Statistica Sinica **32** (2022), 613-633
doi:https://doi.org/10.5705/ss.202019.0408

HYPOTHESIS TESTING IN HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES REGRESSION WITH AN APPLICATION TO GENOMICS DATA

Jiarui Lu and Hongzhe Li

University of Pennsylvania

Abstract: Gene expression and phenotype association can be affected by potential unmeasured confounders from multiple sources, leading to biased estimates of the associations. Because genetic variants largely explain gene expression variations, they can be used as instrumental variables (IVs) when studying the association between gene expressions and phenotypes in a high-dimensional IV regression framework. Because the dimensions of both genetic variants and gene expressions are often larger than the sample size, statistical inferences (e.g., hypothesis testing) for such high-dimensional IV models are not trivial, and have not been investigated in the literature. The problem is made more challenging because the IVs (e.g., genetic variants) have to be selected from a large set of genetic variants. This study considers the problem of hypothesis testing for sparse IV regression models, and presents methods for testing a single regression coefficient and for multiple testing of multiple coefficients, where the test statistic for each single coefficient is constructed based on an inverse regression. A multiple testing procedure is developed for selecting variables, and is shown to control the false discovery rate. Simulations are conducted to evaluate the performance of our proposed methods. Lastly, we apply the proposed methods by analyzing a yeast data set in order to identify genes that are associated with growth in the presence of hydrogen peroxide.

Key words and phrases: Debiased estimation, FDR control, genetical genomics, inverse regression, multiple testing.

1. Introduction

Many genomic studies collect both germline genetic variants and tissue-specific gene expression data on the same set of individuals in order to understand how genetic variants perturb gene expressions that lead to clinical phenotypes. Here, popular methods include association analyses between gene expressions and phenotypes such as the differential gene expression analysis. Such studies have shown that gene expressions are associated with many common human diseases, such as liver disease (Romeo et al. (2008); Speliotes et al. (2011)) and heart fail-

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

ure (Liu et al. (2015)). However, there may be many unmeasured factors that affect both gene expressions and phenotypes of interest (Leek and Storey (2007); Hoggart et al. (2003)). Unmeasured confounding variables can cause a correlation between the error term and one or more of the independent variables, causing us to identify false associations. In particular, the independence assumption between gene expressions and errors is required in a linear regression in order to obtain a valid statistical inference of the effects of gene expressions on phenotypes. If this assumption is violated, standard methods can lead to biased estimates (Lin, Feng and Li (2015); Fan and Liao (2014)).

One way to deal with unmeasured confounding is to apply an instrumental variable (IV) regression, which has been studied extensively in low-dimensional settings (Imbens (2014)). In our applications, we treat genetic variants as IVs when studying the association between gene expressions and phenotypes. The standard method used to fit IV models involves applying two-stage regressions to obtain valid estimations of the true parameters. Using genetic variants as IVs has attracted much interest, because these variants can be considered as randomly assigned to individuals, owing to Mendelian segregation. In genetics, owing to the availability of large-scale genetics data, Mendelian randomization has been proposed for investigating the causal relationship between two variables. Kang et al. (2016) studied the problem of IV estimation by allowing for some invalid instruments. Zhao et al. (2019) considered the statistical inference of Mendelian randomization using summary statistics from two genome-wide association studies, focusing on how to deal with bias in the causal estimate due to weak or invalid instruments.

While the methods of Kang et al. (2016) and Zhao et al. (2019) can be applied to study the effect of one given gene on the response, it is also important to consider many genes jointly in high-dimensional IV regressions. In such models, the dimensions of the genetic variants and gene expressions are much larger than their respective sample sizes, making the classic two-stage regression methods of fitting IV models infeasible. To account for high dimensionality, penalized regression methods have been developed that select the instruments in the first stage, and then select gene expressions in the second stage (Lin, Feng and Li (2015)). Lin, Feng and Li (2015) provided the estimation error bounds of the proposed two-stage estimators, but did not study the related problem of statistical inference.

Here, we presents hypothesis testing methods for high-dimensional IV models, including a statistical test of a single regression coefficient and a false discovery rate (FDR) controlling method for simultaneously testing each of the

coefficients. For linear regression models in high-dimensional settings, Javanmard and Montanari (2014) developed a debiased procedure to construct an asymptotically normally distributed estimator based on the original biased Lasso estimator. The asymptotic results can be used for hypothesis testing. Zhang and Zhang (2014) proposed a low-dimensional projection estimator to correct the bias, sharing a similar idea to that of Javanmard and Montanari (2014). In a more general framework, Ning and Liu (2017) considered the hypothesis testing problem for the general penalized M-estimator, constructing a decorrelated score statistic in a high-dimensional setting. All these methods for high-dimensional linear regression inferences require the critical assumption that the error terms are independent of the covariates, and therefore cannot be applied to IV models directly.

Our proposed inference methods build on the work of Lin, Feng and Li (2015) to obtain consistent estimators of the regression coefficients, and on the work of Liu (2013) to construct the bias-corrected test statistics using an inverse regression. Inverse regressions were first used to study the Gaussian graphical model, and have been extended to perform hypothesis testing in high-dimensional linear regression models (Liu and Luo (2014)). The procedure uses information from the precision matrix to quantify the correlations between the test statistics. We combine this inverse regression procedure with the estimation methods in Lin, Feng and Li (2015) to propose a test statistic with desired properties. In addition, in a high-dimensional setting, the sparsity assumption on the true regression coefficient results in a small number of alternatives, which can lead to conservative FDR control. A less conservative approach is to control the number of falsely discovered variables (FDV) (Liu and Luo (2014)). The proposed test statistic for a single regression coefficient in IV models is shown to be asymptotically normal, and the proposed multiple testing procedure is shown to control the FDR or FDV.

The proposed two-stage regression can be used to identify gene expressions that cause diseases by jointly analyzing genotype and gene expression data. This is similar in spirit to transcriptome-wide association studies (TWAS) (Wainberg et al. (2019); Gamazon et al. (2015)) that aim to identify the molecular mechanisms through which genetic variations affect phenotypes. Most TWAS are performed based on two independent data sets, where the expression panel or reference eQTL data are used to learn per-gene predictive models of expression level. These models are used to predict a gene expression for each individual in a separate genome-wide association study (GWAS) data set. Finally, statistical associations are tested between the predicted gene expression and the trait

(Wainberg et al. (2019)). Such analyses can also be performed using summary statistics (Gusev et al. (2016); Barbeira et al. (2018)). Another related topic is the two-sample Mendelian randomization (MR) from gene expression to trait, using genetic variants as possible IVs (Zhu et al. (2016, 2018); Sanderson et al. (2018)). Most of these MR methods are based on two sets of GWAS summary statistics, and are performed for one gene at a time. In contrast, our method focuses on identifying possible causal genes by considering all genes jointly when trait, genotype, and gene expression data are measured on the same set of individuals. Our method allows multiple causal genes for a given trait, some of which can be associated with the same set of genetic variants through trans-regulation. In this scenario, the key assumption of an absence of pleiotropy in a single-gene MR analysis is violated, which can lead to false causal association (Wainberg et al. (2019)). Finally, our proposed l_1 penalized estimation method allows us to select the highest number of possible causal genes from among those that are highly correlated.

