Honglang Wang
Depart. of Stat. & Prob.
wangho16@msu.edu

Omics Data Integration

Statistical Genetics/Genomics Journal Club

Summary and discussion of "Regularization Methods for High Dimensional Instrumental Variables Regression with an Application to Genetical Genomics"

Omics data integration @ Statistical Genetics/Genomics Journal Club @ MSU

Abstract *This is for the discussion of the paper (Lin et al., 2014). 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.*

Contents

1 Preliminary

Alcohol assumption has been found in observational studies to have a positive effect on coronary heart disease and negative effects on liver cirrhosis, some cancers and mental health problems. These findings, however, are strongly suspected to be confounded by factors like diet, lifestyle and socioeconomic factors. Thus, in order to inform public health recommendations on alcohol intake, for example, it's important to verify which, if any, of these observed associations is in fact causal for the relevant health outcome.

Suppose we are interested in the causal effect of cholesterol on coronary heart disease, and then what can we do?

1. Randomized controlled trials (RCTs), rendering all other explanations unlikely by design.

2. What if we are in the observational study situation? We need to have some other information to help to overcome the problem of unobserved confounding. Basically, we need to have another variable that is predictive of cholesterol but has no effect on coronary heart disease and is independent of the unobserved confounders.

3. But in general it's hard to find a variable that can be verified as a suitable IV for any particular problem.

4. Generic variants that are associated with cholesterol will be good candidates as IV because of Mendelian randomization.

Now let's introduce formally what is instrumental variables through conditional independence: (X, Y, U, G) where X is the predictor, Y is the response, U is unobservable confounder between X and Y, G is the instrument,

1. $G \perp\!\!\!\perp U$, that is G must be marginally independent of the confounder.

2. $G \not\perp\!\!\!\perp X$, that is G must not be marginally independent of X.

3. $G \perp\!\!\!\perp Y|(X, U)$, that is, conditionally on X and the confounder U, the instrument and the response are independent.

In the linear model

$$Y = \mathbf{X}^\intercal \boldsymbol{\beta}^0 + \epsilon,$$

we usually assume that $\text{Cov}(\mathbf{X}, \epsilon) = 0$ or more stringently $\mathbb{E}(\epsilon|\mathbf{X}) = 0$, which is called exogeneity. Fan et al. (2014) discussed this issue in high dimensional setting. The exogeneity can leads to inconsistency for the OLS. One classical way to solve this issue is to introduce the instrumental variables.

A good reference for this is Didelez et al. (2010).

2 Summary of the paper

In this article, the authors focus on the application of high dimensional sparse Instrumental Variables (IV) models to genetical genomics, where they are interested in associating gene expression data with a complex trait to identify potentially causal genes by using genetic variants as instruments.

Although this paper explained from the aspect that genetic variants can be used as instruments to help solve the issue of confounding when discovering the associations of gene expressions with the response of interest (Figure 2), we are interested in that this IV model can jointly analyze gene expression data and genetic variants in exploring their associations with complex traits.

