R2IVE: a user's manual

Qingliang Fan[*] Yaqian Wu[†]

December 3, 2021

1 Introduction

Instrumental variable (IV) regression is prevalent in economic studies involving endogenous treatment effects. The lack of practical guidance for instrument choice is a big challenge in research with observational data, especially when facing a large and mixed set of candidate IVs. It is usually unknown to the empirical researchers regarding the true identity of the high-dimensional covariates: some are excluded instruments, some are useful control variables (and among those, some are relevant for the treatment variable while some are not), and the rest are just noise. Thus, Fan and Wu (2020) developed an IV estimator that utilizes the rich dataset while robust to unqualified instruments and unknown control variables.

The method is a three-step procedure to select the desired instruments and useful control variables. First, the adaptive LASSO is used to select the relevant variables in the reduced form model for the endogenous treatment variable. Second, the treatment variable is replaced by its post-adaptive Lasso predicted value, and the useful controls are selected. In the third step, the treatment effect estimator is obtained via least squares by taking the selected controls and the predicted treatment variable as predictors. It is shown that this procedure has the desired oracle property: it can select the targeted instruments and controls consistently in the first and second steps, respectively. Therefore, it is called **R**obust **IV E**stimator (R2IVE) to both the **I**rrelevant instrument and uncertain **I**ncluded controls. The "2" in R2IVE denotes both types (reduced form and structural equation) of model uncertainty.

This document details an R function named `R2IVE` to implement the proposed method with demonstrative data. To be specific, the demonstrative data is generated as the following:

The structure equation we consider here is

$$Y_i = D_i\beta^* + \mathbf{Z}_{i.}^\top \boldsymbol{\alpha}^* + \varepsilon_i \tag{1.1}$$

where $\beta^* = 0.75$; $\mathbf{Z}_{i.} = (Z_{i1}, Z_{i2}, \ldots, Z_{i100})^\top$ is generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = (\rho_{j_1 j_2})_{100\times 100}$ with $\rho_{j_1 j_2} = 0.5^{|j_1 - j_2|}$, for $j_1, j_2 = 1, \ldots, 100$, and $i = 1, \ldots, 500$; $\boldsymbol{\alpha}^* = (\mathbf{0}_{14}, \boldsymbol{\iota}_{20}, \mathbf{0}_{66})^\top$ and $\boldsymbol{\iota}_{20}$ is a 1×20 vector of 1's, which means there are 20

[*] Department of Economics, The Chinese University of Hong Kong. E-mail: michaelqfan@gmail.com.

[†] School of Economics, Xiamen University. E-mail: yaqian2018@stu.xmu.edu.cn.

useful controls and the first 14 and the last 66 covariates are excluded IV. The endogenous variable D_i is generated based on the following reduced form model,

$$D_i = \mathbf{Z}_{i.}^{\top}\boldsymbol{\gamma}^* + \xi_i \tag{1.2}$$

where, $\boldsymbol{\gamma}^* = (2, 0.75, 1.5, 1, \ldots, \mathbf{0}_{80})^{\top}$ and $\mathbf{0}_{80}$ is a 1×80 vector of zeros, which means the first 20 covariates are relevant IVs. We fill in the values of non-zero elements in $\boldsymbol{\gamma}^*$ by replicating the non-zero elements of $(2, 0.75, 1.5, 1)$ until its length is 20. The error terms in the structural model and reduced form models are generated by

$$\begin{pmatrix} \varepsilon_i \\ \xi_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right)$$

2 Main functions

There are two main functions as described in Table 1. `R2IVE` implements the three-step aforementioned procedure, and `best.tuning` is used to choose the regularization parameter in each step. In this section, we start with the details of the R2IVE procedure and then present the input and output of the function `R2IVE` and `best.tuning`.

Table 1: Description of main functions

Functions	Description
`R2IVE`	Implement Robust IV Estimator (R2IVE) to both the Irrelevant instrument and uncertain Included controls.
`best.tuning`	Choose the optimal regularization parameter in each regularization estimate step and output the corresponding estimated coefficients.