The remainder of the paper is organized as follows. Section 2 presents the high-dimensional IV model, the test statistics for single hypothesis, and a multiple testing procedure that controls the FDR or FDV. Section 3 provides the theoretical results of the single coefficient test statistic and the multiple testing procedure. Simulation results are presented in Section 4. An analysis of a yeast data set using the proposed methods is given in Section 5. A discussion and suggestions for future work are provided in Section 6. Proofs of the theorems are included as online Supplementary Material.

2. IV Models and Proposed Methodology

We first introduce the notation used in the paper. For any set S, $|S|$ denotes its cardinality. For a vector x, $\text{supp}(x)$ is its support, $\|x\|_p$ is the standard ℓ_p-norm, and $\|x\|_0$ is defined as $|\text{supp}(x)|$. For any matrix $A = (a_{ij})$, for $i \in I$ and $j \in J$, and subsets $S \subset I$ and $R \subset J$, $A_{S,R}$ denotes the submatrix $\{(a_{ij}) : i \in S, j \in R\}$, and $A_{-S,-R}$ denotes the submatrix $\{(a_{ij}) : i \notin S, j \notin R\}$. For a matrix A, $A_{\cdot,j}$ represents the jth column of this matrix. For a sequence of random variables x_n and a random variable x, $x_n \rightsquigarrow x$ implies x_n converges weakly to x as $n \to \infty$. Finally, $a \wedge b$ represents the minimum value between a and b, and $a \lesssim b$ if there exists some constant C such that $a \leq Cb$, and $a \lesssim_p b$ if the inequality $a \leq Cb$ holds with probability going to one.

2.1. Sparse IV model

Denote $Y \in \mathbb{R}^n$ as the n-dimension phenotype vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ as the gene expression matrix of p genes, and $\mathbf{Z} \in \mathbb{R}^{n\times q}$ as the matrix of q possible IVs such as the genotypes of q genetic variants. Lin, Feng and Li (2015) considered the following high-dimensional IV regression model:

$$Y = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\eta}, \tag{2.1}$$

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}_0 + \mathbf{E}, \tag{2.2}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the vector of regression coefficients that reflects the association between phenotype Y and gene expression $\mathbf{X}$, and $\boldsymbol{\Gamma}_0$ reveals the relationships between the gene expressions $\mathbf{X}$ and the genetic variants $\mathbf{Z}$. Without loss of generality, we assume $\mathbf{Z}$ is centered and standardized. The error terms $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$ and $\mathbf{E} = (\boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_n})^\top$ are an n-dimensional vector and an n-by-p matrix, respectively. We assume that the error $\left(\boldsymbol{\varepsilon_i}^\top, \eta_i\right)$ is element-wise sub-Gaussian with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_e$, and is independent of $\mathbf{Z}$. To emphasize the correlation between Y and $\mathbf{X}$, we assume that the correlation between $\boldsymbol{\varepsilon_i}$ and η_i is not zero. We are interested in the high-dimensional setting where the dimensions of the covariates p and the potential IVs q can both be larger than n.

As suggested by Lin, Feng and Li (2015), we can estimate $\boldsymbol{\beta}_0$ in a sparse setting using a two-stage penalized least squares method. Specifically, we first estimate the coefficient matrix $\boldsymbol{\Gamma}_0$ in (2.2) column by column, as the follows:

$$\widehat{\boldsymbol{\Gamma}}_{\cdot,j} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^q}{\operatorname{argmin}} \left(\frac{1}{2n} \|\mathbf{X}_{\cdot,j} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda_{2j}\|\boldsymbol{\gamma}\|_1 \right), \quad j = 1, 2, \ldots, p, \tag{2.3}$$

where λ_{2j} is a tuning parameter. After obtaining an estimate of $\boldsymbol{\Gamma}_0$, we plug the predicted value of $\mathbf{X}$, which is $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}}$, into the second-stage model (2.1) to obtain an estimator of $\boldsymbol{\beta}_0$:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{2n} \|Y - \widehat{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \right), \tag{2.4}$$

where λ_1 is a tuning parameter.

The focus of this paper is to develop a statistical test of a single null hypothesis $\mathcal{H}_{0i} : \beta_{0i} = 0$ vs. $\mathcal{H}_{1i} : \beta_{0i} \neq 0$ for a given i, and to develop an FDR controlling procedure for simultaneously testing p such null hypotheses, for $i = 1, \ldots, p$.

2.2. Hypothesis testing for a single hypothesis using inverse regression

Denote $\mathbf{D} = \mathbf{Z}\boldsymbol{\Gamma}_0$. From models (2.1) and (2.2), we have

$$Y = \mu + \mathbf{D}\boldsymbol{\beta}_0 + \boldsymbol{\xi}, \tag{2.5}$$

where $\boldsymbol{\xi} = \boldsymbol{\eta} + \mathbf{E}\boldsymbol{\beta}_0$. When $\mathbf{Z}$ consists of all valid instruments, $\mathbf{D}$ and $\boldsymbol{\xi}$ are independent by the causal assumptions for valid IVs and (2.5) can be treated as a standard linear regression, but with correlated errors. Owing to the dependent errors, the debiased method of Javanmard and Montanari (2014) cannot be applied directly to this linear model, even when $\mathbf{D}$ is known. Instead, we use an inverse regression (Liu and Luo (2014); Liu (2013)) to construct our test statistics. For each $i = 1, 2, \ldots, p$, $\mathbf{D}_i$ is regressed on $(Y, \mathbf{D}_{\cdot,-i})$ as

$$\mathbf{D}_{\cdot,i} = a_i + (Y, \mathbf{D}_{\cdot,-i})\,\boldsymbol{\theta}_i + \boldsymbol{\zeta}_i, \tag{2.6}$$

where $\boldsymbol{\zeta}_i$ satisfies $\mathbb{E}\boldsymbol{\zeta}_i = 0$ and is uncorrelated with $(Y, \mathbf{D}_{\cdot,-i})$. The regression coefficient $\boldsymbol{\theta}_i$ is chosen so that $\mathbb{E}\boldsymbol{\zeta}_i = 0$, and $\boldsymbol{\zeta}_i$ is uncorrelated with $(Y, \mathbf{D}_{\cdot,-i})$ and has the smallest variance. It is easy to check that such a $\boldsymbol{\theta}_i$ is related to the target parameter $\boldsymbol{\beta}_0$, using the following equality:

$$\boldsymbol{\theta}_i = -\sigma_{\zeta_i}^2 \left(-\frac{\beta_{0i}}{\sigma_\xi^2}, \frac{\beta_{0i}\boldsymbol{\beta}_{-0i}^\top}{\sigma_\xi^2} + \boldsymbol{\Omega}_{-i,i}^{\mathbf{D}} \right), \tag{2.7}$$

where $\sigma_{\zeta_i}^2$ and σ_ξ^2 denote the variances of ζ_i and ξ, respectively, and $\boldsymbol{\Omega}^{\mathbf{D}} = \boldsymbol{\Sigma}_{\mathbf{D}}^{-1}$ is the precision matrix for $\mathbf{D}$. Because $\mathrm{Cov}(\mathbf{D}, \xi) = 0$, we have $\sigma_{\zeta_i}^2 \beta_{0i} = \sigma_\xi^2 \boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i1}\mathrm{Cov}(\xi, y) = -\mathrm{Cov}(\xi, \zeta_i)$; therefore, the null hypothesis $\mathcal{H}_{0i} : \beta_{0i} = 0$ is equivalent to

$$\mathcal{H}_{0i} : \ \mathrm{Cov}(\xi, \zeta_i) = 0 \quad \text{vs.} \quad \mathcal{H}_{1i} : \ \mathrm{Cov}(\xi, \zeta_i) \neq 0,$$

for $i = 1, 2, \ldots, p$.

Because the data observed are $\{y_k, \mathbf{X}_k, \mathbf{Z}_k, k = 1, 2, \ldots n\}$, the vector $\mathbf{D}_i$ in (2.6) is not observed for any $i = 1, 2, \ldots, p$. One can estimate $\boldsymbol{\theta}_i$ via regularization by replacing $\mathbf{D}$ with its estimated value $\widehat{\mathbf{D}} = \widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}}$,

$$\widehat{\boldsymbol{\theta}}_i = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \left\{ \frac{1}{2n} \|\widehat{\mathbf{D}}_{\cdot,i} - \left(Y, \widehat{\mathbf{D}}_{\cdot,-i}\right) \boldsymbol{\theta}_i\|_2^2 + \mu_i \|\boldsymbol{\theta}\|_1 \right\}, \tag{2.8}$$

for $i = 1, 2, \ldots, p$, where μ_i is a tuning parameter.

The sample correlation between ξ and ζ_i is then used to construct the test statistic for $\mathcal{H}_{0i}$ (Liu (2013)). Using the estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{D}}$, and $\widehat{\boldsymbol{\theta}}_i$, the estimated

residuals are

$$\widehat{\xi}_k = y_k - \overline{Y} - \left(\widehat{\mathbf{D}}_k - \overline{\widehat{\mathbf{D}}}\right)^\top \widehat{\boldsymbol{\beta}},$$
$$\widehat{\zeta}_{k,i} = \widehat{\mathbf{D}}_{k,i} - \overline{\widehat{\mathbf{D}}}_i - \left\{ y_k - \overline{Y}, \left(\widehat{\mathbf{D}}_{k,-i} - \overline{\widehat{\mathbf{D}}}_{-i}\right)^\top \right\} \widehat{\boldsymbol{\theta}}_i,$$

for $k = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, p$, where

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} y_k, \quad \overline{\widehat{\mathbf{D}}} = \frac{1}{n}\sum_{k=1}^{n} \widehat{\mathbf{D}}_k, \quad \overline{\widehat{\mathbf{D}}}_i = \frac{1}{n}\sum_{k=1}^{n} \widehat{\mathbf{D}}_{k,i}, \quad \overline{\widehat{\mathbf{D}}}_{-i} = \frac{1}{n}\sum_{k=1}^{n} \widehat{\mathbf{D}}_{k,-i}.$$

Using the bias-correction formula in Liu (2013), for each i, define the test statistic as

$$T_i = \sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} \widehat{\xi}_k \widehat{\zeta}_{k,i} + \frac{1}{n}\sum_{k=1}^{n} \widehat{\xi}_k^2 \widehat{\boldsymbol{\theta}}_{1,i} + \frac{1}{n}\sum_{k=1}^{n} \widehat{\zeta}_{k,i}^2 \widehat{\boldsymbol{\beta}}_i\right) \Big/ \widehat{\sigma}_\xi \widehat{\sigma}_{\zeta_i},$$

where $\widehat{\sigma}_\xi^2 = n^{-1}\sum_{k=1}^{n} \widehat{\xi}_k^2$ and $\widehat{\sigma}_{\zeta_i}^2 = n^{-1}\sum_{k=1}^{n} \widehat{\zeta}_{k,i}^2$.

The bias-correction formula adds two extra terms to the original sample correlation in order to eliminate the higher-order bias resulting from the bias of the Lasso-type estimator. Using the transformation theorem in Anderson (2003), the final test statistic for testing $\mathcal{H}_{0i} : \mathrm{Cov}(\xi, \zeta_i) = 0$ is defined as

$$\widehat{T}_i = \frac{T_i}{1 - (T_i^2/n)\mathbf{1}\left((T_i^2/n) < 1\right)},$$

which has an asymptotic $N(0,1)$ distribution under the null (see Theorem 1).

2.3. Rejection regions for multiple testing procedure with FDR and FDV control

After obtaining the test statistic $\widehat{T}_i$ for $\mathcal{H}_{0i}$, we determine the rejection region for multiple tests of $\widehat{T}_i$ for $\mathcal{H}_{0i}$, for $i = 1, \ldots, p$. Recall that the definitions of FDR and FDV are:

$$FDR = \mathbb{E}\left\{\frac{\sum_{i\in\mathcal{H}_0} \mathbf{1}\left(|\widehat{T}_i| \geq t\right)}{\sum_{i=1}^{p} \mathbf{1}\left(|\widehat{T}_i| \geq t\right) \vee 1}\right\}, \quad FDV = \mathbb{E}\left\{\sum_{i\in\mathcal{H}_0} \mathbf{1}\left(|\widehat{T}_i| \geq t\right)\right\}.$$

Suppose the rejection region for each $\mathcal{H}_{0i}$ is $\{|\widehat{T}_i| \geq t\}$. By the definitions of false discovery proportion and FDR, an ideal choice of t that controls the FDR below

a certain level α is

$$t_0 = \inf\left\{0 \le t \le \sqrt{2\log p} : \frac{\sum_{i\in\mathcal{H}_0} \mathbf{1}\left(|\widehat{T}_i| \ge t\right)}{\sum_{i=1}^{p} \mathbf{1}\left(|\widehat{T}_i| \ge t\right) \vee 1} \le \alpha\right\}. \tag{2.9}$$

In practice, the quantity $\sum_{i\in\mathcal{H}_0} \mathbf{1}\left(|\widehat{T}_i| \ge t\right)$ can be estimated by $2p\,(1-\Phi(t))$, which gives the critical value $\widehat{t}_0$, where $\Phi(t)$ is the cumulative distribution function of the standard normal distribution. We reject the hypothesis $\mathcal{H}_{0i}$ if $|\widehat{T}_i| \ge \widehat{t}_0$, for $i = 1, 2, \ldots, p$.

