Constrained Instrumental Variables in Mendelian Randomization with Pleiotropy (CIVMR)

Lai Jiang

11/09/2017

CIVMR is a package developed to carry out Mendelian Randomization (MR) analyses for individual level data with observed pleiotropy using constrained instrumental variables (CIV) and/or constrained instrumental variables with smoothed L0 penalty (CIV_smooth). It also includes various other methods, such as 2SLS, Allele scores and their variants, to assess the causal effect of a phenotype/exposure X on an outcome Y when pleiotropy phenotype(s) Z are observed using genotypes G as instruments.

The rest of this vignette is structured as follows:

- The solution and motivation of CIV/CIV_smooth.
- Brief description of the contents of the package CIVMR.
- An example of MR data analysis using this package.

1. Constrained Instrumental Variable Methods

Assume we have an individual level dataset of n samples. Let X be an exposure of interest (e.g. methylation levels at a specific CpG site) with dimension $n \times 1$. G is a selected $n \times p$ matrix containing p genotypes (e.g. SNPs) that are associated with X. Y is the outcome (e.g. blood lipid levels) with dimension $n \times 1$.

The Constraine instrumental Variable (**CIV**) for MR analyses $(\mathbf{G}, \mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is defined as $\mathbf{G}^* = \mathbf{G}\mathbf{c}$, where $\mathbf{c}$ is a weight vector $\mathbf{c} \in \mathbb{R}^p$, s.t.

$$\max_{\mathbf{c} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^r} \mathbf{c}^\top \mathbf{G}^\top \mathbf{X} \mathbf{v}$$

subject to conditions:

$$\begin{aligned} \mathbf{c}^\top \mathbf{G}^\top \mathbf{G} \mathbf{c} &= 1, \\ \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} &= 1, \\ \mathbf{c}^\top \mathbf{G}^\top \mathbf{Z} &= \mathbf{0}. \end{aligned}$$

The new CIV is then used directly to infer causal effect of $\mathbf{X}$ on $\mathbf{Y}$ from $(\mathbf{G}^*, \mathbf{X}, \mathbf{Y})$ using 2SLS. CIV can be embedded inside a bootstrap or cross-validation to find a sample bias-corrected CIV weight. We refer to these two methods as **"CIV_boot"** and **"CIV_cv"**.

Alternatively we can also impose smoothed L_0 penalty on the weight $\mathbf{c}$ to obtain an approximate sparse solution of $\mathbf{c}$, s.t.

$$\max_{\mathbf{c} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^r} \mathbf{c}^\top \mathbf{G}^\top \mathbf{X} \mathbf{v} - \lambda (p - \sum_j f_\sigma(\mathbf{c}_j))$$

subject to conditions:

$$\begin{aligned} \mathbf{c}^\top \mathbf{G}^\top \mathbf{G} \mathbf{c} &\leq 1, \\ \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} &\leq 1, \end{aligned}$$

$$\mathbf{c}^\top \mathbf{G}^\top \mathbf{Z} = \mathbf{0},$$

given a specific value of regularization parameter λ.

In our algorithm, the value of regularization parameter λ is chosen to minimizes the projected prediction error $||\mathbf{P}_{\mathbf{G}^*}(\mathbf{Y} - \mathbf{X}\beta_*)||$, where $\mathbf{P}_{\mathbf{G}^*} = \mathbf{G}^{*\mathbf{T}}(\mathbf{G}^{*\mathbf{T}}\mathbf{G}^*)^{-1}\mathbf{G}^*$. β_* is the corresponding causal effect estimation using new instruments $\mathbf{G}^* = \mathbf{Gc}$, which is referred to as **"CIV_smooth"** method in this package.

The motivation of **CIV** is to construct a relatively strong instrumental variable **Gc** (a linear combination of variables in **G**) that are uncorrelated with **Z**. In addition, **CIV_smooth** is designed to incorporate smooth L0 penalty within the same framework, while achiving approximate sparse solution of **c** and deliver robust MR analyses. More technical details of **CIV** and **CIV_smooth** can be found in our paper[1].

