# Regularization Methods for High-Dimensional Instrumental Variables Regression With an Application to Genetical Genomics

**Wei Lin [Postdoctoral Researcher]**, **Rui Feng [Assistant Professor]**, and **Hongzhe Li [Professor]**

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

Wei Lin: weilin1@mail.med.upenn.edu; Rui Feng: ruifeng@mail.med.upenn.edu; Hongzhe Li: hongzhe@upenn.edu

## Abstract

In genetical genomics studies, it is important to jointly analyze gene expression data and genetic variants in exploring their associations with complex traits, where the dimensionality of gene expressions and genetic variants can both be much larger than the sample size. Motivated by such modern applications, we consider the problem of variable selection and estimation in high-dimensional sparse instrumental variables models. To overcome the difficulty of high dimensionality and unknown optimal instruments, we propose a two-stage regularization framework for identifying and estimating important covariate effects while selecting and estimating optimal instruments. The methodology extends the classical two-stage least squares estimator to high dimensions by exploiting sparsity using sparsity-inducing penalty functions in both stages. The resulting procedure is efficiently implemented by coordinate descent optimization. For the representative $L_1$ regularization and a class of concave regularization methods, we establish estimation, prediction, and model selection properties of the two-stage regularized estimators in the high-dimensional setting where the dimensionality of co-variates and instruments are both allowed to grow exponentially with the sample size. The practical performance of the proposed method is evaluated by simulation studies and its usefulness is illustrated by an analysis of mouse obesity data. Supplementary materials for this article are available online.

### Keywords

Causal inference; Confounding; Endogeneity; Sparse regression; Two-stage least squares; Variable selection

## 1 Introduction

Genome-wide studies have been widely conducted to search tens of thousands of gene expressions or hundreds of thousands of single nucleotide polymorphisms (SNPs) to detect

---

**Supplementary Materials**

The supplementary material contains the proof of Theorem 4.

associations with complex traits. By measuring and analyzing gene expressions and genetic variants on the same subjects, genetical genomics studies provide an integrative and powerful approach to addressing fundamental questions in genetics and genomics at the functional level. In these studies, gene expression levels are viewed as quantitative traits that are subject to genetic analysis for identifying expression quantitative trait loci (eQTLs), in order to understand the genetic architecture of gene expression variation. The increasing availability of high-throughput genetical genomics data sets opens up the possibility of jointly analyzing gene expression data and genetic variants in exploring their associations with complex traits, with the goal of identifying key genes and genetic markers that are potentially causal for complex human diseases such as obesity, heart disease, and cancer (Emilsson et al. 2008).

Although in the past decade gene expression profiling has led to the discovery of many gene signatures that are highly predictive for clinical outcomes, the effort of using these findings to dissect the genetics of complex traits and diseases is often compromised by the critical issue of confounding. It is well known that many factors, such as unmeasured variables, experimental conditions, and environmental perturbations, may exert pronounced influences on gene expression levels, which may in turn induce spurious associations and/or distort true associations of gene expressions with the response of interest (Leek and Storey 2007; Fusi, Stegle, and Lawrence 2012). Moreover, due to the difficulty of high dimensionality, empirical studies are mostly based on marginal models, which are especially prone to variability caused by pleiotropic effects and dependence among genes. Ignoring these confounding issues tends to produce results that are both biologically less interpretable and less reproducible across independent studies.

Instrumental variables (IV) methods provide a practical and promising approach to control for confounding in genetical genomics studies, with genetic variants playing the role of instruments. This approach exploits the reasonable assumption that the genotype is assigned randomly, given the parents' genes, at meiosis and independently of possible confounding factors, and affects a clinical phenotype only indirectly through some intermediate phenotypes. In observational epidemiology, Mendelian randomization has been proposed as a class of methods for using genetic variants as instruments to assess the causal effect of a modifiable phenotype or exposure on a disease outcome; see, for example, Lawlor et al. (2008) for a review. The primary scenario considered in this context, however, involves only one exposure variable and requires the existence of a genetic variant whose relationship with the exposure has been well established. Thus, the methodology intended for Mendelian randomization is typically not applicable to genetical genomics studies, where the number of expression phenotypes is exceedingly large and the genetic architecture of each phenotype may be complex and unknown.

IV models and methods have been extensively studied in the econometrics literature, where the problem is often cast in the framework of simultaneous equation models (Heckman 1978). It has been shown that classical IV estimators such as the two-stage least squares (2SLS) estimator and the limited information maximum likelihood (LIML) estimator are consistent only when the number of instruments increases slowly with the sample size (Chao and Swanson 2005; Hansen, Hausman, and Newey 2008). Recent developments have

introduced regularization methods to mitigate the overfitting problem in high-dimensional feature space by exploiting the sparsity of important covariates, thereby improving the performance of IV estimators substantially. Caner (2009) considered penalized generalized method of moments (GMM) with the bridge penalty for variable selection and estimation in the classical setting of fixed dimensionality. Gautier and Tsybakov (2011) developed a Dantzig selector–type procedure to select important covariates and estimate the noise level simultaneously in high-dimensional IV models where the dimensionality may be much larger than the sample size. Under the assumption that the important covariates are uncorrelated with the regression error, Fan and Liao (2012) proposed a penalized focused GMM method based on a nonsmooth loss function to perform variable selection and achieve oracle properties in high dimensions. All the aforementioned methods, however, do not exploit the sparsity of instruments and hence are still facing the dimensionality curse of many instruments. Another active line of research in the econometrics literature has been concerned with the use of regularization and shrinkage methods for estimating optimal instruments in the context of estimating a low-dimensional parameter; see, for example, Okui (2011) and Carrasco (2012). Of particular interest is the recent work of Belloni et al. (2012), where Lasso-based methods were applied to form first-stage predictions and estimate optimal instruments in an IV model with many instruments but itself of fixed dimensionality.

In this article, we focus on the application of high-dimensional sparse IV models to genetical genomics, where we are interested in associating gene expression data with a complex trait to identify potentially causal genes by using genetic variants as instruments. Motivated by this important application, we propose a two-stage regularization (2SR) methodology for identifying and estimating important covariate effects while selecting and estimating optimal instruments. Our approach merges the two independent lines of research mentioned above and provides a regularization framework for IV models that accommodate covariates and instruments both of high dimensionality. Specifically, the proposed procedure consists of two stages: In the first stage the covariates are regressed on the instruments in a regularized multivariate regression model and predictions are obtained, and in the second stage the response of interest is regressed on the first-stage predictions in a regularized regression model to perform final variable selection and estimation. In each stage, a sparsity-inducing penalty function is employed to yield desirable statistical properties and practical performance. The methodology can be viewed a high-dimensional extension of the 2SLS method, allowing the use of regularization methods to address the high-dimensional challenge in both stages.

Several key features make the proposed methodology especially appealing for the kind of applications we consider in this article. First, unlike marginal regression models commonly used in empirical studies that analyze a few variables at a time, our method allows for the joint modeling and inference of high-dimensional genetical genomics data. In view of the fact that many genes interact with each other and contribute together to a complex trait or disease, joint modeling is crucial for correcting bias and controlling false positives due to possible confounding. Second, our method requires neither a specification of a small set of important instruments nor an importance ranking among the instruments; instead, we consider the estimation of optimal instruments as a variable selection problem and allow the

procedure to choose important instruments based on the data. Third, the proposed implementation by coordinate descent optimization is computationally very efficient and has provable convergence properties, therefore bypassing the computational obstacles faced by traditional model selection methods. Finally, we rigorously justify our method for the representative $L_1$ regularization and a class of concave regularization methods in the high-dimensional setting where the dimensionality of covariates and instruments are both allowed to grow exponentially with the sample size. Through the theoretical analysis, we explicate the impact of dimensionality and the role of regularization, and provide strong performance guarantees for the proposed method.

The remainder of this article is organized as follows. Section 2 introduces the high-dimensional sparse IV model. The 2SR methodology and implementation are presented in Section 3. Theoretical properties of the regularized estimators are investigated in Section 4. We illustrate our method by simulation studies in Section 5 and an analysis of mouse obesity data in Section 6. We conclude with some discussion in Section 7. Proofs are relegated to the Appendix and Supplementary Material.

## 2 Sparse Instrumental Variables Models

Suppose we have a quantitative trait or clinical phenotype *y*, a *p*-vector of gene expression levels **x**, and a *q*-vector of numerically coded genotypes **z**. In reality, there may be a sufficient set of unobserved confounding phenotypes **w** that act as proxies for the long-term effects of environmental exposures and/or the state of the microenvironment of the cells or tissues within which the biological processes occur. These phenotypes are likely to have strong influences on gene expression levels while contributing substantially to the clinical phenotype. Figure 1(a) illustrates the confounding between **x** and *y* with an example of six variables. If an ordinary regression analysis is to be applied, the effects of $x_1$ and $x_2$ on *y* would be seriously confounded by *w*, resulting in a spurious association or effect modification.

One way of controlling for the confounding due to **w** is through the use of the genotype **z** as instruments. In order for **z** to be valid instruments, the following conditions must be satisfied (Didelez, Meng, and Sheehan 2010):

1. The genotype **z** is (marginally) independent of the confounding phenotype **w** between **x** and *y*;

2. The genotype **z** is not (marginally) independent of the intermediate phenotype **x**;

3. Conditionally on **x** and **w**, the genotype **z** and the clinical phenotype *y* are independent.

The above conditions are not easily testable from the observed data, but can often be justified on the basis of plausible biological assumptions. Condition 1 is ensured by the usual assumption that the genotype is assigned at meiosis randomly, given the parents' genes, and independently of any confounding phenotype. Condition 2 requires that the genetic variants be reliably associated with the gene expression levels, which is often demonstrated by *cis*-eQTLs with strong regulatory signals. Condition 3 requires that the

genetic variants have no direct effects on the clinical phenotype and can affect the latter only indirectly through the gene expression levels. Owing to the large pool of gene expressions included in genetical genomics studies, the possibility of a strong indirect effect is greatly reduced and hence this condition is also mild and tends to be satisfied in practice.

We discuss here more on these assumptions for genetical genomics data and possible biological complications. Population stratification is a major concern in genome-wide association studies, where the presence of subpopulations with different allele frequencies and different distributions of quantitative traits or risks of disease can lead to spurious associations (Lin and Zeng 2011). Two typical scenarios for the impact of population stratification are illustrated in Figure 1(b) and (c). In Figure 1(b), all three conditions for valid IVs are still satisfied, although the population substructure, represented by an unobserved variable $u$, may strengthen or weaken the associations between the genotype $\mathbf{z}$ and gene expression levels $\mathbf{x}$ required by Condition 2. In Figure 1(c), Condition 3 is violated because conditioning on $\mathbf{x}$ and $w$ alone is insufficient to guarantee the independence of the genotype $\mathbf{z}$ and the clinical phenotype $y$. To deal with possible population stratification, one can regress out the principal components calculated from the genotype data in clinical phenotype regression and gene expression regressions. We also require that the tissue where the gene expressions are measured be relevant to the clinical phenotype. Condition 3 assumes that the genetic variants have no direct effects on the clinical phenotype but manifest their effects through expressions in the relevant tissue. Using a phenotype-irrelevant tissue can potentially lead to violation of Condition 3. It is important, however, to note that strong instruments, a majority of which are most likely *cis*-eQTLs, play a predominant role in our methodology. Recent studies have revealed that these *cis*-eQTLs and their effect sizes are highly conserved across human tissues and populations (Göring 2012; Stranger et al. 2012). This fact helps to lessen the risks of potential assumption violations, although great care should be exercised in justifying the assumptions on a case-by-case basis. See, for example, Didelez and Sheehan (2007) and Lawlor et al. (2008) for more discussion on the complications in Mendelian randomization studies.

Suppose we have $n$ independent observations of $(y, \mathbf{x}, \mathbf{z})$. Denote by $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$, respectively, the $n \times 1$ response vector, the $n \times p$ covariate matrix, and the $n \times q$ genotype matrix. Using the genotypes as instruments, we consider the following linear IV model for the joint modeling of the data $(\mathbf{y}, \mathbf{X}, \mathbf{Z})$:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\eta}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma}_0 + \mathbf{E}, \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\Gamma}_0$ are a $p \times 1$ vector and a $q \times p$ matrix, respectively, of regression coefficients, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ and $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^T$ are an $n \times 1$ vector and an $n \times p$ matrix, respectively, of random errors such that the $(p + 1)$-vector $(\boldsymbol{\varepsilon}_i^T, \eta_i)$ is multivariate normal conditional on $\mathbf{Z}$ with mean zero and covariance matrix $\Sigma = (\sigma_{jk})$. We write $\sigma_{jj} = \sigma_j^2$. Without loss of generality, we assume that each variable is centered about zero so that no intercept terms appear in (1), and that each column of $\mathbf{Z}$ is standardized to have $L_2$ norm $\sqrt{n}$. We emphasize that $\varepsilon_i$ and $\eta_i$ may be correlated because of the arbitrary covariance structure.

In contrast to the usual linear model regressing $\mathbf{y}$ on $\mathbf{X}$, model (1) does not require that the covariate $\mathbf{X}$ and the error $\boldsymbol{\eta}$ be uncorrelated, thus substantially relaxing the assumptions of ordinary regression models and being more appealing in data analysis.

We are interested in making inference for the IV model (1) in the high-dimensional setting where the dimensions $p$ and $q$ can both be much larger than the sample size $n$. In addition to selecting and estimating important covariate effects, since the identities of optimal instruments are unknown, we also regard the identification and estimation of optimal instruments as a variable selection and estimation problem. As is typical in high-dimensional sparse modeling, we assume that model (1) is sparse in the sense that only a small subset of the regression coefficients in $\boldsymbol{\beta}_0$ and $\Gamma_0$ are nonzero. Our goal is, therefore, to identify and estimate the nonzero coefficients in both $\boldsymbol{\beta}_0$ and $\Gamma_0$.

## 3 Regularization Methods and Implementation

In this section, we first study the suboptimality of penalized least squares (PLS) estimators for the causal parameter $\boldsymbol{\beta}_0$. We then propose the 2SR methodology and present an efficient coordinate descent algorithm for implementation. Finally, strategies for tuning parameter selection are discussed.

### 3.1 Suboptimality of Penalized Least Squares

In the classical setting where no regularization is needed, it is well known that the ordinary least squares estimator is inconsistent in the presence of endogeneity, that is, when some of the covariates are correlated with the error term. In high dimensions, without using the instruments, a direct application of one-stage regularization leads to the PLS estimator

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{arg\ min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_\mu\left(|\beta_j|\right) \right\},$$

where $\beta_j$ is the $j$th component of $\boldsymbol{\beta}$ and $p_\mu(\cdot)$ is a penalty function that depends on a tuning parameter $\mu > 0$. With appropriately chosen penalty functions, the PLS estimator has been shown to enjoy superior performance and theoretical properties; see, for example, Fan and Lv (2010) for a review. When the data are generated from the linear IV model (1), however, the usual linear model that assumes the covariates to be uncorrelated with the error term is misspecified, and the PLS estimator $\hat{\boldsymbol{\beta}^*}$ is no longer a reasonable estimator of $\boldsymbol{\beta}_0$. In fact, theoretical results in Lu, Goldberg, and Fine (2012) and Lv and Liu (2013) on misspecified generalized linear models imply that, under some regularity conditions, the PLS estimator $\hat{\boldsymbol{\beta}^*}$ is consistent for the least false parameter $\boldsymbol{\beta}^*$ that minimizes the Kullback–Leibler divergence from the true model, which satisfies the equation

$$\mathbf{X}^T \left(\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\eta}_0 - \mathbf{X}\boldsymbol{\beta}^*\right) = 0, \quad (2)$$

where $\boldsymbol{\eta}_0 = E(\boldsymbol{\eta} \mid \mathbf{X})$. The following proposition shows that there is a nonnegligible gap between $\boldsymbol{\beta}^*$ and the true parameter $\boldsymbol{\beta}_0$.