Similarly, to control the FDV, the rejection region $|\widehat{T}_i| \ge \widehat{t}_0$ is given by

$$\widehat{t}_0 = G^{-1}\left(\frac{k}{p}\right), \tag{2.10}$$

where $G(t) = 2(1-\Phi(t))$.

2.4. Implementation

The construction of the test statistics involves a set of convex optimizations and selecting the tuning parameters in order to solve the Lasso regressions (2.3), (2.4), and (2.8). The optimizations can be implemented efficiently using the coordinate descent (CD) algorithm (Friedman, Hastie and Tibshirani (2010); Lin, Feng and Li (2015)). The CD algorithm is a well known and widely used convex optimization algorithm for penalized regressions, so we omit the details here.

For the tuning parameter selection, we have separate strategies for the two groups of tuning parameters λ and μ. For the optimization problems (2.3) and (2.4), the tuning parameters λ_1 and λ_{2j}, for $j = 1, 2, \ldots, p$, can be chosen using K-fold cross-validation (CV), for $K = 5$ or 10, where λ_1^{opt} and $\lambda_{2j}^{\text{opt}}$, for $j = 1, 2, \ldots, p$, are determined by minimizing the CV errors of the corresponding optimization problem. When both p and q are very large, performing CV can be time consuming. Therefore, in our simulations and real-data applications, we applied an alternative method to select these two groups of tuning parameters that relies on the scaled Lasso (Sun and Zhang (2012)), which is computationally more efficient.

The tuning parameters for the inverse regression (2.8) are selected using a data-driven procedure, as suggested by Liu (2013) and Liu and Luo (2014). Specifically, let $\delta_j = j$, for $j = 1, 2, \ldots, 100$, and $\mu_j = 0.02\delta_j\sqrt{\widehat{\mathbf{\Sigma}}^D_{i,i}\log p/n}$, where

$\widehat{\mathbf{\Sigma}}^D$ is the sample covariance matrix of $\widehat{\mathbf{D}}$. The choice of δ is determined by

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}} \sum_{k=30}^{90} \left\{ \frac{\sum_{i=1}^{p} \mathbf{1}\left(|\widehat{T}_i| \geq \Phi^{-1}\left(1 - k/200\right)\right)}{kp/100} - 1 \right\}^2 .$$

The tuning parameter μ_i in (2.8) is chosen as $\hat{\mu}_i = 0.02\hat{\delta}\sqrt{\widehat{\mathbf{\Sigma}}^D_{i,i} \log p/n}$.

3. Theoretical Results

Here, we provide some theoretical results for the proposed methods. Because our proposed hypothesis test is based on the two-stage penalized estimation of Lin, Feng and Li (2015), we first briefly state the estimation error bounds for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{\Gamma}}_{\cdot,j}$. Under assumptions similar to those of Bickel, Ritov and Tsybakov (2009) and Lin, Feng and Li (2015) to ensure that the matrices $\mathbf{Z}$ and $\mathbf{D}$ are well-behaved and that the ℓ_1 norms of the true parameters $\boldsymbol{\beta}_0$ and $\mathbf{\Gamma}_0$ are bounded away from infinity, Lin, Feng and Li (2015) showed that when the tuning parameters are chosen appropriately, we have

$$\|\mathbf{Z}\left(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\right)\|_F^2 \leq \frac{16\widetilde{C}^2 C^2}{\kappa^2(s_2, \mathbf{Z})} s_2 p \left(\log p + \log q\right), \tag{3.1}$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq C_3 s_1 \sqrt{\frac{s_2\left(\log p + \log q\right)}{n}}, \tag{3.2}$$

where $\widetilde{C}, C$, and C_3 are some constants, $s_1 = \|\boldsymbol{\beta}_0\|_0$, $s_2 = \max_j \|\mathbf{\Gamma}_{\cdot,j}\|_0$, and $\kappa^2(s_2, \mathbf{Z})$ is the constant in the restricted eigenvalue condition for $\mathbf{Z}$. See the Supplementary Material (S1) for detailed conditions (A1) - (A3).

3.1. Asymptotic distribution of the test statistic for single null hypothesis

In order to make an inference for the parameter $\boldsymbol{\beta}_0$ using an inverse regression, three additional assumptions are needed.

(B1) In the inverse regression model (2.6), denote $\mathbf{M}_i = (Y, \mathbf{D}_{\cdot,-i})$, for $i = 1, \ldots, p$; then $\mathbf{M}_i$ satisfies the restricted eigenvalue condition with some constant $\kappa(r, \mathbf{M}_i)$. In addition, assume that there exists a positive constant $\kappa(Y, \mathbf{D})$ such that $\min_i \kappa(r, \mathbf{M}_i) \geq \kappa(Y, \mathbf{D})$.

(C1) The precision matrix $\mathbf{\Omega}^{\mathbf{D}}$ and covariance matrix $\Sigma_{\mathbf{D}}$ satisfy $\max_{1 \leq j \leq p}\left(\mathbf{\Omega}^{\mathbf{D}}_{j,j}, \Sigma^{\mathbf{D}}_{j,j}\right) \leq C$, for some constant C and $\mathrm{Var}(Y_i) \leq C$.

(C2) The dimensional parameters n, p, q, s_1, s_2, and r satisfy the following asymptotic scaling condition as $n \to \infty$:

$$\max\{r\sqrt{s_2}, s_1, s_2\}\sqrt{\frac{\log p\,(\log p + \log q)}{n}} = o(1).$$

Assumption (B1) guarantees that $\boldsymbol{\theta}_i$ is well estimated. This assumption is implicitly assumed, though not stated, in Liu and Luo (2014). Assumptions (C1) and (C2) are needed to obtain the asymptotic distribution of $\widehat{T}_i$. In particular, assumption (C1) bounds the entries of the covariance matrix $\Sigma_{\mathbf{D}}$ and the precision matrix $\Omega^{\mathbf{D}}$, and assumption (C2) provides the relation between the dimension and the sparsity parameters n, p, q, s_1, s_2, and r, where s_1, s_2, and r control the sparsity of $\boldsymbol{\beta}_0$, $\boldsymbol{\Gamma}_0$, and $\boldsymbol{\theta}_i$ respectively. In addition, if we fix q, which is the number of instruments, then assumption (C2) is equivalent to $\log p = o(\sqrt{n})$. This assumption is often made in inference results related to the Lasso and other high-dimensional models (Gold, Lederer and Tao (2017); Javanmard and Montanari (2014); Ning and Liu (2017)).

