times p} = [x_1, x_2, ..., x_n]^T \), \( \Sigma \) is the covariance structure of \( X \), \( \beta = (\beta_1, \beta_2, ..., \beta_p)^T \) is a \( p \times 1 \) vector of unknown regression coefficients, \( \epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T \) is an \( n \times 1 \) unobservable the random errors. The errors are assumed to have i.i.d. normal distribution of \( \epsilon_i \sim N(0, \sigma^2) \).

There are typically two ways to estimate the unknown regression coefficients \( \beta \). They are: least square estimate (LSE) and maximum likelihood estimate (MLE). Defined in Equation 2, LSE is the estimate that make the \( X\hat{\beta} \) closest to \( Y \). MLE is the estimate that maximize the log-likelihood function with respect to \( \beta \) (Equation 3). In the previous assumed ordinary linear model, when the order of data matrix \( r(X) < p \), LSE of the unknown regression coefficients \( \beta \) is identifiable and is shown to be best linear unbiased estimates, minimum variance unbiased estimates and equivalent to the maximum likelihood estimates\(^{45}\), which is \( \hat{\beta}_{LS} = \hat{\beta}_{ML} = (X'X)^{-1}X'Y \). In addition to the estimate of \( \beta \), we are also interested in the estimate of \( \sigma^2 \), which is often provided by \( \hat{\sigma}^2 = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-p} \) when \( n > p \). It is noticeable that the MLE of \( \hat{\sigma}^2_{ML} = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n} \) is not used because it is biased. However, \( \hat{\sigma}^2_{ML} \) is still applicable when \( n \leq p \).
After the parameter inference, three general approaches are considered to testing the null hypothesis $H_0: \beta_i = 0$. First, the Score test is well studied for the low-dimensional (p<n) data analysis. Provided the log likelihood function $L(\beta)$, the score function $S$ and Fisher information matrix $I$ is defined as the first and second derivative of $L(\beta)$: $S = \nabla L(\beta)$, $I = -E_\beta(\nabla^2 L(\beta))$. Suppose we are interested in testing the null hypothesis $\beta_i = 0$, let $\beta_{-i}$ denotes the other regression coefficients, it is well known that

$$n\hat{S}_{\beta_i}^2 \hat{I}_{\beta_i|\beta_{-i}}^{-1} \sim \chi^2_{df=1}$$

where $\hat{I}_{\beta_i|\beta_{-i}} = \hat{I}_{\beta_i|\beta_{-i}} - \hat{I}_{\beta_i|\beta_{-i}}^{-1} \hat{I}_{\beta_{-i} \beta_i} \hat{I}_{\beta_{-i} \beta_i}$ is evaluated at the MLE under the null hypothesis. In the Score test, $p$-value $= 1 - CDF_{\chi^2_{df=1}}(n\hat{S}_{\beta_i}^2 \hat{I}_{\beta_i|\beta_{-i}}^{-1}) = \mathbb{P}(\chi^2_{df=1} > n\hat{S}_{\beta_i}^2 \hat{I}_{\beta_i|\beta_{-i}}^{-1})$. The null hypothesis is rejected when the $p$-value is less than the significance threshold.

Second, the Wald test is based on the assumption that the difference between $\hat{\beta}_i$ and the proposed value $\beta$ will be asymptotically normally distributed. Under the previously defined LM, $\hat{\beta}_{LS} \overset{d}{\rightarrow} N(\beta, \text{var}(\hat{\beta}_{LS}))$. In practice, the variance of $\hat{\beta}_{LS}$ is estimated by the Fisher information matrix $\text{var}(\hat{\beta}_{LS}) \approx \frac{1}{n_l}$. The $p$-value of Wald test for $H_0(\beta_i = 0)$ is

$$1 - CDF_{\chi^2_{df=1}}\left(\frac{(\hat{\beta}_i - \beta_i)^2}{\text{var}(\hat{\beta}_i)}\right) = \mathbb{P}(\chi^2_{df=1} > \frac{(\hat{\beta}_i - \beta_i)^2}{\text{var}(\hat{\beta}_i)})$$. Additionally, a confidence interval can be constructed for $\hat{\beta}_{LS}$ as $[\hat{\beta}_{LS} - Z_{\alpha/2} \sqrt{\text{var}(\hat{\beta}_{LS})}, \hat{\beta}_{LS} + Z_{1-\alpha/2} \sqrt{\text{var}(\hat{\beta}_{LS})}]$, where $\alpha$ is
the significance level, $Z_{\alpha}$ is the quantile function of the Normal distribution. If $Z$ follows a Normal distribution, $Z_{\alpha}$ is the value such that $\Pr\left(Z \leq Z_{\alpha}\right) = \alpha/2$.

In hypothesis testing, a large number of simultaneous tests conducted under the same model will lead to the multiple testing problem. It means, by a commonly used significance level $\alpha = 0.05$, the probability of getting at least one significant test result due to chance is far greater than 0.05, which is the probability for the single hypothesis test. The probability of getting at least one significant test in a set of simultaneous tests is also defined as the family wise error rate (FWER).

Several methods have been studied and widely used to deal with multiple testing by adjusting $\alpha$ in some way. For example, the Bonferroni correction and FDR (false discovery rate) estimation. The Bonferroni correction sets the significance level at $\alpha/n$, where $\alpha$ is the desired significance level for a single test, such as the 0.05. After Bonferroni correction, we have the FWER $\leq \alpha^{47}$. The false discovery rate is defined as the proportion of false significant results among all the significant results. Instead of using the strict FWER control, it is sufficient in practice to identify a set of significant findings with the estimated FDR $\leq \alpha^{48}$.

**LASSO regression**

Given similar assumptions about the linear model in Equation 1, we further define the sparsity $s_0$ as the number of non-zero effects in $\beta$; $Y = (y_1, y_2, ..., y_n)^T$ is centered to simplify the model by removing the intercept for continuous $Y$. The LASSO solution is:
\[ \hat{\beta} = \text{argmin} \left\{ \frac{1}{2n} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}, \]  

(Equation 4)

where \( \lambda \geq 0 \) is the tuning parameter controlling the sparsity of \( \beta \) – the proportion of non-zero elements in \( \beta \). When \( \lambda = 0 \), the LASSO solution reduces to the ordinary least solution \( \hat{\beta}_{LS} \) in Equation 2. When \( \lambda \to +\infty \), all \( \beta \) are forced to be 0, which gives the sparsest solution.

There are two major categories of ways to find the optimum \( \lambda \). First is the information criteria (IC), including but not limited to Akaike’s Information Criterion (AIC)\textsuperscript{49}, Bayesian Information Criterion (BIC)\textsuperscript{50}, and a few recently proposed high-dimensional information criteria\textsuperscript{51,52}. The IC is mostly defined as the following format:

\[ IC = -2 \times \log(\text{likelihood}) + k \times \lambda_c, \]  

(Equation 5)

where \( k \) is the number of estimated non-zero parameters in the model. With different \( \lambda_c \), there are different types of ICs. For example, Equation 5 turns to BIC when \( \lambda_c = \log n \), \( n \) is the sample size. When \( \lambda_c = p^{1/3} \), it is another high-dimensional information criterion (HDIC).

The other major ways for tuning parameter is the cross-validation\textsuperscript{53}. In cross-validation, the total data points \((x_i, y_i)\) are randomly separated into \( h \) equal sized subsamples. Given a specific \( \lambda \), the LASSO regression model in Equation 4 is fitted \( h \) times with each subsample serves as a testing sample exactly once and the rest \( h - 1 \) subsamples serves as training samples for parameter inference. After the \( h \) times training and testing procedures, a model fit performance measure is averaged to choose the optimum \( \lambda \) from a
series of $\lambda$, such as the mean squared error \( \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \). In practice, $h = 10$ is usually picked.

For non-continuous $Y$, which is common in genetic analysis, the likelihood function is used to replace the least square function in Equation 4. For example, the response variable $Y$ is binary for a case-control study. Let $p_i = \Pr(Y_i = 1|x_i\beta) = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$, the logistic regression model is used for the binary outcome:

$$\log \frac{p_i}{1-p_i} = x_i\beta + \epsilon, \quad \text{(Equation 6)}$$

The corresponding likelihood based LASSO solution is:

$$\hat{\beta} = \arg\min \left\{-\frac{1}{n} \sum [y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \lambda \|\beta\|_1 \right\}, \quad \text{(Equation 7).}$$

A main weakness of LASSO is that the LASSO estimator is biased, although the LASSO is able to handle $p>n$ problems and reduce the overfitting of the linear model caused by a large number of predictors. A simple way to get a bias reduced estimator is the Lasso-OLS hybrid strategy\(^{30}\). It uses LASSO to select variables but not to estimate their effects. Instead, the linear model with the variables selected from LASSO is fitted to get the estimate of the de-biased effects. Another way is that, in some recent studies, they decomposed the LASSO estimator into unbiased and bias components\(^{39,40}\). The bias part is estimated at the same time as the variable selection. Then a de-biased estimator $\hat{\beta}_d$ is derived by removing the bias term $\hat{\beta}_d = \hat{\beta} + \frac{MX^T(Y-X\hat{\beta})}{n}$.

Compared with the LASSO modeling and estimation, the significance testing in LASSO is less understood. Recently several studies have proposed the hypothesis testing and/or
confidence intervals for LASSO under different conditions\textsuperscript{36,37}. Lockhart et al. developed a test for the coefficient that newly added each time along the LASSO solution path\textsuperscript{36}. For testing the arbitrary coefficients without order, three methods at a given value of $\lambda$ were proposed with different restrictions based on the same idea that decomposing the biased LASSO estimator. Javanmard et al. developed a hypothesis testing requiring less sample size but assuming sparse covariance structures $\Sigma$ and Gaussian distributed design matrix $X$\textsuperscript{39}. van de Geer et al. derived the test in a similar way assuming a sparse $\Sigma^{-1}$\textsuperscript{40}. Without the sparsity assumptions on $\Sigma$ and normality on $X$, a hypothesis testing was proposed by Javanmard et al.\textsuperscript{38}. Different from the hypothesis testing method developed from the de-biased estimator, Ning and Liu proposed a general testing framework based on a decorrelated score function approach and applied it to the linear regression and other generalized linear model\textsuperscript{42,43}. In spite of these advancements, the statistical inference and testing for the sparse regression model under EPS remain unknown.

**LASSO in genetic data analyses**

Since it was proposed in 1996, LASSO and its extensions have been increasingly studied for the transcriptomic gene expression analysis, GWAS, and other various genetic data analyses. The application of LASSO in gene expression analyses mainly includes candidate gene selection and sample classification/prediction\textsuperscript{54,55}. The next year after LASSO was proposed, Tibshirani (1997) provided a version of LASSO for Cox’s regression model with survival data, which gained a lot of applications in cancer survival analysis from microarray data\textsuperscript{56,57}. The trait value, for example, the censored survival time, is modeled as $Y$ in Equation 1. The gene expression value is modeled as $X$ after
some necessary pre-processing, such as the centering, standardization. The simulation and real data analysis shows LASSO is comparable to other standard methods, for instance, the partial least squares regression, and can identify known cancer activities in the context of gene expression data\textsuperscript{58}.

In the traditional GWAS, the LASSO regression is employed to identify disease-relevant SNPs and even their interactions\textsuperscript{26, 59}. Before implementing the LASSO, the SNP genotypes are recoded as the input of the data matrix $X$. If we consider the additive model, the SNPs are coded as three values 0, 1 or 2 corresponding to the three SNPs genotypes A/A, A/a and a/a, respectively. Dominant and recessive model are using similar coding scheme by collapsing some genotypes.