2. Package Contents

2.1 Datasets

CIVMR provides two data files: 'sim.rda' and 'ADNI.rda' for illustration of MR analysis. The first dataset is generated within the framework of simulation series I in the CIV paper. It contains 500 subjects with one phenotype (X) of interest, one pleiotropic phenotype (Z) and one outcome (Y) simulated for each subject. 9 SNPs were generated and they were associated with both X and Z. The true causal effect of X on Y is 1. This data file is simulated to serve as an example to estimate the causal effect of X on Y while accounting for potential pleiotropic effect from Z. Users can compare the performance of different MR methods on this simulation dataset since the true causal effect is known.

'ADNI.rda' is a subset of the Alzheimer's Disease Neuroimaging Initiative (ADNI) project[2]. It contains 491 subjects, whose phenotypes ($A\beta$, Ptau, Ttau, Glucose levels) and Alzheimer's status were collected. 20 SNPs associated with Amyloid beta ($A\beta$) were selected and their dosage of the 491 subjects were recorded. The target of this data is to serve as an example of estimating the causal effect of Amyloid beta on progression of Alzheimer's disease while accounting for potential pleiotropic effect from other phenotypes (Ptau, Ttau, Glucose levels). More details of this dataset can be found in "ADNI" session of reference manual inside package.

The default format of input data is data.frame() for all functions of instrumental variable methods in CIVMR. An example of the SNP data in the 'simulation' file is shown as below:

```r
library(CIVMR)
data(simulation)
G <- simulation$G
colnames(G) <- paste("rs", seq(1:9), sep="")
rownames(G) <- paste("sub", seq(1:nrow(G)), sep="")
knitr::kable(G[1:5,])
```

	rs1	rs2	rs3	rs4	rs5	rs6	rs7	rs8	rs9
sub1	2	1	0	0	1	0	1	1	1
sub2	2	0	2	2	0	0	1	1	0
sub3	0	0	1	0	1	1	0	0	2
sub4	1	1	1	2	2	0	1	2	0
sub5	0	2	1	2	2	2	1	1	1

[1] "Constrained Instruments and its Application to Mendelian Randomization with Pleiotropy"

[2] Mueller, Susanne G., et al. "Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI)." Alzheimer's & Dementia 1.1 (2005): 55-66.

2.2 MR Methods

CIVMR contains functions to implement several instrumental variable methods, including CIV_boot, CIV_cv, CIV_smooth, 2SLS and Allele scores. Given pleiotropic phenotypes Z, CIV/CIV_smooth methods construct new instruments to adjust causal effect estimation automatically within the algorithm; while Allele score method and 2SLS method need extra instructions to account for pleiotropic Z.

There are two different ways to account for Z in 2SLS and Allele score method. The first method is to replace G with residuals after regressing each of these genetic variants on Z. This is known as "adjusting" according to Z. The second way is to apply multiple linear regression of Y on $\hat{X}$ and $\hat{Z}$ jointly, where $\hat{X}$ and $\hat{Z}$ are the predicted phenotypes using G (or corresponding Allele score) as the instruments. This is known as "multiple" MR analyses with respect to Z. The functions in CIVMR implementing instrumental variable methods are listed here.