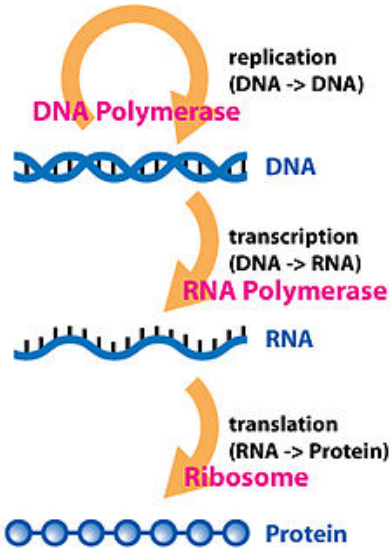

Figure 1: **Biology Dogma.**

2.1 Models

Consider the following linear IV model

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\eta}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma}^0 + \mathbf{E} \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}^0 \in \mathbb{R}^p, \boldsymbol{\Gamma}^0 := (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_p) \in \mathbb{R}^{q\times p}$ are unknown parameters, $\boldsymbol{\eta} \in \mathbb{R}^n, \mathbf{E} := (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n)^\intercal \in \mathbb{R}^{n\times p}$ are errors, and $(\boldsymbol{\epsilon}_i^\intercal, \eta_i) \in \mathbb{R}^{p+1} \big| \mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. $\mathbf{Y} \in \mathbb{R}^n$ are responses (disease traits), $\mathbf{X} \in \mathbb{R}^{n\times p}$ are predictors (expression levels) and $\mathbf{Z} \in \mathbb{R}^{n\times q}$ are instruments (generic variants). Without loss of generality, we assume the each variable is centered about zero and each column of $\mathbf{Z}$ is standardized to have L_2 norm $\sqrt{n}$. We assume that this model is sparse in the sense that only a small subset of the regression coefficients in $\boldsymbol{\beta}^0$ and $\boldsymbol{\Gamma}^0$ are nonzero.

We are not only interested in selecting and estimating important covariate effects, but also interested in identification and estimation of optimal instruments. That is our goal is to identify and estimate the nonzero coefficients in both $\boldsymbol{\beta}^0$ and $\boldsymbol{\Gamma}^0$.

Remark. 1. In order for $\mathbf{Z}$ to be valid instruments, we have to check the three conditions listed in the preliminary section, which can not be easily testable from the observed data, but can often be justified on the basis of plausible biological assumption.

2. Under the above model assumption, one important thing to notice is that we are not assuming that $\boldsymbol{\eta}$ and $\mathbf{X}$ be uncorrelated ($\mathbb{E}(\eta_i\mathbf{X}) = \mathbb{E}(\eta_i(\mathbf{Z}_i\boldsymbol{\Gamma}^0 + \boldsymbol{\epsilon}_i)) = \mathbb{E}(\eta_i\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{1:p,p+1}$). One standard way of eliminating such endogeneity issue is to replace the covariates by their expectations conditional on the instruments. This

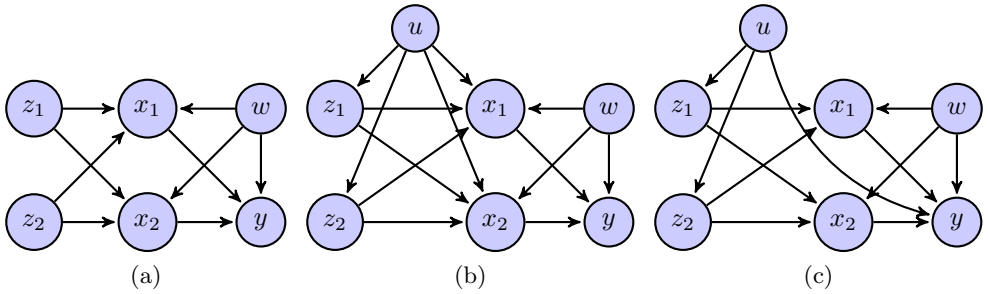

Figure 2: **Casual diagrams.** Causal diagrams showing the relationships between two genotypes z_1 and z_2, two gene expression levels x_1 and x_2, a clinical phenotype y, an unobserved phenotype w that confounds the associations between gene expression levels and the clinical phenotype, and an unobserved 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.

idea leads to the classical two-state least squares method, in which the covariates are first regressed on the instruments and the response is then regressed on the first-state predictions of the covariates.