We have the following lemma on the estimation error bound of $\boldsymbol{\theta}_i$ in the inverse regression.

Lemma 1 (Estimation error bounds of $\boldsymbol{\theta}_i$)**.** *Under assumptions* (A1)–(A3) *(Supplementary Material) and* (B1)*, for each $i = 1, 2, \ldots, p$, there exists some positive constants C_4, C_5, C_5^*. If the tuning parameter μ_i is chosen as*

$$\mu_i = \frac{C_4^*}{\kappa(s_2, \mathbf{Z})}\sqrt{\frac{s_2(\log p + \log q)}{n}},$$

with $C_4^* = C_5^* \max(C, \sigma_{\zeta_i})$, then with probability at least $1 - C_4\,(pq)^{-C_5}$, $\widehat{\boldsymbol{\theta}}_i$ in (2.8) *satisfies*

$$\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 \le \frac{64C_4^*}{\kappa^2(Y, \mathbf{D})\kappa(s_2, \mathbf{Z})} r\sqrt{\frac{s_2(\log p + \log q)}{n}}.$$

This Lemma 1, together with the estimation bounds (3.1) and (3.2) can be used to derive the asymptotic distribution of the test statistic $\widehat{T}_i$ under the null $\mathcal{H}_{0i}$.

Theorem 1 (Asymptotic distribution of $\widehat{T}_i$)**.** *Under assumptions* (A1)–(A3)*,* (B1)*, and* (C1)–(C2)*, with proper choices of the tuning parameters λ_1, λ_{2j}, and μ, as stated in Lemma* 1 *and Lemma* 1 *in the Supplementary Material, for each*

$i = 1, 2. \ldots, p$, *under the null* $\mathcal{H}_{0i} : \beta_{0i} = 0$,

$$\widehat{T}_i \rightsquigarrow N(0, 1).$$

3.2. Theoretical results on FDR and FDV

For the proposed multiple testing procedure, in order to control the FDR or FDV, one additional condition on the precision matrix is needed to ensure that the test statistics are not too highly correlated.

(C3) The precision matrix $\mathbf{\Omega^D}$ satisfies the following condition: for some $\varepsilon > 0$ and $\delta > 0$,

$$\sum_{(i,j)\in\mathcal{A}(\varepsilon)} p^{2|\rho_{ij,\omega_\mathbf{D}}|/(1+|\rho_{ij,\omega_\mathbf{D}}|)+\delta} = \mathcal{O}\bigg(\frac{p^2}{(\log p)^2}\bigg),$$

where $\rho_{ij,\omega_\mathbf{D}} = \Omega^\mathbf{D}_{ij}/(\Omega^\mathbf{D}_{ii}\Omega^\mathbf{D}_{jj})^{1/2}$ and $\mathcal{A}(\varepsilon) = \mathcal{B}((\log p)^{-2-\varepsilon})$, with $\mathcal{B}(\delta) = \{(i, j) : |\rho_{ij,\omega_\mathbf{D}}| \geq \delta, i \neq j\}$.

The next theorem shows that the proposed multiple testing procedure controls the FDR at a prespecified level.

Theorem 2 (Asymptotic result for multiple testing procedure)**.** *Denote FDR*= *FDR*$(\widehat{t}_0)$. *Assuming* (A1)–(A3), (B1),(C1), *and* (C3) *hold,* $p \leq n^c$, *for some* $c > 0$. *We further assume a condition stronger than* (C2) *such that the quantities in the left of assumption* (C2) *are of order* $o((\log p)^{-1/2})$ *instead of* $o(1)$, *and for some* $\widetilde{c} > 2$,

$$\sum_{i\in\mathcal{H}_1} \mathbf{1}\left(\frac{\boldsymbol{\beta}_i}{\sqrt{\sigma^2_\xi \Omega^\mathbf{D}_{i,i}}} \geq \sqrt{\frac{\widetilde{c}\log p}{n}}\right) \to \infty, \tag{3.3}$$

as $(n, p) \to \infty$. *Then, with a proper choice of all tuning parameters and the threshold* $\widehat{t}_0$, *with a prespecified level* α, *we have*

$$\lim_{n,p\to\infty} \frac{FDR}{\alpha p_0/p} = 1 \text{ and } \lim_{n,p\to\infty} \frac{FDV}{kp_0/p} = 1.$$

This theorem indicates that under proper conditions, the empirical FDR(eFDR) is controlled under a prespecified level. In addition to the assumptions mentioned previously, we require a stronger condition (3.3), which requires that the number of true alternatives tends to infinity. This condition is also required in Liu and Luo (2014).

4. Simulation Studies

We evaluate the performance of the proposed methods using a set of simulations. Following models (2.1) and (2.2), we first generate the instruments matrix $\mathbf{Z}$, where $\mathbf{Z}_i \sim N(0, \mathbf{\Sigma}_z)$. The covariance matrix $\mathbf{\Sigma}_z$ satisfies $(\mathbf{\Sigma}_z)_{ij} = 0.5^{|i-j|}$. For each $\mathbf{\Gamma}_{\cdot,j}$, we first randomly pick s_2 out of q nonzero entries, and then each entry is generated randomly from a uniform distribution $U([-b,-a] \cup [a,b])$, with $a = 0.75, b = 1$. The parameter $\boldsymbol{\beta}_0$ is generated similarly, where we pick s_1 out of p nonzero entries, and each entry is generated randomly from $U([-0.3,-0.1] \cup [0.1,0.3])$. For the joint distribution of $\left(\boldsymbol{\varepsilon}_{\boldsymbol{i}}^{\top}, \eta_i\right)$, the covariance matrix $\mathbf{\Sigma}_e$ is generated by $(\mathbf{\Sigma}_e)_{ij} = 0.2^{|i-j|}$, for $1 \le i,j \le p$ and $(\mathbf{\Sigma}_e)_{p+1,p+1} = 1$, and from among $(\mathbf{\Sigma}_e)_{i,p+1}$, where $i = 1,\ldots,p$, 10 entries are picked randomly and set to be 0.3. We impose this structure so that η_i is correlated with $\boldsymbol{\varepsilon}_{\boldsymbol{i}}$. The covariates $\mathbf{X}$ and response Y are generated based on our model.

We consider different values of (n,p,q), with $(n,p,q) = (200,100,100), (400, 200,200)$, and $(200,500,500)$, and $(s_1,s_2) = (10,10)$. Because we focus on inferences on $\boldsymbol{\beta}$, we omit the estimation comparisons here; see Lin, Feng and Li (2015) for an extensive study. We observed similar improvements over the estimates from the naive Lasso estimate. We compare our methods with the test developed in Liu and Luo (2014) for a high-dimensional regression analysis linking Y to $\mathbf{X}$, ignoring the fact that $\mathbf{X}$ and $\boldsymbol{\eta}$ are correlated. It should be noted that the independent error assumption is necessary for the method in Liu and Luo (2014) to work. We evaluate the performances of the hypothesis testing procedures by calculating the empirical type-I errors for testing single regression coefficients, and the eFDR, empirical FDV(eFDV) for multiple testing procedures.