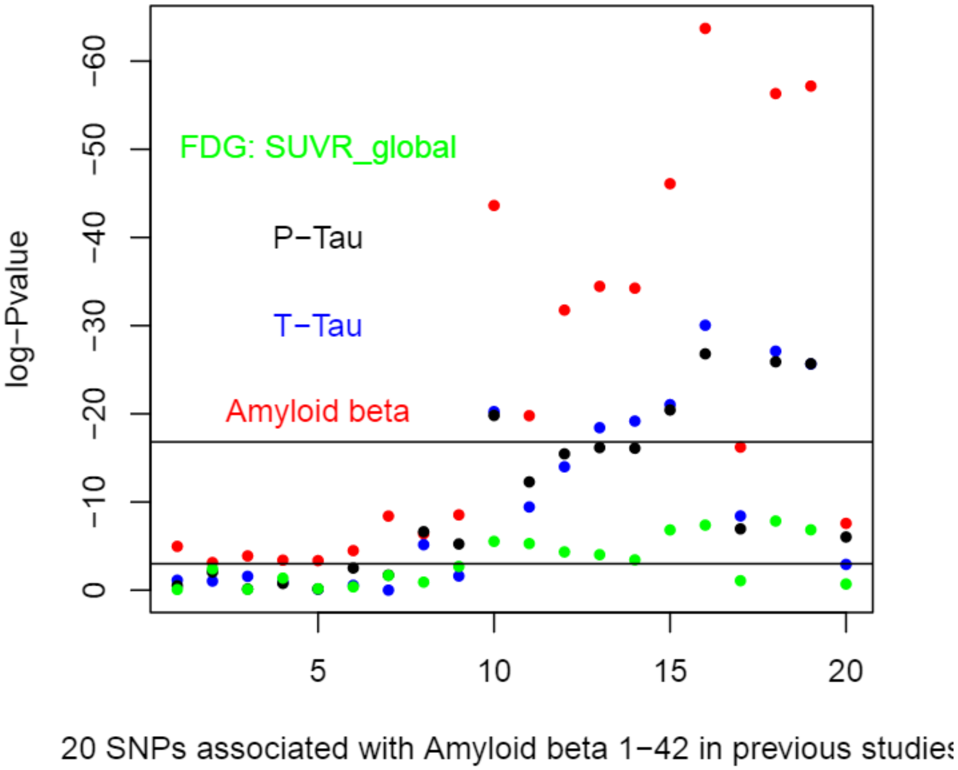
- CIV(): Constrained instrumental variable method.
- cv_CIV(): Cross-validated Constrained instrumental variable method.
- boot_CIV(): Bootstrap corrected constrained instrumental variable method.
- smooth_CIV(): Constrained instrumental variable method with smooth L0 penalty.
- TSLS_IV(): Two stage least square regression method.
- allele(): Cross-validated Allele score method.

More details of these functions can be found in the reference manual. Technical details of CIV and CIV_smooth can be found in the CIV paper.

3. Example

Finally, we show an example of using this package to conduct MR analysis on real data analysis. First we load the ADNI data and evaluate the $G - X$ associations. The pvalues of univariate T-test for all SNPs in G with respect to Amyloid β (X) is shown in the figure:

```r
library(CIVMR)
data(list=ADNI)
list_mat <- NULL
list_mat <- rbind(list_mat, lmPvalue(ADNI$X,ADNI$G) )
list_mat <- rbind(list_mat, lmPvalue(ADNI$Z[,1],ADNI$G) )
list_mat <- rbind(list_mat, lmPvalue(ADNI$Z[,2],ADNI$G) )
list_mat <- rbind(list_mat, lmPvalue(ADNI$Z[,3],ADNI$G) )
list_mat <- log(list_mat)

plot(list_mat[1,],col="red", ylab ="log-Pvalue",  pch =20, ylim=c(0,min(list_mat)),
     xlab="20 SNPs associated with Amyloid beta 1-42 in previous studies")
text(5, -20, "Amyloid beta" ,col="red" )
text(5, -30, "T-Tau" ,col="blue" )
text(5, -40, "P-Tau" ,col="black" )
text(5, -50, "FDG: SUVR_global" ,col="green" )
points(list_mat[2,], col="blue", pch =20 )
points(list_mat[3,], col="black", pch =20)
points(list_mat[4,], col="green", pch =20)
abline(h=log(0.05))
abline(h=log(5*10^(-8)))
```



This Figure shows the association test result between the 4 phenotypes of interest and 20 selected SNPs. The X-axis corresponds to the 20 selected SNPs, the y-axis shows the log-pvalues of univariate t-test for each of these SNPs. The different colors denote the association with different phenotypes. The two horizental line corresponds to the threshold 0.05 and 5×10^{-8}.