2.2 Two-Stage Regularization

According to the above remark, the authors proposed the following 2SR methodology:

1. Regress the predictors on the instruments: prediction of the covariates $\hat{\mathbf{X}} = \mathbf{Z}\hat{\mathbf{\Gamma}}$:

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}\in\mathbb{R}^{q\times p}} \Big\{ \|\mathbf{X} - \mathbf{Z}\mathbf{\Gamma}\|_F^2/(2n) + \sum_{k=1}^{q}\sum_{j=1}^{p} p_{\lambda_j}(|\gamma_{kj}|) \Big\}. \tag{2}$$

2. Regress the response on the predicted covariates:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p}} \Big\{ \|\mathbf{Y} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2/(2n) + \sum_{j=1}^{p} p_{\mu}(|\beta_j|) \Big\}. \tag{3}$$

The penalty function $p_\lambda(t), \lambda > 0, t \geq 0$ could be chosen as the Lasso penalty, SCAD penalty, or MCP penalty. 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 algorithm (Mazumder et al., 2011).

And the $p + 1$ tuning parameters $\{\lambda_j, \mu, j = 1, 2, \cdots, p\}$ are selected by K-fold cross validation.

2.3 Theory—Under LASSO Penalty

To derive nonasymptotic bounds on the estimation and prediction loss of the regularized estimator $\hat{\mathbf{\Gamma}}$ and $\hat{\boldsymbol{\beta}}$, we impose the following conditions through the restricted eigenvalue condition which is defined in general for $\mathbf{A} \in \mathbb{R}^{n\times m}$ and $1 \leq s \leq m$

$$\kappa^2(\mathbf{A}, s) := \min_{S:|S|\leq s} \min_{\boldsymbol{\beta}\neq\mathbf{0}:\boldsymbol{\beta}\in\mathscr{C}(S,3)} \frac{\boldsymbol{\beta}^\intercal(\mathbf{A}^\intercal\mathbf{A}/n)\boldsymbol{\beta}}{\boldsymbol{\beta}^\intercal\boldsymbol{\beta}} \tag{4}$$

where $\mathscr{C}(S,3) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_{S^c}\|_1 \leq 3\|\boldsymbol{\beta}_S\|_1\}$. And we say that $\mathbf{A}$ satisfies RE$(s,3)$.

1. (C1) There exists $\kappa_1 > 0$ such that $\kappa(\mathbf{Z}, r) \leq \kappa_1$ with $r = \max_{1\leq j\leq p} \|\boldsymbol{\gamma}_j^0\|_0$.
2. (C2) There exists $\kappa_2 > 0$ such that $\kappa(\mathbf{Z}\mathbf{\Gamma}^0, s) \leq \kappa_2$ with $s = \|\boldsymbol{\beta}^0\|_0$.

We also assume that $\|\mathbf{\Gamma}^0\|_1 \leq L$ and $\|\boldsymbol{\beta}^0\|_1 \leq M$ for some constants L, M.

We have to recall several important results in Theorem 7.2 from Bickel et al. (2009):

Theorem 1 (Bickel et al. (2009))**.** *For the following linear model*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{5}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a response vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a fixed design matrix with columns standardized such that the diagonal elements of $\mathbf{\Sigma} = \mathbf{X}^\intercal\mathbf{X}/n$ are all equal 1, and $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is a vector of unknown regression coefficients with $\|\boldsymbol{\beta}^0\|_0 \leq s$ where $1 \leq s \leq p$. And we assume that $p \geq 2, n \geq 1$. Assume that $\mathbf{X}$ satisfies RE$(s,3)$. Consider the lasso solution $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1, \lambda > 0, \tag{6}$$

with $\lambda = C\sigma\sqrt{\frac{\log p}{n}}$ and $C > 2\sqrt{2}$. Then with probability at least $1 - p^{1-C^2/8}$, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq \frac{16C}{\kappa^2(\mathbf{X}, s)}\sigma s\sqrt{\frac{\log p}{n}}, \tag{7}$$

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n \leq \frac{16C^2}{\kappa^2(\mathbf{X}, s)}\sigma^2 s \log p/n. \tag{8}$$

Remark. In raw form, we have with probability at least $1 - p\exp(-n\lambda^2/(8\sigma^2))$, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq \frac{16}{\kappa^2(\mathbf{X}, s)}s\lambda, \tag{9}$$

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n \leq \frac{16}{\kappa^2(\mathbf{X}, s)}s\lambda^2. \tag{10}$$