**Proposition 1**—(Gap between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}_0$). *If* $\|n^{-1} \mathbf{X}^T \boldsymbol{\eta}_0\|_\infty \quad o_P(1)$ *and* $max_{1 \le j \le p}\|\mathbf{x}_j\|_2 = O_p(\sqrt{n})$, *where* $\mathbf{x}_j$ *is the jth column of* $\mathbf{X}$, *then* $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_1 \quad o_P(1)$.

It is interesting to compare the gap $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_1 \quad o_P(1)$ with the minimax optimal rate $O(s\sqrt{(log\, p)/n})$ for high-dimensional linear regression in $L_1$ loss over the $L_0$ ball $\boldsymbol{B}_0(s) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \quad s\}$ (Ye and Zhang 2010; Raskutti, Wainwright, and Yu 2011): the former dominates the latter if $s^2 \log p = o(n)$. Thus, Proposition 1 entails that, in the presence of endogeneity, any optimal procedure for estimating $\boldsymbol{\beta}^*$, such as the PLS estimator with $L_1$ or other sparsity-inducing penalties, is suboptimal for estimating $\boldsymbol{\beta}_0$ as long as $s^2 \log p = o(n)$. Moreover, since by definition $\mathbf{X}\boldsymbol{\beta}^*$ is the orthogonal projection of $E(\mathbf{y}|\mathbf{X})$ onto the column space of $\mathbf{X}$, the component $\mathbf{X}\boldsymbol{\beta}^*$ is generally closer to the expected response than $\mathbf{X}\boldsymbol{\beta}_0$. This will likely lead to a larger proportion of variance explained for the PLS method. Hence, to assess how well the fitted model predicts the response in model (1), it is more meaningful to compare the predicted values $\mathbf{X}\hat{\boldsymbol{\beta}}$ to the causal component $\mathbf{X}\boldsymbol{\beta}_0$.

### 3.2 Two-Stage Regularization

One standard way of eliminating endogeneity is to replace the covariates by their expectations conditional on the instruments. This idea leads to the classical two-stage least squares (2SLS) method (Anderson 2005), in which the covariates are first regressed on the instruments and the response is then regressed on the first-stage predictions of the covariates. The performance of the 2SLS method deteriorates drastically or become inapplicable, however, as the dimensionality of covariates and instruments increase. We thus propose to apply regularization methods to cope with the high dimensionality in both stages of the 2SLS method, resulting in the following 2SR methodology.

**Stage 1**—The goal of the first stage is to identify and estimate the nonzero effects of the instruments and obtain the predicted values of the covariates. Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix. The first-stage regularized estimator is defined as

$$\hat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{\Gamma} \in \mathbb{R}^{q \times p}}{arg\, min} \left\{ \frac{1}{2n}\|\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma}\|_F^2 + \sum_{i=1}^{q}\sum_{j=1}^{p} p_{\lambda_j}(|\gamma_{ij}|) \right\}, \quad (3)$$

where $\gamma_{ij}$ is the $(i, j)$th entry of the matrix $\boldsymbol{\Gamma}$, $p_{\lambda_j}(\cdot)$ is a sparsity-inducing penalty function to be discussed later, and $\lambda_j > 0$ are tuning parameters that control the strength of the first-stage regularization. After the estimate $\hat{\boldsymbol{\Gamma}}$ is obtained, the predicted value of $\mathbf{X}$ is formed by $\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\Gamma}}$.

**Stage 2**—Substituting the first-stage prediction $\hat{\mathbf{X}}$ for $\mathbf{X}$, we proceed to identify and estimate the nonzero effects of the covariates. The second-stage regularized estimator is defined as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{arg\, min} \left\{ \frac{1}{2n}\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_\mu(|\beta_j|) \right\}, \quad (4)$$

where $\beta_j$ is the $j$th component of $\boldsymbol{\beta}$, $p_\mu(\cdot)$ is a sparsity-inducing penalty function as before, and $\mu > 0$ is a tuning parameter that controls the strength of the second-stage regularization. We thus obtain the pair $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}})$ as our final estimator for the regression parameter $(\boldsymbol{\beta}_0, \boldsymbol{\Gamma}_0)$ in model (1).

We consider the following three choices of the penalty function $p_\lambda(t)$ for $t \geq 0$: (a) the $L_1$ penalty or Lasso (Tibshirani 1996), $p_\lambda(t) = \lambda t$; (b) the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001),

$$p_\lambda(t) = \lambda \int_0^t \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\} d\theta, \quad a > 2;$$

and (c) the minimax concave penalty (MCP) (Zhang 2010),

$$p_\lambda(t) = \int_0^t \frac{(a\lambda - \theta)_+}{a} d\theta, \quad a > 1.$$

The SCAD and MCP penalties have an additional tuning parameter $a$ to control the shape of the function. These penalty functions have been widely used in high-dimensional sparse modeling and their properties are well understood in ordinary regression models (e.g., Fan and Lv 2010). Moreover, the fact that these penalties belong to the class of quadratic spline functions on $[0, \infty)$ allows for a closed-form solution to the corresponding penalized least squares problem in each coordinate, leading to very efficient implementation via coordinate descent (e.g., Mazumder, Friedman, and Hastie 2011).

### 3.3 Implementation

We now present an efficient coordinate descent algorithm for solving the optimization problems (3) and (4) with the Lasso, SCAD, and MCP penalties. We first note that the matrix optimization problem (3) can be decomposed into $p$ penalized least squares problems,

$$\hat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}_j \in \mathbb{R}^q}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{x}_j - \boldsymbol{Z}\boldsymbol{\gamma}_j\|_2^2 + \sum_{i=1}^q p_{\lambda_j}(|\gamma_{ij}|) \right\}, \quad (5)$$

where $\mathbf{x}_j$ is the $j$th column of the covariate matrix $\mathbf{X}$ and $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{qj})^T$. The univariate solution to the unpenalized least squares problem (5) is given by

$\tilde{\gamma}_{ij} = n^{-1} \left( \mathbf{x}_j - \sum_{k \neq i} \gamma_{kj} \mathbf{z}_k \right)^T \mathbf{z}_i = n^{-1} \mathbf{r}_j^T \mathbf{z}_i + \gamma_{ij}$, where $\mathbf{z}_j$ is the $j$th column of the instrument matrix $\mathbf{Z}$, $\mathbf{r}_j = \mathbf{x}_j - \sum_{k=1}^q \gamma_{kj} \mathbf{z}_k$ is the current residual, and we have used the fact $n^{-1} \mathbf{z}_j^T \mathbf{z}_j = 1$ due to standardization. The penalized univariate solution, then, can be obtained by $\gamma_{ij} = S(\tilde{\gamma}_{ij}; \lambda)$, where $S(\cdot; \lambda)$ is a thresholding operator defined for Lasso, SCAD, and MCP, respectively, as $S_{\text{Lasso}}(t; \lambda) = \text{sgn}(t)(|t| - \lambda)_+$,

$$S_{SCAD}(t;\lambda) = \begin{cases} sgn(t)(|t|-\lambda)_+, & if & |t| \le 2\lambda, \\ sgn(t)\frac{|t|-\lambda a/(a-1)}{1-1/(a-1)}, & if & 2\lambda < |t| \le a\lambda, \\ t, & if & |t| > a\lambda, \end{cases}$$

and

$$S_{MCP}(t;\lambda) = \begin{cases} sgn(t)\frac{(|t|-\lambda)_+}{1-1/a}, & if & |t| \le a\lambda, \\ t, & if & |t| > a\lambda. \end{cases}$$

Similarly, if the $j$th column $\hat{\mathbf{x}_j}$ of the first-stage prediction matrix $\hat{\mathbf{X}}$ is standardized to have $L_2$ norm $\sqrt{n}$, the penalized univariate solution for the optimization problem (4) is given by $\beta_j = S(\tilde{\beta}_j; \mu)$, where $\tilde{\beta}_j = n^{-1}\mathbf{r}^T\hat{\mathbf{x}}_j + \beta_j$ is the unpenalized univariate solution and $\mathbf{r} = \mathbf{y} - \sum_{k=1}^{p}\beta_k\hat{\mathbf{x}}_k$ is the current residual. We summarize the coordinate descent algorithm for computing the 2SR estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}})$ in Algorithm 1.

---

**Algorithm 1**. Coordinate descent for the 2SR estimator

---

Initialize: $\boldsymbol{\beta}, \boldsymbol{\Gamma} \leftarrow \mathbf{0}$ or warm starts, $\lambda_1, \dots, \lambda_p, \mu > 0$

**for** $j = 1, \dots, p$ **do**

    **while** $\boldsymbol{\gamma}_j$ not convergent **do**

        **for** $i = 1, \dots, q$ **do**

            $\gamma_{ij} \leftarrow S(\tilde{\gamma}_{ij}; \lambda_j)$

$\hat{\boldsymbol{\Gamma}} \leftarrow (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p), \hat{\mathbf{X}} \leftarrow \mathbf{Z}\hat{\boldsymbol{\Gamma}}$

**while** $\boldsymbol{\beta}$ not convergent **do**

    **for** $j = 1, \dots, p$ **do**

        $\beta_j \leftarrow S(\tilde{\beta}_j; \mu)$

$\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}$

---

The convergence of Algorithm 1 to a local minimum for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Gamma}}$ follows from the convergence properties of coordinate descent algorithms for penalized least squares; see, for example, Lin and Lv (2013). Since the SCAD and MCP penalties are nonconvex, convergence to a global minimum is not guaranteed in general. In practice, coordinate descent algorithms are often used to produce a solution path over a grid of regularization parameter values, with warm starts from nearby solutions. In this case, the algorithm tends to find a sparse local solution with superior performance.

### 3.4 Tuning parameter selection

The 2SR method has $p + 1$ regularization parameters $\lambda_1, \dots, \lambda_p$ and $\mu$ to be tuned. We propose to choose the optimal tuning parameters by $K$-fold cross-validation. Specifically, we define the cross-validation error for $\lambda_j$ and $\mu$ by

$$CV(\lambda_j) = \frac{1}{K} \sum_{k=1}^{K} \left\| \mathbf{x}_j^{(k)} - \mathbf{Z}^{(k)} \hat{\boldsymbol{\gamma}}_j^{(-k)}(\lambda_j) \right\|_2^2 \quad (6)$$

and

$$CV(\mu) = \frac{1}{K} \sum_{k=1}^{K} \left\| \mathbf{y}^{(k)} - \hat{\mathbf{X}}^{(k)} \hat{\boldsymbol{\beta}}^{(-k)}(\mu) \right\|_2^2, \quad (7)$$

respectively, where $\mathbf{x}_j^{(k)}$, $\mathbf{Z}^{(k)}$, $\mathbf{y}^{(k)}$, and $\hat{\mathbf{X}}^{(k)}$ are vectors/matrices for the $k$th part of the sample, and $\hat{\boldsymbol{\Gamma}}^{(-k)}(\lambda_j)$ and $\hat{\boldsymbol{\beta}}^{(-k)}(\mu)$ are the estimates obtained with the $k$th part removed. In view of the fact that in typical genetical genomics studies, both $p$ and $q$ can be in the thousands, it is necessary to reduce the search dimension of tuning parameters. To this end, we propose to first determine the optimal $\lambda_j$ that minimizes the criterion (6), for $j = 1,\ldots, p$, and then, with $\lambda_1, \ldots, \lambda_p$ fixed, find the optimal $\mu$ that minimizes the criterion (7). The practical performance of this search strategy proves to be very satisfactory.

## 4 Theoretical Properties

In this section, we investigate the theoretical properties of the 2SR estimators. Through our theoretical analysis, we wish to understand (a) the impact of the dimensionality of covariates and instruments as well as other factors on the quality of the regularized estimators, and (b) the role of the two-stage regularization in providing performance guarantees for the regularized estimators, especially for the second-stage estimators. To address (a), we adopt a *nonasymptotic* framework that allows the dimensionality of covariates and instruments to vary freely and thus can both be much larger than the sample size; to address (b), we impose conditions only on the instrument matrix $\mathbf{Z}$, and treat the covariate matrix $\mathbf{X}$ and the first-stage prediction $\hat{\mathbf{X}}$ as *nondeterministic*. The major challenge arises in the characterization of the second-stage estimation, where the "design matrix" $\hat{\mathbf{X}}$ is neither fixed nor a random design sampled from a known distribution. Therefore, existing formulations for the high-dimensional analysis of ordinary regression models are inapplicable to our setting. We also stress that our theoretical analysis is essentially different from the recent developments in sparse IV models. The methods and results developed by Gautier and Tsybakov (2011) and Fan and Liao (2012) involve only one-stage estimation and regularization. The secondstage estimation considered by Belloni et al. (2012) is of fixed dimensionality, which allows them to focus on estimation efficiency based on standard asymptotic analysis. Owing to the complications involved in the analysis of a general penalty, we first consider the representative case of $L_1$ regularization in Section 4.1, which allows us to obtain clean conditions providing important insights. We then present in Section 4.2 a generalization of the theory, which is applicable to a much broader class of regularization methods.

### 4.1 $L_1$ Regularization

We begin by introducing some notation. Let $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the matrix 1-norm and $\infty$-norm, respectively, that is, $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$ for any matrix $\mathbf{A} = (a_{ij})$. For any vector $\boldsymbol{a}$, matrix $\mathbf{A}$, and index sets $I$ and $J$, let $\boldsymbol{a}_J$ denote the subvector formed

by the $j$th components of $\boldsymbol{a}$ with $j \in J$, and $\mathbf{A}_{IJ}$ the submatrix formed with the $(i, j)$th entries of $\mathbf{A}$ with $i \in I$ and $j \in J$. Also, denote by $J^c$ the complement of $J$ and $|J|$ the number of elements in $J$. Following Bickel, Ritov, and Tsybakov (2009), define the restricted eigenvalue for an $n \times m$ matrix $\mathbf{A}$ and $1 \le s \le m$ by

$$\kappa(\mathbf{A}, s) = \min_{|J| \le s} \quad \min_{\substack{\boldsymbol{\delta} \neq 0 \\ \|\boldsymbol{\delta}_{J^c}\|_1 \le 3\|\boldsymbol{\delta}_J\|_1}} \frac{\|\mathbf{A}\boldsymbol{\delta}\|_2}{\sqrt{n}\|\boldsymbol{\delta}_J\|_2}.$$

Let supp($\boldsymbol{a}$) denote the support of a vector $\boldsymbol{a} = (a_j)$, that is, supp($\boldsymbol{a}$) = $\{j: a_j \neq 0\}$. Define the sparsity levels $r = \max_{1 \le j \le p} |\mathrm{supp}(\boldsymbol{\gamma}_{0j})|$ and $s = |\mathrm{supp}(\boldsymbol{\beta}_0)|$, and the first-stage noise level $\sigma_{\max}$ = $\max_{1 \le j \le p} \sigma_j$, where $\boldsymbol{\gamma}_{0j}$ is the $j$th column of $\boldsymbol{\Gamma}_0$. We consider the parameter space with $\|\boldsymbol{\Gamma}_0\|_1 \le L$ and $\|\boldsymbol{\beta}_0\|_1 \le M$ for some constants $L, M > 0$.