2.1 Procedure of R2IVE

Step 1: Selection of relevant IVs

Consider the following objective function with an adaptive Lasso penalty

$$\ddot{\boldsymbol{\gamma}}_n = \arg\min_{\boldsymbol{\gamma}} \left\{ \|\mathbf{D} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda_n \sum_{j=1}^{L_n} \omega_j |\gamma_j| \right\} \tag{2.1}$$

where the adaptive weights are defined by $\omega_j = |\widetilde{\gamma}_j|^{-1}$ and $\widetilde{\boldsymbol{\gamma}}_n = (\widetilde{\gamma}_1, \ldots, \widetilde{\gamma}_{L_n})^{\top}$ is obtained from the least squares estimator $\widetilde{\boldsymbol{\gamma}}_n(\text{ols})$ when $L_n \ll n$ and elastic-net estimator $\widetilde{\boldsymbol{\gamma}}_n(\text{enet})$ when L_n is relatively large but no more than n. Specifically, the elastic-net estimator as initial $\widetilde{\boldsymbol{\gamma}}_n$ is defined as

$$\widetilde{\boldsymbol{\gamma}}_n(\text{enet}) = \left\{ \arg\min_{\boldsymbol{\gamma}} \|\mathbf{D} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda_2\|\boldsymbol{\gamma}\|_2^2 + \lambda_1\|\boldsymbol{\gamma}\|_1 \right\}, \tag{2.2}$$

Denote $\widehat{\mathcal{A}}_R = \{j : |\ddot{\gamma}_j| > 0\}$ as the adaptive Lasso selected instruments and then run the least squares of D_i on the selected IVs to obtain the refitted estimator $\widehat{\boldsymbol{\gamma}}_n$ and the predicted D_i: $\widehat{D}_i = \sum_{j \in \widehat{\mathcal{A}}_R} \widehat{\gamma}_j Z_{ij}$. Denote $\widehat{\mathbf{D}} = (\widehat{D}_1, \cdots, \widehat{D}_n)^{\top}$.

Step 2: Selection of useful controls

Before selecting the useful controls, the initial estimator of $\boldsymbol{\alpha}$ satisfying the adaptive irrepresentation condition needed to be constructed.

Step 2.1 : Initial estimator for $\boldsymbol{\alpha}$

Plugging (1.2) into (1.1), we have

$$Y_i = \mathbf{Z}_{i.}^{\top}\boldsymbol{\Gamma}^* + u_i \tag{2.3}$$

where, $\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^*\boldsymbol{\gamma}^*$ and $u_i = \varepsilon_i + \beta^*\xi_i$. Similar to (2.1), consider the following objective function to estimate $\boldsymbol{\Gamma}$

$$\ddot{\boldsymbol{\Gamma}}_n = \arg\min_{\boldsymbol{\gamma}} \left\{ \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\Gamma}\|_2^2 + \lambda_n'' \sum_{j=1}^{L_n} \omega_j''|\Gamma_j| \right\} \tag{2.4}$$

where the adaptive weights are defined by $\omega_j'' = |\widetilde{\Gamma}_j|^{-1}$ and $\widetilde{\boldsymbol{\Gamma}}_n = \left(\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_{L_n}\right)^{\top}$ is obtained from the least squares estimator $\widetilde{\boldsymbol{\Gamma}}_n(\text{ols})$ when $L_n \ll n$ and elastic-net estimator $\widetilde{\boldsymbol{\Gamma}}_n(\text{enet})$ when L_n is relatively large but no more than n similar to (2.2).

Run the least square of Y_i on the selected IVs from (2.4) to obtain the refitted estimator $\widehat{\boldsymbol{\Gamma}}_n$ and define

$$\widetilde{\beta} = \text{median}\left(\left\{\frac{\Gamma_j}{\gamma_j},\ j \in \widehat{\boldsymbol{\mathcal{A}}}_R\right\}\right) \tag{2.5}$$

Then, the initial estimator for $\boldsymbol{\alpha}$ is estimated by

$$\widetilde{\boldsymbol{\alpha}}_n(\text{enet}) = \left\{ \arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{Y} - \boldsymbol{D}\tilde{\beta} - \boldsymbol{Z}\boldsymbol{\alpha} \right\|_2^2 + \lambda_2'\|\boldsymbol{\alpha}\|_2^2 + \lambda_1'\|\boldsymbol{\alpha}\|_1 \right\}, \tag{2.6}$$

and let $\omega_j' = |\widetilde{\alpha}_j|^{-1}$.