It can bee seen that most of the 20 SNPs are significantly associated with Amyloid beta under threshold 0.05, this is expected since they are chosen because of their strong associations in previous studies. However, we can also see some of these SNPs are also strongly associated with Ptau/Ttau. This raises questions about pleiotropy, and also the appropriateness of using these SNPs as valid instruments G to infer causal effect of Amyloid beta (X) on AD progression (Y). To assess the validity of using these SNPs as instruments, we run the Sargan test:

```
library(AER)
fm <- ivreg(Y ~ X | G, data = ADNI)
summary(fm, vcov = sandwich, df = Inf, diagnostics = TRUE)

##Call:
##ivreg(formula = Y ~ X | G, data = ADNI)

##Residuals:
##     Min      1Q  Median      3Q     Max
##-0.9314 -0.4743  0.1744  0.3129  0.7423

##Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
##(Intercept) -1.513e-17  1.958e-02    0.00        1
##X           -2.501e-03  5.777e-04   -4.33 1.49e-05 ***

##Diagnostic tests:
##                  df1 df2 statistic p-value
##Weak instruments  20 470    17.375  <2e-16 ***
##Wu-Hausman         1 488     0.001   0.976
##Sargan            19  NA    23.312   0.224
##---
##Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##Residual standard error: 0.4347 on Inf degrees of freedom
##Multiple R-Squared: 0.09217,  Adjusted R-squared: 0.09031
##Wald test: 18.75 on 1 DF,  p-value: 1.494e-05
```

The Sargan test gives Pvalue = **0.224**, it is not significant to reject the null hypothesis that G are valid instruments for Mendelian randomization for Amyloid beta (X) on AD progression (Y). However, this does not rule out the pleiotropy effect since all 20 SNPs are also associated with Ptau and Ttau. Here, we conduct MR analysis using instrumental variable methods within CIVMR to account for potential pleiotropy:

```
G <- lm(G~Z,data=ADNI)$residuals
X <- lm(X~Z,data=ADNI)$residuals
Y <- lm(Y~Z,data=ADNI)$residuals
ADNI.adj <- list(G=G,X=X,Y=Y)

ADNI.mul <- list(G=ADNI$G, X=cbind(ADNI$X,ADNI$Z),Y=ADNI$Y)

#2SLS method (naive)
tsls.naive <- TSLS_IV(ADNI, Fstats=TRUE, var_cal=TRUE)
#2SLS method (adjusted)
tsls.adj <- TSLS_IV(ADNI.adj, Fstats=TRUE, var_cal=TRUE)
#2SLS method (multiple regression)
ADNI.mul <- list(G=ADNI$G, X=cbind(ADNI$X,ADNI$Z),Y=ADNI$Y)
tsls.mul <- TSLS_IV(ADNI.mul, Fstats=TRUE, var_cal=TRUE)
```

First we run two-stage least square regression in three ways to account for Z: (1) naive way: do nothing; (2) Adjust G, X, Y according to Z first; (3) multiple regression for using (X, Z) as a combined phenotype.

Then we can implement allele score method in two ways: (1) naive Allele score method with cross-validated weights; (2) adjust for Z before running allele score method:

```
#1. naive allele score method.
allele.est <- allele(ADNI)

#2. allele score method while adjusting for pleiotropic phenotypes Z
allele.adj <- allele(ADNI.adj)
```

Then we implement Constrained instrumental variable (CIV) methods. We first remove the highly correlated SNPs with a certain criteria (cor>0.99) to avoid singularity of CIV algorithm. The function CIV() is then applied to obtain CIV instrument, which will be used to infer causal effect of X on Y with TSLS_IV(). Two variants of CIV methods: CIV_CV (CIV with cross-validation) and CIV_boot (bootstrap corrected CIV) are also available with function cv_CIV() and boot_CIV():

```
#this step is not mandatory
sel.G <- IV_reduction(ADNI$G, crit_high_cor = 0.8)$sel_snp
```