To derive nonasymptotic bounds on the estimation and prediction loss of the regularized estimators $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\beta}}$, we impose the following conditions:

    **(C1)**    There exists $\kappa_1 > 0$ such that $\kappa(\mathbf{Z}, r) \ge \kappa_1$.

    **(C2)**    There exists $\kappa_2 > 0$ such that $\kappa(\mathbf{Z}\boldsymbol{\Gamma}_0, s) \ge \kappa_2$.

We emphasize that dimensions $p$ and $q$, sparsity levels $r$ and $s$, and lower bounds $\kappa_1$ and $\kappa_2$ may all depend on the sample size $n$; we have suppressed the dependency for notational simplicity. Conditions (C1) and (C2) are analogous to those in Bickel, Ritov, and Tsybakov (2009) for usual linear models, and require that the matrices $\mathbf{Z}$ and $\mathbf{Z}\boldsymbol{\Gamma}_0$ be well behaved over some restricted sets of sparse vectors. One important difference, however, is that Condition (C2) is imposed on the conditional expectation matrix $\mathbf{Z}\boldsymbol{\Gamma}_0$ of $\mathbf{X}$, rather than the first-stage prediction matrix, or the second-stage design matrix, $\hat{\mathbf{X}}$. This condition is more natural in our context, but poses new challenges for the analysis.

The estimation and prediction quality of the first-stage estimator $\hat{\boldsymbol{\Gamma}}$ is characterized by the following result.

**Theorem 1**—(Estimation and prediction loss of $\hat{\boldsymbol{\Gamma}}$). *Under Condition (C1), if we choose*

$$\lambda_j = C\sigma_j \sqrt{\frac{\log p + \log q}{n}} \quad (8)$$

*with a constant $C \ge 2\sqrt{2}$, then with probability at least $1 - (pq)^{1 - C^2/8}$, the regularized estimator $\hat{\boldsymbol{\Gamma}}$ defined by* (3) *with the $L_1$ penalty satisfies*

$$\|\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\|_1 \le \frac{16C}{\kappa_1^2}\sigma_{max}r\sqrt{\frac{\log p + \log q}{n}}$$

*and*

$$\|\boldsymbol{Z}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0)\|_F^2 \leq \frac{16C^2}{\kappa_1^2}\sigma_{max}^2\, pr\,(log\,p + log\,q).$$

Using the nonasymptotic bounds provided by Theorem 1, we can show that Condition (C2) also holds with high probability for the matrix $\mathbf{X} = \mathbf{Z}\hat{\boldsymbol{\Gamma}}$ with a smaller $\kappa_2$; see Lemma A.1 in the Appendix. This allows us to establish the following result concerning the estimation and prediction quality of the second-stage estimator $\hat{\boldsymbol{\beta}}$.

**Theorem 2**—(Estimation and prediction loss of $\hat{\boldsymbol{\beta}}$). *Under Conditions (C1) and (C2), if the regularization parameters $\lambda_j$ are chosen as in* (8) *and satisfy*

$$\lambda_{max}(2L + \lambda_{max}) \leq \frac{\kappa_1^2\kappa_2^2}{32^2 rs}, \quad (9)$$

*where $\lambda_{max} = \max_{1 \leq j \leq p} \lambda_j$, then there exist constants $c_0, c_1, c_2 > 0$ such that, if we choose*

$$\mu = \frac{C_0}{\kappa_1}\sqrt{\frac{r\,(log\,p + log\,q)}{n}}, \quad (10)$$

*where $C_0 = c_0 L \max(\sigma_{p+1}, M\,\sigma_{max})$, then with probability at least $1 - c_1(pq)^{-c_2}$, the regularized estimator $\hat{\boldsymbol{\beta}}$ defined by* (4) *with the $L_1$ penalty satisfies*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq \frac{64C_0}{\kappa_1\kappa_2^2}s\sqrt{\frac{r\,(log\,p + log\,q)}{n}}$$

*and*

$$\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq \frac{64C_0^2}{\kappa_1^2\kappa_2^2}rs(log\,p + log\,q).$$

We now turn to the model selection consistency of $\hat{\boldsymbol{\beta}}$. Let $\mathbf{C} = n^{-1}\,(\mathbf{Z}\boldsymbol{\Gamma}_0)^T\,\mathbf{Z}\boldsymbol{\Gamma}_0$, $S = \mathrm{supp}(\boldsymbol{\beta}_0)$, and $\phi = \|(\mathbf{C}_{SS})^{-1}\|_\infty$. Define the minimum signal $b_0 = \min_{j \in S}|\beta_{0j}|$, where $\beta_{0j}$ is the $j$th component of $\boldsymbol{\beta}_0$. To study the model selection consistency, we replace Condition (C2) by the following condition:

**(C3)** There exists a constant $0 < a \leq 1$ such that $\|\mathbf{C}_{S^c S}(\mathbf{C}_{SS})^{-1}\|_\infty \leq 1 - a$.

Condition (C3) is in the same spirit as the irrepresentability condition in Zhao and Yu (2006) for the ordinary Lasso problem. Although Condition (C3) is placed on the covariance matrix of $\mathbf{Z}\boldsymbol{\Gamma}_0$, we can apply Theorem 1 to show that it also holds with high probability for the covariance matrix of $\mathbf{X} = \mathbf{Z}\hat{\boldsymbol{\Gamma}}$ with a smaller $a$; see Lemma A.3 in the Appendix. The model selection consistency of $\hat{\boldsymbol{\beta}}$, along with a closely related $L_\infty$ bound, is established by the following result.

**Theorem 3**—(*Model selection consistency of* $\hat{\beta}$). *Under Conditions (C1) and (C3), if the regularization parameter* $\lambda_j$ *are chosen as in* (8) *and satisfy*

$$\frac{16\varphi}{\kappa_1^2} rs\lambda_{max}(2L+\lambda_{max}) \leq \frac{\alpha}{4-\alpha}, \quad (11)$$

*then there exist constants* $c_0, c_1, c_2 > 0$ *such that, if the regularization parameter* $\mu$ *is chosen as in* (10) *and the minimal signal satisfies*

$$b_0 > \frac{2}{2-\alpha}\varphi\mu,$$

*then with probability at least* $1 - c_1 (pq)^{-c_2}$, *there exists a regularized estimator* $\hat{\beta}$ *defined by* (4) *with the* $L_1$ *penalty that satisfies*

**a.** *(Sign consistency)* $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta_0)$, *and*

**b.** *($L_\infty$ loss)*

$$\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_\infty \leq \frac{2C_0\varphi}{(2-\alpha)\kappa_1}\sqrt{\frac{r(logp+logq)}{n}}.$$

Theorem 3 shows that the second-stage estimator $\hat{\beta}$ has the weak oracle property in the sense of Lv and Fan (2009). Two remarks are in order. First, the validity of our arguments for Theorems 2 and 3 relies on the first-stage regularization only through the estimation and prediction bounds given in Theorem 1; this allows the arguments to be generalized to a generic class of regularization methods for the first stage, which will be explored in Section 4.2. Second, a key difference from the high-dimensional analysis of the usual linear model is that $\mathbf{X}$ and $\eta$ may be correlated, and we have to make good use of the assumption that $\mathbf{E}$ and $\eta$ are mean zero conditional on $\mathbf{Z}$; see Lemma A.2 in the Appendix.

Theorems 1–3 deliver the important message that dimensions $p$ and $q$ contribute only a logarithmic factor to the estimation and prediction loss, and thus are both allowed to grow exponentially with the sample size $n$. Note that (9) and (11) are critical assumptions relating the first-stage regularization parameter $\lambda_{max}$ to the key quantities in the second stage. To gain further insight into the dimension restrictions, suppose for simplicity that $\kappa_1$, $\kappa_2$, and $\varphi$ are constants; then (9) and (11) hold for sufficiently large $n$ provided that

$$r^2 s^2 (logp + logq) = o(n). \quad (12)$$

This implies that dimensions $p$ and $q$ can grow at most as $e^{o(n)}$ and sparsity levels $r$ and $s$ can grow as $o(\sqrt{n})$, if the other quantities are fixed. Moreover, when $q$ and $r$ are also fixed, the relation (12) reduces to $s^2 \log p = o(n)$. In view of the remark following Proposition 1 and the $L_1$ bound given in Theorem 2, we see that the 2SR estimator achieves the optimal rate for estimating $\beta_0$, which is asymptotically faster than that of the PLS estimator.

## 4.2 General Regularization

We next present a theory for the second-stage estimator $\hat{\beta}$ that generalizes the results in Section 4.1 in two aspects. First, we allow the first-stage regularization to be arbitrary provided that certain nonasymptotic bounds are satisfied. Second, we allow the secondstage regularization to adopt a generic form of sparsity-inducing penalties, thus including the Lasso, SCAD, and MCP as special cases. Specifically, we impose the following conditions:

**(C4)** There exist $e_1$, $e_2$, and probability $\pi_0$, which may depend on $(n, p, q, r)$, such that the first-stage estimator $\hat{\Gamma}$ satisfies $\|\hat{\Gamma} - \Gamma_0\|_1 \leq e_1$ and

$$max_{1 \leq j \leq p} n^{-1}\|\boldsymbol{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})\|_2^2 \leq e_2 \text{ with probability } 1 - \pi_0.$$

**(C5)** The penalty function $\rho_\mu(\cdot)$ is increasing and concave on $[0, \infty)$, and has a continuous derivative $\rho'_\mu(\cdot)$ on $(0, \infty)$. In addition, $\rho'_\mu(\cdot)$ is increasing in $\mu$, and $\rho'_\mu(0+) \equiv \rho'(0+) \in (0, \infty)$ is independent of $\mu$.

Moreover, we replace Condition (C3) by the weaker assumption:

**(C6)** There exist constants $0 < \alpha \leq 1, 0 \leq v \leq 1/2$, and $c \geq 1$ such that

$$\|\boldsymbol{C}_{S^cS}(\boldsymbol{C}_{SS})^{-1}\|_\infty \leq \left\{(1 - \alpha)\frac{\rho'(0+)}{\rho'_\mu(b_0/2)}\right\} \wedge (cn^v).$$

The family of penalty functions in Condition (C5) and a similar condition to (C6) were studied by, for example, Fan and Lv (2011) for generalized linear models; see the discussion therein for the motivation of these conditions. In particular, Condition (C5) captures several desirable properties of commonly used sparsity-inducing penalties, and allows us to establish a unified theory for these penalties. Condition (C6) is generally weaker than Condition (C3), since concavity implies that $\rho'(0+) \geq \rho'_\mu(b_0/2)$ and the right-hand side can be much larger than $1 - \alpha$. Note that for the $L_1$ penalty, $\rho'_\mu(\cdot) \equiv 1$ and this condition reduces to Condition (C3). For SCAD and MCP, when the signals are sufficiently strong such that $b_0/2 \geq a\mu$, we have $\rho'_\mu(b_0/2) = 0$ and the right-hand side can grow at most as $O(\sqrt{n})$.

Following Lv and Fan (2009), for any vector $\boldsymbol{\theta} = (\theta_j)$ with $\theta_j \geq 0$ for all $j$, define the local concavity of the penalty function $\rho_\mu(\cdot)$ at point $\boldsymbol{\theta}$ by

$$\tau(\rho_\mu; \boldsymbol{\theta}) = \lim_{\varepsilon \to 0+} max_j \sup_{|\theta_j| - \varepsilon < t_1 < t_2 < |\theta_j| + \varepsilon} \left\{-\frac{\rho'_\mu(t_2) - \rho'_\mu(t_1)}{t_2 - t_1}\right\}.$$

Further, define

$$\tau_0 = \sup\{\tau(\rho_\mu; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^s, \|\boldsymbol{\theta} - \boldsymbol{\beta}_{0S}\|_\infty \leq b_0/2\}$$

and

$$\mu_0 = \Lambda_{min}(\mathbf{C}_{SS}) - \mu\tau_0.$$

The following result generalizes Theorem 3 and establishes the model selection consistency and weak oracle property of $\hat{\beta}$.

**Theorem 4**—(Weak oracle property of $\hat{\beta}$). *Under Conditions (C4)–(C6), if $\mu_0 > 0$ and the first-stage error bounds $e_1$ and $e_2$ satisfy*

$$s(2Le_1 + e_2) \leq \frac{\alpha}{(4 - \alpha)\varphi} \wedge \frac{(\mu_0/2)^2}{s}, \quad (13)$$

*then there exist constants $c_0$, $c_1$, $c_2 > 0$ such that, if we choose*

$$\mu \geq C_0 n^\nu \sqrt{\frac{\log p + \log q}{n}} \vee e_2,$$

*where $C_0 = c_0 L \max(\sigma_{p+1}, M\sigma_{\max}, M)$, and the minimum signal satisfies*

$$b_0 \geq \frac{7}{2}\varphi\mu\rho'(0+), \quad (14)$$

*then with probability at least $1 - \pi_0 - c_1(pq)^{-c_2}$, there exists a regularized estimator $\hat{\beta}$ defined by* (4) *that satisfies*

a. *(Sign consistency)* $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta_0)$, *and*

b. *($L_\infty$ loss)*

$$\|\hat{\beta}_S - \beta_{0S}\|_\infty \leq \frac{7}{4}\varphi\mu\rho'(0+).$$

Compared with Theorem 3, Theorem 4 justifies the advantages of concave penalties such as SCAD and MCP in that model selection consistency and weak oracle property are established under substantially relaxed conditions. To understand the implications of the assumption (13), note that, for the $L_1$ penalty, Theorem 1 gives $e_1 = O(r\sqrt{\log p + \log q})/n)$, and $e_2 = O(r(\log p + \log q)/n)$, and the term involving $\mu_0$ is not needed. Assuming for simplicity that $\varphi$ is constant, (13) reduces to the dimension restriction (12). Therefore, (13) plays essentially the same role as the assumption (11), but applies to a generic first-stage estimator. Moreover, taking $e_2$ as above and $\nu = 0$, we obtain the same rate of convergence for the $L_\infty$ loss as in Theorem 3.

## 5 Simulation Studies

In this section, we report on simulation studies to evaluate the performance of the proposed 2SR method with the Lasso, SCAD, and MCP penalties. We compare the proposed method with the PLS estimators with the same penalties that do not utilize the instruments, as well

as the PLS and 2SR oracle estimators that knew the relevant covariates and instruments in advance. We are particularly interested in investigating how the PLS and 2SR methods perform differently in relation to the sample size and how the dimensionality and instrument strength affect the performance of the 2SR method.

We first consider the case where the dimensions $p$ and $q$ are moderately high and smaller than the sample size $n$. Four models were examined, with $(n, p, q) = (200, 100, 100)$ in Model 1 and $(400, 200, 200)$ in Models 2–4. We first generated the coefficient matrix $\Gamma_0$ by sampling $r = 5$ nonzero entries of each column from the uniform distribution $U([-b, -a] \cup [a, b])$. To represent different levels of instrument strength, we took $(a, b) = (0.75, 1)$ for strong instruments in Models 1 and 2, and $(a, b) = (0.5, 0.75)$ for weak instruments in Model 3. In Model 4, which reflects a more realistic setting, we sampled $r = 50$ nonzero entries, consisting of 5 strong/weak instruments with $(a, b) = (0.5, 1)$ and 45 very weak instruments with $(a, b) = (0.05, 0.1)$. Similarly, we generated the coefficient vector $\beta_0$ by sampling $s = 5$ nonzero components from $U([-1, -0.5] \cup [0.5, 1])$. The covariance matrix $\Sigma = (\sigma_{ij})$ was specified as follows: We first set $\sigma_{ij} = (0.2)^{|i-j|}$ for $i, j = 1, \ldots, p$, and $\sigma_{p+1, p+1} = 1$; in addition to the five $\sigma_{j,p+1}$'s corresponding to the nonzero components of $\beta_0$, we sampled another five entries from the last column of $\Sigma$; we then set these ten entries to 0.3 and let $\sigma_{p+1, j} = \sigma_{j, p+1}$ for $j = 1, \ldots, p$. Note that the nonzero $\sigma_{j, p+1}$'s were intended to cause both effect modifications and spurious associations for the PLS method. Finally, the instrument matrix $\mathbf{Z}$ was generated by sampling each entry independently from Bernoulli$(p_0)$, where $p_0 = 0.5$ in Models 1–3 and $p_0 \sim U([0, 0.5])$ in Model 4, and the covariate matrix $\mathbf{X}$ and the response vector y were generated accordingly.