Step 2.2: Selection of useful controls

There are two different methods to obtain the final estimator for $\boldsymbol{\alpha}$. One is via

$$\widehat{\boldsymbol{\alpha}}_n = \left\{ \arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{Y} - \boldsymbol{D}\tilde{\beta} - \boldsymbol{Z}\boldsymbol{\alpha} \right\|_2^2 + \lambda_n' \sum_{j=1}^{L_n} \omega_j'|\alpha_j| \right\} \tag{2.7}$$

The other is via

$$\widehat{\boldsymbol{\alpha}}_n = \left\{ \arg\min_{\boldsymbol{\alpha}} \left\| \widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\alpha} \right\|_2^2 + \lambda_n' \sum_{j=1}^{L_n} \omega_j'|\alpha_j| \right\} \tag{2.8}$$

where, $\widetilde{\mathbf{Y}} = \mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Y}$ and $\widetilde{\mathbf{Z}} = \mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Z}$. Here, $\mathcal{M}_{\mathbf{X}} = \mathbf{I}_n - \mathcal{P}_{\mathbf{X}}$, $\mathcal{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}$ is the projection matrix onto the column space of $\mathbf{X}$, and $\mathbf{I}_n$ is the n-dimensional identity matrix. See Fan and Wu (2020) for more details. Denote $\widehat{\boldsymbol{\mathcal{A}}}_C = \{j : |\hat{\alpha}_j| > 0\}$ as the selected set of included controls in the structural equation.

Step 3: Treatment effect estimation

The proposed IV estimator for the treatment effect β^*, R2IVE, is the least squares solution

$$\widehat{\beta} = \left(\widehat{\mathbf{D}}^\top \mathcal{M}_{\widehat{\mathcal{A}}_C} \widehat{\mathbf{D}}\right)^{-1} \widehat{\mathbf{D}}^\top \mathcal{M}_{\widehat{\mathcal{A}}_C} \mathbf{Y} \tag{2.9}$$

which is robust to the irrelevant IVs and unknown controls.

2.2 Function R2IVE

Tables 2 and 3 provide the input arguments and output values for function R2IVE. Note that the input lambda11, lambda12, lambda21 and lambda22 are regularization parameters in (2.1) and (2.7) (or 2.8). Other regularization parameters used in constructing initial estimators are selected by default criterion.

Table 2: Input arguments of the R2IVE function

Arguments	Description
y	A numeric vector of outcomes.
D	A numeric vector of observed treatment variable.
Z	A numeric matrix of instruments, with each column referring to one instrument.
intercept	A logical declaring whether the intercept is included in the structure model. Default is FALSE.
IV.intercept	A logical declaring whether the intercept is included in the reduced form. Default is FALSE.
lambda11	A user-specified regularization parameter for l_1 penalty in Step 1, $0 \leq$ lambda11 ≤ 1. Default value is determined by “BIC” criterion.
lambda12	A user-specified regularization parameter for l_2 penalty in Step 1, $0 \leq$ lambda12 ≤ 1. Default is 0. lambda11 $\neq 0$ and lambda12 $= 0$ are adaptive LASSO penalty, lambda11 $\neq 0$ and lambda12 $\neq 0$ are adaptive elastic-net penalty.
lambda21	A user-specified regularization parameter for l_1 penalty in Step 2, $0 \leq$ lambda21 ≤ 1. Default value is selected by “BIC” criterion.
lambda22	A user-specified regularization parameter for l_2 penalty in Step 2, $0 \leq$ lambda22 ≤ 1. Default is 0. Refer to the discussions in lambda12.
criterion	The criterion to select the regularization parameters: lambda11 in Step 1 and lambda21 in Step 2. Can choose of “CV”, “BIC” and “EBIC”, default is “BIC”.
nfolds	Number of cross-validation folds when criterion = “CV”.
tau	The significance level of the confidence interval. τ= 0.95 if 95% confidence interval is needed.
type	The algorithm type in step 2. type=1 if we use (2.7) to select controls and type=2 if (2.8) is used.

Table 3: Output values of the R2IVE function

Values	Description
coef	The coefficient of endogenous variable D.
ste	The standard deviation of the endogenous variables' coefficients.
whichrelevant	Estimated set of relevant instruments.
whichcontrol	Estimated set of useful controls.
Dhat	The predicted values of endogenous variable D.
upper	The upper bound of $\tau \times 100$ percent confidence interval for endogenous variables.
lower	The lower bound of $\tau \times 100$ percent confidence interval for endogenous variables.

2.3 Function best.tuning

Tables 4 and 5 provide the input arguments and output values for function best.tuning.