4.1. Type-I error of single hypothesis test

Figure 1 shows box plots of the empirical type-I errors for testing the single null hypothesis for the variables with a zero coefficient, based on IV models and the standard Lasso regression. When the errors and covariates are correlated owing to unobserved confounding, the naive Lasso regression may fail to control the type I error for some null coefficients, leading to inflated type-I errors. This indicates that the naive method may falsely select some unrelated variables. As a comparison, the test based on the IV regression controls the type-I errors below the specified significance level.

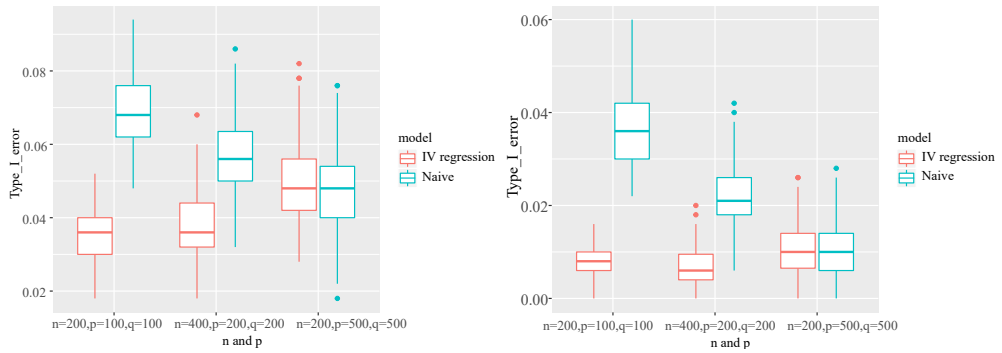

Figure 1. Box plots of the empirical type-I errors for single hypothesis testing based on an IV regression and a naive Lasso regression under different settings for α-level of 0.05 (left) and 0.01 (right).

4.2. FDR controlling for multiple testing

To examine the performance of the proposed multiple testing procedure, for a given $\alpha = 0.05, 0.1, 0.2$, we calculate the eFDR defined as the average of the observed FDR:

$$\text{FDR} = \frac{\sum_{i\in\mathcal{H}_0} \mathbf{1}\left(|\widehat{T}_i| \geq \widehat{t}_0\right)}{\sum_{i=1}^{p} \mathbf{1}\left(|\widehat{T}_i| \geq \widehat{t}_0\right) \vee 1}.$$

Similarly, the power of the multiple testing procedure is defined as $\sum_{i\in\mathcal{H}_1} \mathbf{1}(|\widehat{T}_i| \geq \widehat{t}_0)/|\mathcal{H}_1|$.

Table 1 shows the eFDR for the proposed procedure for the IV regression and the method of Liu and Luo (2014) for the standard high-dimensional regression, without including the IVs. We observe that the proposed multiple test procedure for the IV regression can indeed control the FDR at the correct level. In contrast, the procedure based on the standard high-dimensional regression fails to control the FDR, unless the model is very sparse, as when $p = q = 500$.

We similarly evaluate the procedure for controlling the number of falsely discovered variables k=2,3,4. The eFDV is defined as average of the observed FDV=$\sum_{i\in\mathcal{H}_0} \mathbf{1}(|\widehat{T}_i| \geq \widehat{t}_{FDV})$, and its power is given by $\sum_{i\in\mathcal{H}_1} \mathbf{1}(|\widehat{T}_i| \geq \widehat{t}_{FDV})$. Table 1 shows that the proposed procedure also controls the FDV at the specified level. However, the method based on the standard regression can result in a larger eFDV than the prespecified number.

Note that for $p = q = 500$, the performance of our proposed method is very similar to that of the naive test. The reason is that, by our construction of the covariance matrix of the error terms, the dependency between the covariates and

Table 1. Simulation results based on 500 replications. The eFDR/eFDV and power for the multiple testing procedure based on the IV regression and the naive high-dimensional linear regression (eFDR* and eFDV*) for different combinations of (n, p, q) and different $\alpha/k-$levels.

FDR				FDV			
α	eFDR	power (sd)	eFDR*	k	eFDV	power (sd)	eFDV*
(n,p,q)=(200, 100, 100)							
0.05	0.044	0.547 (0.15)	0.198	2	1.35	6.35 (1.5)	4.11
0.10	0.075	0.58 (0.15)	0.239	3	1.94	6.57 (1.4)	4.87
0.20	0.134	0.622 (0.15)	0.296	4	2.49	6.71 (1.4)	5.55
(n,p,q)=(400, 200, 200)							
0.05	0.026	0.752 (0.13)	0.153	2	1.27	8.16 (1.1)	4.18
0.10	0.060	0.781 (0.12)	0.197	3	1.94	8.31 (1.1)	5.13
0.20	0.124	0.814 (0.12)	0.268	4	2.59	8.42 (1.1)	5.96
(n,p,q)=(200, 500, 500)							
0.05	0.074	0.390 (0.12)	0.055	2	2.21	4.93 (1.3)	2.04
0.10	0.129	0.427 (0.13)	0.103	3	3.19	5.17 (1.4)	3.01
0.20	0.224	0.472 (0.14)	0.197	4	4.13	5.39 (1.4)	3.98

the errors becomes very weak for large p, in which case, the two methods are expected to perform similarly.

4.3. Sensitivity to model assumptions

We perform a sensitivity analysis to determine how non-normal errors impact the performance of our method. We consider the same model parameters as in the previous section, but assume that $\left(\boldsymbol{\varepsilon_i}^\top, \eta_i\right)$ have a multivariate t-distribution with mean $\mathbf{0}$, covariance $\boldsymbol{\Sigma}_e$, and degree of freedom three, where $\boldsymbol{\Sigma}_e$ is the same as in Section 4. Table 2 shows that eFDR and eFDV are, in general, controlled below the specified values, indicating that the method is not too sensitive to the distribution of the error terms.

We further examine the performance of our method when the IVs have direct effects on the outcome, a violation of being a valid IV. We generate the data by $Y_i = \mathbf{X}_i\boldsymbol{\beta}_0 + \mathbf{Z}_i\boldsymbol{\tau} + \varepsilon_i$. In the setting of weak or moderate direct effects, we assume two IVs to have nonzero coefficients of 0.2 and -0.2 (weak), and 0.5 and -0.5 (moderate), respectively. Table 2 shows that under the weak or moderate direct effects, the proposed method can still control the FDR or FDV well. However, if the IV have very strong effects with a coefficient of $(1, 1, 0.5, 0.5, -0.5)$, we observe over-inflated eFDRs and eFDVs.