Since the PLS method provides no estimates for the coefficient matrix $\Gamma_0$ and our main interest is in how the estimation of $\beta_0$ can be improved by the 2SR method, we focus our comparisons on the second-stage estimation. Five measures on estimation, prediction, and model selection qualities were used to assess the performance of each method. The $L_1$ estimation loss $\|\hat{\beta} - \beta_0\|_1$ and the prediction loss $n^{-1/2}\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2$ quantify the estimation and prediction performance, respectively. The model selection performance is characterized by the number of true positives (TP), the model size, and the Matthews correlation coefficient (MCC). Here, positives refer to nonzero estimates. The MCC is a measure on the correlation between the observed and predicted binary classifications and is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TN, FP, and FN denote the number of true negatives, false positives, and false negatives, respectively; a larger MCC indicates a better variable selection performance. In all simulations, we applied ten-fold cross-validation to choose the optimal tuning parameters and averaged each performance measure over 50 replicates.

The simulation results for Models 1–4 are summarized in Table 1. From the table we see that the 2SR method improved on the performance of the PLS method substantially in all cases. The improvement on model selection performance was most remarkable. The PLS

method selected an exceedingly large model with many false positives because of its failure in distinguishing between the true and confounding effects, whereas the 2SR method resulted in a much sparser model and controlled the number of false positives at a reasonable level. As a result, the 2SR method had a much higher MCC than the PLS method, indicating a superior variable selection performance. The estimation and prediction performance of the PLS method was also greatly compromised by the confounding effects, and the 2SR method achieved a dramatic improvement on the $L_1$ estimation loss and a considerable improvement on the prediction loss. The comparisons between Model 2 and the weaker instrument settings, Models 3 and 4, suggest that a weaker instrument strength tends to decrease the performance of the 2SR method, as expected, especially on the estimation and prediction quality. We observe, however, that the model selection quality was only slightly affected and the overall performance of the 2SR method was still very satisfactory.

To facilitate performance comparisons among different methods with varying sample size, Figure 2 depicts the trends in three performance measures with the dimensions $p = q = 100$ fixed and the sample size $n$ varying from 200 to 1500. It is clear from Figure 2 that the performance of the 2SR method improves consistently as the sample size increases, whereas the PLS method does not in general see performance gain and may even deteriorate. Moreover, a closer look at the tails of the curves for the 2SR method with different penalties reveals certain advantages of SCAD and MCP over the Lasso. There seems to be a nonvanishing gap between the Lasso and oracle estimators, which agrees with the existing theory in the context of linear regression that the Lasso does not possess the oracle property (Zou 2006).

We further study the case where the dimensions $p$ and $q$ are ultrahigh and larger than the sample size $n$. We considered four models with the same settings as in Models 1–4, except that $(n, p, q) = (300, 600, 600)$ in Model 5 and $(500, 1000, 1000)$ in Models 6–8. Table 2 summarizes the simulation results for Models 5–8, and Figure 3 shows the performance curves with $p = q = 600$ fixed and $n$ varying from 200 to 1500. Trends in performance comparisons among different methods are similar to those in Table 1 and Figure 2, demonstrating the advantages of the 2SR method over the PLS method. We observe that, although the ultrahigh dimensionality caused the 2SR method to select a larger model and resulted in a slightly lower MCC than in the previous settings, the performance of the 2SR method still compared favorably to the PLS method and the difference was pronounced for moderate sample sizes. These results suggest that the dimensionality has only mild impact on the performance of the 2SR method compared with the sample size, in agreement with our theoretical results in Section 4.

## 6 Analysis of Mouse Obesity Data

To illustrate the application of the proposed method, in this section we present results from the analysis of a mouse obesity data set described by Wang et al. (2006). The study includes an F2 intercross of 334 mice derived from the inbred strains C57BL/6J and C3H/HeJ on an apolipoprotein E (ApoE) null background, which were fed a high-fat Western diet from 8 to 24 weeks of age. The mice were genotyped using 1327 SNPs at an average density of 1.5 cM across the whole genome, and the gene expressions of the liver tissues of these mice

were profiled on microarrays that include probes for 23,388 genes. Data on several obesity-related clinical traits were also collected on the animals. The genotype, gene expression, and clinical data are available for download, respectively, at http://www.genetics.org/cgi/content/full/genetics.110.116087/DC1, ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/SeriesMatrix/GSE2814/, and http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/MouseWeight/. Since the mice came from the same genetic cross, population stratification is unlikely an issue. Also, a study using a superset of these data demonstrated that most *cis*-eQTLs were highly replicable across mouse crosses, tissues, and sexes (van Nas et al. 2010). Therefore, the three assumptions for valid IVs seem to be plausible.

After the individuals, SNPs, and genes with a missing rate greater than 0.1 were removed, the remaining missing genotype and gene expression data were imputed using the Beagle approach (Browning and Browning 2007) and nearest neighbor averaging (Troyanskaya et al. 2001), respectively. Merging the genotype, gene expression, and clinical data yielded a complete data set with $q = 1250$ SNPs and 23,184 genes on $n = 287$ (144 female and 143 male) mice. To enhance the interpretability and stability of the results, we focus on the $p = 2825$ genes that can be mapped to the Mouse Genome Database (MGD) (Eppig et al. 2012) and have standard deviation of gene expression levels greater than 0.1. The latter criterion is reasonable because gene expressions of too small variation are typically not of biological interest and suggest that the genetic perturbations may not be sufficiently strong for the genetic variants to be used as instruments.

Our goal is to jointly analyze the genotype, gene expression, and clinical data to identify important genes related to body weight. In order to utilize data from both sexes, we first adjusted the body weight for sex by fitting a marginal linear regression model with sex as the covariate and subtracting the estimated sex effect from the body weight. We then applied the proposed 2SR method with the Lasso, SCAD, and MCP penalties to the data set with adjusted body weight as the response. For comparison, we also applied the PLS method to the same data set, and used ten-fold cross-validation to choose the optimal tuning parameters for both methods. The models selected by cross-validation include 110 (Lasso), 49 (SCAD), and 16 (MCP) genes for the PLS method, and include 37 (Lasso), 15 (SCAD), and 9 (MCP) genes for the 2SR method. The selected models resulted in an adjusted $R^2$ of 0.894 (Lasso), 0.833 (SCAD), and 0.820 (MCP) for the PLS method, and 0.594 (Lasso), 0.581 (SCAD), and 0.579 (MCP) for the 2SR method, which is consistent with our remark following Proposition 1. Since we have no knowledge of the causal component $\mathbf{X}\boldsymbol{\beta}_0$ for the real data, a direct comparison between the PLS and 2SR methods in assessing the model fit is not possible. Nevertheless, we observe that the 2SR method produced a much sparser model with reasonably high proportion of variance explained.

To gain insight into the stability of the selection results, we followed the idea of stability selection (Meinshausen and Bühlmann 2010) to compute the selection probability of each gene over 100 subsamples of size $\lfloor n/2 \rfloor$ for each fixed value of the regularization parameter $\mu$. The resulting stability paths for different methods are displayed in Figure 4. It is interesting to observe that, among the genes with maximum selection probability at least 0.4, only 5 (Lasso), 3 (SCAD), and 0 (MCP) genes are common to both the PLS and 2SR methods. As can be seen from Figure 4, these few genes, which are reasonably conjectured

to be among the truly relevant ones, stand out more clearly in the stability paths of the 2SR method. Moreover, the overall stability paths of the 2SR method seem less noisy and hence can be more useful for distinguishing the most important genes from the irrelevant ones.

Table 3 lists the genes that were chosen by stability selection with maximum selection probability at least 0.5 using the 2SR method with three different penalties. Among these 17 genes, only three were also selected by the PLS method. This includes insulin-like growth factor binding protein 2 (Igfbp2), which has been shown to protect against the development of obesity (Wheatcroft et al. 2007). Among the genes identified only by the 2SR method, apolipoprotein A-IV (Apoa4) plays an important role in lipoprotein metabolism and has been implicated in the control of food intake in rodents (Tso, Sun, and Liu 2004); it inhibits gastric emptying and serves as a satiety factor in response to ingestion of dietary fat. Apoa4 also acts as an enterogastrone, a humoral inhibitor of gastric acid secretion and motility (Okumura et al. 1994), and is regulated by leptin, a major component of energy homeostasis (Doi et al. 2001). These previous findings suggest a potential role of Apoa4 in the regulation of food intake and, consequently, body weight. Suppressor of cytokine signaling 2 (Socs2) is a negative regulator in the growth hormone/insulin-like growth factor (IGF)-I signaling pathway (Metcalf et al. 2000), which is directly related to obesity. Slc22a3 is a downstream gene of the IGF signaling pathway. Recent studies have showed that the Gpld1 gene is associated with serum glycosylphosphatidylinositol-specific phospholipase D (GPI-PLD) levels, which predict changes in insulin sensitivity in response to a low-fat diet in obese women (Gray et al. 2008). The IGF-binding protein also induces laminin gamma 1 (Lamc1) transcription (Abrass and Hansen 2010). These identified genes clearly point out the importance of the IGF signaling pathway in the development of obesity in mice.

Table 3 also presents the *cis*-SNPs, which are defined to be the SNPs within a 10 cM distance of each gene, that are associated with the selected genes. These *cis*-SNPs are likely to play a critical role in the regulation of the target genes and serve as strong instruments in statistical analysis. Not all selected genes have *cis*-SNPs identified, partly due to the nonuniform, relatively sparse distribution of genotyped SNPs. If the criterion is relaxed to within 25 cM of each gene, we find that 13 of the 17 genes in the table have at least one *cis*-SNP identified. Many of these *cis*-SNPs coincide with QTLs detected for body weight traits in previous studies; for example, rs3663003 (Chr 1, 46.1 cM), rs4136518 (Chr 3, 54.6 cM), rs3694833 (Chr 10, 47.7 cM), rs4231406 (Chr 17, 12.0 cM), and rs3661189 (Chr 18, 27.5 cM) fall in previously detected QTL regions (Rocha et al. 2004). Moreover, rs4231406 was previously identified as a QTL for atherosclerosis, which is strongly associated with body weight and adiposity (Wang et al. 2007). These results demonstrate that our method can provide a more integrative, comprehensive understanding of the genetic architecture of complex traits than classical QTL analysis and gene expression studies, and would be useful for prioritizing candidate genes for complex diseases.

## 7 Discussion

We have proposed a 2SR method for variable selection and estimation in sparse IV models where the dimensionality of covariates and instruments can both be much larger than the sample size. We have developed a high-dimensional theory that supports the theoretical

advantages of our method and sheds light on the impact of dimensionality in the resulting procedure. We have applied our method to genetical genomics studies for jointly analyzing gene expression data and genetic variants to explore their associations with complex traits. The proposed method provides a powerful approach to effectively integrating and utilizing genotype, gene expression, and clinical data, which is of great importance for large-scale genomic inference. We have demonstrated on simulated and real data that our method is less affected by confounding and can lead to more reliable and biologically more interpretable results. Although we are primarily motivated by genetical genomics applications, the methodology is in fact very general and likely to find a wide range of applications in epidemiology, econometrics, and many other fields.

In our analysis of genetical genomics data, only genetic variants are used as instruments, and gene expressions that fail to be associated with any genetic variants in the first stage of the 2SR method have to be excluded at the second stage, which may comprise the inferences for genes with weak genetic effects. Epigenetic processes, such as DNA methylation, histone modification, and various RNA-mediated processes, are also known to play an essential role in the regulation of gene expression, and their influences on the gene expression levels may be profound (Jaenisch and Bird 2003). Thus, when epigenetic data are also collected on the same subjects, they can be similarly treated as potential instruments in the sparse IV model. The joint consideration of genetic and epigenetic variants as instruments is likely to yield stronger instruments than using the genetic variants alone, which may lead to more reliable genomic inference.

Several extensions and improvements of the methodology are worthwhile to pursue. We have applied regularization methods to exploit the sparsity of individual coefficients, allowing the first stage to be decomposed into $p$ regression problems. While our general theory in Section 4.2 applies to a generic first-stage estimator, the first-stage estimation and prediction could be improved by taking into account the correlations among the covariates and borrowing information across the $p$ subproblems. Two possibilities would be to exploit the structural sparsity of the coefficient matrix through certain matrix decompositions (e.g., Chen, Chan, and Stenseth 2012; Chen and Huang 2012), and to jointly estimate the coefficient matrix and the covariance structure (e.g., Rothman, Levina, and Zhu 2010; Cai et al. 2013). Moreover, since the 2SR method is a high-dimensional extension of the classical 2SLS method, it would be natural to ask whether other IV estimators such as the LIML and GMM estimators can also be extended to our high-dimensional setting. Although asymptotic efficiency would not be a primary concern in high dimensions, certain advantages of these estimators in low dimensions may carry over and lead to performance improvement.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

## Appendix

## Appendix: Proofs

### A.1 Proof of Proposition 1

We prove the result by contradiction. From (2) we have $\mathbf{X}^T \boldsymbol{\eta}_0 = \mathbf{X}^T \mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)$. If $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_1 = o_P(1)$, then

$$\|\frac{1}{n}\mathbf{X}^T \boldsymbol{\eta}_0\|_\infty \leq \max_{1 \leq i,j \leq p} \frac{1}{n} \mathbf{x}_i^T \mathbf{x}_j \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_1 \leq \max_{1 \leq i,j \leq p} \frac{1}{n} \|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2 \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_1 = o_P(1).$$

This yields a contradiction and completes the proof.

### A.2 Proof of Theorem 1

Since the optimization problem (3) can be decomposed into $p$ penalized least squares problems, the result is a straightforward extension of Theorem 7.2 of Bickel, Ritov, and Tsybakov (2009) to the multivariate regression case. From Condition (C1) and the aforementioned result it follows that, with probability at least $1 - q\,exp(-n\lambda_j^2/(8\sigma_j^2))$, the regularized estimator $\hat{\boldsymbol{\gamma}_j}$ defined by (5) with the $L_1$ penalty satisfies

$$\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j}\|_1 \leq 16r\lambda_j/\kappa_1^2 \quad \text{(A.1)}$$

and

$$\|\boldsymbol{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})\|_2^2 \leq 16nr\lambda_j^2/\kappa_1^2. \quad \text{(A.2)}$$

Using the union bound, we have, with probability at least $1 - \sum_{j=1}^{p} q\,exp(-n\lambda_j^2/(8\sigma_j^2))$, the regularized estimator $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}_1}, \ldots, \hat{\boldsymbol{\gamma}_p})$ satisfies $\|\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\|_1 \leq 16r\lambda_{max}/\kappa_1^2$ and $\|\boldsymbol{Z}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0)\|_F^2 \leq 16npr\lambda_{max}^2/\kappa_1^2$. Now, if we choose $\lambda_j = C\sigma_j \sqrt{(log\,p + log\,q)/n}$ with a constant $C \geq 2\sqrt{2}$, then with probability at least $1 - (pq)^{1-C^2/8}$, the desired inequalities hold.

### A.3 Proof of Theorem 2

The proof of Theorem 2 relies on two key lemmas. Lemma A.1 shows that Condition (C2), which is imposed on the matrix $\mathbf{Z}\boldsymbol{\Gamma}_0$, also holds with high probability for the matrix $\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\Gamma}}$. Lemma A.2 establishes a fundamental inequality that is essential to the proof. To avoid repeatedly stating the probability bounds for certain inequalities to hold, we will occasionally condition on the events that these inequalities hold, and incorporate the probability bounds into the result by the union bound.