Table 4: Input arguments of the best.tuning function

Arguments	Description
X	A numeric matrix of predictors (of dimension $N * p$); each row is an observation vector.
y	A numeric vector of the response variable.
lambda	A user-specified regularization parameter for l_1 penalty, $0 \leq$ lambda ≤ 1. Default is decided by criterion "BIC".
lambda2	A user-specified regularization parameter for l_2 penalty, $0 \leq$ lambda2 ≤ 1. Default is 0. lambda $\neq 0$ and lambda2 $= 0$ are adaptive LASSO penalty, lambda $= 0$ and lambda $\neq 0$ and lambda2 $\neq 0$ are adaptive elastic-net penalty.
criterion	The criterion to select the regularization parameter lambda. One of "CV", "BIC" and "EBIC", default is "BIC".
nfolds	Number of cross-validation folds when criterion = "CV".
cons	A user-specified parameter to adjust "BIC" and "EBIC" criteria when type=2 in R2IVE function. Default is 1.
pf	l_1 penalty factor of length p used for adaptive LASSO or adaptive elastic net. Default is 1 for all variables.

Table 5: Output values of the `best.tuning` function

Values	Description
`beta`	The coefficient of predictors X.
`best.lambda`	lambda used in the algorithm.
`criterion`	The criterion to select the lambda.

3 An R code implementation of a demonstrative data

In this section, we give an R code implementation using demonstrative data. The data generation method is described in Section 1.

```
###GDP###
s1=20;s2=20;q=14;p=100;n=500
beta=0.75
gamma<-c(rep(c(2,0.75,1,1.5),len=s1),rep(0,p-s1))
alpha<-c(rep(0,q),rep(1,s2),rep(0,p-s2-q))
rele.index<-which(gamma!=0) #the set of true relevant IVs
col.index<-which(alpha!=0)  #the set of true useful controls
sigma_errors <- matrix(c(1,0.8,0.8,1),nrow=2,ncol=2,byrow =
   T)
sigma_iv<-matrix(0,p,p)
for(i in 1:p){
 for(j in 1:p){
  sigma_iv[i,j]<-0.5^abs(i-j)
 }
}
errors <- MASS::mvrnorm(n,c(0,0),sigma_errors)
epsilon<-errors[,1]; xi<-errors[,2]
Z<-MASS::mvrnorm(n,rep(0,p),sigma_iv)
D<-Z%*%gamma+xi
y<-Z%*%alpha+D*beta+epsilon
```

```
###tpye=1###
myfit1<-R2IVE(y,D,Z,intercept=FALSE,IV.intercept=FALSE,
criterion=c("BIC"),nfolds=10,tau = 0.95,type = 1)
results<-c(myfit1$coef,myfit1$ste,myfit1$lower,myfit1$upper)
names(results)<-c('coef','ste','lower','upper')
results
```

```
##      coef         ste        lower        upper
## 0.742069370  0.009437101  0.723572992  0.760565748
```

```r
length(myfit1$whichrelevant) #the number of selected relevant IVs
```

```
## [1] 20
```

```r
all(rele.index %in% myfit1$whichrelevant) #whether all true relevant IVs are selected
```

```
## [1] TRUE
```

```r
length(myfit1$whichrelevant) #the number of selected useful controls
```

```
## [1] 20
```

```r
all(col.index %in% myfit1$whichcontrol) #whether all true useful controls are selected
```

```
## [1] TRUE
```

```r
###type=2###
myfit2<-R2IVE(y,D,Z,intercept=FALSE,IV.intercept=FALSE,
criterion=c("BIC"),nfolds=10,tau = 0.95,type = 2)
results<-c(myfit2$coef,myfit2$ste,myfit2$lower,myfit2$upper)
names(results)<-c('coef','ste','lower','upper')
results
```

```
##        coef         ste       lower       upper
## 0.742069370 0.009437101 0.723572992 0.760565748
```

```r
length(myfit2$whichrelevant)
```

```
## [1] 20
```

```r
all(rele.index %in% myfit2$whichrelevant)
```

```
## [1] TRUE
```

```r
length(myfit2$whichcontrol)
```

```
## [1] 20
```

```r
all(col.index %in% myfit2$whichcontrol)
```

```
## [1] TRUE
```

References

Fan, Q., Wu, Y., 2020. Endogenous treatment effect estimation with some invalid and irrelevant instruments. arXiv preprint arXiv:2006.14998 .