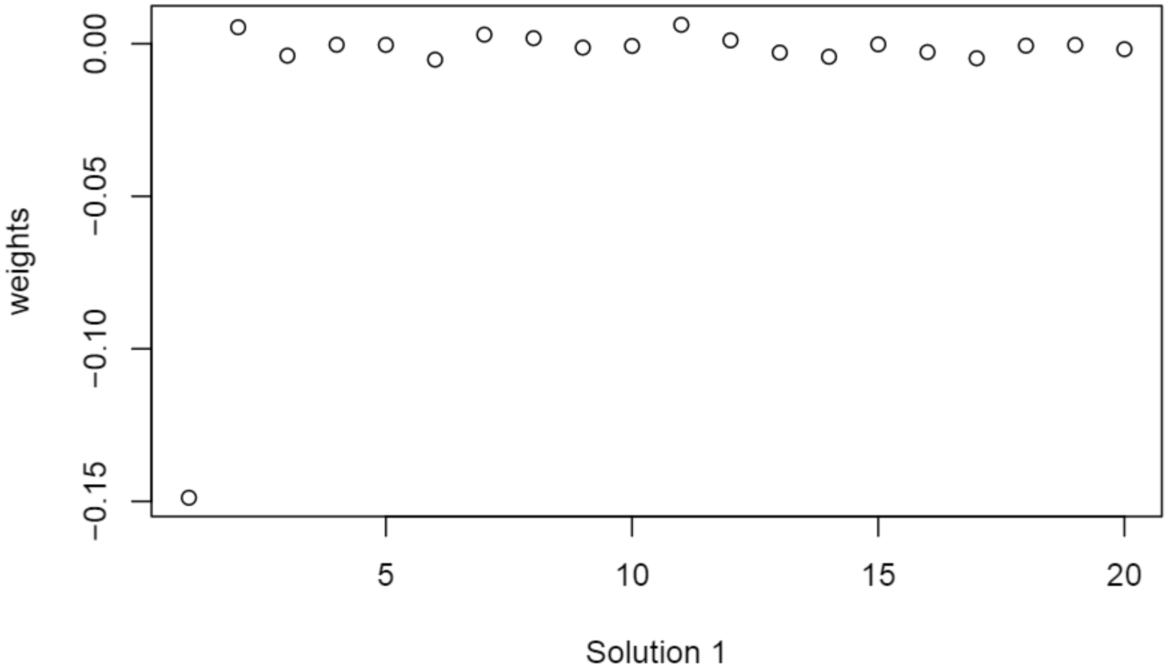
```
#then calculate the CIV instrument
civ.G <- CIV(list(G=sel.G,X=ADNI$X,Z=ADNI$Z,Y=ADNI$Y) )$CIV
#then run the TSLS estimation method
civ.est <- TSLS_IV(list(G=civ.G, X = ADNI$X, Y=ADNI$Y, y=ADNI$Y),
                   Fstats=TRUE, var_cal=TRUE)

#2. then the bootstrapped CIV instrument
civ.boot.G <- sel.G %*% (boot_CIV(list(G=sel.G,X=ADNI$X,Z=ADNI$Z,Y=ADNI$Y) )$boots.cor.u)
#civ.boot estimation.
civ.boot.est <- TSLS_IV(list(G= civ.boot.G, X = ADNI$X, Y=ADNI$Y, y=ADNI$Y),
                        Fstats=TRUE, var_cal=TRUE)

#3. then the croos-validated CIV instrument.
civ.cv.G <- cv_CIV(list(G=sel.G,X=ADNI$X,Z=ADNI$Z,Y=ADNI$Y))
```

At last we implement CIV_smooth() algorithm. Note that some of the CIV_smooth solutions are only slightly different from each other, which is expected since the algorithm converges to limited number of local maximum solutions even starting from multiple initial points. We then extract distinct solutions by removing the highly correlated/repeated IVs (keep representaives).