Then by decomposing the optimization problem (2) into p penalized least squares problems, with $\|\mathbf{A}\|_1 := \max_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$, we have

Theorem 2. *Under condition (C1), if we choose* $\lambda_j = C\sqrt{\mathbf{\Sigma}_{j,j}}\sqrt{\frac{\log p+\log q}{n}}$ *with a constant* $C > 2\sqrt{2}$*, then with probability at least* $1-(pq)^{1-C^2/8}$*, we have with* $\sigma_{\max} := \max_{1\le j\le p}\sqrt{\mathbf{\Sigma}_{j,j}}$

$$\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^0\|_1 \le \frac{16C}{\kappa_1^2}\sigma_{\max} r\sqrt{\frac{\log p+\log q}{n}}, \tag{11}$$

$$\|\mathbf{Z}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^0)\|_F^2 \le \frac{16C}{\kappa_1^2}\sigma_{\max}^2 rp(\log p+\log q). \tag{12}$$

Proof. This follows from the union bound and that with probability $1-q\exp(-n\lambda_j^2/(8\mathbf{\Sigma}_{j,j}))$

$$\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^0\|_1 \le \frac{16}{\kappa_1^2} r\lambda_j, \tag{13}$$

$$\|\mathbf{Z}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^0)\|_2^2/n \le \frac{16}{\kappa_1^2} r\lambda_j^2, \tag{14}$$

by taking $\lambda_j = C\sqrt{\mathbf{\Sigma}_{j,j}}\sqrt{(\log p+\log q)/n}$. □

Theorem 3. *Under conditions (C1) and (C2), if we choose* $\lambda_j = C\sqrt{\mathbf{\Sigma}_{j,j}}\sqrt{\frac{\log p+\log q}{n}}$ *with a constant* $C > 2\sqrt{2}$*, and* $\lambda_{\max}(2L+\lambda_{\max}) \le \frac{\kappa_1^2\kappa_2^2}{32^2 rs}$*, where* $\lambda_{\max} = \max_{1\le j\le p}\lambda_j$*, then there exits 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* $C_0 = c_0 L\max(\sqrt{\mathbf{\Sigma}_{p+1,p+1}}, M\sigma_{\max})$*, then with probability at least* $1-c_1(pq)^{c_2}$*, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \le \frac{64C_0}{\kappa_1\kappa_2^2} s\sqrt{\frac{r(\log p+\log q)}{n}}, \tag{15}$$

$$\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \le \frac{64C_0^2}{\kappa_1^2\kappa_2^2} rs(\log p+\log q). \tag{16}$$

Proof. This proof follows that if we choose $\lambda_j = C\sqrt{\mathbf{\Sigma}_{j,j}}\sqrt{\frac{\log p+\log q}{n}}$ with a constant $C > 2\sqrt{2}$, and $\lambda_{\max}(2L+\lambda_{\max}) \le \frac{\kappa_1^2\kappa_2^2}{32^2 rs}$, where $\lambda_{\max} = \max_{1\le j\le p}\lambda_j$, with probability at least $1-\sum_{j=1}^p q\exp(-n\lambda_j^2/(8\mathbf{\Sigma}_{j,j}))$, the matrix $\hat{\mathbf{X}} = \mathbf{Z}\hat{\mathbf{\Gamma}}$ satisfies

$$\kappa(\hat{\mathbf{X}}, s) \ge \kappa_2/2.$$

This one actually just follows from Theorem 2 by comparing $\mathbf{Z}\hat{\mathbf{\Gamma}}$ and $\mathbf{Z}\mathbf{\Gamma}^0$.