Not only the application in GWAS, LASSO has also been extended to rare variant analyses. The low MAF of rare variant provides very limited information for the effect of the single rare variant. Consequently, numerous statistical methods testing accumulative effects of a group of rare variants have been developed. There are several approaches to aggregate the effects of the rare variants. The collapsing method collapses genotypes across all rare variants into a so-called super allele\textsuperscript{60}. For example, an individual is coded as 1 if a rare variant is observed at any of the variant sites in the region and as 0 otherwise. The weighted and un-weighted sum methods summarize the rare allele count in case and control groups considering some weights from data or prior knowledge\textsuperscript{61, 62}. The permutation or rank sum test is usually used to test the sum statistic. The variance component method detects the group of rare variants based on the variance of the allele frequency or the variance of the random effects of the allele frequency\textsuperscript{63, 64}. Larson et al. simultaneously evaluated the effects of rare variants in multiple genes using LASSO and
Group LASSO⁴⁹. In their analyses, the effects of rare variants in each gene were aggregated as a genetic score by the Madsen and Browning weighted sum method⁶². Then the genetic score for candidate genes was plugged into the LASSO regression model as the design matrix.

Although LASSO regression is widely studied and applied in genetic data analyses, most of them didn’t include a hypothesis testing procedure, which cannot evaluate the confidence of the findings. Meanwhile, a few methods used the permutation test, it requires exchangeable observations under the null hypothesis and great computational resources, which may not valid in practice.
III. Research Question

We are interested in simultaneously testing the associations of multiple genes/variants with phenotypes under EPS. In random sampling data analysis, the linear regression model is commonly used in genetic data analyses. It can model both continuous gene expression levels and categorical SNP genotypes.

In a typical linear model shown in Equation 1, \( Y \) are the phenotypical value, \( X \) are the genetic data and \( \beta \) are the regression coefficients. The phenotypic data \( Y \) given the \( X \) and \( \beta \) are supposed to be normally distributed, so are the noise term in the linear regression model. We rewrote the Equation 1 again for a quick reference:

\[
Y = X\beta + \varepsilon, \quad (\text{Equation 1})
\]

where \( \varepsilon \sim N(0, \sigma^2) \), and \( Y \sim N(X\beta, \sigma^2) \).

In the extreme sampling data analysis, the same regression model (Equation 1) can be fitted. However, the noise and phenotypic value are no longer normally distributed due to the extreme sampling. Suppose the lower and upper threshold for the subject selection is \( c_1 \) and \( c_2 \), the phenotypic values follow a truncated normal distribution with the probability density function (pdf) of:

\[
f(y_i) = \frac{\phi(x_i\beta, \sigma^2)}{\Phi(c_1, \sigma^2) + 1 - \Phi(c_2, \sigma^2)}, \quad y_i \leq c_1 \text{ or } y_i \geq c_2, \quad (\text{Equation 8})
\]

where \( \Phi(x_i\beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i-x_i\beta)^2}{2\sigma^2}\right] \), \( \Phi(c, \sigma^2) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-c\beta)^2}{2\sigma^2}\right] \).

The log-likelihood function is:
\[
\log \mathcal{L}(\beta, \sigma | X) = -\frac{n}{2} \log(2\pi \sigma^2) - \frac{\sum (y_i - x_i' \beta)^2}{2\sigma^2} - \sum \log[\Phi(c_2, x_i' \beta, \sigma^2) + 1 - \Phi(c_1, x_i' \beta, \sigma^2)]
\]

(Equation 9)

If the extreme samples are selected by top and bottom percent denoted as \(pc_1\) and \(pc_2\), the pdf in Equation 8 can be simplified to:

\[
f(y_i) = \frac{\phi(x_i \beta, \sigma^2)}{pc_1 + pc_2}, \quad y_i \leq c_1 \text{ or } y_i \geq c_2.
\]

(Equation 10)

In view of the challenges in EPS data analyses and the advantages and current research progress of the LASSO regression, the research question is how to simultaneously identify the associations of multiple genetic factors with phenotypes under EPS and test their confidences. As a typical example, the proposed model is applied to the gene expression and rare variant analyses. The data matrix \(X\) could be gene expression values for transcriptomic studies, or the aggregated effects of the rare variants in some regions for rare variant analysis. The model is also appropriate for other genetic data, such as the methylation or metabolomics data. The regression coefficients \(\beta\) are tested to evaluate if the genetic factors have some significant effects on the phenotype. The performance of the proposed model will be evaluated by the comparison with the existing methods for EPS.
IV. Methods and Data

Sparse regression for EPS genetic data analysis

Given the research question described in the previous section, we propose to employ the decorrelated score function based decorrelated score test for LASSO regression (EPS-LASSO) considering extremely sampled traits to simultaneously detect the disease-associated genetic factors.

For the low-dimensional (p<n) EPS data analysis, several methods have studied the maximum likelihood estimate (MLE) of $\beta$ and $\sigma$ by the Newton Raphson (N-R) procedure$^{14,21}$. However, these methods fail to obtain estimates for high-dimensional data.

To address this challenge, we propose a penalized maximum likelihood estimate (PMLE) for EPS:

$$\hat{\beta}, \hat{\sigma} = \arg\min \left( -\log L(\beta, \sigma|X) + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad \text{(Equation 11)}$$

where $\lambda$ is a tuning parameter controlling the sparsity ($s_0$) of $\beta$.

Given $\lambda$, we use an iterative algorithm (Algorithm 1) akin to the scaled sparse regression to infer $\hat{\beta}$ and $\hat{\sigma}$ by solving the optimization problem in Equation 11. The initial estimate of $\hat{\beta}^{(0)}$ is from classical Lasso. $\hat{\sigma}_{MLE} > 0$ is guaranteed when $(Y - X\beta)^T(Y - X\beta) \neq 0$.

**Proof 1**: In Algorithm 1 $\sigma$ is the maximum likelihood estimate (MLE). When we solve the MLE, the equation that the first derivative equals 0 can be simplified to be a quadratic equation as below.
Let \( l = - \log L(\beta, \sigma) = \frac{n}{2} \log(2\pi \sigma^2) + \frac{\sum(y_i - x_i'\beta)^2}{2\sigma^2} + \sum \log[\Phi(c_2, x_i'\beta, \sigma^2) + 1 - \Phi(c_1, x_i'\beta, \sigma^2)], \) \( t_{i1} = \frac{c_1 - x_i'\beta}{\sigma}, t_{i2} = \frac{c_2 - x_i'\beta}{\sigma}. \)

we have \( \frac{\partial l}{\partial \sigma} = \frac{n}{\sigma} - \frac{(Y - \mathbf{x}\beta)'(Y - \mathbf{x}\beta)}{\sigma^3} - \frac{1}{\sigma^2} \sum \frac{\phi(t_{i2})(c_2 - x_i'\beta) - \phi(t_{i1})(c_1 - x_i'\beta)}{\phi(t_{i2}) + 1 - \phi(t_{i1})}. \)

Let \( \frac{\partial l}{\partial \sigma} = 0, \) we have:

\[
n \sigma_{\text{MLE}}^2 - \sigma_{\text{MLE}} \sum \frac{\phi(t_{i2})(c_2 - x_i'\beta) - \phi(t_{i1})(c_1 - x_i'\beta)}{\phi(t_{i2}) + 1 - \phi(t_{i1})} - (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) = 0.
\]

Let \( b = \sum \frac{\phi(t_{i2})(c_2 - x_i'\beta) - \phi(t_{i1})(c_1 - x_i'\beta)}{\phi(t_{i2}) + 1 - \phi(t_{i1})}, c = (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) \geq 0, \)

then \( n \sigma_{\text{MLE}}^2 - b \sigma_{\text{MLE}} - c = 0. \)

\[
\hat{\sigma}_{\text{MLE}} = \frac{b \pm \sqrt{b^2 + 4nc}}{2n}
\]

\[
b^2 + 4nc \geq b^2 \Rightarrow \sqrt{b^2 + 4nc} \geq |b| \geq \pm b
\]

\[
\Rightarrow b + \sqrt{b^2 + 4nc} \geq 0, b - \sqrt{b^2 + 4nc} \leq 0
\]

It is easy to show that \( \sigma_{\text{MLE}} \) has and only has a positive solution when \( c = (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) \neq 0. \) The positive \( \hat{\sigma}_{\text{MLE}} \) is picked as the estimate of \( \sigma. \) The calculation process is the same as that used in the CEP-SKATO\(^{21}.\)

We use BIC\(^{50}\) considering both model fit and number of estimated parameters \( (k_{\lambda}) \) to select the optimum \( \lambda^* \) from a series of \( \lambda \) values. The reason we choose BIC is because of its high power and low FDR compared to cross-validation\(^{53}\) and other high-dimensional information criterion (Appendix X.I), which also agrees with previous studies\(^{65, 66}.\)
\[ \lambda^* = \text{arg \min} (-2 \log L + k_\lambda \log n) \quad \text{(Equation 11)} \]

At Algorithm 1 step (iii), we take the support set of converged \( \hat{\bm{\beta}}^{(k)} \) at step (ii) and on its support re-estimate \( \hat{\bm{\beta}} \) and \( \hat{\sigma} \) through the N-R procedure. The refitted estimator is motivated by the LARS-OLS hybrid strategy\(^{30}\) and numerical studies from\(^{43}\), in which the refitted estimator leads to better finite sample performance (Appendix X.II).

---

**Algorithm 1 Estimate of \((\bm{\beta}, \hat{\sigma})\) in Equation 11**

Require: Data set of \( n \) i.i.d. pairs \((x_i, y_i)\), and tuning parameter \( \lambda \)

(i): Initialize at \( k=0 \):

\[
\begin{align*}
\hat{\bm{\beta}}^{(k)} &= \text{arg \ min} \left( \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i' \bm{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \\
\hat{\sigma}^{(k)} &= \text{arg \ min} \left( -\log L(\hat{\bm{\beta}}^{(k)}, \sigma) \right)
\end{align*}
\]

(ii): For \( k=k+1 \) until convergence:

\[
\begin{align*}
\hat{\bm{\beta}}^{(k)} &= \text{arg \ min} \left( -\log L(\bm{\beta}, \hat{\sigma}^{(k-1)}) + \lambda \sum_{j=1}^{p} |\beta_j| \right), \\
\hat{\sigma}^{(k)} &= \text{arg \ min} \left( -\log L(\hat{\bm{\beta}}^{(k)}, \sigma) \right)
\end{align*}
\]

(iii): Let \( \hat{S} = \{ m \in \{1, \ldots, p\}: \hat{\beta}_m^{(k)} \neq 0 \} \), \( \hat{S}^c \) is the complement set.

Refit \((\hat{\bm{\beta}}_{\hat{S}}, \hat{\sigma}) = \text{arg \ min} \left( -\log L(\bm{\beta}_{\hat{S}}, \sigma) \right), \hat{\bm{\beta}}_{\hat{S}^c} = 0\)

(iv): Return \( \hat{\bm{\beta}} = (\hat{\bm{\beta}}_{\hat{S}}, \hat{\bm{\beta}}_{\hat{S}^c}), \hat{\sigma} \)

---

To test the rare variant (RV) effects simultaneously, the RVs are first collapsed into regions or genes, which extends EPS-LASSO to EPS-LASSO-RV for RV analysis.

Suppose there are \( n \) samples, \( s \) regions and \( m_j \) RVs within each region \((j = 1, 2, \ldots, s)\).

The \( k \)-th mutation allele count of the \( i \)-th sample \((i = 1, 2, \ldots, n)\) is denoted as

\[ g_{ik} \sim \text{Binom}(2, f_k), \text{ where } f_k = \frac{1 + \sum_{i=1}^{n} g_{ik}}{2(n+1)} \quad \text{is the estimated MAF of the} \ k \text{-th mutation.} \]

The aggregated signal of the \( m_j \) RVs in the \( j \)-th region for the \( i \)-th sample is calculated as

\[ x_{ij} = \frac{1}{2m_j} \sum_{k=1}^{m_j} w_k g_{ik}. \] The weight \( w_k = \frac{1}{\sqrt{2f_k(1-f_k)}} \) is the inverse of the standard
deviation of $g_{ik}$, which adapts to the empirical rarity of the variant by up weighting rare variants\textsuperscript{62}. Consider a vector of covariates $x_{cl}$ and a continuous phenotype $y_i$:

$$y_i = z_i'\beta + \varepsilon_i,$$

(Equation 12)

where $y_i$ is centered to simplify the model by removing the intercept; $z_i = (x_i', x_{cl}')$ is the combined vector of collapsed RV effects in regions $x_i' = \{x_{ij}|1 \leq j \leq s\}$ and covariates $x_{cl}'; \beta = (\alpha', \alpha_c')'$ is the vector of regression coefficients for the regions and covariates; other assumptions are the same as Equation 1. Consequently, the Algorithm 1 can also be applied to the gene based RV association model.