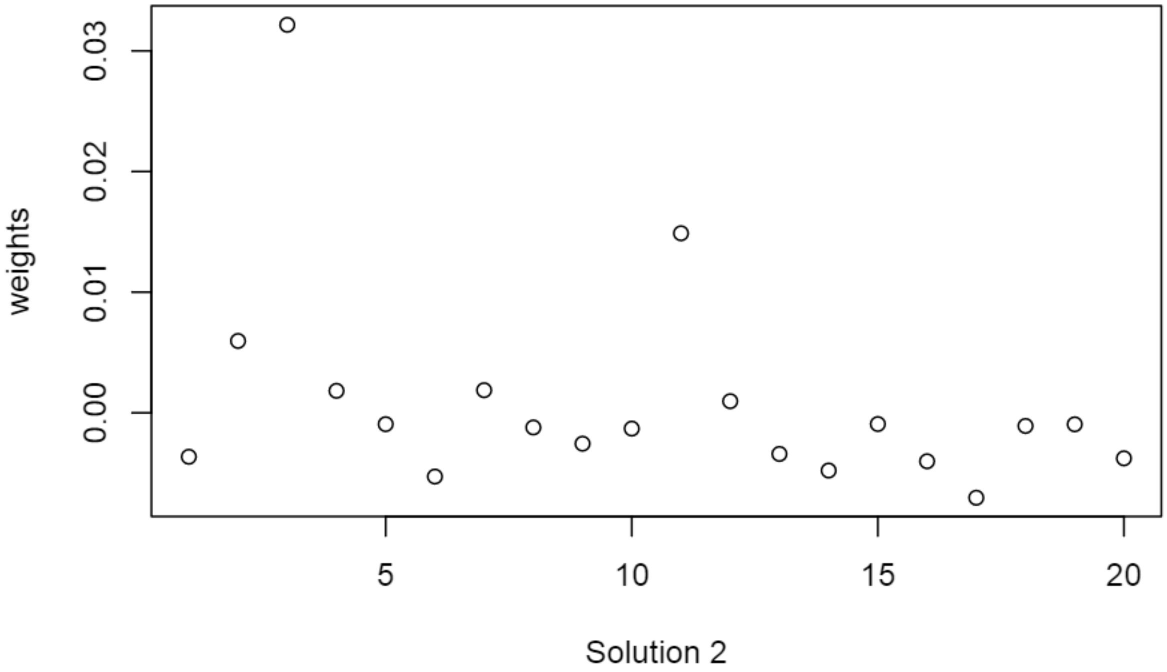
```
#Find CIV_smooth solution.
civ.smooth <- smooth_CIV(ADNI$G,ADNI$X,ADNI$Z,ADNI$Y,n_IV = 10)

#this step is not mandatory
civ.iv <- IV_reduction(civ.smooth$IV_mat, crit_high_cor = 0.9)

#now we obtain the causal effect estimation.
civ.smooth.est <- TSLS_IV(list(G=civ.iv$sel_snp, X = ADNI$X, Y=ADNI$Y, y=ADNI$Y),
                          Fstats=TRUE, var_cal=TRUE)
```

There are distinct solutions/weights been found from smooth_CIV(). The figure of these two vectors of weight are draw here:

```
plot(civ.smooth$u_mat[,civ.iv$id_snp[1]],xlab="Solution 1",ylab="weights")
```



```r
plot(civ.smooth$u_mat[,civ.iv$id_snp[2]],xlab="Solution 2",ylab="weights")
```



As we can see, the CIV_smooth solutions are approximately sparse. Extra selection procedure can be implemented based on these solutions. The consolidated causal effect estimation of all instrumental variable methods from CIVMR are shown as follows:

	causal effect estimation	s.d estimation
2SLS.naive	-0.0025	0.0006
2SLS.adj	-0.0015	0.0008
2SLS.mul	-0.0010	0.0015
Allele.naive	-0.0023	NA
Allele.adj	-0.0013	NA
CIV	-0.0024	0.0011
CIV_boot	-0.0006	0.0012
CIV_cv	-0.0014	NA
CIV_smooth	-0.0021	0.0020

Note that the variance estimation for CIV_cv and Allele score does not exist because they all rely on the cross-validation. The variance estimation for CIV, CIV_boot and CIV_smooth only captures the variance of the causal effect in the second stage regression and does not include the variance of the instrumental variable itself. Because of the limited the computational efficiency of CIV_smooth(), it is recommended for users to use the bootstrapped method with parallel computing techniques to obtain more accurate estimation of CIV_smooth causal effect estimator variance.