**Lemma A.1**—*Under Conditions (C1) and (C2), if the regularization parameters $\lambda_j$ are chosen to satisfy*

$$\lambda_{max}(2L + \lambda_{max}) \leq \frac{\kappa_1^2 \kappa_2^2}{32^2 rs}, \quad \text{(A.3)}$$

*then with probability at least* $1 - \sum_{j=1}^{p} q\,exp(-n\lambda_j^2/(8\sigma_j^2))$, *the matrix* $\hat{\mathbf{X}} = \mathbf{Z}\hat{\mathbf{\Gamma}}$, *where* $\hat{\mathbf{\Gamma}}$ *is defined by* (3) *with the* $L_1$ *penalty, satisfies*

$$\kappa(\hat{\mathbf{X}}, s) \geq \frac{\kappa_2}{2}.$$

**Proof**—For any subset $J \subset \{1, \dots, p\}$ with $|J| \leq s$ and any $\boldsymbol{\delta} \in \mathbb{R}^p$ with $\boldsymbol{\delta} \neq \mathbf{0}$ and $\|\boldsymbol{\delta}_{J^c}\|_1 \leq 3\|\boldsymbol{\delta}_J\|_1$, we have

$$\frac{\boldsymbol{\delta}^T (\hat{\mathbf{X}}^T\hat{\mathbf{X}} - (\mathbf{Z}\mathbf{\Gamma}_0)^T\mathbf{Z}\mathbf{\Gamma}_0)\boldsymbol{\delta}}{n\|\boldsymbol{\delta}_J\|_2^2} \leq \frac{\|\boldsymbol{\delta}\|_1^2 \, max_{1\leq i,j\leq p}|\hat{\mathbf{x}}_i^T\hat{\mathbf{x}}_j - (\mathbf{Z}\boldsymbol{\gamma}_{0i})^T\mathbf{Z}\boldsymbol{\gamma}_{0j}|}{n\|\boldsymbol{\delta}_J\|_2^2}.$$

Since $\|\boldsymbol{\delta}_{J^c}\|_1 \leq 3\|\boldsymbol{\delta}_J\|_1$ we have $\|\boldsymbol{\delta}\|_1^2 = (\|\boldsymbol{\delta}_J\|_1 + \|\boldsymbol{\delta}_{J^c}\|_1)^2 \leq 16\|\boldsymbol{\delta}_J\|_1^2 \leq 16s\|\boldsymbol{\delta}_J\|_2^2$. To bound the entrywise maximum, we write

$$\hat{\mathbf{x}}_i^T\hat{\mathbf{x}}_j - (\mathbf{Z}\boldsymbol{\gamma}_{0i})^T\mathbf{Z}\boldsymbol{\gamma}_{0j}$$
$$= (\hat{\mathbf{x}}_i - \mathbf{Z}\boldsymbol{\gamma}_{0i})^T(\hat{\mathbf{x}}_j - \mathbf{Z}\boldsymbol{\gamma}_{0j}) + (\hat{\mathbf{x}}_i - \mathbf{Z}\boldsymbol{\gamma}_{0i})^T\mathbf{Z}\boldsymbol{\gamma}_{0j} + (\mathbf{Z}\boldsymbol{\gamma}_{0i})^T(\hat{\mathbf{x}}_j - \mathbf{Z}\boldsymbol{\gamma}_{0j})$$
$$= (\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{0i})^T\mathbf{Z}^T\mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j}) + (\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{0i})^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}_{0j} + (\mathbf{Z}\boldsymbol{\gamma}_{0i})^T\mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})$$
$$\equiv T_1 + T_2 + T_3.$$

We now condition on the event that (A.1) and (A.2) in the proof of Theorem 1 hold for $j = 1$, $\dots, p$, which occurs with probability at least $1 - \sum_{j=1}^{p} q\,exp(-n\lambda_j^2/(8\sigma_j^2))$. Then, by the Cauchy–Schwarz inequality,

$$|T_1| \leq \|\mathbf{Z}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{0i})\|_2 \|\mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})\|_2 \leq 16nr\lambda_{max}^2/\kappa_1^2.$$

Also, noting that $\|\mathbf{z}_j\|_2 = \sqrt{n}$ by standardization and $\|\mathbf{\Gamma}_0\|_1 \leq L$, we have

$$
\begin{aligned}
|T_2| &\leq \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{0i}\|_1 \max_{1\leq k,l\leq q}|\mathbf{z}_k^T\mathbf{z}_l| \|\boldsymbol{\gamma}_{0j}\|_1 \\
&\leq \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{0i}\|_1 \max_{1\leq k,l\leq q}\|\mathbf{z}_k\|_2\|\mathbf{z}_l\|_2 \|\boldsymbol{\gamma}_{0j}\|_1 \leq 16Lnr\lambda_{max}/\kappa_1^2.
\end{aligned}
$$

Similarly, $|T_3| \leq 16Lnr\lambda_{max}/\kappa_1^2$. Combining these bounds and using the assumption (A.3), we obtain

$$\frac{\boldsymbol{\delta}^T (\hat{\mathbf{X}}^T\hat{\mathbf{X}} - (\mathbf{Z}\mathbf{\Gamma}_0)^T\mathbf{Z}\mathbf{\Gamma}_0)\boldsymbol{\delta}}{n\|\boldsymbol{\delta}_J\|_2^2} \leq \frac{16^2 rs}{\kappa_1^2}\lambda_{max}(2L + \lambda_{max}) \leq \frac{16^2 rs}{\kappa_1^2}\cdot\frac{\kappa_1^2\kappa_2^2}{32^2 rs} = \left(\frac{\kappa_2}{2}\right)^2.$$

This, together with Condition (C2), proves the lemma.

**Lemma A.2**—*Under Conditions (C1) and (C2), if $\kappa_1^{-2} r\,(\log p + \log q) = O(n)$ and the regularization parameters $\lambda_j$ are chosen as in (8), then there exist constants $c_0$, $c_1$, $c_2 > 0$ such that, if we choose*

$$\mu = \frac{C_0}{\kappa_1}\sqrt{\frac{r(\log p + \log q)}{n}},$$

*where $C_0 = c_0\,L\,\max(\sigma_{p+1}, M\sigma_{\max})$, then with probability at least $1 - c_1(pq)^{-c_2}$, the regularized estimator $\hat{\boldsymbol{\beta}}$ defined by (4) with the $L_1$ penalty satisfies*

$$\frac{1}{2n}\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 + \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 2\mu\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|.$$

**Proof**—By the optimality of $\hat{\boldsymbol{\beta}}$, we have

$$\frac{1}{2n}\|\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}\|_2^2 + \mu\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{2n}\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}_0\|_2^2 + \mu\|\boldsymbol{\beta}_0\|_1.$$

Substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\eta}$, we write

$$
\begin{aligned}
\|\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}\|_2^2 &= \|\boldsymbol{\eta} - (\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0)\|_2^2 \\
&= \|\boldsymbol{\eta}\|_2^2 + \|\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2 - 2\boldsymbol{\eta}^T(\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0) \\
&= \|\boldsymbol{\eta}\|_2^2 + \|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\|_2^2 - 2\boldsymbol{\eta}^T(\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0) \\
&= \|\boldsymbol{\eta}\|_2^2 + \|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 + \|(\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\|_2^2 - 2\boldsymbol{\eta}^T(\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0) \\
&\quad + 2\boldsymbol{\beta}_0^T(\hat{\mathbf{X}} - \mathbf{X})^T\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)
\end{aligned}
$$

and

$$\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}_0\|_2^2 = \|\boldsymbol{\eta} - (\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\|_2^2 = \|\boldsymbol{\eta}\|_2^2 + \|(\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\|_2^2 - 2\boldsymbol{\eta}^T(\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0.$$

Combining the last three displays yields

$$
\begin{aligned}
\frac{1}{2n}\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 &\leq \mu\|\boldsymbol{\beta}_0\|_1 - \mu\|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{n}\boldsymbol{\eta}^T\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \frac{1}{n}\boldsymbol{\beta}_0^T(\hat{\mathbf{X}} - \mathbf{X})^T\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \quad \text{(A.} \\
&\leq \mu\|\boldsymbol{\beta}_0\|_1 - \mu\|\hat{\boldsymbol{\beta}}\|_1 + \left\|\frac{1}{n}\hat{\mathbf{X}}^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}^T(\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\right\|_\infty\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1. \quad \text{4)}
\end{aligned}
$$

Next, we condition on the event that (A.1) and (A.2) in the proof of Theorem 1 hold for $j = 1, \ldots, p$, which occurs with probability at least $1 - (pq)^{1 - C^2/8}$, and find a probability bound for the event that

$$\|\frac{1}{n}\hat{\mathbf{X}}^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}^T(\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\|_\infty \le \frac{\mu}{2}. \quad \text{(A.5)}$$

Substituting $\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\Gamma}}$ and $\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}_0 + \mathbf{E}$, we write

$$
\begin{aligned}
&\frac{1}{n}\hat{\mathbf{X}}^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}^T(\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\\
&= \frac{1}{n}\hat{\boldsymbol{\Gamma}}^T\mathbf{Z}^T\boldsymbol{\eta} - \frac{1}{n}\hat{\boldsymbol{\Gamma}}^T\mathbf{Z}^T(\mathbf{Z}\hat{\boldsymbol{\Gamma}} - \mathbf{Z}\boldsymbol{\Gamma}_0 - \mathbf{E})\boldsymbol{\beta}_0\\
&= \frac{1}{n}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0)^T\mathbf{Z}^T\boldsymbol{\eta} + \frac{1}{n}\boldsymbol{\Gamma}_0^T\mathbf{Z}^T\boldsymbol{\eta} + \frac{1}{n}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0)^T\mathbf{Z}^T\mathbf{E}\boldsymbol{\beta}_0 + \frac{1}{n}\boldsymbol{\Gamma}_0^T\mathbf{Z}^T\mathbf{E}\boldsymbol{\beta}_0\\
&\quad - \frac{1}{n}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0)^T\mathbf{Z}^T\mathbf{Z}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0)\boldsymbol{\beta}_0 - \frac{1}{n}\boldsymbol{\Gamma}_0^T\mathbf{Z}^T\mathbf{Z}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0)\boldsymbol{\beta}_0\\
&\equiv T_1 + T_2 + T_3 + T_4 + T_5 + T_6.
\end{aligned}
$$

To bound term $T_1$, it follows from (A.1), the union bound, and the classical Gaussian tail bound that

$$P\left(\|T_1\|_\infty \ge \frac{\mu}{12}\right) \le P\left(\|\frac{1}{n}\mathbf{Z}^T\boldsymbol{\eta}\|_\infty \ge \frac{\kappa_1^2}{16r\lambda_{max}} \cdot \frac{\mu}{12}\right) \le q\,exp\left\{-\frac{n}{2\sigma_{p+1}^2}\left(\frac{\kappa_1^2}{16r\lambda_{max}} \cdot \frac{\mu}{12}\right)^2\right\}.$$

Noting that $\|\boldsymbol{\Gamma}_0\|_1 \quad L$, we have

$$P\left(\|T_2\|_\infty \ge \frac{\mu}{12}\right) \le P\left(\|\frac{1}{n}\mathbf{Z}^T\boldsymbol{\eta}\|_\infty \ge \frac{\mu}{12L}\right) \le q\,exp\left\{-\frac{n}{2\sigma_{p+1}^2}\left(\frac{\mu}{12L}\right)^2\right\}.$$

To bound term $T_3$, using (A.1) and $\|\boldsymbol{\beta}_0\|_1 \quad M$, we obtain

$$
\begin{aligned}
P\left(\|T_3\|_\infty \ge \frac{\mu}{12}\right) &\le P\left(\max_{1 \le i \le q, 1 \le j \le p}\left|\frac{1}{n}\mathbf{z}_i^T\boldsymbol{\varepsilon}_j\right| \ge \frac{\kappa_1^2}{16r\lambda_{max}} \cdot \frac{\mu}{12M}\right)\\
&\le pq\,exp\left\{-\frac{n}{2\sigma_{max}^2}\left(\frac{\kappa_1^2}{16r\lambda_{max}} \cdot \frac{\mu}{12M}\right)^2\right\},
\end{aligned}
$$

where $\boldsymbol{\varepsilon}_j$ is the $j$th column of the matrix $\mathbf{E}$. Similarly,

$$P\left(\|T_4\|_\infty \ge \frac{\mu}{12}\right) \le P\left\{\max_{1 \le i \le q, 1 \le j \le p}\left|\frac{1}{n}\mathbf{z}_i^T\boldsymbol{\varepsilon}_j\right| \ge \frac{\mu}{12LM}\right\} \le pq\,exp\left\{-\frac{n}{2\sigma_{max}^2}\left(\frac{\mu}{12LM}\right)^2\right\}.$$

To bound term $T_5$, it follows from (A.2), $\|\boldsymbol{\beta}_0\|_1 \quad M$, and the Cauchy–Schwarz inequality that

$$
\begin{aligned}
\|T_5\|_\infty &\le M\max_{1 \le i, j \le p}\left|\frac{1}{n}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{0i})^T\mathbf{Z}^T\mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})\right|\\
&\le M\max_{1 \le i, j \le p}\frac{1}{n}\|\mathbf{Z}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{0i})\|_2\|\mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})\|_2 \le M\frac{16r\lambda_{max}^2}{\kappa_1^2}.
\end{aligned}
$$

Noting that $\|\mathbf{z}_j\|_2 = \sqrt{n}$ by standardization, we have

$$\begin{aligned}
\|T_6\|_\infty &\leq & LM \max_{1\leq i\leq q, 1\leq j\leq p}\left|\frac{1}{n}\mathbf{z}_i^T \mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})\right| \\
&\leq & LM \max_{1\leq j\leq p}\frac{1}{\sqrt{n}}\|\mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j})\|_2 \leq LM\frac{4\sqrt{r}\lambda_{max}}{\kappa_1}.
\end{aligned}$$

Combining these bounds and in view of the assumption $\kappa_1^{-2} r(\log p + \log q) = O(n)$, there exist constants $c_0, c_1, c_2 > 0$ such that, if we choose

$$\mu = \frac{C_0}{\kappa_1}\sqrt{\frac{r(\log p + \log q)}{n}},$$

where $C_0 = c_0 L \max(\sigma_{p+1}, M\sigma_{max})$, then (A.5) holds with probability at least $1 - c_1(pq)^{-c_2}$. This, together with (A.4), implies

$$\frac{1}{2n}\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq \mu\|\boldsymbol{\beta}_0\|_1 - \mu\|\hat{\boldsymbol{\beta}}\|_1 + \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1.$$

Adding $\mu\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1/2$ to both sides yields

$$\begin{aligned}
\frac{1}{2n}\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 + \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 &\leq \mu(\|\boldsymbol{\beta}_0\|_1 - \|\hat{\boldsymbol{\beta}}\|_1 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1) \\
&= \mu(\|\boldsymbol{\beta}_{0S}\|_1 - \|\hat{\boldsymbol{\beta}}_S\|_1 + \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_1) \leq 2\mu\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_1.
\end{aligned}$$

This completes the proof of the lemma.