**Hypothesis testing for EPS-LASSO(-RV)**

After the estimate of the regression coefficients, we develop a hypothesis testing procedure to control the uncertainty of the regression estimate for the proposed EPS sparse regression model by following Ning and Liu’s general framework for high-dimensional models \textsuperscript{43}. The null hypothesis is $H_0: \beta_j = 0$ vs $H_a: \beta_j \neq 0$. For general purpose, the null hypothesis is testing the $j$-th predictor, which represents the overall effects of the RVs within a gene for RV analysis.

Provided the null hypothesis $H_0: \beta_j = 0$, and let $\beta_{-j}$ denote the other regression coefficients, $\hat{S}_{\beta_j} = \nabla_{\beta_j} L(\hat{\beta}, \hat{\sigma}|X)$ and $\hat{I} = -E_{\beta} \left( \nabla_{\beta\beta}^2 L(\hat{\beta}, \hat{\sigma}|X) \right)$ be the score function and Fisher information matrix respectively, it is widely known that \textsuperscript{46}:

$$n \hat{S}_{\beta_j}^2 \hat{I}_{\beta_j|\beta_{-j}}^{-1} \sim \chi^2_{df=1},$$
where $\hat{\beta}_{j|\beta_{-j}} - \hat{\beta}_{j|\beta_{-j}} = \hat{\beta}_{j|\beta_{-j}} - w\hat{\beta}_{j|\beta_{-j}}$, $w = \frac{\hat{\beta}_{j|\beta_{-j}}}{\hat{\beta}_{j|\beta_{-j}} - \hat{\beta}_{j|\beta_{-j}}}$ is evaluated at the MLE under the null hypothesis. However, Ning and Liu demonstrated that the classical score statistic does not work in the high-dimensional situation, because of the asymptotically ignorable remainder converges to some intractable limiting distribution. Instead, they proposed a revised score statistic based on a decorrelated score function for a broad class of high-dimensional generalized linear model:

$$S_{\beta_j}^* = \hat{S}_{\beta_j} - \hat{w}^T \hat{S}_{\beta_{-j}}, \quad \text{(Equation 12)}$$

where $\hat{w}$ is estimated by the best sparse linear combination of $\hat{S}_{\beta_{-j}}$ to approximate $\hat{S}_{\beta_j}$.

For the EPS, given the PDF and log-likelihood function in Equations (8) and (9), we have the score function $S = \nabla_{\beta} L(\beta|X) = \frac{x'y-x'\beta}{\sigma^2} - \frac{x'M}{\sigma}$, where $M$ is a vector of length $n$ with

$$m_i = \frac{\Phi(\phi_i-x_i'\beta, \sigma^2) - \Phi(\phi_i-x_i'\beta, \sigma^2)}{\Phi(\phi_i-x_i'\beta, \sigma^2)+1-\Phi(\phi_i-x_i'\beta, \sigma^2)}.$$

and the Fisher information matrix $I = \frac{x'Vx}{\sigma^2}$, in which $V$ is a $n$-dimensional diagonal matrix with the $i$-th diagonal element

$$v_i = 1 - \frac{(c_2-x_i'\beta)\Phi(\phi_i-x_i'\beta, \sigma^2)-(c_1-x_i'\beta)\Phi(\phi_i-x_i'\beta, \sigma^2)}{[\Phi(\phi_i-x_i'\beta, \sigma^2)+1-\Phi(\phi_i-x_i'\beta, \sigma^2)]\sigma} - m_i^2.$$

Then, the decorrelated score function is $S_{\beta_j}^*$ defined as (Equation 12) with the $w$ solved by the following Lasso type estimator for the high-dimensional setting:

$$\hat{w} = \text{argmin}_w \frac{1}{2n} \|\hat{S}_{\beta_j} - w^T \hat{S}_{\beta_{-j}}\|^2_2 + \lambda' \sum_{d=1}^n |w_d|,$$

where $\lambda'$ is selected by cross validation using the R package GLMNET.
With several assumptions that are commonly made on asymptotic of LASSO type estimator \(40, 43, 67\), we prove the following fact for the estimate from EPS-LASSO:

\[
\sqrt{n} \hat{\beta}_j \sim N(0, 1).
\]

The detailed proof is shown in the Appendix X.III. Finally, with the parameter estimate of \(\hat{\beta}\) and \(\hat{\sigma}\) in Algorithm 1, we implement the Algorithm 2 to test the null hypothesis that \(\beta_j = 0\).

Algorithm 2. Hypothesis test of \(\beta_j = 0\)

Require: \((\hat{\beta}, \hat{\sigma})\) from Algorithm 1

(i): Set \(\tilde{\beta}_j = 0\) and calculate \(\tilde{S} = \frac{x'y - x'\hat{\beta}}{\hat{\sigma}^2} - \frac{x'M}{\hat{\sigma}^2}, \tilde{I} = \frac{x'\tilde{S}}{\hat{\sigma}^2}\)

(ii): Solve \(\tilde{w} = argmin 1 \quad \frac{1}{2n} \left\vert \hat{S}_{\beta_j} - \hat{w}^T \hat{S}_{\beta_j} \right\vert^2_2 + \lambda' \sum_{d=1}^{n} |w_d|\)

(iii): Calculate \(S_{\beta_j} = \hat{S}_{\beta_j} - \tilde{w}^T \hat{S}_{\beta_j}, \hat{I}_{\beta_j|\beta_{-j}} = \tilde{I}_{\beta_j|\beta_{-j}} - \tilde{w} \hat{I}_{\beta_j|\beta_{-j}}\)

\[p\text{-value} = 1 - CDF_{\chi^2_{df=1}}(nS_{\beta_j^2\hat{I}_{\beta_j|\beta_{-j}}^{-1}}) = \mathbb{P}(\chi^2_{df=1} > nS_{\beta_j^2\hat{I}_{\beta_j|\beta_{-j}}^{-1}})\]

In our method, the decorrelated score function can be regarded as an approximately unbiased estimation function for \(\beta^{68}\). Given the decorrelated score function, a de-biased estimator (\(\tilde{\beta}_j\)) can be derived by solving \(\hat{S}_{\beta_j} + (\tilde{\beta}_j - \hat{\beta}_j)\hat{I}_{\beta_j|\beta_{-j}} = 0\), which is similar to Huber’s One-step estimate in the linear model\(^{69}\). Further, \(\tilde{\beta}_j\) is proved to be approximately normal \(\sqrt{n}(\tilde{\beta}_j - \beta_j) \sim \tilde{I}_{\beta_j|\beta_{-j}}^{1/2} N(0, 1)\) in EPS-LASSO (Appendix X.III).

Then, we have the Wald test statistic for the \(H_0\) as: \(n\tilde{\beta}_j^2 \hat{I}_{\beta_j|\beta_{-j}}\) with the \(p\)-value = 1 – \(CDF_{\chi^2_{df=1}}(n\tilde{\beta}_j^2 \hat{I}_{\beta_j|\beta_{-j}}) = \mathbb{P}(\chi^2_{df=1} > n\tilde{\beta}_j^2 \hat{I}_{\beta_j|\beta_{-j}})\).
Simulation design

To validate the proposed sparse regression model with hypothesis testing for EPS data, we simulate several typical scenarios of subjects with extreme phenotypes and genetic predictors. We assume the extreme sampling is conducted by selecting top and bottom 20% subjects from a random sampling size of 500, which results in a sample size of 200 for EPS. The number of the predictors ($p$) could be 100 for low-dimensional, and 200, 400 or 800 for high-dimensional situations. For each scenario, the phenotype of the $i$-th individual in the random sampling pool is generated by a linear model:

$$y_i = \sum_{s=1}^{p} \beta_s G_{is} + \epsilon_i,$$

where $\epsilon_i \sim N(0,1)$ is the random noise; and $\beta_s$ is the effect size of the corresponding predictor. $G_{is}$ is the $s$-th predictor value for the $i$-th individual, which are generated from a multivariate normal distribution $G_i \sim N(\mathbf{0}, \Sigma)$ with the covariance matrix $\Sigma$:

$$\Sigma_{jk} = \begin{cases} \rho^{k-j} & \text{if } k \in \{j, j+1, \ldots, j+5\} \\ 0 & \text{for all other } j \leq k \end{cases}.$$

$\Sigma_{jk} = \Sigma_{kj}$, for $j > k$, and $\rho$ is chosen from $(0, 0.2, 0.4)$.

We design a set of null effects scenarios to examine the type I error via setting all $\beta_s = 0$. For effective scenarios to examine the power and FDR, we randomly pick 10 non-correlated predictors as causal factors with same non-zero $\beta_s$ selected from $(0.1, 0.15, 0.2, 0.25, 0.3)$. The type I error for those non-causal causal predictors is also summarized to compare the type I error control when correlations among non-causal and causal factors are present.
Using the simulated data sets, we compare our model, named EPS-LASSO, with several commonly used methods for hypothesis testing including the ordinary linear model (LM), logistic regression model (LGM), linear model based on EPS likelihood (EPS-LM), and a high-dimensional Lasso testing method assuming random sampling (SSLASSO) \(^{38}\). LM, LGM, and EPS-LM are applied to test the predictors individually. In LGM, samples at the bottom and up percentiles are treated as two groups. They are all implemented in R. We prepare the EPS-LM source code based on the R package CEPSKAT. The source code for SSLASSO is downloaded from the author’s website (https://web.stanford.edu/~montanar/sslasso/code.html).

After 500 replications, the type I error, power and FDR are assessed at original and Bonferroni corrected \(\alpha = 0.05\) level. The type I error is defined as the proportion of significant non-causal predictors among all the non-causal predictors. Since two significance level are considered, to make a straightforward comparison, we define another index – type I error inflation rate as the ratio of observed type I error versus expected type I error. The power is defined as the proportion of significant causal predictors among all the causal predictors. The FDR is defined as the proportion of significant non-causal predictors among all the significant predictors. Additionally, the absolute bias (\(|\hat{\sigma} - \sigma|\)) of the estimate of the noise standard deviation (SD) from EPS-LASSO is compared to those from LM, EPS-LM and Scaled Lasso, respectively. To show the advantage of EPS-LASSO using the true distribution to infer \(\beta\) and \(\sigma\), we run a few scenarios to compare EPS-LASSO to LASSO following the same decorrelated score test (LASSO-DST). In LASSO-DST, the ordinary LASSO and the sample variance estimate of the mean residual square was used for the input of the following decorrelated
score test. The variance estimate from Scaled LASSO is not used because of the large bias in the estimate of \( \sigma \).

These simulation scenarios work for general genetic data, like the transcriptomic gene expression data. We simulate more scenarios to test EPS-LASSO-RV for the rare variant genetic study. First, a haplotype pool of 10,000 subjects was generated using an efficient whole genome-wide SNP simulator – GeneEvolve\(^7\) based on the 1000 Genomes Project Phase 3 human reference genome (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) to make data with the properties of the real genome. The rare variant effects in gene regions of chromosome 20 are simulated to assess the model performance. The total gene number in chromosome 20 is 721.