Table 2. Sensitivity analysis for non-normal errors and invalid IVs with weak, moderate, or strong direct effects on the outcome. Simulation results are based on 500 replications.

α-level	eFDR	k	eFDV	α-level	eFDR	k	eFDV
(n,p,q)=(200, 100, 100)				(n,p,q)=(400, 200, 200)			
t-distribution of the errors							
0.05	0.09	2	1.5	0.05	0.057	2	1.62
0.1	0.11	3	2.19	0.1	0.09	3	2.36
0.2	0.17	4	2.90	0.2	0.16	4	3.2
Weak direct effects							
0.05	0.050	2	1.42	0.05	0.035	2	1.38
0.1	0.081	3	2.03	0.1	0.068	3	2.05
0.2	0.14	4	2.69	0.2	0.13	4	2.77
Moderate direct effects							
0.05	0.13	2	2.44	0.05	0.09	2	2.23
0.1	0.16	3	3.08	0.1	0.13	3	3.04
0.2	0.23	4	3.82	0.2	0.21	4	3.76
Strong direct effects							
0.05	0.51	2	8.50	0.05	0.56	2	13.10
0.1	0.54	3	9.51	0.1	0.60	3	14.49
0.2	0.58	4	10.30	0.2	0.65	4	15.68

5. Application to a Yeast Data Set

We demonstrate our method using a data set collected on 102 yeast segregants by crossing two genetically diverse strains (Brem and Kruglyak (2005)). The data set includes the growth yields of individual segregants grown in the presence of different chemicals (Perlstein et al. (2007)). These segregants have different genotypes, represented by 585 markers, after removing the markers that are in almost complete linkage disequilibrium. The genotype differences in these strains contribute to rich phenotypic diversity in the segregants. In addition, 6,189 yeast genes were profiled in rich media and in the absence of any chemical or drug using expression arrays (Brem and Kruglyak (2005)). Using the same data preprocessing steps as Chen et al. (2009), we compiled a list of candidate gene expression features based on their potential regulatory effects, including transcription factors, signaling molecules, chromatin factors, and RNA factors and genes involved in vacuolar transport, endosome, endosome transport, and vesicle-mediated transport. We further filter out genes with $s.d \leq 0.2$ in expression level, yielding a total of 813 genes in our analysis.

We are interested in identifying genes with expression levels that are associated with yeast growth yield after being treated with hydrogen peroxide by

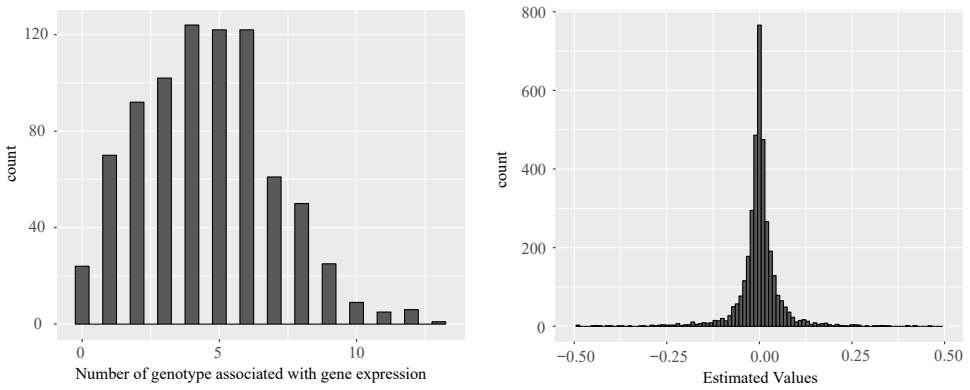

Figure 2. Analysis of yeast eQTL data sets, showing a histogram of the number of genotypes associated with each gene expression (left plot), and a histogram of the estimated regression coefficients in the first stage ($\widehat{\mathbf{\Gamma}}$) based on Lasso regressions (right plot).

fitting the proposed two-stage sparse IV model. Figure 2 shows a histogram of the number of SNPs selected for each gene expression, and a histogram of the estimated regression coefficients ($\mathbf{\Gamma}_0$) from the Lasso. These results show that genetic variants are strongly associated with gene expressions, and therefore, can be used as IVs for gene expressions.

Using these selected genotypes as the IVs for each of the gene expressions, we obtain the fitted expression values. Then we apply the Lasso with these fitted expressions as the predictors and the yeast growth yield as the response. For each gene j, we test the null of $\beta_j = 0$ and obtain its p-value. The 15 significant genes at a nominal $p < 0.05$ are presented in Table 3. At FDR< 0.10, three genes are selected. These genes are related to resistance to chemicals, competitive fitness, and cell growth, partially explaining their association with yeast growth in the presence of hydrogen peroxide. For example, among the genes with negative coefficients, over-expression indicates decreased yeast growth. The RRM3 gene is involved in DNA replication, and over-expression of the gene leads to abnormal budding and decreased resistance to chemicals. Over-expression of the POP5 and FUN26 genes causes a decreased vegetative growth rate of the yeast (`https://www.yeastgenome.org`).

The three selected genes using FDR< 0.10 all have positive coefficients, indicating that an over-expression of these genes led to increased yeast growth in the presence of hydrogen peroxide. Among these, BDP1 is a general activator of RNA polymerase III transcription, and is required for transcription from all three types of polymerase III promoters (Ishiguro, Kassavetis and Geiduschek (2002)).

Table 3. Results from the analysis of the yeast growth yield data set. The table shows the selected genes using a single test statistic ($p < 0.05$) and the multiple testing procedure with FDR< 0.10 and FDV< 2(marked by $*$). The gene names and estimated regression coefficients $\hat{\boldsymbol{\beta}}$ and refitted values $\hat{\boldsymbol{\beta}}^*$ are listed.

Gene id	Gene	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^*$	Gene id	Gene	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^*$
Negative coefficient				Negative coefficient			
YHR031C	RRM3	-3.82	-5.00	YNL331C	AAD14	0.07	0.17
YAL033W	POP5	-0.22	-0.69	YHR014W	SPO13	0.47	2.20
YLR275W	SMD2	-0.20	-0.31	YNR045W*	PET494	0.70	0.86
YNL236W	SIN4	-4.67	-5.63	YHR018C*	ARG4	0.22	0.34
YNL138W	SRV2	-0.63	-1.68	YHR097C	YHR097C	0.06	0.15
YNL146W	YNL146W	-0.24	-0.12	YNL039W*	BDP1	1.82	3.96
YAR035W	YAT1	-1.74	-2.79				
YAL022C	FUN26	-2.89	-4.79				
YHL018W	YHL018W	-0.79	-2.29				

Here, an over-expression of this gene is expected to increase the yeast viability and growth. PET494 is a mitochondrial translational activator specific to mitochondrial mRNA encoding cytochrome c oxidase subunit III (coxIII) (Marykwas and Fox (1989)). Finally, a null mutant of the ARG4 gene shows decreased resistance to chemicals (`https://www.yeastgenome.org`); therefore, segregants with higher expressions of this gene are expected to have increased resistance to chemicals and increased growth yield.