**Proof of Theorem 2**—We first note that (8) and (9) imply that the condition $\kappa_1^{-2} r(\log p + \log q) = O(n)$ is satisfied. Then it follows from Lemma A.2 that, with probability at least $1 - c_1(pq)^{-c_2}$, we have

$$\frac{1}{2n}\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq 2\mu\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_1 \leq 2\mu\sqrt{s}\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_2 \quad \text{(A.6)}$$

and

$$\frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 2\mu\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_1, \quad or \quad \|\hat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}_{S^c}^0\|_1 \leq 3\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_1.$$

The last inequality, the definition of $\kappa(\hat{\mathbf{X}, s})$, Lemma 1 together imply

$$\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_2 \leq \frac{\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2}{\sqrt{n}\kappa(\hat{\mathbf{X}}, s)} \leq \frac{2\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2}{\sqrt{n}\kappa_2}. \quad \text{(A.7)}$$

Combining (A.6) and (A.7), we obtain

$$\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq \frac{64}{\kappa_2^2} ns\mu^2$$

and

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 4\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_1 \leq 4\sqrt{s}\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_2 \leq \frac{64}{\kappa_2^2}s\mu.$$

Substituting (10) for $\mu$ concludes the proof.

### A.4 Proof of Theorem 3

Central to the proof of Theorem 3 is the following lemma, which shows that Condition (C3) also holds with high probability for the matrix $\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\Gamma}}$ and gives a useful bound for the inverse matrix norm $\|(\hat{\mathbf{C}}_{SS})^{-1}\|_\infty$.

**Lemma A.3**—*Under Condition (C3), if the regularization parameters $\lambda_j$ are chosen to satisfy*

$$\frac{16\varphi}{\kappa_1^2} rs\lambda_{max}(2L + \lambda_{max}) \leq \frac{\alpha}{4 - \alpha}, \quad \text{(A.8)}$$

*then with probability at least $1 - \sum_{j=1}^p q\, exp(-n\lambda_j^2/(8\sigma_j^2))$, the matrix $\hat{\mathbf{C}} = n^{-1}\hat{\mathbf{X}}^T \hat{\mathbf{X}} = n^{-1}(\mathbf{Z}\hat{\boldsymbol{\Gamma}})^T \mathbf{Z}\hat{\boldsymbol{\Gamma}}$, where $\hat{\boldsymbol{\Gamma}}$ is defined by (3) with the $L_1$ penalty, satisfies*

$$\|(\hat{\mathbf{C}}_{SS})^{-1}\|_\infty \leq \frac{4 - \alpha}{2(2 - \alpha)}\varphi \quad \text{(A.9)}$$

*and*

$$\|\hat{\mathbf{C}}_{S^c S} (\hat{\mathbf{C}}_{SS})^{-1}\|_\infty \leq 1 - \frac{\alpha}{2}. \quad \text{(A.10)}$$

**Proof**—We condition on the event that (A.1) and (A.2) in the proof of Theorem 1 hold for $j = 1, \ldots, p$, which occurs with probability at least $1 - \sum_{j=1}^p q\, exp(-n\lambda_j^2/(8\sigma_j^2))$. From the proof of Lemma A.1, we have

$$\max_{1 \leq i,j \leq p} \frac{1}{n}|\hat{\mathbf{x}}_i^T \hat{\mathbf{x}}_j - (\mathbf{Z}_{\gamma_{0i}})^T \mathbf{Z}_{\gamma_{0j}}| \leq \frac{16}{\kappa_1^2} r\lambda_{max}(2L + \lambda_{max}).$$

This, along with the assumption (A.8), gives

$$\varphi\|\hat{\mathbf{C}}_{SS} - \mathbf{C}_{SS}\|_\infty \leq \frac{16\varphi}{\kappa_1^2} rs\lambda_{max}(2L + \lambda_{max}) \leq \frac{\alpha}{4 - \alpha}, \quad \text{(A.11)}$$

and similarly,

$$\varphi\|\hat{\mathbf{C}}_{S^c S} - \mathbf{C}_{S^c S}\|_\infty \le \frac{\alpha}{4 - \alpha}. \quad \text{(A.12)}$$

Then, by an error bound for matrix inversion (Horn and Johnson 1985, p. 336), we have

$$\|(\hat{\mathbf{C}}_{SS})^{-1} - (\mathbf{C}_{SS})^{-1}\|_\infty \le \frac{\varphi\|\hat{\mathbf{C}}_{SS} - \mathbf{C}_{SS}\|_\infty}{1 - \varphi\|\hat{\mathbf{C}}_{SS} - \mathbf{C}_{SS}\|_\infty}\varphi \le \frac{\alpha}{2(2 - \alpha)}\varphi.$$

The triangle inequality implies

$$\|(\hat{\mathbf{C}}_{SS})^{-1}\|_\infty \le \|(\mathbf{C}_{SS})^{-1}\|_\infty + \|(\hat{\mathbf{C}}_{SS})^{-1} - (\mathbf{C}_{SS})^{-1}\|_\infty \le \varphi + \frac{\alpha}{2(2 - \alpha)}\varphi = \frac{4 - \alpha}{2(2 - \alpha)}\varphi,$$

which proves (A.9).

To show inequality (A.10), we write

$$\hat{\mathbf{C}}_{S^c S}(\hat{\mathbf{C}}_{SS})^{-1} - \mathbf{C}_{S^c S}(\mathbf{C}_{SS})^{-1} = (\hat{\mathbf{C}}_{S^c S} - \mathbf{C}_{S^c S})(\hat{\mathbf{C}}_{SS})^{-1} - \mathbf{C}_{S^c S}(\mathbf{C}_{SS})^{-1}(\hat{\mathbf{C}}_{SS} - \mathbf{C}_{SS})(\hat{\mathbf{C}}_{SS})^{-1}.$$

Then it follows from (A.9), (A.11), (A.12), and Condition (C3) that

$$\|\hat{\mathbf{C}}_{S^c S}(\hat{\mathbf{C}}_{SS})^{-1} - \mathbf{C}_{S^c S}(\mathbf{C}_{SS})^{-1}\|_\infty$$
$$\le \|\hat{\mathbf{C}}_{S^c S} - \mathbf{C}_{S^c S}\|_\infty\|(\hat{\mathbf{C}}_{SS})^{-1}\|_\infty + \|\mathbf{C}_{S^c S}(\mathbf{C}_{SS})^{-1}\|_\infty\|\hat{\mathbf{C}}_{SS} - \mathbf{C}_{SS}\|_\infty\|(\hat{\mathbf{C}}_{SS})^{-1}\|_\infty$$
$$\le \frac{\alpha}{(4-\alpha)\varphi}\frac{4-\alpha}{2(2-\alpha)}\varphi + (1 - \alpha)\frac{\alpha}{(4-\alpha)\varphi}\frac{4-\alpha}{2(2-\alpha)}\varphi = \frac{\alpha}{2},$$

which, along with Condition (C3), proves (A.10). This completes the proof of the lemma.

**Proof of Theorem 3**—For an index set $J$, let $\mathbf{X}_J$ and $\hat{\mathbf{X}_J}$ denote the submatrices formed by the $j$th columns of $\mathbf{X}$ and $\hat{\mathbf{X}}$ with $j \in J$, respectively. The optimality conditions for $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ to be a solution to problem (4) with the $L_1$ penalty can be written as

$$\frac{1}{n}\hat{\mathbf{X}}_{\hat{S}}^T(\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}) = \mu\, sgn(\hat{\boldsymbol{\beta}}_{\hat{S}}) \quad \text{(A.13)}$$

and

$$\|\frac{1}{n}\hat{\mathbf{X}}_{\hat{S}^c}^T(\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}})\|_\infty \le \mu. \quad \text{(A.14)}$$

It suffices to find a $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ with the desired properties such that (A.13) and (A.14) hold. Let $\hat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$. The idea of the proof is to first determine $\hat{\boldsymbol{\beta}}_S$ from (A.13), and then show that thus obtained $\hat{\boldsymbol{\beta}}$ also satisfies (A.14).

Using similar arguments to those in the proof of Lemma A.2, we can show that, there exist constants $c_0$, $c_1$, $c_2 > 0$ such that, if we can choose $\mu$ as before, then with probability at least $1 - c_1(pq)^{-c_2}$, it holds that

$$\|\frac{1}{n}\hat{\mathbf{X}}^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}^T(\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta}_0\|_\infty \leq \frac{\alpha}{4-\alpha}\mu. \quad \text{(A.15)}$$

From now on, we condition on the event that (A.15) holds and analyze conditions (A.13) and (A.14).

We first determine $\hat{\boldsymbol{\beta}_S}$ from (A.13). By substituting

$$\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}} = \mathbf{X}_S\boldsymbol{\beta}_{0S} + \boldsymbol{\eta} - \hat{\mathbf{X}}_S\hat{\boldsymbol{\beta}}_S = \boldsymbol{\eta} - (\hat{\mathbf{X}}_S - \mathbf{X}_S)\boldsymbol{\beta}_{0S} - \hat{\mathbf{X}}_S(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}), \quad \text{(A.16)}$$

we write (A.13) with $\hat{S}$ replaced by $S$ in the form

$$\frac{1}{n}\hat{\mathbf{X}}_S^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}_S^T(\hat{\mathbf{X}}_S - \mathbf{X}_S)\boldsymbol{\beta}_{0S} - \hat{\mathbf{C}}_{SS}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}) = \mu\, sgn(\hat{\boldsymbol{\beta}}_S),$$

or

$$\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S} = (\hat{\mathbf{C}}_{SS})^{-1}\left\{\frac{1}{n}\hat{\mathbf{X}}_S^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}_S^T(\hat{\mathbf{X}}_S - \mathbf{X}_S)\boldsymbol{\beta}_{0S} - \mu\, sgn(\hat{\boldsymbol{\beta}}_S)\right\}. \quad \text{(A.17)}$$

This, along with (A.9) and (A.15), leads to

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_\infty &\leq \|(\hat{\mathbf{C}}_{SS})^{-1}\|_\infty\left\{\|\frac{1}{n}\hat{\mathbf{X}}_S^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}_S^T(\hat{\mathbf{X}}_S - \mathbf{X}_S)\boldsymbol{\beta}_{0S}\|_\infty + \mu\right\}\\
&\leq \frac{4-\alpha}{2(2-\alpha)}\varphi\left(\frac{\alpha}{4-\alpha}\mu + \mu\right) = \frac{2}{2-\alpha}\varphi\mu < b_0
\end{aligned}$$

by assumption, which entails that sgn $(\hat{\boldsymbol{\beta}_S})$ = sgn $(\boldsymbol{\beta}_{0S})$. Since $\hat{\boldsymbol{\beta}}_{S^c} = \boldsymbol{\beta}_{S^c}^0 = 0$ by definition, we have $\hat{S} = S$. Let $\hat{\boldsymbol{\beta}_S}$ be defined by (A.17) with sgn$(\hat{\boldsymbol{\beta}_S})$ replaced by sgn$(\boldsymbol{\beta}_{0S})$. Clearly, thus defined $\hat{\boldsymbol{\beta}}$ satisfies the desired properties and (A.13).

It remains to show that $\hat{\boldsymbol{\beta}}$ also satisfies (A.14). By substituting (A.16) and (A.17), we write

$$\begin{aligned}
\frac{1}{n}\hat{\mathbf{X}}_{S^c}^T(\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}) = {} & \frac{1}{n}\hat{\mathbf{X}}_{S^c}^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}_{S^c}^T(\hat{\mathbf{X}}_S - \mathbf{X}_S)\boldsymbol{\beta}_{0S}\\
& - \hat{\mathbf{C}}_{S^cS}(\hat{\mathbf{C}}_{SS})^{-1}\left\{\frac{1}{n}\hat{\mathbf{X}}_S^T\boldsymbol{\eta} - \frac{1}{n}\hat{\mathbf{X}}_S^T(\hat{\mathbf{X}}_S - \mathbf{X}_S)\boldsymbol{\beta}_{0S} - \mu\, sgn(\hat{\boldsymbol{\beta}}_S)\right\}.
\end{aligned}$$

Then it follows from (A.10) and (A.15) that

$$\|\tfrac{1}{n}\hat{\mathbf{X}}_{S^c}^T(\mathbf{y}-\hat{\mathbf{X}}\hat{\boldsymbol{\beta}})\|_\infty \;\leq\; \|\tfrac{1}{n}\hat{\mathbf{X}}_{S^c}^T\boldsymbol{\eta}-\tfrac{1}{n}\hat{\mathbf{X}}_{S^c}^T(\hat{\mathbf{X}}_S-\mathbf{X}_S)\boldsymbol{\beta}_{0S}\|_\infty$$
$$+\|\hat{\mathbf{C}}_{S^c S}(\hat{\mathbf{C}}_{SS})^{-1}\|_\infty\left\{\|\tfrac{1}{n}\hat{\mathbf{X}}_S^T\boldsymbol{\eta}-\tfrac{1}{n}\hat{\mathbf{X}}_S^T(\hat{\mathbf{X}}_S-\mathbf{X}_S)\boldsymbol{\beta}_{0S}\|_\infty+\mu\right\}$$
$$\leq\; \tfrac{\alpha}{4-\alpha}\mu+\left(1-\tfrac{\alpha}{2}\right)\left(\tfrac{\alpha}{4-\alpha}\mu+\mu\right)=\mu.$$

Since $\hat{S}=S$, we see that $\hat{\boldsymbol{\beta}}$ also satisfies (A.14), which concludes the proof.

## References

Abrass CK, Hansen KM. Insulin-Like Growth Factor-Binding Protein-5-Induced Laminin $\gamma$1 Transcription Requires Filamin A. Journal of Biological Chemistry. 2010; 285:12925–12934. [PubMed: 20167606]

Anderson TW. Origins of the Limited Information Maximum Likelihood and Two-Stage Least Squares Estimators. Journal of Econometrics. 2005; 127:1–16.

Belloni A, Chen D, Chernozhukov V, Hansen C. Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. Econometrica. 2012; 80:2369–2429.

Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous Analysis of Lasso and Dantzig Selector. The Annals of Statistics. 2009; 37:1705–1732.

Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering. American Journal of Human Genetics. 2007; 81:1084–1097. [PubMed: 17924348]

Cai TT, Li H, Liu W, Xie J. Covariate-Adjusted Precision Matrix Estimation With an Application in Genetical Genomics. Biometrika. 2013; 100:139–156.

Caner M. Lasso-Type GMM Estimator. Econometric Theory. 2009; 25:270–290.

Carrasco M. A Regularization Approach to the Many Instruments Problem. Journal of Econometrics. 2012; 170:383–398.

Chao JC, Swanson NR. Consistent Estimation With a Large Number of Weak Instruments. Econometrica. 2005; 73:1673–1692.

Chen K, Chan K-S, Stenseth NC. Reduced Rank Stochastic Regression With a Sparse Singular Value Decomposition. Journal of the Royal Statistical Society. 2012; 74:203–221. Series B.

Chen L, Huang JZ. Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection. Journal of the American Statistical Association. 2012; 107:1533–1545.

Didelez V, Meng S, Sheehan NA. Assumptions of IV Methods for Observational Epidemiology. Statistical Science. 2010; 25:22–40.

Didelez V, Sheehan N. Mendelian Randomization as an Instrumental Variable Approach to Causal Inference. Statistical Methods in Medical Research. 2007; 16:309–330. [PubMed: 17715159]

Doi T, Liu M, Seeley RJ, Woods SC, Tso P. Effect of Leptin on Intestinal Apolipoprotein AIV in Response to Lipid Feeding. American Journal of Physiology Regulatory, Integrative and Comparative Physiology. 2001; 281:R753–R759.

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. Genetics of Gene Expression and Its Effect on Disease. Nature. 2008; 452:423–428. [PubMed: 18344981]

Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. and the Mouse Genome Database Group. The Mouse Genome Database (MGD): Comprehensive Resource for Genetics and Genomics of the Laboratory Mouse. Nucleic Acids Research. 2012; 40:D881–D886. [PubMed: 22075990]

Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Fan J, Liao Y. Endogeneity in Ultrahigh Dimension. 2012 unpublished manuscript. Available at arXiv: 1204.5536.