Like the simulation study for EPS-LASSO, we design null effects scenarios with all genetic effects to be 0 and effective scenarios. In the effective scenarios, we randomly pick 10 genes as causal genes carrying rare variant effects. Each of the 10 genes contains ~130 rare variants and is ~50,000 bases in length, which is about the average gene length in the genome. The phenotype of the \( i \)-th individual was produced by a linear model:

\[
y_i = x_{1i} + x_{2i} + \sum_{s=1}^{10} G_{si} + \varepsilon_i, \tag{Equation 13}
\]

where \( x_{1i} \sim N(20,1) \) and \( x_{2i} \sim \text{Bernoulli}(0.5) \) are two covariates; \( \varepsilon_i \sim N(0,1) \) is the random noise; \( G_{si} = \sum_{t=1}^{t_s} \beta_{st} g_{it} \) is the aggregated effect of the total \( t_s \) rare variants within the \( s \)-th causal gene. \( g_{it} \) is the genotype of the \( t \)-th rare variant in the \( i \)-th subject. \( \beta_{st} = u | \log_{10} \text{MAF}_t | \) is the corresponding effect size determined by a constant \( u \) and the variant’s MAF, akin to settings in existing studies\(^21, 29\). All \( \beta_{st} = 0 \) in the null effects.
scenarios. The proportion of causal rare variants can be controlled by varying the constant $u$. We tested the proportion of 20% and 40% respectively. Only deleterious effects are considered. Within each causal gene, rare variants are sampled until the cumulative variance explained (VE) by the gene reached $\sim$1.5%, which is a common finding in previous rare variant analyses\textsuperscript{71}. From the randomly generated samples using Equation 13, we selected top and bottom 20% individuals by their phenotypes. The samples size after selection includes 200, 500, 800, and 1000 covering both low- and high-dimensional situations. Using the simulated data sets, we compared our model EPS-LASSO-RV to SSLASSO and CEP-SKATO\textsuperscript{21}. CEP-SKATO is a state-of-the-art method for detecting rare variants under EPS. SSLASSO is applied after the RV effects were collapsed into gene regions. After 100 replications, the type I error, power and FDR defined in the previous paragraph are assessed at original and Bonferroni corrected $\alpha = 0.05$ level.

**Real data analyses**

Further, we would like to apply the model to the two types of genetic data analyses: EPS-LASSO for the transcriptomic gene expression and EPS-LASSO-RV for the rare variant analyses using real data.

The transcriptomic gene expression data is downloaded from a sub-study of Framingham Cohort project (dbGaP: phs000363)\textsuperscript{15,16}, which includes a profiling of 17,621 genes for 2,442 Framingham Heart Study offspring subjects using the Affymetrix Human Exon 1.0
ST Gene Chip platform. The gene expression values were normalized with quality control measures as previously reported \(^7\). We pick the BMI as the interesting trait, which is a major characteristic of obesity. Gender, age, drinking, and smoking status are considered as potential covariates. After removing the missing value in phenotypes, 972 subjects with the highest 20% or lowest 20% of BMI are selected as the EPS sample. The Bonferroni corrected significance level of 0.05 is used to claim the significance.

The real data for rare variant analyses is downloaded from the Framingham Heart Study Allelic Spectrum Sequencing Discovery project (dbGaP: phs000307)\(^{15, 16}\), which includes a deep coverage targeted re-sequencing and variant identification for 216 genes of 1,635 unrelated subjects. The SNPs are re-mapped into 251 gene regions based on a gene annotation file provided by Plink (https://www.cog-genomics.org/plink/1.9/resources#genelist). Here, we also analyze the obesity-related phenotype – BMI. After removing the missing value in phenotypes, 540 subjects with highest 20% and lowest 20% of BMI are selected as the EPS sample. Gender, age, age squared, and smoking status are considered as potential covariates. Only rare variants with MAF<0.01 were analyzed. Genes not containing any rare variants in the extreme samples are excluded from the analysis. The Bonferroni corrected significance level of 0.05 is used to claim the significance.
Source code

All the simulation and real data analyses were conducted using R packages and in-house scripts. We developed an R package called EPSLASSO to implement our method. It is free to download and install at https://github.com/xu1912/EPSLASSO.
V. Simulation Result

Performance of EPS-LASSO

We first assess the type I error of EPS-LASSO under various null scenarios at original and Bonferroni corrected significance level $\alpha = 0.05$. The type I error rates of EPS-LASSO at a significance level of 0.05 under various null scenarios are compared with those from SSLASSO, LM, EPS-LM, and LGM (Fig. V-1). The ideal type I error after Bonferroni correction is $5E-04$, $2.5E-04$, $1.25E-04$ and $6.25E-05$ for 100, 200, 400 and 800 predictors respectively.

The type I error of EPS-LASSO and other methods after Bonferroni correction are listed in Table V-1. EPS-LASSO controls type I error well at both original and Bonferroni corrected significance level in all the scenarios, so does LM and EPS-LM. Conversely, LGM results in a deflated type I error, while another high-dimensional method SSLASSO inflates the type I error.

The Quantile-Quantile plot of observed p-values against expected p-values for all the scenarios is shown in Fig. V-2. The points from EPS-LASSO align close to the diagonal line and all fall into the 95% confidence region, which indicates that EPS-LASSO has well-controlled type I error rates for both low- and high-dimensional situations. Conversely, LM, EPS-LM, and LGM result in a slightly deflated type I error for high-dimensional and high-correlation scenarios, while another high-dimensional method SSLASSO inflates the type I error in some scenarios.

We further compare EPS-LASSO to LASSO-DST in terms of the type I error under a few scenarios. As a result, LASSO-DST cannot produce correct type I error. In the null model,
at a significance level of 0.05, the type I error of LASSO-DST for the high-dimensional scenarios is inflated (Fig. V-3). Considering the Bonferroni correction, the inflation is even higher (Table V-2). Meanwhile, the estimate of $\sigma$ in LASSO-DST is also more biased than EPS-LASSO (Fig. V-4), which may partially explain the incorrect type I error. On the other hand, EPS-LASSO yields a type I error close to the expected level due to using the true distribution to infer $\beta$ and $\sigma$.

![Fig. V-1. Type I error comparison at a significance level of 0.05 for null scenarios. Different colors indicate different methods. The horizontal axis represents $\rho$ (rho): 0, 0.2, and 0.4. Number of predictors $p$: 100, 200, 400, and 800.](image)

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\rho$</th>
<th>EPS-LASSO</th>
<th>LM</th>
<th>LGM</th>
<th>EPS-LM</th>
<th>SSLASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.00</td>
<td>4.80E-04</td>
<td>4.80E-04</td>
<td>2.60E-04</td>
<td>4.80E-04</td>
<td>7.00E-04</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>4.80E-04</td>
<td>4.80E-04</td>
<td>2.60E-04</td>
<td>4.60E-04</td>
<td>6.60E-04</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>4.60E-04</td>
<td>5.20E-04</td>
<td>2.60E-04</td>
<td>5.20E-04</td>
<td>6.40E-04</td>
</tr>
<tr>
<td>200</td>
<td>0.00</td>
<td>2.20E-04</td>
<td>2.40E-04</td>
<td>1.10E-04</td>
<td>2.40E-04</td>
<td>7.90E-04</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>2.10E-04</td>
<td>2.40E-04</td>
<td>1.10E-04</td>
<td>2.30E-04</td>
<td>8.40E-04</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>2.50E-04</td>
<td>3.00E-04</td>
<td>1.00E-04</td>
<td>2.90E-04</td>
<td>7.10E-04</td>
</tr>
<tr>
<td>400</td>
<td>0.00</td>
<td>1.45E-04</td>
<td>1.15E-04</td>
<td>6.00E-05</td>
<td>1.05E-04</td>
<td>7.35E-04</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>1.20E-04</td>
<td>1.05E-04</td>
<td>5.50E-05</td>
<td>1.05E-04</td>
<td>7.05E-04</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>1.50E-04</td>
<td>1.40E-04</td>
<td>5.50E-05</td>
<td>1.25E-04</td>
<td>5.90E-04</td>
</tr>
<tr>
<td>800</td>
<td>0.00</td>
<td>5.25E-05</td>
<td>5.00E-05</td>
<td>1.75E-05</td>
<td>4.75E-05</td>
<td>9.03E-04</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>7.00E-05</td>
<td>5.25E-05</td>
<td>1.50E-05</td>
<td>5.00E-05</td>
<td>8.48E-04</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>9.50E-05</td>
<td>5.25E-05</td>
<td>2.25E-05</td>
<td>5.25E-05</td>
<td>6.43E-04</td>
</tr>
</tbody>
</table>
Fig. V-2. Quantile-Quantile plot for null models without causal predictors. Different colors indicate different methods.

Table V-2. Type I error inflation rate of EPS-LASSO and LASSO-DST

<table>
<thead>
<tr>
<th>n</th>
<th>( \rho = 0 )</th>
<th>( \rho = 0.4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPS-LASSO</td>
<td>LASSO-DST</td>
</tr>
<tr>
<td>100</td>
<td>1.12</td>
<td>1.88</td>
</tr>
<tr>
<td>200</td>
<td>0.68</td>
<td>1.76</td>
</tr>
<tr>
<td>400</td>
<td>1.32</td>
<td>5.72</td>
</tr>
<tr>
<td>800</td>
<td>1.28</td>
<td>32.64</td>
</tr>
</tbody>
</table>
Then, we examine the type I error in the scenarios with causal predictors using the scheme of 400 predictors as an example. When the causal factors are added into the null model, EPS-LASSO still controls the type I error in the range between 0.045 and 0.055, so does the SSLASSO (Fig. V-5). On the other hand, the three low-dimensional methods LM, LGM and EPS-LM yield inflated type I error as great as ~0.074 with the increase of the correlation between predictors ($\rho$) and the magnitude of the causal effect size ($\beta_s$). After multiple testing correction, the type I error of EPS-LASSO is slightly inflated for large effect size (Fig. V-5). SSLASSO inflates type I error under all conditions. The type I error of LM and EPS-LM are controlled when the correlation is weak ($\rho \leq 0.2$), while
the type I error of LGM is deflated. When \( \rho \) increases to 0.4, severe inflation occurs for LM and EPS-LM. It shows the potential advantage of EPS-LASSO in practice, where the correlation among predictors and multiple causal genetic factors are present.

Fig. V-5. Type I error at raw (top) and after Bonferroni correction (bottom) significance level of 0.05 for effective scenarios with 400 predictors. Red dashed line is the expected level. The x-axis represents \( \rho \) (rho).
We compare the power and FDR of EPS-LASSO to other methods at the Bonferroni corrected significance level of 0.05 in Fig. V-6. In all the scenarios, EPS-LASSO outperforms other low-dimensional methods (LM, LGM, and EPS-LM) by a faster-growing power with the increase of the causal effect size (Fig. V-6). When the causal effect size is less than 0.2, EPS-LASSO yields a power close to EPS-LM and LM. Then the power of EPS-LASSO exceeds others for effect size $\geq 0.2$, while LGM results in the worst power due to the least information used. In addition, increasing the number of predictors led a power loss of all these methods, but EPS-LASSO is less sensitive compared with other methods. For example, for the scenarios of $\beta_s = 0.3$, $\rho = 0$, the power of EPS-LASSO decreases from 86.9% to 76.5% when the number of predictors increases from 200 to 800. The relative decline of 12.0% is less than LM (13.6%), EPS-LM (14.0%) and LGM (25.4%) (Fig. V-7). The other high-dimensional method SSLASSO fails to work for low-dimensional setting ($p=100$). Although the power of SSLASSO raises for high-dimensional scenarios, the FDR boosts to as great as $\sim 0.3$ for a large number of predictors and small effect sizes. On the other side, EPS-LASSO produces a stable FDR across all scenarios and is robust to the effect size, number of predictors, and correlation among predictors than other methods. The FDR of low-dimensional methods are lower than EPS-LASSO for weak correlation scenarios, however, their FDR greatly uplifted and exceeded EPS-LASSO when $\rho > 0.2$. 
Fig. V-6. Power and FDR for effective scenarios with 100, 200, 400 and 800 ($p$) predictors. Different colors indicate different methods. The lines aligned to the left vertical axis shows the power. The histograms aligned to the right vertical axis show the FDR. The horizontal axis represents $\rho$ (rho).