As a comparison, we also apply a Lasso regression with 813 gene expressions as the predictors, without using the genotype data. The same statistical test is applied to each of the genes. At a nominal p-value of 0.05, 34 genes are selected by the Lasso. However, no gene is selected after adjusting for multiple comparisons with FDR< 0.10. This suggests that by effectively using the genotype data, we are able to identify biologically meaningful genes that are associated with yeast growth in the presence of hydrogen peroxide. We further compare the fitted versus the observed yeast growth yields for different models (see Figure S3). Overall, we observe that the proposed IV regression yields a better fit than those based on linear regressions with gene expressions as the covariates.

6. Discussion

We have developed methods for exploring the association between gene expressions and phenotypes in an IV regression framework when there are possible unmeasured confounders. Here, the genetic variants are used as possible IVs. We

have constructed a test statistic using an inverse regression and derived its asymptotic null distribution. We have also developed a multiple testing procedure for high-dimensional two-stage least-square methods. Both our theoretical results and our simulations show the correctness of our procedure and its improved performance over that of the standard Lasso regression when the covariates and errors are correlated.

For the yeast genotype and gene expression data, our two-stage regression method was able to identify three yeast genes whose expressions were associated with growth in the presence of hydrogen peroxide. In contrast, using gene expression data alone, and the Lasso regression did not identify any growth associated genes. Because growth yield is highly inheritable (Perlstein et al. (2007)), using genotype-predicted gene expressions in our two-stage estimation can help to identify gene expressions that might be causal to the phenotype. For model organisms such as yeast, the conditional independence assumption between the genotypes and the outcome, given gene expression levels, is expected to hold. However, for human studies, one should be cautious of such an assumption, because genetic variants can affect phenotype via other mechanisms such as changing protein structures.

In stage 1 of our method, we used the Lasso to identify the genetic variants associated with gene expressions. Alternatively, we can use a ridge regression as suggested by one reviewer. Our simulations, shown in Table S1 of the Supplementary Material, indicate that such an approach might lead to a conservative inference. In addition, as shown in Shao and Deng (2012), sparsity on the regression coefficients is still required in order to have prediction consistency. Without such a sparsity assumption, the expected L_2 norm $\|\mathbf{Z}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0)\|_F^2/np$ may not converge to zero, in which case, it is not clear how one can develop methods for inferences on $\boldsymbol{\beta}$.

Our simulations showed that when some of the IVs have strong direct effects on the outcome, the proposed test is not valid. One important future research direction is to detect and account for the existence of weak or invalid IVs in high-dimensional IV regression analysis framework. The problem of incorporating weak/invalid instruments has been studied extensively (Kang et al. (2016); Guo et al. (2018); Zhao et al. (2019)). However, these studies only consider one or a few covariates, in contrast to our setup. It would be interesting to extend these recent methods to high-dimensional covariates and the multiple testing problem.

Supplementary Material

The online Supplemental Material includes proofs of Lemma 1 and Theorem 1 and 2, as well as additional simulations, real-data analysis results, and the Matlab code to implement the algorithm. The real data sets used in this study are available upon request.

Acknowledgments

The authors are grateful to the editor, associate editor, and two anonymous referees for their helpful comments and suggestions. This research was supported by NIH grants GM129781 and GM123056.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, Hoboken.

Barbeira, A., Dickinson, S., Bonazzola, R., Zheng, J., Wheeler, H., Torres, J. et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communication* **9**, 1825.

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

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences* **102**, 1572–1577.

Chen, B.-J., Causton, H. C., Mancenido, D., Goddard, N. L., Perlstein, E. O. and Pe'er, D. (2009). Harnessing gene expression to identify the genetic basis of drug resistance. *Molecular Systems Biology* **5**, 310.

Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. *The Annals of Statistics* **42**, 872–917.

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

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

Gold, D., Lederer, J. and Tao, J. (2017). Inference for high-dimensional nested regression. *arXiv preprint arXiv:1708.05499*.

Guo, Z., Kang, H., Cai, T. T. and Small, D. S. (2018). Testing endogeneity with high dimensional covariates. *Journal of Econometrics* **207**, 175–187.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252.

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G. et al. (2003). Control of confounding of genetic associations in stratified populations. *The American Journal of Human Genetics* **72**, 1492–1504.

Imbens, G. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science* **27**, 323–358.

Ishiguro, A., Kassavetis, G. A. and Geiduschek, E. P. (2002). Essential roles of Bdp1, a subunit of RNA polymerase III initiation factor TFIIIB, in transcription and tRNA processing. *Molecular and Cellular Biology* **22**, 3264–3275.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15**, 2869–2909.

Kang, H., Zhang, A., Cai, T. T. and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association* **111**, 132–144.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**, e161.

Lin, W., Feng, R. and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association* **110**, 270–288.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics* **41**, 2948–2978.

Liu, W. and Luo, S. (2014). Hypothesis testing for high-dimensional regression models †. Technical report.

Liu, Y., Morley, M., Brandimarto, J., Hannenhalli, S., Hu, Y., Ashley, E. A. et al. (2015). Rna-seq identifies novel myocardial gene expression signatures of heart failure. *Genomics* **105**, 83–89.

Marykwas, D. and Fox, T. (1989). Control of the saccharomyces cerevisiae regulatory gene PET494: Transcriptional repression by glucose and translational induction by oxygen. *Molecular and Cellular Biology* **9**, 484–491.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45**, 158–195.

Perlstein, E. O., Ruderfer, D. M., Roberts, D. C., Schreiber, S. L. and Kruglyak, L. (2007). Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nature Genetics* **39**, 496–502.

Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L. A. et al. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature Genetics* **40**, 1461–1465.

Sanderson, E., Davey Smith, G., Windmeijer, F. and Bowden, J. (2018). An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology* **48**, 713–727.

Shao, J. and Deng, X. (2012). Estimation in high dimensional linear models with deterministic design matrix. *The Annals of Statistics* **40**, 812–831.

Speliotes, E. K., Yerges-Armstrong, L. M., Wu, J., Hernaez, R., Kim, L. J., Palmer, C. D. et al. (2011). Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genetics* **7**, e1001324.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–898.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D. et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics* **51**, 592–599.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 217–242.

Zhao, Q., Wang, J., Hemani, G., Bowden, J. and Small, D. (2019). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *The Annals of Statistics* **48**, 1742–1769.

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

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R. et al.(2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224.

Jiarui Lu

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

E-mail: jiaruilu@pennmedicine.upenn.edu

Hongzhe Li

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

E-mail: hongzhe@pennmedicine.upenn.edu

(Received November 2019; accepted March 2021)