Fan J, Lv J. A Selective Overview of Variable Selection in High Dimensional Feature Space. Statistica Sinica. 2010; 20:101–148. [PubMed: 21572976]

Fan J, Lv J. Nonconcave Penalized Likelihood With NP-Dimensionality. IEEE Transactions on Information Theory. 2011; 57:5467–5484. [PubMed: 22287795]

Fusi N, Stegle O, Lawrence ND. Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies. PLoS Computational Biology. 2012; 8:e1002330. [PubMed: 22241974]

Gautier E, Tsybakov A. High-Dimensional Instrumental Variables Regression and Confidence Sets. 2011 unpublished manuscript. Available at arXiv:1105.2454.

Göring HHH. Tissue Specificity of Genetic Regulation of Gene Expression. Nature Genetics. 2012; 44:1077–1078. [PubMed: 23011224]

Gray DL, O'Brien KD, D'Alessio DA, Brehm BJ, Deeg MA. Plasma Glycosylphosphatidylinositol-Specific Phospholipase D Predicts the Change in Insulin Sensitivity in Response to a Low-Fat But Not a Low-Carbohydrate Diet in Obese Women. Metabolism Clinical and Experimental. 2008; 57:473–478. [PubMed: 18328347]

Hansen C, Hausman J, Newey W. Estimation With Many Instrumental Variables. Journal of Business and Economic Statistics. 2008; 26:398–422.

Heckman JJ. Dummy Endogenous Variables in a Simultaneous Equation System. Econometrica. 1978; 46:931–959.

Horn, RA.; Johnson, CR. Matrix Analysis. New York: Cambridge University Press; 1985.

Jaenisch R, Bird A. Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrisic and Environmental Signals. Nature Genetics. 2003; 33(Suppl):245–254. [PubMed: 12610534]

Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology. Statistics in Medicine. 2008; 27:1133–1163. [PubMed: 17886233]

Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genetics. 2007; 3:e161.

Lin DY, Zeng D. Correcting for Population Stratification in Genomewide Association Studies. Journal of the American Statistical Association. 2011; 106:997–1008. [PubMed: 22467997]

Lin W, Lv J. High-Dimensional Sparse Additive Hazards Regression. Journal of the American Statistical Association. 2013; 108:247–264.

Lu W, Goldberg Y, Fine JP. On the Robustness of the Adaptive Lasso to Model Misspecification. Biometrika. 2012; 99:717–731. [PubMed: 25294946]

Lv J, Fan Y. A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares. The Annals of Statistics. 2009; 37:3498–3528.

Lv J, Liu JS. Model Selection Principles in Misspecified Models. Journal of the Royal Statistical Society. 2013 Series B, to appear.

Mazumder R, Friedman JH, Hastie T. SparseNet: Coordinate Descent With Nonconvex Penalties. Journal of the American Statistical Association. 2011; 106:1125–1138. [PubMed: 25580042]

Meinshausen N, Bühlmann P. Stability Selection. Journal of the Royal Statistical Society. 2010; 72:417–473. (with discussion), Series B.

Metcalf D, Greenhalgh CJ, Viney E, Willson TA, Starr R, Nicola NA, Hilton DJ, Alexander WS. Gigantism in Mice Lacking Suppressor of Cytokine Signalling-2. Nature. 2000; 405:1069–1073. [PubMed: 10890450]

Okui R. Instrumental Variable Estimation in the Presence of Many Moment Conditions. Journal of Econometrics. 2011; 165:70–86.

Okumura T, Fukagawa K, Tso P, Taylor IL, Pappas T. Intracisternal Injection of Apolipoprotein A-IV Inhibits Gastric Secretion in Pylorus-Ligated Conscious Rats. Gastroenterology. 1994; 107:1861–1864. [PubMed: 7958703]

Raskutti G, Wainwright MJ, Yu B. Minimax Rates of Estimation for High-Dimensional Linear Regression Over $\ell_q$-Balls. IEEE Transactions on Information Theory. 2011; 57:6976–6994.

Rocha JL, Eisen EJ, Van Vleck LD, Pomp D. A Large-Sample QTL Study in Mice: I. Growth. Mammalian Genome. 2004; 15:83–99. [PubMed: 15058380]

Rothman AJ, Levina E, Zhu J. Sparse Multivariate Regression With Covariance Estimation. Journal of Computational and Graphical Statistics. 2010; 19:947–962. [PubMed: 24963268]

Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Davey Smith G, Evans D, Gutierrez-Arcelus M, Price A, Raj T, Nisbett J, Nica AC, Beazley C, Durbin R, Deloukas P, Dermitzakis ET. Patterns of *Cis* Regulatory Variation in Diverse Human Populations. PLoS Genetics. 2012; 8:e1002639. [PubMed: 22532805]

Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. 1996; 58:267–288. Series B.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing Value Estimation Methods for DNA Microarrays. Bioinformatics. 2001; 17:520–525. [PubMed: 11395428]

Tso P, Sun W, Liu M. Gastrointestinal Satiety Signals IV. Apolipoprotein A-IV. American Journal of Physiology Gastrointestinal and Liver Physiology. 2004; 286:G885–G890. [PubMed: 15132947]

van Nas A, Ingram-Drake L, Sinsheimer JS, Wang SS, Schadt EE, Drake T, Lusis AJ. Expression Quantitative Trait Loci: Replication, Tissue- and Sex-Specificity in Mice. Genetics. 2010; 185:1059–1068. [PubMed: 20439777]

Wang S, Yehya N, Schadt EE, Wang H, Drake TA, Lusis AJ. Genetic and Genomic Analysis of a Fat Mass Trait With Complex Inheritance Reveals Marked Sex Specificity. PLoS Genetics. 2006; 2:e15. [PubMed: 16462940]

Wang SS, Schadt EE, Wang H, Wang X, Ingram-Drake L, Shi W, Drake TA, Lusis AJ. Identification of Pathways for Atherosclerosis in Mice: Integration of Quantitative Trait Locus Analysis and Global Gene Expression Data. Circulation Research. 2007; 101:e11–e30. [PubMed: 17641228]

Wheatcroft SB, Kearney MT, Shah AM, Ezzat VA, Miell JR, Modo M, Williams SCR, Cawthorn WP, Medina-Gomez G, Vidal-Puig A, Sethi JK, Crossey PA. IGF-Binding Protein-2 Protects Against the Development of Obesity and Insulin Resistance. Diabetes. 2007; 56:285–294. [PubMed: 17259371]

Ye F, Zhang CH. Rate Minimaxity of the Lasso and Dantzig Selector for the $\ell_q$ Loss in $\ell_r$ Balls. Journal of Machine Learning Research. 2010; 11:3519–3540.

Zhang C-H. Nearly Unbiased Variable Selection Under Minimax Concave Penalty. The Annals of Statistics. 2010; 38:894–942.

Zhao P, Yu B. On Model Selection Consistency of Lasso. Journal of Machine Learning Research. 2006; 7:2541–2563.

Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association. 2006; 101:1418–1429.
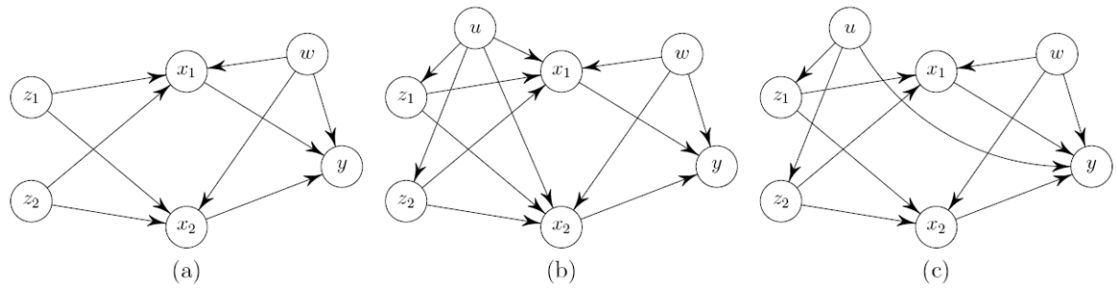
**Figure 1.**
Causal diagrams showing the relationships between two genotypes $z_1$ and $z_2$, two gene expression levels $x_1$ and $x_2$, a clinical phenotype $y$, an unobserved phenotype $w$ that confounds the associations between gene expression levels and the clinical phenotype, and an unobserved variable $u$ representing possibly present population substructure. The population substructure (a) is not present, (b) affects genotypes and gene expression levels, or (c) affects genotypes and the clinical phenotype.
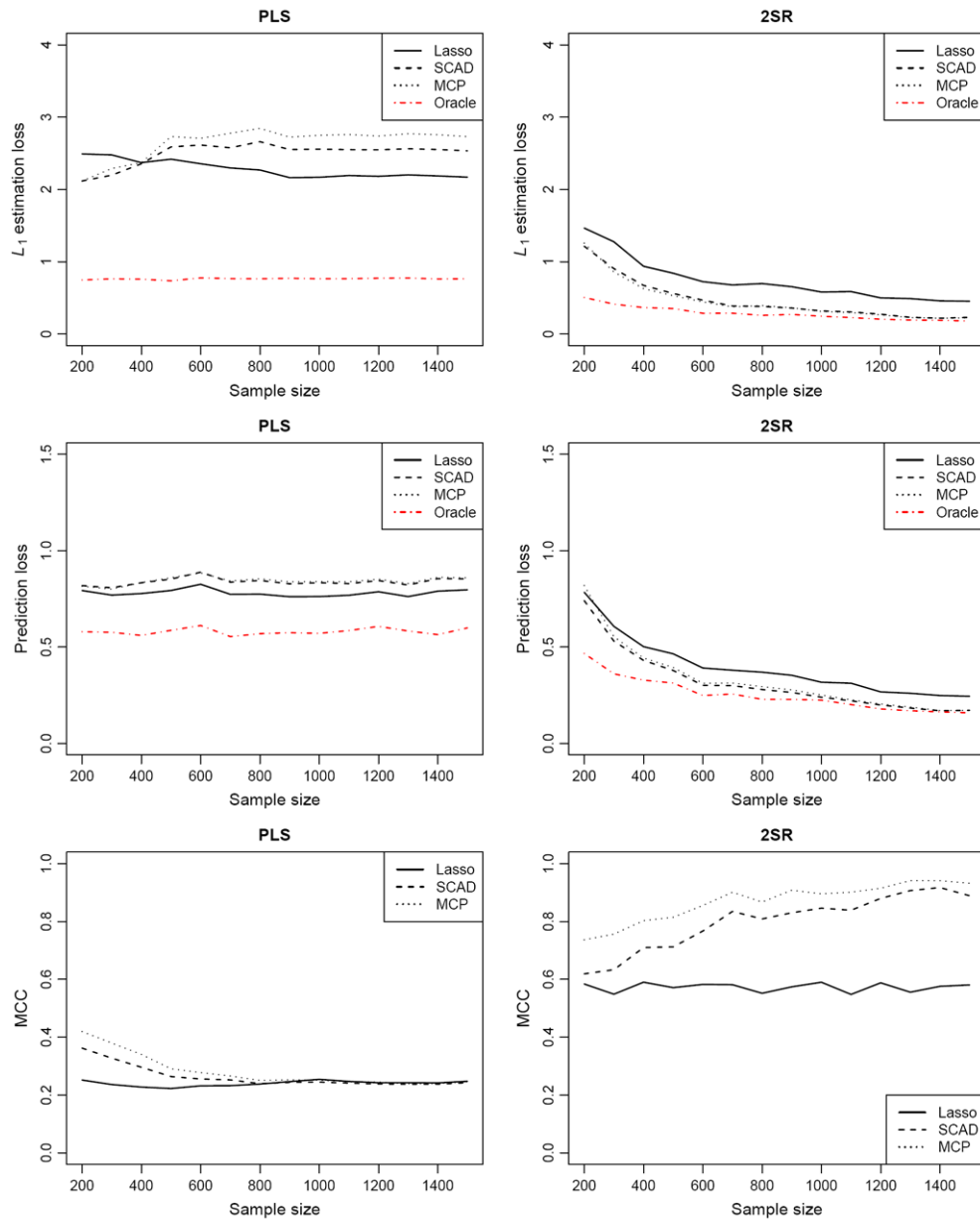
**Figure 2.**
Performance curves for different methods with the dimensions $p = q = 100$ fixed and the sample size $n$ varying from 200 to 1500.
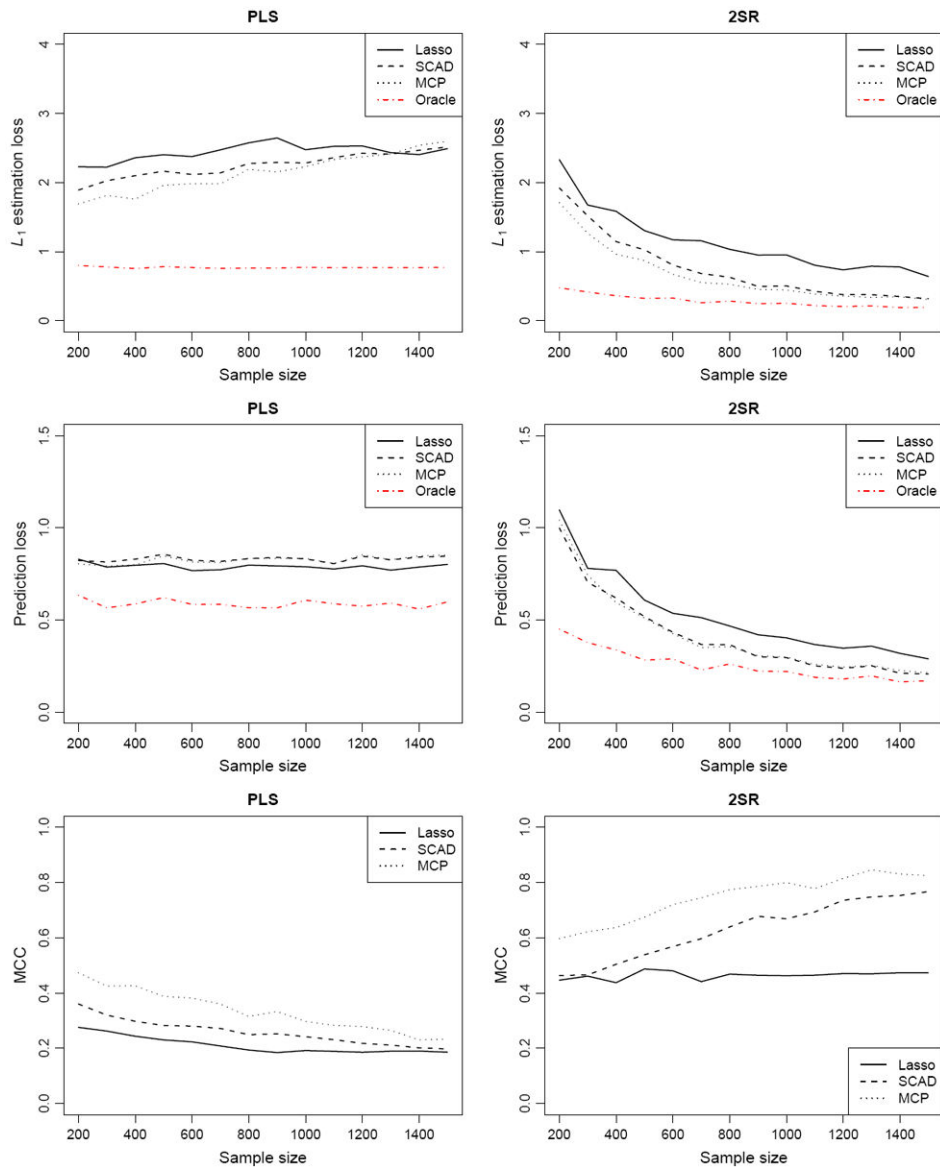
**Figure 3.**
Performance curves for different methods with the dimensions $p = q = 600$ fixed and the sample size $n$ varying from 200 to 1500.
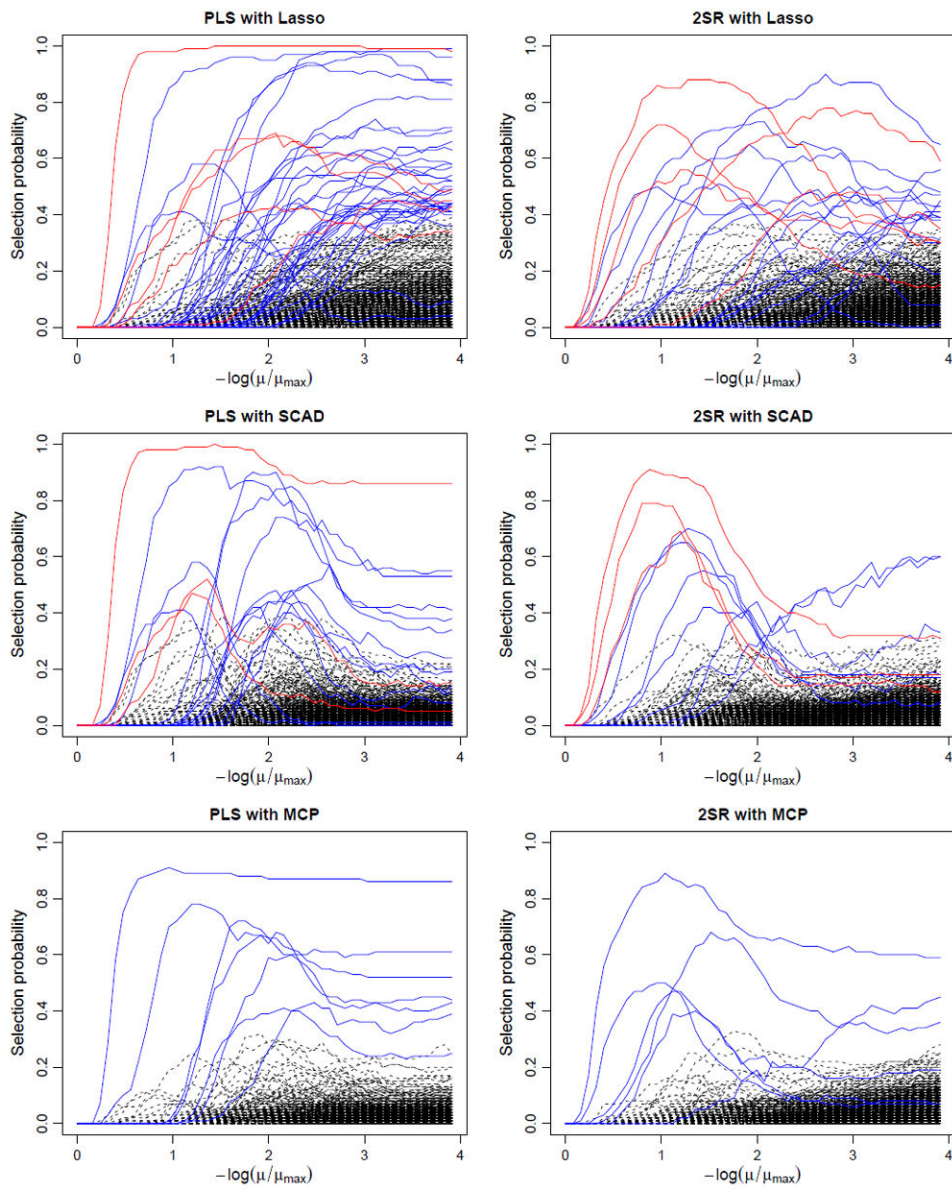
**Figure 4.**
Stability paths for different methods applied to the mouse obesity data based on 100 subsamples. Genes with maximum selection probability at least 0.4 are displayed in solid lines, among which genes common to both the PLS and 2SR methods are shown in red and the distinct ones in blue, and the remaining genes are displayed in dashed lines.