Fig. V-7. Power decrease with the increase of predictors. The x-axis represents the number of predictors ($p$).
Moreover, EPS-LASSO is superior to LM and EPS-LM with respect to the estimate of the noise SD across all the scenarios (Table V-3). The absolute bias of EPS-LASSO ranges from 0.051 to 0.085 with a slightly increasing trend about the effect size. The absolute bias of EPS-LM is comparable to EPS-LASSO when the effect size was small, but increases to ~0.372 with the increase of the effect size. The ordinary method – LM resulted in greater bias as large as ~1.029. Another sparse model providing variance estimate – Scaled-Lasso gives an even worse result, which is unstable and much greater for most cases.
Table V-3. The absolute bias of the noise SD estimate in all scenarios

<table>
<thead>
<tr>
<th>Effect size ($\beta_s$)</th>
<th>Correlation ($\rho$)</th>
<th>EPS-LASSO</th>
<th>LM</th>
<th>EPS-LM</th>
<th>Scaled Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=100</td>
<td>200</td>
<td>400</td>
<td>800</td>
<td>100</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0.051</td>
<td>0.053</td>
<td>0.053</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.050</td>
<td>0.052</td>
<td>0.051</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.050</td>
<td>0.052</td>
<td>0.053</td>
<td>0.051</td>
</tr>
<tr>
<td>0.15</td>
<td>0</td>
<td>0.060</td>
<td>0.070</td>
<td>0.071</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.061</td>
<td>0.068</td>
<td>0.071</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.062</td>
<td>0.072</td>
<td>0.074</td>
<td>0.077</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0.056</td>
<td>0.062</td>
<td>0.064</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.058</td>
<td>0.063</td>
<td>0.071</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.059</td>
<td>0.065</td>
<td>0.069</td>
<td>0.080</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0.067</td>
<td>0.068</td>
<td>0.071</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.067</td>
<td>0.069</td>
<td>0.073</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.067</td>
<td>0.071</td>
<td>0.073</td>
<td>0.074</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0.078</td>
<td>0.075</td>
<td>0.080</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.077</td>
<td>0.079</td>
<td>0.080</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.077</td>
<td>0.081</td>
<td>0.080</td>
<td>0.085</td>
</tr>
</tbody>
</table>
As described in the Method part, a Wald test statistic of $\hat{\beta}_j$ can be constructed in addition to the Score test. We do more simulations using null model and causal model with an effect size of 0.2 to assess the performance of the Wald test. Shown in Table V-4, the Wald test acts similarly but slightly liberally relative to the score test regarding the type I error. Especially when the Bonferroni correction is applied, the type I error inflation rate of Wald test is always greater than that of Score test. Using the effective scenarios, the power of Wald test is a little bit greater than the power of Score test, while the FDR of Wald test is usually greater than that of Score test too (Fig. V-8).

![Fig. V-8](image_url)

Fig. V-8. Power and FDR of EPS-LASSO using Score and Wald test. Line aligned to the left axis represents power. Histogram aligned to the right axis represents FDR. The horizontal axis is the number of predictors.
Table V-4. Type I error inflation rate of EPS-LASSO using Score and Wald test

<table>
<thead>
<tr>
<th>$\rho(rho)$</th>
<th>$p$</th>
<th>Score test</th>
<th>Wald test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha=0.05$</td>
<td>$\alpha=0.05/p$</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>1.00</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.01</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.01</td>
<td>1.28</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>1.02</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.99</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.99</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.99</td>
<td>1.08</td>
</tr>
</tbody>
</table>
Performance of EPS-LASSO-RV

We examine the type I error of EPS-LASSO for rare variant analysis (EPS-LASSO-RV) under various scenarios at significance level $\alpha = 0.05$. Shown in Table V-5, EPS-LASSO-RV yields a stable type I error of $\sim 0.05$. So does SSSLASSO and CEP-SKATO. When Bonferroni corrected significance level is considered, all the methods inflate the type I error. But EPS-LASSO-RV is nearest to the expected value with the least inflation rate (Table V-6). SSSLASSO results in the highest inflation rate. EPS-LASSO-RV can better control the type I error than SSSLASSO and CEP-SKATO when considering the multiple testing. The QQ plot of the P-value shows all the methods fell into the 95% confidence region for most cases (Fig. V-9). The EPS-LASSO-RV using Wald statistic acts almost the same as the Score statistic (Table V-5 and V-6). The QQ plot does not include the EPS-LASSO-RV using Wald statistic because of the high overlapping between its result and that from the Score statistic.

<table>
<thead>
<tr>
<th>$p$</th>
<th>SSLASSO</th>
<th>CEP-SKATO</th>
<th>EPS-LASSO-RV</th>
<th>EPS-LASSO-RV(Wald)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TIE</td>
<td>IR</td>
<td>TIE</td>
<td>IR</td>
</tr>
<tr>
<td>200</td>
<td>5.32E-02</td>
<td>1.064</td>
<td>5.55E-02</td>
<td>1.111</td>
</tr>
<tr>
<td>500</td>
<td>5.17E-02</td>
<td>1.034</td>
<td>5.13E-02</td>
<td>1.027</td>
</tr>
<tr>
<td>800</td>
<td>5.13E-02</td>
<td>1.027</td>
<td>5.46E-02</td>
<td>1.092</td>
</tr>
<tr>
<td>1000</td>
<td>5.23E-02</td>
<td>1.047</td>
<td>5.09E-02</td>
<td>1.018</td>
</tr>
</tbody>
</table>

Table V-6. TIE and IR of EPS-LASSO-RV using Score and Wald test after Bonferroni correction

<table>
<thead>
<tr>
<th>$p$</th>
<th>SSLASSO</th>
<th>CEP-SKATO</th>
<th>EPS-LASSO-RV</th>
<th>EPS-LASSO-RV(Wald)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TIE</td>
<td>IR</td>
<td>TIE</td>
<td>IR</td>
</tr>
<tr>
<td>500</td>
<td>3.17E-04</td>
<td>4.221</td>
<td>2.07E-04</td>
<td>2.761</td>
</tr>
<tr>
<td>800</td>
<td>2.86E-04</td>
<td>3.811</td>
<td>1.18E-04</td>
<td>1.575</td>
</tr>
<tr>
<td>1000</td>
<td>2.85E-04</td>
<td>3.805</td>
<td>2.07E-04</td>
<td>2.757</td>
</tr>
</tbody>
</table>
Fig. V-9. Quantile-Quantile plot for null scenarios without causal predictors. Different colors indicate different methods.

In addition to the complete null scenarios, we also check the type I error of the three methods in the effective scenarios. As shown in Fig. V-10, for a raw significance level of 0.05, SSLASSO and EPS-LASSO resulted in a type I error of ~0.05, while CEP-SKATO greatly inflates the type I error to be as high as 0.25. When Bonferroni correction is applied, all the methods inflate the type I error. But EPS-LASSO-RV produces the least type I error while CEP-SKATO produces the highest. Again, the Wald statistic of EPS-LASSO-RV behaves closely to the Score statistic of EPS-LASSO-RV.
Fig. V-10. Type I error for effective scenarios. a. at a raw significance level of 0.05; b. at Bonferroni corrected significance level of 0.05.

We compare the power and FDR of EPS-LASSO-RV to SSLASSO and CEP-SKATO at the Bonferroni corrected significance level of 0.05 (Fig. V-11). CEP-SKATO results in greater power than other methods. However, similar to the inflation of type I error, their FDR are surprisingly high while the FDR of EPS-LASSO-RV is stable and low. Taking the scenario of 20% causal variants all carrying deleterious effects as an example (Fig. V-11), the FDR of CEP-SKATO increases from ~50% to ~90% with the sample size
varying from 200 to 1000. To the contrary, the FDR of EPS-LASSO-RV fluctuates around 10%. The other high-dimensional method SSLASSO yields a higher power than EPS-LASSO-RV with a higher FDR too. Provided the sample size of 200, the FDR of SSLASSO is ~20%, which is two times greater than the FDR of EPS-LASSO-RV. When the sample size is large, the power and FDR of SSLASSO and EPS-LASSO-RV performs close to each other. The Wald statistic of EPS-LASSO-RV results in a slightly greater power relative to the Score statistic, while the FDR is very close but higher than that of Score statistic. Additionally, the power of all the methods increases with the increase of the proportion of causal RV in the gene region. However, the trends of FDR vary for different methods. The FDR of high-dimensional methods EPS-LASSO-RV and SSLASSO increase for low sample size but decrease for sample sizes of 500, 800 and 1000. Conversely, the FDR of low-dimensional method CEP-SKATO is always increasing with the increase of the proportion of causal RVs. The full result containing scenarios with 40% causal RVs is shown in Table V-7.

![Image](image_url)

**Fig. V-11. Power and FDR for effective scenarios.** Different colors indicate different methods. The columns aligning to the left axis indicate power. The lines aligning to the right axis indicate FDR. The x-axis represents sample size.
Fig. V-12. Power and FDR change with the proportion of causal RV. a. CEP-SKATO; b. SSLASSO; c. EPS-LASSO-RV. The lines aligning to the left axis indicate power. The columns aligning to the right axis indicate FDR. The x-axis represents sample size.
Table V-7. Full result of effective scenarios

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>SSLASSO</th>
<th>SSLASSO_mt</th>
<th>CEPSKATO</th>
<th>CEPSKATO_mt</th>
<th>EPS-LASSO-RV</th>
<th>EPS-LASSO-RV_mt</th>
<th>EPS-LASSO-RV (Wald)</th>
<th>EPS-LASSO-RV (Wald)_mt</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>Power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.539</td>
<td>0.203</td>
<td>0.702</td>
<td>0.112</td>
<td>0.486</td>
<td>0.095</td>
<td>0.489</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.773</td>
<td>0.348</td>
<td>0.940</td>
<td>0.510</td>
<td>0.686</td>
<td>0.252</td>
<td>0.694</td>
<td>0.303</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.866</td>
<td>0.485</td>
<td>0.993</td>
<td>0.771</td>
<td>0.770</td>
<td>0.409</td>
<td>0.778</td>
<td>0.469</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.920</td>
<td>0.599</td>
<td>0.995</td>
<td>0.884</td>
<td>0.849</td>
<td>0.523</td>
<td>0.852</td>
<td>0.565</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.824</td>
<td>0.176</td>
<td>0.903</td>
<td>0.476</td>
<td>0.822</td>
<td>0.089</td>
<td>0.822</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.778</td>
<td>0.141</td>
<td>0.921</td>
<td>0.678</td>
<td>0.787</td>
<td>0.085</td>
<td>0.786</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.772</td>
<td>0.102</td>
<td>0.933</td>
<td>0.763</td>
<td>0.778</td>
<td>0.088</td>
<td>0.777</td>
<td>0.108</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.761</td>
<td>0.093</td>
<td>0.940</td>
<td>0.796</td>
<td>0.765</td>
<td>0.068</td>
<td>0.764</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>200</td>
<td>5.32E-02</td>
<td>1.59E-03</td>
<td>1.03E-01</td>
<td>2.25E-03</td>
<td>5.26E-02</td>
<td>5.21E-04</td>
<td>5.27E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>5.20E-02</td>
<td>1.49E-03</td>
<td>1.70E-01</td>
<td>2.12E-02</td>
<td>5.15E-02</td>
<td>7.08E-04</td>
<td>5.18E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>5.50E-02</td>
<td>1.25E-03</td>
<td>2.12E-01</td>
<td>4.16E-02</td>
<td>5.26E-02</td>
<td>9.92E-04</td>
<td>5.28E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>5.43E-02</td>
<td>1.32E-03</td>
<td>2.38E-01</td>
<td>5.48E-02</td>
<td>5.29E-02</td>
<td>8.41E-04</td>
<td>5.29E-02</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>Power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.654</td>
<td>0.260</td>
<td>0.776</td>
<td>0.157</td>
<td>0.533</td>
<td>0.132</td>
<td>0.543</td>
<td>0.183</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.883</td>
<td>0.493</td>
<td>0.968</td>
<td>0.623</td>
<td>0.787</td>
<td>0.357</td>
<td>0.794</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.956</td>
<td>0.664</td>
<td>0.996</td>
<td>0.846</td>
<td>0.896</td>
<td>0.576</td>
<td>0.903</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.981</td>
<td>0.763</td>
<td>1.000</td>
<td>0.940</td>
<td>0.934</td>
<td>0.690</td>
<td>0.939</td>
<td>0.723</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.803</td>
<td>0.202</td>
<td>0.904</td>
<td>0.513</td>
<td>0.801</td>
<td>0.113</td>
<td>0.801</td>
<td>0.138</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.758</td>
<td>0.079</td>
<td>0.923</td>
<td>0.710</td>
<td>0.745</td>
<td>0.070</td>
<td>0.745</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.753</td>
<td>0.089</td>
<td>0.937</td>
<td>0.786</td>
<td>0.724</td>
<td>0.053</td>
<td>0.724</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.749</td>
<td>0.075</td>
<td>0.944</td>
<td>0.823</td>
<td>0.743</td>
<td>0.053</td>
<td>0.743</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>200</td>
<td>5.33E-02</td>
<td>2.00E-03</td>
<td>1.15E-01</td>
<td>4.58E-03</td>
<td>4.81E-02</td>
<td>8.66E-04</td>
<td>4.85E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>5.17E-02</td>
<td>9.18E-04</td>
<td>1.80E-01</td>
<td>2.79E-02</td>
<td>4.52E-02</td>
<td>6.93E-04</td>
<td>4.55E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>5.36E-02</td>
<td>1.32E-03</td>
<td>2.26E-01</td>
<td>5.10E-02</td>
<td>4.58E-02</td>
<td>7.07E-04</td>
<td>4.60E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>5.37E-02</td>
<td>1.25E-03</td>
<td>2.56E-01</td>
<td>6.79E-02</td>
<td>5.11E-02</td>
<td>7.81E-04</td>
<td>5.12E-02</td>
</tr>
</tbody>
</table>