And similar to the proof of Theorem (1) together with the result from Theorem (2), we have the fundamental inequalities under our conditions that

$$\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + \mu\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1/2 \le 2\mu\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1.$$

Once we have these two, the result follows directly from the definition of $\kappa(\hat{\mathbf{X}}, s)$. In fact since

$$\begin{aligned}
\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) \leq 2\mu\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 \leq 2\mu\sqrt{s}\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_2 \\
\mu\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1/2 \leq 2\mu\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 \Rightarrow \|\hat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}_{S^c}^0\| \leq 3\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 \\
\Rightarrow \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_2 \leq \frac{\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2}{\sqrt{n}\kappa(\hat{\mathbf{X}}, s)} \leq \frac{2\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2}{\sqrt{n}\kappa_2}
\end{aligned}$$

together we have

$$\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \leq \frac{64}{\kappa^2} ns\mu^2$$

and

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

□

Under some conditions, with the strong irrepresentable condition holds, the authors also proved the model selection consistency.

3 Simulation

We tried the simulation for the following date generation:

```
n=200;p=100;q=100
Gamma matrix: for each column, it has r=5 non-zero entries which are
IID from U([-1,-0.75],[0.75,1])
r=5
gam <- matrix(0, q, p)
for (j in 1:p)
        gam[sample(1:q, r), j] <- (2*rbinom(r, 1, 0.5) - 1)*runif(r, 0.75, 1)

beta vector: it has s=5 non-zero entries which are IID from
U([-1,-0.5],[0.5,1])
s=5
bet <- rep(0, p)
ind <- sample(1:p, s)
bet[ind] <- (2*rbinom(s, 1, 0.5) - 1)*runif(s, 0.5, 1)

Sigma matrix: for the first p*p submatrix, it's AR 1 structure with
rho=0.2, and for the last row (and then by symmetry also last column),
the (p+1, p+1) element is 1, and there are other s0=10 entries with
value 0.3, in particular, 5 of them are corresponding to the places with
non-zero beta's.
s0=10
```

```
sig <- matrix(, p + 1, p + 1)
sig <- 0.2^abs(row(sig) - col(sig))
sig[1:p, p + 1] <- 0
sig[c(ind, sample(setdiff(1:p, ind), s0 - s)), p + 1] <- 0.3
sig[p + 1, 1:p] <- sig[1:p, p + 1]
sig[,1+p]

e <- mvrnorm(n, rep(0, p + 1), sig)
z <- matrix(rbinom(n*q, 1, 0.5), n, q)
x <- z %*% gam + e[, 1:p]
y <- drop(x %*% bet) + e[, p + 1]
```

Note that although the parameters are generated randomly (and the authors did the simulation 50 times with randomly generated parameters for each time), we are going to fix the parameters and only randomly generated $\mathbf{Z}, \mathbf{E}$ and $\boldsymbol{\eta}$ for 100 repetitions.

We compare the penalized least squares (PLS) method without using the instruments and the 2 stage regularization (2SR) method for the five measures:

1. L_1 estimation error: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1$;
2. Prediction error: $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2/\sqrt{n}$;
3. True positives: number of successfully recovered signals;
4. Model size: number of selected variables;
5. Matthews correlation coefficient (MCC): a larger MCC indicates a better variable selection performance.

Table 1: Simulation results for $(n = 200, p = 100, q = 100, nsim = 100)$. Each performance measure was averaged over nsim=100 replicates with standard deviation shown in parentheses.

Method		L_1 estimation loss	Prediction loss	True Positive	Model size	MCC
Lasso	PLS	2.36 (0.44)	0.74 (0.09)	5 (0)	43.9 (9.11)	0.27 (0.05)
	2SR	1.79 (0.72)	0.89 (0.33)	5 (0)	15.89 (4.70)	0.55 (0.09)
SCAD	PLS	2.06 (0.52)	0.76 (0.10)	5 (0)	28.43 (7.68)	0.38 (0.08)
	2SR	1.26 (0.46)	0.73 (0.26)	5 (0)	12.92 (3.24)	0.61 (0.10)
MCP	PLS	2.01 (0.56)	0.75 (0.10)	5 (0)	22.67 (7.23)	0.44 (0.09)
	2SR	1.35 (0.62)	0.83 (0.33)	5 (0)	9.59 (2.90)	0.73 (0.12)
Oracle	PLS	0.80 (0.10)	0.54 (0.08)	5 (0)	5 (0)	1 (0)
	2SR	0.62 (0.24)	0.49 (0.18)	5 (0)	5 (0)	1 (0)

Table 2: Simulation results for $(n = 200, p = 100, q = 100, nsim = 100)$ without endogeneity. Each performance measure was averaged over nsim=100 replicates with standard deviation shown in parentheses.