**Table 1**

Simulation results for Models 1–4. Each performance measure was averaged over 50 replicates with standard deviation shown in parentheses

| Method | PLS | | | | | 2SR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ est. loss | Pred. loss | TP | Model size | MCC | $L_1$ est. loss | Pred. loss | TP | Model size | MCC |
| Model 1: $(n, p, q) = (200, 100, 100)$, $(a, b) = (0.75, 1)$ | | | | | | | | | | |
| Lasso | 2.49 (0.57) | 0.79 (0.16) | 5.0 (0.0) | 46.9 (11.2) | 0.25 (0.06) | 1.47 (0.69) | 0.78 (0.26) | 5.0 (0.2) | 14.5 (5.6) | 0.58 (0.12) |
| SCAD | 2.12 (0.53) | 0.82 (0.17) | 5.0 (0.0) | 29.6 (6.0) | 0.36 (0.06) | 1.21 (0.55) | 0.74 (0.32) | 5.0 (0.2) | 12.9 (4.3) | 0.62 (0.12) |
| MCP | 2.12 (0.59) | 0.82 (0.17) | 5.0 (0.0) | 24.3 (6.6) | 0.42 (0.08) | 1.26 (0.66) | 0.82 (0.34) | 4.9 (0.2) | 9.5 (3.8) | 0.74 (0.16) |
| Oracle | 0.75 (0.11) | 0.58 (0.12) | 5 (0) | 5 (0) | 1 (0) | 0.51 (0.20) | 0.47 (0.22) | 5 (0) | 5 (0) | 1 (0) |
| Model 2: $(n, p, q) = (400, 200, 200)$, $(a, b) = (0.75, 1)$ | | | | | | | | | | |
| Lasso | 2.71 (0.44) | 0.81 (0.14) | 5.0 (0.0) | 74.5 (14.3) | 0.21 (0.03) | 1.16 (0.52) | 0.57 (0.18) | 5.0 (0.0) | 18.1 (7.0) | 0.54 (0.11) |
| SCAD | 2.32 (0.36) | 0.84 (0.14) | 5.0 (0.0) | 47.2 (10.2) | 0.29 (0.05) | 0.86 (0.43) | 0.49 (0.17) | 5.0 (0.0) | 14.0 (5.5) | 0.62 (0.13) |
| MCP | 2.28 (0.45) | 0.84 (0.15) | 5.0 (0.0) | 36.2 (11.5) | 0.36 (0.08) | 0.76 (0.39) | 0.51 (0.19) | 5.0 (0.0) | 9.3 (3.3) | 0.76 (0.14) |
| Oracle | 0.76 (0.07) | 0.58 (0.11) | 5 (0) | 5 (0) | 1 (0) | 0.41 (0.16) | 0.37 (0.15) | 5 (0) | 5 (0) | 1 (0) |
| Model 3: $(n, p, q) = (400, 200, 200)$, $(a, b) = (0.5, 0.75)$ | | | | | | | | | | |
| Lasso | 3.04 (0.39) | 0.86 (0.11) | 5.0 (0.0) | 72.3 (12.5) | 0.22 (0.03) | 1.72 (0.73) | 0.73 (0.24) | 5.0 (0.0) | 18.3 (6.3) | 0.53 (0.10) |
| SCAD | 2.64 (0.36) | 0.89 (0.11) | 5.0 (0.0) | 43.3 (11.9) | 0.32 (0.06) | 1.50 (0.65) | 0.69 (0.25) | 5.0 (0.1) | 16.8 (6.7) | 0.56 (0.12) |
| MCP | 2.61 (0.42) | 0.89 (0.12) | 5.0 (0.0) | 33.5 (11.9) | 0.38 (0.09) | 1.36 (0.67) | 0.71 (0.26) | 5.0 (0.2) | 11.0 (4.4) | 0.69 (0.14) |
| Oracle | 1.00 (0.08) | 0.63 (0.09) | 5 (0) | 5 (0) | 1 (0) | 0.57 (0.23) | 0.43 (0.17) | 5 (0) | 5 (0) | 1 (0) |
| Model 4: $(n, p, q) = (400, 200, 200)$, $(a, b) = (0.5, 1)$ or $(0.05, 0.1)$ | | | | | | | | | | |
| Lasso | 2.88 (0.36) | 0.79 (0.08) | 5.0 (0.0) | 71.4 (13.4) | 0.22 (0.04) | 1.68 (0.72) | 0.72 (0.21) | 5.0 (0.0) | 18.8 (7.2) | 0.52 (0.10) |
| SCAD | 2.49 (0.29) | 0.83 (0.08) | 5.0 (0.0) | 42.2 (10.2) | 0.32 (0.05) | 1.54 (0.85) | 0.68 (0.24) | 5.0 (0.0) | 17.4 (7.5) | 0.55 (0.11) |
| MCP | 2.46 (0.40) | 0.82 (0.08) | 5.0 (0.0) | 32.6 (10.6) | 0.38 (0.08) | 1.52 (0.88) | 0.73 (0.25) | 5.0 (0.1) | 12.8 (5.7) | 0.65 (0.14) |
| Oracle | 0.94 (0.07) | 0.58 (0.08) | 5 (0) | 5 (0) | 1 (0) | 0.42 (0.17) | 0.30 (0.12) | 5 (0) | 5 (0) | 1 (0) |

## Table 2

Simulation results for Models 5–8. Each performance measure was averaged over 50 replicates with standard deviation shown in parentheses

| Method | PLS | | | | | 2SR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ est. loss | Pred. loss | TP | Model size | MCC | $L_1$ est. loss | Pred. loss | TP | Model size | MCC |
| Model 5: $(n,p,q) = (300,600,600)$, $(a,b) = (0.75,1)$ | | | | | | | | | | |
| Lasso | 2.22 (0.44) | 0.79 (0.15) | 5.0 (0.0) | 69.1 (20.0) | 0.26 (0.04) | 1.67 (0.81) | 0.78 (0.29) | 5.0 (0.0) | 25.5 (10.5) | 0.46 (0.10) |
| SCAD | 2.03 (0.38) | 0.81 (0.17) | 5.0 (0.0) | 48.6 (14.1) | 0.32 (0.05) | 1.52 (0.74) | 0.70 (0.25) | 5.0 (0.0) | 26.3 (12.2) | 0.47 (0.12) |
| MCP | 1.81 (0.34) | 0.79 (0.17) | 5.0 (0.0) | 28.7 (9.3) | 0.43 (0.07) | 1.26 (0.68) | 0.74 (0.30) | 5.0 (0.1) | 14.7 (7.1) | 0.62 (0.14) |
| Oracle | 0.78 (0.08) | 0.57 (0.10) | 5 (0) | 5 (0) | 1 (0) | 0.42 (0.15) | 0.38 (0.15) | 5 (0) | 5 (0) | 1 (0) |
| Model 6: $(n,p,q) = (500,1000,1000)$, $(a,b) = (0.75,1)$ | | | | | | | | | | |
| Lasso | 2.21 (0.46) | 0.80 (0.19) | 5.0 (0.0) | 87.1 (29.3) | 0.24 (0.04) | 1.28 (0.78) | 0.56 (0.23) | 5.0 (0.0) | 26.8 (13.3) | 0.46 (0.11) |
| SCAD | 2.05 (0.36) | 0.83 (0.20) | 5.0 (0.0) | 61.2 (21.0) | 0.29 (0.06) | 0.93 (0.56) | 0.48 (0.23) | 5.0 (0.0) | 21.9 (12.3) | 0.54 (0.16) |
| MCP | 1.82 (0.36) | 0.80 (0.20) | 5.0 (0.0) | 33.7 (15.0) | 0.41 (0.09) | 0.84 (0.63) | 0.49 (0.27) | 5.0 (0.0) | 14.1 (8.9) | 0.66 (0.16) |
| Oracle | 0.76 (0.07) | 0.55 (0.10) | 5 (0) | 5 (0) | 1 (0) | 0.29 (0.11) | 0.26 (0.11) | 5 (0) | 5 (0) | 1 (0) |
| Model 7: $(n,p,q) = (500,1000,1000)$, $(a,b) = (0.5,0.75)$ | | | | | | | | | | |
| Lasso | 2.65 (0.42) | 0.86 (0.16) | 5.0 (0.0) | 86.3 (27.1) | 0.24 (0.04) | 2.06 (1.32) | 0.77 (0.29) | 5.0 (0.0) | 27.6 (15.1) | 0.47 (0.14) |
| SCAD | 2.39 (0.26) | 0.90 (0.17) | 5.0 (0.0) | 47.2 (19.0) | 0.34 (0.07) | 1.79 (0.80) | 0.70 (0.23) | 5.0 (0.0) | 28.9 (13.7) | 0.46 (0.15) |
| MCP | 2.26 (0.27) | 0.89 (0.17) | 5.0 (0.0) | 28.3 (13.1) | 0.45 (0.11) | 1.52 (1.02) | 0.68 (0.31) | 5.0 (0.2) | 16.5 (10.5) | 0.61 (0.16) |
| Oracle | 1.00 (0.07) | 0.61 (0.09) | 5 (0) | 5 (0) | 1 (0) | 0.40 (0.15) | 0.30 (0.12) | 5 (0) | 5 (0) | 1 (0) |
| Model 8: $(n,p,q) = (500,1000,1000)$, $(a,b) = (0.5,1)$ or $(0.05,0.1)$ | | | | | | | | | | |
| Lasso | 2.68 (0.58) | 0.78 (0.09) | 5.0 (0.0) | 95.4 (37.8) | 0.23 (0.05) | 2.01 (0.80) | 0.73 (0.21) | 5.0 (0.0) | 29.5 (13.0) | 0.44 (0.09) |
| SCAD | 2.38 (0.31) | 0.81 (0.08) | 5.0 (0.0) | 56.8 (20.2) | 0.30 (0.06) | 1.58 (0.63) | 0.65 (0.23) | 5.0 (0.1) | 25.7 (10.6) | 0.46 (0.10) |
| MCP | 2.16 (0.31) | 0.80 (0.08) | 5.0 (0.0) | 30.3 (12.9) | 0.43 (0.10) | 1.45 (0.83) | 0.66 (0.24) | 5.0 (0.2) | 16.7 (9.1) | 0.59 (0.15) |
| Oracle | 0.95 (0.07) | 0.58 (0.06) | 5 (0) | 5 (0) | 1 (0) | 0.41 (0.14) | 0.29 (0.11) | 5 (0) | 5 (0) | 1 (0) |

**Table 3**

Genes chosen by stability selection with maximum selection probability (values shown) at least 0.5 and *cis*-SNPs (SNPs within 10 cM of each gene) identified by applying the 2SR method with different penalties to the mouse obesity data. Asterisks indicate genes that overlap those selected by the PLS method.

| Gene | Lasso | SCAD | MCP | *cis*-SNPs |
|---|---|---|---|---|
| Igfbp2* | 0.56 | 0.69 | | rs3663003 |
| Lamc1 | | 0.70 | | |
| Sirpa | 0.51 | 0.55 | | |
| Gstm2* | 0.88 | 0.91 | 0.89 | rs4136518 |
| Ccnl2 | 0.50 | | | rs3720634 |
| Glcci1 | 0.56 | | | |
| Vwf* | 0.72 | 0.79 | 0.50 | |
| Irx3 | 0.62 | | | |
| Apoa4 | 0.65 | | | |
| Socs2 | | | 0.68 | rs3694833 |
| Avpr1a | 0.78 | | | |
| Abca8a | 0.50 | | | |
| Gpld1 | 0.50 | | | |
| Fam105a | | 0.60 | | |
| Dscam | | 0.60 | | |
| Slc22a3 | 0.90 | | | rs4137196, rs3722983, rs4231406 |
| 2010002N04Rik | 0.73 | 0.65 | | rs3661189, rs3655324 |