Note: columns with “mt” show result after Bonferroni correction; the expected type I error is ~7.50E-05.
VI. Application to Gene Expression Analysis

The real data is downloaded from a sub-study of Framingham Cohort project (dbGaP: phs000363)\textsuperscript{15, 16}, which includes a profiling of 17,621 genes for Framingham study offspring subjects from the Affymetrix Human Exon 1.0 ST Gene Chip platform. The gene expression values were normalized with quality control measures as previously reported\textsuperscript{72}. We pick the body mass index (BMI) as the interested trait, which is a major characteristic of obesity. Gender, age, drinking, and smoking status are considered as potential covariates. After removing the missing value in phenotypes, 972 subjects with the highest 20\% or lowest 20\% of BMI are selected as the EPS sample. The Bonferroni corrected significance level of 0.05 is used to claim the significance.

From the total 17,621 genes, EPS-LASSO identifies 10 significant genes after multiple testing correction using the significance level of $2.84 \times 10^{-6}$ (Table VI-1). Meanwhile, SSLASSO, LM, LGM, and EPS-LM find 14, 576, 468 and 600 significant genes respectively. The three low-dimensional methods result in a large number of significant findings, which may include plentiful fake candidates and need further analyses to get promising targets in practice. Eight of the EPS-LASSO significant genes are also significant in at least one of the other methods. Five of the 8 genes are well supported by previous studies on obesity or obesity-related diseases. They are \textit{MMP8}, \textit{CX3CR1}, \textit{UBE2J1}, \textit{ARL6IP1}, \textit{DAAM2}. The most significant gene \textit{MMP8} (p-value $1.64 \times 10^{-11}$), has been widely studied for its role in human obesity as early as 2009\textsuperscript{73, 74}. Polymorphisms in \textit{CX3CR1}, \textit{UBE2J1}, \textit{DAAM2} have been associated with obesity in GWAS\textsuperscript{75-77}. \textit{ARL6IP1} have been linked to the nonalcoholic fatty liver disease (NFLD), which is in close relation with obesity. There are 2 genes (\textit{TMEM56} and \textit{IRS2}) significant
in EPS-LASSO, but not significant in any of other methods. One of them – the IRS2 gene is another well-known obesity gene since 2003. IRS2 was found to be an influential gene in severe obesity and glucose intolerance in a study population. The dysregulation of IRS2 in beta cells causes obesity. As shown above, the EPS-LASSO can find true genetic effects.

Table VI-1. P-value of identifying significant genes in EPS-LASSO and other methods for the real data analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>EPS-LASSO</th>
<th>EPS-LM</th>
<th>LM</th>
<th>LGM</th>
<th>SSLASSO</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMP8</td>
<td>1.64E-11</td>
<td>1.31E-26*</td>
<td>4.22E-26*</td>
<td>1.43E-18*</td>
<td>&lt;2E-32*</td>
<td>Y</td>
</tr>
<tr>
<td>CX3CR1</td>
<td>5.01E-09</td>
<td>1.48E-10*</td>
<td>1.99E-10*</td>
<td>1.69E-09*</td>
<td>6.99E-05</td>
<td>Y</td>
</tr>
<tr>
<td>TMEM56</td>
<td>3.16E-08</td>
<td>7.39E-03</td>
<td>7.90E-03</td>
<td>3.38E-03</td>
<td>1.39E-02</td>
<td>N</td>
</tr>
<tr>
<td>IRS2</td>
<td>1.56E-07</td>
<td>2.68E-02</td>
<td>2.81E-02</td>
<td>5.09E-03</td>
<td>9.93E-03</td>
<td>Y</td>
</tr>
<tr>
<td>UBE2J1</td>
<td>2.73E-07</td>
<td>1.25E-18*</td>
<td>1.96E-18*</td>
<td>7.47E-18*</td>
<td>&lt;2E-32*</td>
<td>Y76</td>
</tr>
<tr>
<td>GCET2</td>
<td>2.87E-07</td>
<td>3.43E-11*</td>
<td>4.69E-11*</td>
<td>6.38E-09*</td>
<td>7.08E-08*</td>
<td>N</td>
</tr>
<tr>
<td>ARL6IP1</td>
<td>3.79E-07</td>
<td>1.49E-09*</td>
<td>1.99E-09*</td>
<td>1.20E-09*</td>
<td>2.84E-03</td>
<td>Y</td>
</tr>
<tr>
<td>TMEM111</td>
<td>5.88E-07</td>
<td>2.36E-12*</td>
<td>3.25E-12*</td>
<td>1.98E-09*</td>
<td>9.25E-03</td>
<td>N</td>
</tr>
<tr>
<td>DAAM2</td>
<td>1.02E-06</td>
<td>2.40E-13*</td>
<td>3.52E-13*</td>
<td>3.56E-10*</td>
<td>1.66E-06*</td>
<td>Y</td>
</tr>
<tr>
<td>TPST1</td>
<td>1.20E-06</td>
<td>2.04E-13*</td>
<td>2.92E-13*</td>
<td>2.52E-11*</td>
<td>9.32E-05</td>
<td>N</td>
</tr>
</tbody>
</table>

Note: (*) indicates significance by EPS-LM, LM, LGM, and SSLASSO.

There are 4 significant genes reported by EPS-LASSO, but we fail to find direct literature support. However, they are potentially obesity-associated genes. For example, the TMEM56 gene, which is not significant in other methods either, exhibited a highly significant difference in allelic and expression distributions between hypertensive and normal subjects. It is noted that being overweight or obese, especially when associated with increased visceral adiposity, is a major cause of hypertension. Additionally, we find TMEM56 gene was one of the top differentially expressed genes in a coronary heart disease (CHD) study also using participants from Framingham Heart Study. Obesity is an independent risk factor for cardiovascular disease, including CHD. Given the
association of $TMEM56$ gene with two diseases may be caused by obesity, $TMEM56$ is well worth further investigation of its role in obesity.
VII. Application to Rare Variant Analysis

We apply SSLASSO, CEP-SKATO, and EPS-LASSO-RV to analyze the effects of the rare variants on BMI using the EPS sample made from the Framingham Heart Study (see Chapter IV). Shown in Table VII-1, the two high-dimensional methods SSLASSO and EPS-LASSO-RV identify 12 significant genes carrying RV effects at the significance level of 0.05, which is more than the significant genes identified by CEP-SKATO. Unfortunately, no genes reach the Bonferroni corrected significance level by any methods, which may due to the small sample size and a small number of genes considered. SSLASSO and EPS-LASSO-RV did find a significant factor – gender, for BMI after Bonferroni correction. On the other hand, the covariates in CEP-SKATO are adjusted before the test of genetic factors. The significance of gender by CEP-SKATO could not be concluded directly, which shows another advantage of our high-dimensional regression model in covariate adjustment.

Under the same framework of EPS-LASSO-RV, Score test and Wald test result in different significant genes. In fact, compared with 12 significant genes from Wald test, Score test only find 3 genes with P-value < 0.05, which were \( CYP7A1 \), \( RPL36A-HNRNPH2 \), and \( WNK4 \). They are all significant in Wald test with a less P-value (Table VII-1). \( CYP7A1 \) and \( WNK4 \) are also significant by SSLASSO and CEP-SKATO, while \( RPL36A-HNRNPH2 \) is significant by SSLASSO but not by CEP-SKATO. \( CYP7A1 \) is a protein-coding gene encoding the enzyme Cholesterol 7 alpha-hydroxylase. Contained in Metabolic and Statin pathway, \( CYP7A1 \) was shown to play a critical role in maintaining whole body lipid, glucose, and energy homeostasis\(^8\). The induction of \( CYP7A1 \) expression under certain conditions could be used for treating obesity in humans. \( WNK4 \)
is a gene encoding a member of the WNK family of serine-threonine protein kinases. A recent study found WNK4 is an adipogenic factor and deleting it reduces diet-induced obesity in mice\textsuperscript{84}. However, the relation between \textit{RPL36A-HNRNPH2} gene and obesity is not well studied.

Table VII-1. Significant genes and their p-value identified

<table>
<thead>
<tr>
<th>SSLASSO</th>
<th>EPS-LASSO-RV</th>
<th>Score</th>
<th>Wald</th>
<th>CEP-SKATO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCG5</td>
<td>1.8E-02</td>
<td>ABCG5</td>
<td>6.05E-02</td>
<td>BBS4</td>
</tr>
<tr>
<td>BBS4</td>
<td>4.17E-02</td>
<td>BBS4</td>
<td>5.69E-02</td>
<td>CYP7A1</td>
</tr>
<tr>
<td>CYP11B2</td>
<td>3.79E-02</td>
<td>CYP7A1</td>
<td>1.77E-02</td>
<td>LIG3</td>
</tr>
<tr>
<td>CYP7A1</td>
<td>1.53E-02</td>
<td>GLA</td>
<td>7.44E-02</td>
<td>NCAPH2</td>
</tr>
<tr>
<td>FOXO1</td>
<td>1.17E-02</td>
<td>LTBR</td>
<td>7.42E-02</td>
<td>PPARG</td>
</tr>
<tr>
<td>NR3C1</td>
<td>3.46E-02</td>
<td>NCAPH2</td>
<td>8.16E-02</td>
<td>WNK4</td>
</tr>
<tr>
<td>POMC</td>
<td>4.39E-02</td>
<td>POMC</td>
<td>7.20E-02</td>
<td></td>
</tr>
<tr>
<td>PPARG</td>
<td>2.27E-02</td>
<td>RPL36A-HNRNPH2</td>
<td>4.36E-02</td>
<td>2.21E-02</td>
</tr>
<tr>
<td>PSRC1</td>
<td>4.10E-02</td>
<td>SCNN1A</td>
<td>7.84E-02</td>
<td></td>
</tr>
<tr>
<td>RPL36A-HNRNPH2</td>
<td>7.07E-03</td>
<td>gender</td>
<td>9.58E-05</td>
<td>1.42E-04</td>
</tr>
<tr>
<td>SCN4B</td>
<td>3.00E-02</td>
<td>SIM1</td>
<td>7.24E-02</td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>0.00E+00</td>
<td>TMEM67</td>
<td>5.14E-02</td>
<td></td>
</tr>
<tr>
<td>WNK4</td>
<td>6.04E-03</td>
<td>WNK4</td>
<td>1.24E-02</td>
<td>4.43E-03</td>
</tr>
</tbody>
</table>

Note: each group ranked by gene name.
VIII. Discussion

Based on the decorrelated score function, we develop Score test (default) and Wald test under the high-dimensional regression model for EPS genetic data. Both the simulation and real data analysis for EPS-LASSO and EPS-LASSO-RV show that the Wald test is slightly liberal relative to the Score test. In practice, the choice of the two tests will depend on the demands and specific situations. In addition to the Score and Wald test, the likelihood ratio test is also shown can be applied in the high-dimensional regression\(^4\), which needs more theoretical and simulation study for its application in EPS-LASSO.