Method		L_1 estimation loss	Prediction loss	True Positive	Model size	MCC
Lasso	PLS	0.91 (0.33)	0.38 (0.08)	5 (0)	21.98 (8.73)	0.46 (0.11)
	2SR	1.62 (0.59)	0.82 (0.26)	5 (0)	15.66 (4.63)	0.55 (0.09)
SCAD	PLS	0.30 (0.23)	0.20 (0.10)	5 (0)	7.63 (5.26)	0.88 (0.19)
	2SR	1.12 (0.54)	0.67 (0.26)	5 (0)	12.33 (3.63)	0.63 (0.10)
MCP	PLS	0.29 (0.19)	0.20 (0.08)	5 (0)	6.8 (3.69)	0.91 (0.16)
	2SR	1.23 (0.67)	0.78 (0.29)	5 (0)	9.51 (3.19)	0.74 (0.12)
Oracle	PLS	0.20 (0.07)	0.17 (0.05)	5 (0)	5 (0)	1 (0)
	2SR	0.58 (0.19)	0.48 (0.16)	5 (0)	5 (0)	1 (0)

4 Discussion

1. Here we introduce the instrumental variables in such a linear way. Any other way to introduce the instrumental variables?

2. How to study the uncertainty of the estimation and significance assignment?

3. For the model $Y = \alpha^0 Z + \mathbf{X}^\intercal \boldsymbol{\beta}^0 + \epsilon$, we want to do inference for α^0. One way to try is to do penalized least squares without penalty for α_0, i.e. to select important variables from $\mathbf{X}$ and then do post selection inference if we can insure that the selection procedure will include all of the important variables. But the issue here is that since most of the selection procedures are designed for prediction instead of learning about model parameters, any variable that is highly correlated with Z will tend to be dropped since including such a variable will tend not to add much predictive power for the outcome given that Z is already in the model.

   Since we never know whether we observed enough covariates or not and we may have lots of other unobserved predictors in ϵ, why not we also put $\mathbf{X}$ back to ϵ since we are only interested in Z itself? And then the collinearity issue between Z and $\mathbf{X}$ becomes the endogeneity problem between Z and the new error ϵ. Now in order to deal with the endogeneity issue, what if we regard $\mathbf{X}$ as instruments for Z? We then regress Z on $\mathbf{X}$ first to select the important instruments $Z = \mathbf{X}\boldsymbol{\gamma} + \eta$ by assuming sparsity for $\boldsymbol{\gamma}$. This is exactly the idea of Zhang and Zhang (2014).

   Then what if consider

$$Y = \alpha^0 Z + \mathbf{X}^\intercal \boldsymbol{\beta}^0 + \epsilon \tag{17}$$

$$Z = \mathbf{X}\boldsymbol{\gamma} + \eta \tag{18}$$

and post-double-selection method for inference proceeds by applying model selection methods to both equations and taking the selected controls as the union of controls selected from each equation? This selection is than followed by applying OLS to the selected controls. This is considered in Belloni et al. (2014).

References

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Vanessa Didelez, Sha Meng, and Nuala A Sheehan. Assumptions of iv methods for observational epidemiology. *Statistical Science*, pages 22–40, 2010.

Jianqing Fan, Yuan Liao, et al. Endogeneity in high dimensions. *The Annals of Statistics*, 42(3):872–917, 2014.

Wei Lin, Rui Feng, and Hongzhe Li. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, (just-accepted):00–00, 2014.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106 (495), 2011.