Our method is motivated by the general theory of hypothesis test for high-dimensional models, which answers the question by dealing with the score statistic in high-dimension. There is another commonly used technique that decomposes the estimate of regression coefficients into a bias term and a normally distributed term. Then a hypothesis testing procedure can be developed based on the normal term. Although we use de-biased estimate for testing in our method, this technique is different. It literally incorporates the de-bias process into the hypothesis test. SSLASSO is a typical example derived based on this technique and shows higher power than EPS-LASSO for high-dimensional settings in our simulation result. Although the statistical properties of this technique remain unknown under extreme sampling, it is another promising approach to detect genetic associations under EPS.

In practice, the data type and dimension in the genetic study is different by the research target and platform. Here, the straightforward application of our method in gene expression analysis shows the feasibility of analyzing several thousands of continuous
Determined by the practical computing capability, dimension reduction is still necessary for candidate genetic factors in millions, such as the genome-wide, epigenome-wide, and metagenome-wide association study. A frequently used method is region based analysis by collapsing effects or hierarchical modeling.

Our current work on RV analysis is an example of collapsing effects. However, there are some weaknesses in the collapsing methods. For example, when there are both protective and deleterious RV effects in the region, the power of EPS-LASSO-RV will not be as competent as CEP-SKATO, which is based on a mixed model testing the variance of the regression coefficients. The collapsing method we used would cancel out the bi-directional effects to some degree. This is a limitation of the current model and other existing uni-directional collapsing methods. To amend this caveat, a sign assignment procedure proposed by Hoffman et al. could be used in the future work, where an estimated direction of a variant can be added in the weighted sum collapsing measure. Other high-order information collapsing method and structured sparse regression methods may also be explored, like the group LASSO.

Re-use of the data is an issue when the same data are used multiple times for model selection and further statistical inferences. Our method contains two model selection steps and one hypothesis testing step, which has the problem of using the data multiple times. Using the same data, the statistical inference/test based on the selected model may be misleading. But in certain applications, re-use of data may only explain little on the deficient coverage of confidence intervals after model selection. In view of the potential issues, we further examine the impact of using data multiple times in our method. We did more simulations using 3 different data sets for tuning parameters $\lambda$, $\lambda'$,
and the hypothesis testing. Similar to Kabaila’s latest finding (Kabaila and Mainzer, 2017), re-use of data in our method has little effect on the performance of power and error control (Appendix X.IV).

When we do the real data analysis, we try to search the natural dataset with extreme phenotype sampling design in the largest public transcriptome data repository – Gene Expression Omnibus database. Because most of the datasets from EPS design were analyzed as categorical data, the quantitative trait information is not provided. Inspired by the strategy used in another EPS methodology article, we select samples with extreme traits from the widely-used Framingham Cohort dataset. The major goal of our real data analysis is to evaluate the performance of EPS-LASSO and compare it with existing methods for samples with extreme sampling. In practice, we can certainly apply the proposed method to natural datasets with extreme phenotype sampling design, where the truncated quantitative trait info is complete.

Additionally, using the full dataset, we look into the significant genes identified by EPS-LASSO (in Table VI-1). The ordinary linear regression is applied to analyze the full dataset (LM-FULL) with the same covariates adjusted. In Table VIII-1, for genes significant in both EPS-LASSO and LM-FULL, LM-FULL produces a lower p-value. The two genes significant in EPS-LASSO but not in other methods fail to gain significance in LM-FULL too. It shows EPS can boost the power as concluded in a previous study in some situations, where samples in the middle of the distribution may introduce more noises than signals.
Table VIII-1. P-value of identifying significant genes in EPS-LASSO and other methods for the real data analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>EPS-LASSO</th>
<th>EPS-LM</th>
<th>LM</th>
<th>LGM</th>
<th>SSLASSO</th>
<th>Literature</th>
<th>LM-FULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMP8</td>
<td>1.64E-11</td>
<td>1.31E-26*</td>
<td>4.22E-26*</td>
<td>1.43E-18*</td>
<td>&lt;2E-32*</td>
<td>Y</td>
<td>3.94E-29*</td>
</tr>
<tr>
<td>CX3CR1</td>
<td>5.01E-09</td>
<td>1.48E-10*</td>
<td>1.99E-10*</td>
<td>1.69E-09*</td>
<td>6.99E-05</td>
<td>Y</td>
<td>6.52E-11*</td>
</tr>
<tr>
<td>TMEM56</td>
<td>3.16E-08</td>
<td>7.39E-03</td>
<td>7.90E-03</td>
<td>3.38E-03</td>
<td>1.39E-02</td>
<td>N</td>
<td>5.28E-03</td>
</tr>
<tr>
<td>IRS2</td>
<td>1.56E-07</td>
<td>2.68E-02</td>
<td>2.81E-02</td>
<td>5.09E-03</td>
<td>9.93E-03</td>
<td>Y</td>
<td>7.56E-02</td>
</tr>
<tr>
<td>UBE2J1</td>
<td>2.73E-07</td>
<td>1.25E-18*</td>
<td>1.96E-18*</td>
<td>7.47E-18*</td>
<td>&lt;2E-32*</td>
<td>Y</td>
<td>4.55E-21*</td>
</tr>
<tr>
<td>GCET2</td>
<td>2.87E-07</td>
<td>3.43E-11*</td>
<td>4.69E-11*</td>
<td>6.38E-09*</td>
<td>7.08E-08*</td>
<td>N</td>
<td>5.03E-13*</td>
</tr>
<tr>
<td>ARL6IP1</td>
<td>3.79E-07</td>
<td>1.49E-09*</td>
<td>1.99E-09*</td>
<td>1.20E-09*</td>
<td>2.84E-03</td>
<td>Y</td>
<td>1.49E-10*</td>
</tr>
<tr>
<td>TMEM111</td>
<td>5.88E-07</td>
<td>2.36E-12*</td>
<td>3.25E-12*</td>
<td>1.98E-09*</td>
<td>9.25E-03</td>
<td>N</td>
<td>4.86E-11*</td>
</tr>
<tr>
<td>DAAM2</td>
<td>1.02E-06</td>
<td>2.40E-13*</td>
<td>3.52E-13*</td>
<td>3.56E-10*</td>
<td>1.66E-06*</td>
<td>Y</td>
<td>5.49E-16*</td>
</tr>
<tr>
<td>TPST1</td>
<td>1.20E-06</td>
<td>2.04E-13*</td>
<td>2.92E-13*</td>
<td>2.52E-11*</td>
<td>9.32E-05</td>
<td>N</td>
<td>1.84E-12*</td>
</tr>
</tbody>
</table>

Similar to the type I error inflation of low-dimensional methods for general genetic data analysis, we observe that the type I error of CEP-SKATO inflates in the simulation. The phenomenon is different from the previous result\(^{21}\). We speculate that the correlation or similarity among the genes/regions may induce the incorrect type I error of CEP-SKATO. Literally, the test statistic of CEP-SKATO is an aggregation of the individual score statistic for testing the marginal effect of each marker within a region\(^{64}\), which does not consider the correlation among different regions in the whole genome. Other existing individual region based RV detection methods may also have this problem. In contrast, EPS-LASSO-RV is testing the regions simultaneously and can address the possible correlation among regions, which is an advantage in real data analysis for the complex diseases caused by multiple genetic factors.

As has been noted, the advantage of EPS-LASSO-RV is the control of type I error and lower level of FDR, while the weakness is the slightly lower power relative to other methods. The selection of EPS-LASSO-RV Score, Wald test, and other methods in a real application will depend on the specific objectives. For an exploratory study, where a liberal significance threshold is usually used, CEP-SKATO or EPS-LASSO-RV Wald test is preferred because they are powerful in including as many as possible candidate
targets. For a refined study, where the error is barely tolerated, EPS-LASSO-RV Score test is favored. Moreover, investigators may also adopt an aggressive strategy by combining the results from all the methods.

In the end, we find several potential developments may be interesting for future exploration with EPS-LASSO. First is the dimensional reduced EPS-LASSO with the aid of initial screening. In penalized model selection, the feature screening is removing the features with little chance (usually approximate to 0) to carry effects before the analysis. There have been several studies on the feature screening, including the sure screening (SS) given various assumptions based on marginal Pearson correlation or distance correlation. The SS may help in elevating power by reducing the number of predictors considered, so does the burden of multiple testing. We did a rough experiment on the screening aided EPS-LASSO by selecting non-zero predictors using the Algorithm 1 of EPS-LASSO (Appendix X.V). The result shows EPS-LASSO with initial SS is an appealing direction, especially for the ultrahigh-dimensional data.

Second, in this study, the standard Bonferroni correction is applied to address the multiple testing problem. We evaluate the corresponding FDR for all the methods. Taking the general genetic data analysis as an example, the FDR of EPS-LASSO across different scenarios varies from ~2% to ~5%, while the range for other methods is between ~1% and ~7%. Although EPS-LASSO performs better than other methods when the predictors are correlated, in fact, the FDR is not considered in the model. The direct application of Bonferroni correction may result in a power loss in detecting effects, such as the RV effects. Be aware of the latest application of an FDR controlled penalized regression model – SLOPE LASSO – in genetic variants under random sampling.
development of an FDR controlled LASSO under EPS is another direction worth pursuing.
IX. Conclusion

In this study, we have proposed a novel sparse penalized regression model for the continuous trait under extreme phenotype sampling. Based on the decorrelated score function. We develop a hypothesis testing using Score and Wald statistic for the proposed sparse penalized regression model. Through simulation and real data analysis, we evaluate the performance of the sparse regression model for general genetic data analysis and extend it to the rare variant analysis.

For the general genetic data analysis, the comprehensive simulation indicates the model (EPS-LASSO) can control the type I error correctly. The FDR of EPS-LASSO is more stable and robust than existing methods. Especially when the predictors are correlated, EPS-LASSO outperforms other existing EPS methods by lower type I error and FDR. In addition, EPS-LASSO can provide a persistent power for both low- and high-dimensional situations compared with the other high-dimensional method. The power of EPS-LASSO is close to other low-dimensional methods for small effect size and is superior to them when the causal effect is large. EPS-LASSO manages to identify significant BMI associated genes supported by existing studies. Similarly, for the region based rare variants (RV) data analysis, EPS-LASSO-RV collapsing RV effects also yield correct type I error control. The FDR of EPS-LASSO-RV is more stable and lower than existing methods. The power of EPS-LASSO-RV is not superior to other methods but is tightly close. In the real data analysis, EPS-LASSO-RV successfully detects genetic effects with literature support.
Overall, with the study on general genetic data analysis and rare variant analysis, EPS-LASSO is an advanced method for high-dimensional data analysis under EPS, which can account for correlated predictors.
X. Appendix

X.I The selection of criterion for tuning parameter

The information criterion (IC) defined as

\[ IC = -2 \times \log(\text{likelihood}) + k \times \lambda_c, \]

where \( k \) is the number of estimated parameters in the model, is a commonly used method for model selection. With different \( \lambda_c \), there are different types of ICs, such as the classical Bayesian Information Criterion (BIC)\(^{50}\), and various high dimensional information criteria (HDIC).

We considered the cross-validation\(^53\), BIC and another two typically used HDIC\(^{51,52}\):

- **BIC**: \( \lambda_c = \log n \)
- **HDIC-1**: \( \lambda_c = p^{1/3} \)
- **HDIC-2**: \( \lambda_c = 2 \log p \)

where \( n \) is the sample size, \( p \) is the total number of parameters. Using the simulation design of \( p=400 \), correlation among predictors \( \rho = 0 \), causal effect size from 0.2, 0.25 and 0.3 (see Method), we compared their power and false discovery rate (FDR) in Fig. X-1. The FDR of cross-validation is severely uncontrolled (~0.6) compared to BIC and HDICs. BIC produced higher power than the two HDICs with a relative high FDR, but still under control (<0.03). The result also agrees with previous studies\(^{65,66}\). Consequently, we picked BIC to select the optimum \( \lambda \).
Fig. X-1. Power and FDR of EPS-LASSO using 4 different criteria. The columns aligned to the left vertical axis is the power. The lines aligned to the right vertical axis is the FDR. The horizontal axis is the effect size.
X.II Refitted estimator vs non-refitted estimator

Motivated by the LARS-OLS hybrid strategy and numerical experience from previous study, we used a refitted estimator to reduce the bias of the LASSO type estimator. We compared the type I error and power of EPS-LASSO using the refitted or non-refitted estimator. Under the null model scenarios, their type I error rates are both close to the desirable level (Table X-1). When causal predictors are considered with Bonferroni correction, the refitted estimator results in a higher power than the non-refitted estimator (Fig. X-2).

![Fig. X-2. Power of EPS-LASSO using refitted and non-refitted estimator](image)

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$p$</th>
<th>Refitted estimator $\alpha=0.05$</th>
<th>Refitted estimator $\alpha=0.05/p$</th>
<th>Non-refitted estimator $\alpha=0.05$</th>
<th>Non-refitted estimator $\alpha=0.05/p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>1.00</td>
<td>1.12</td>
<td>1.04</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1.00</td>
<td>0.68</td>
<td>0.98</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.01</td>
<td>1.32</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.01</td>
<td>1.28</td>
<td>0.99</td>
<td>0.68</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>1.02</td>
<td>1.04</td>
<td>1.02</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.99</td>
<td>0.72</td>
<td>0.98</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.99</td>
<td>0.88</td>
<td>0.99</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.99</td>
<td>1.08</td>
<td>0.98</td>
<td>0.68</td>
</tr>
</tbody>
</table>
X.III Proof of the asymptotic distribution of the Score and Wald statistics

The proof of the limiting distribution is based on the Ning and Liu’s general results for the high dimensional generalized linear model, specifically the Assumption E.1 and Theorem E.1 in \(^{43}\). We list them below as a quick reference.

Given a generalized linear model with the negative log-likelihood can be expressed as:

\[
\ell(\beta, \beta_{-j}) = -\frac{1}{n} \sum_{i=1}^{n} \left( Y_i (\beta_{ij} X_{ij} + \beta^T_{-j} X_{i,-j}) - b(\beta_{ij} X_{ij} + \beta^T_{-j} X_{i,-j}) \right),
\]

where \( b(\cdot) \) are known functions, then we have the following assumption and theorem.

Notation: \( \beta^*, w^*, I^* \) indicates the true value for \( \beta, w, I \). \( s^* = \|\beta_{-j}^*\|_0 \) and \( s' = \|w^*\|_0 \).

\( \lambda_{\text{min}}(M) \) represents the minimal eigenvalue of the matrix \( M \).

Assumption E.1. Assume (1) \( \lambda_{\text{min}}(I^*) \geq \kappa^2 \) for some constant \( \kappa > 0 \); (2) \( S = \text{supp}(\beta^*) \) and \( S' = \text{supp}(w^*) \) satisfy \( |S| = s^* \) and \( |S'| = s' \); (3) \( \|X_i\|_\infty \leq K \), \( |w^T X_{i,-j} \leq K| \), for some constant \( K \), and \( |Y_i - b'(X_i^T \beta^*)| \) is sub-exponential, (4) there exists an interval \( [K_1, K_2] \) such that \( X_i^T \beta^* \in [K_1, K_2] \). In addition, for any \( t \in [K_1 - \epsilon, K_2 + \epsilon] \) with some constant \( \epsilon > 0 \) and a sequence \( t_1 \) satisfying \( |t_1 - t| = o(1) \), it holds that \( 0 < b''(t) \leq C \) and \( |b''(t_1) - b''(t)| \leq C|t_1 - t|b''(t) \) for some constant \( C > 0 \).

Theorem E.1. Assume that Assumption E.1 holds. With \( \lambda = \sqrt{\frac{\log p}{n}} \) and \( \lambda' = \sqrt{\frac{\log p}{n}} \), if \( \sqrt{n}(s' \vee s^*) \log p = o(1) \), then under the null hypothesis, for each \( t \in \mathbb{R} \),
\[ \lim_{n \to \infty} \left| \mathbb{P}_{\theta'}(\bar{U}_n \leq t) - \Phi(t) \right| = 0, \text{ and } \sqrt{n}(\bar{\beta}_j - \beta_j^*) \xrightarrow{d} \mathcal{N}(0,1), \text{ where } \bar{U}_n = \sqrt{n} \hat{S}_{\beta_j}^{-1/2} \bar{\beta}_j - \hat{S}_{\beta_j}^{-1} \beta_j - \hat{S}_{\beta_j}^{-1} \beta_j. \]

The negative log-likelihood for our model is:

\[ \ell(\beta_j, \beta_{-j}) = -\log L = \frac{n}{2} \log(2\pi \sigma^2) + \frac{\sum (y_i - x_i' \beta)^2}{2\sigma^2} + \sum \log(\Phi(c_2, x_i' \beta, \sigma^2) + 1 - \Phi(c_1, x_i' \beta, \sigma^2)). \]

It can be reformulated as following by removing some nuisance term:

\[ \ell(\beta_j, \beta_{-j}) = -\frac{1}{n} \sum_{i=1}^n \left\{ Y_i (\beta_j X_{i,j} + \beta_{-j}^T X_{i,-j}) - b(\beta_j X_{i,j} + \beta_{-j}^T X_{i,-j}) \right\} \]

with \( b(x_i' \beta) = (x_i' \beta)^2 + 2\sigma^2 \log(\Phi(c_2, x_i' \beta, \sigma^2) + 1 - \Phi(c_1, x_i' \beta, \sigma^2)) \).

We directly assume conditions (1), (2) and (3) in Assumptions E.1, which are also assumed by [6,7]. In fact, the first 3 conditions are basic assumptions. The condition (1) is implied by an independent score function vector. The condition (2) is the consistency of the support of the Lasso type estimator. The condition (3) is assuming \( X \) is bounded and the residual is sub-exponential. For the condition (4), given the \( b(x_i' \beta) \) and

\[ b'' = 2X_{i,j}^2 (1 - \frac{\Phi(c_2 - x_i' \beta, \sigma^2) - \Phi(c_1 - x_i' \beta, \sigma^2)}{\Phi(c_2, x_i' \beta, \sigma^2) + 1 - \Phi(c_1, x_i' \beta, \sigma^2)} - m_i^2), \]

with \( m_i = \frac{\Phi(c_2 - x_i' \beta, \sigma^2) - \Phi(c_1 - x_i' \beta, \sigma^2)}{\Phi(c_2, x_i' \beta, \sigma^2) + 1 - \Phi(c_1, x_i' \beta, \sigma^2)}. \)
if we take \( K_1 \) and \( K_2 \) to be infinity and \( t \in [K_1 - \epsilon, K_2 + \epsilon] \), then \( b''(t) = 2X_{t,j}^2 \). The condition (4) holds for some constant \( C > 0 \).

Consequently, by applying the Theorem E.1, we have the \( \sqrt{n} \hat{\xi}_{j}^{*} \hat{l}_{j}^{-1/2} \overset{d}{\to} N(0,1) \) and

\[
\sqrt{n}(\hat{\beta}_j - \beta_j^*) \hat{l}_{j|\beta_{-j}}^{1/2} \overset{d}{\to} N(0,1),
\]

where \( \hat{\beta}_j = \hat{\beta}_j - \xi_{j|\beta_{-j}}^{*} / \hat{l}_{j|\beta_{-j}} \).
X.IV Using same and using different data

Using the same data, the statistical inference/test based on the selected model can be misleading (Chatfield, 1995; Kabaila and Giri, 2009). In view of the potential issues, we do more simulations to examine the impact of using data multiple times in our method. In the manuscript, we use the same data to find the tuning parameters $\lambda$ and $\lambda'$, then do the hypothesis testing. Therefore, in the new simulations, we use 3 different data sets each having the same sample size as the re-used dataset for tuning parameters $\lambda$, $\lambda'$, and the hypothesis testing. Under the null model with rho=0 and 0.4, the type I error of re-using same data is close to that of using different data (Table X-2). Under the causal effects model with an effect size of 0.2, rho=0 and 0.4, using same data behaves similarly to using different data in power and FDR (Fig. X-3). Consequently, the re-use of data in our method has little effect on the performance of power and error control, which is similar to the Kabaila’s latest finding that re-use of data explains little on the deficient coverage of confidence intervals after model selection (Kabaila and Mainzer, 2017).

<table>
<thead>
<tr>
<th>$\rho$(rho)</th>
<th>$\rho$</th>
<th>Using same</th>
<th>Using different</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha=0.05$</td>
<td>$\alpha=0.05/\rho$</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>1.00</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.01</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.01</td>
<td>1.28</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>1.02</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.99</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.99</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.99</td>
<td>1.08</td>
</tr>
</tbody>
</table>
Fig. X-3. Power and FDR of EPS-LASSO using same vs using different data set. Line aligned to the left axis represents power. Histogram aligned to the right axis represents FDR. The horizontal axis is the number of predictors.
X.V Screening aided EPS-LASSO

An initial screening followed by a dimensional reduced EPS-LASSO may increase the performance of our model. Especially for the power of analyzing ultrahigh-dimensional data. For example, the SNP data in millions, the power may be improved by reducing the burden of multiple testing. The feature screening in penalized model selection has been widely studied, including the sure screening under multiple model assumptions via marginal Pearson correlation or distance correlation. The sure screening means all true predictors can be selected with probability approaching one as the sample size diverges to infinity. It is important for the error control of the analysis after screening. However, the sure screening property of applying these methods for EPS-LASSO is unclear.

Based on the Algorithm 1 of EPS-LASSO, we did more simulations to compare EPS-LASSO and the dimensional reduced EPS-LASSO with the aid of initial screening (EPS-LASSO-Sr) using the scenarios of 400 and 800 predictors. In EPS-LASSO-Sr, the Algorithm 1 of EPS-LASSO is employed to select non-zero predictors. Then the dimension reduced dataset is applied for the hypothesis testing with the de-biased estimate. Under the null model, EPS-LASSO-Sr can still yield correct type I error (Table X-3). After Bonferroni correction for the number of predictors after screening ($p^*$), the power of EPS-LASSO-Sr is elevated while the FDR is increased too (Fig. X-4). If using $p$ in Bonferroni correction, the performance of EPS-LASSO-Sr is tightly close to EPS-LASSO. Consequently, the improved power of EPS-LASSO-Sr is mostly owing to the reduced dimensionality after the initial screening.
Given these points and a lack of study on the sure screening property under EPS, we consider EPS-LASSO with sure screening as a promising future direction, especially for the ultrahigh-dimensional data.

Table X-3. Type I error inflation rate of EPS-LASSO vs EPS-LASSO-Sr

<table>
<thead>
<tr>
<th>$\rho$ (rho)</th>
<th>$p$</th>
<th>EPS-LASSO $\alpha=0.05$</th>
<th>EPS-LASSO $\alpha=0.05/p^*$</th>
<th>EPS-LASSO-Sr $\alpha=0.05$</th>
<th>EPS-LASSO-Sr $\alpha=0.05/p^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>400</td>
<td>1.01</td>
<td>1.28</td>
<td>1.02</td>
<td>1.28</td>
</tr>
<tr>
<td>800</td>
<td></td>
<td>1.01</td>
<td>1.28</td>
<td>1.01</td>
<td>0.68</td>
</tr>
<tr>
<td>0.4</td>
<td>400</td>
<td>0.99</td>
<td>0.88</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>800</td>
<td></td>
<td>0.99</td>
<td>1.08</td>
<td>1.00</td>
<td>0.92</td>
</tr>
</tbody>
</table>

*Inflation rate = (type I error)/(significance level); $p^*$: number of predictors after screening

Fig. X-4. Power and FDR of EPS-LASSO and EPS-LASSO-Sr. Line aligned to the left axis represents power. Histogram aligned to the right axis represents FDR. The horizontal axis indicates $\rho$ (rho): 0 and 0.4. Number of predictors $p=400$. 
References


variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91 (2):224-237


46. Cox DR and Hinkley DV (1979) Theoretical Statistics. CRC Press, Boca Raton, FL


65. Chen J and Chen Z (2012) EXTENDED BIC FOR SMALL-n-LARGE-P SPARSE GLM. Statistica Sinica 22 (2):555-574


susceptibility genes in Han Chinese with a genome-wide gene-based association study. PLoS One 7 (3):e32907


87. Kabaila P and Giri K (2009) UPPER BOUNDS ON THE MINIMUM COVERAGE PROBABILITY OF CONFIDENCE INTERVALS IN


