Example for MVMR Adaptive Lasso Functions

July 12, 2021

There are two functions to implement the instrumental variable (IV) selection and estimation procedure with two endogenous variables based on adaptive Lasso:

- *MVadap.cv*
- *MVadap.dt*

In function ***MVadap.cv***, the penalty parameter λ for adaptive Lasso is chosen by 10-fold cross validation. Results are reported for two values of the penalty parameter: λ_{cv} that gives the minimum cross-validation Sargan statistics, and λ_{cvse} using the one-standard-error rule.

In function ***MVadap.dt***, the penalty parameter λ is chosen by a downward testing procedure based on the Sargan test: start with a large value of λ where all the candidate instruments are selected as valid. Decrease λ so that for each Lasso step, the number of selected valid instruments decreases by 1. For each Lasso step, do Sargan test on the selected valid instruments, and choose the model with the largest degrees of freedom.

The usage of the two functions is illustrated by the following example:
There are 21 candidate instruments with the first 9 being invalid. $\boldsymbol{D}$ consists of two endogenous variables, and $\boldsymbol{X}$ contains two exogenous control variables. $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are the outcome variables and candidate instruments respectively. The causal effects $\boldsymbol{\beta} = (0.5, 1)$. See the end of the file for parameters specification and DGP.

Use ***MVadap.cv*** for IV selection and estimation:

```
source("MVMR Adptive Lasso Functions.R")
output.cv <- MVadap.cv(Y,D,Z,X)
```

$\boldsymbol{X}$ is an optional input. Apart from $\boldsymbol{Y}$, $\boldsymbol{D}$, $\boldsymbol{Z}$ and $\boldsymbol{X}$, another input is ***alpha*** which specifies the the significance level for the confidence intervals for the causal effect estimate (default = 0.05).

Output:

```
Invalid Instruments (minimum cross-validation Sargan statistics, cv):
 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z9

Number of Invalid Instruments (cv):
 9

Invalid Instruments (one-standard-error rule,cvse):
 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z9

Number of Invalid Instruments (cvse):
 9

Coefficients:
          Post 2SLS(cv)     SE(cv) Post 2SLS(cvse)   SE(cvse) ALasso(cv) ALasso(cvse) Median-of-medians
D1           0.53243824 0.04760147      0.53243824 0.04760147  0.7343060    0.8671360         0.6147971
D2           0.96090864 0.04985396      0.96090864 0.04985396  0.7188228    0.6117343         0.9526667
intercept    0.09528516 0.02296877      0.09528516 0.02296877         NA           NA                NA
X1           0.49358278 0.02357374      0.49358278 0.02357374         NA           NA                NA
X2           0.79970313 0.02371402      0.79970313 0.02371402         NA           NA                NA

Confidence Interval (cv): [0.4391,0.6257], [0.8632,1.0586]
Confidence Interval (cvse): [0.4391,0.6257], [0.8632,1.0586]

p-value of Sargan (cv): 0.6316876
p-value of Sargan (cvse): 0.6316876
```

In the output of the function, **all the results are report for λ_{cv} and λ_{cvse} respectively.** It includes the names/identities of the selected invalid instruments, and the number of the selected invalid instruments. Estimation results are reported for the coefficients of $\boldsymbol{D}$ and $\boldsymbol{X}$ (include an intercept). Estimators reported are (1) post-selection 2SLS estimators (*Post 2SLS*) and their standard errors (*SE*), (2) the adaptive Lasso estimators (*ALasso*), and (3) the median-of-medians estimators. Confidence intervals for the coefficients estimates of $\boldsymbol{D}$ are also reported. *p-value of Sargan* is the Sargan test results for the selected valid instruments.

Use ***MVadap.dt*** for IV selection and estimation:

```
source("MVMR Adptive Lasso Functions.R")
output.dt <- MVadap.dt(Y,D,Z,X)
```

Other inputs are ***alpha*** which specifies the the significance level for the confidence intervals for the causal effect estimate (default = 0.05), and ***tuning*** which specifies the threshold p-value for the Sargan test in the downward testing procedure (default = 0.1/log(n)).

Output:

```
Invalid Instruments:
 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z9

Number of Invalid Instruments:
 9

Coefficients:
           Post 2SLS         SE Median-of-medians
D1         0.53243824 0.04760147         0.6147971
D2         0.96090864 0.04985396         0.9526667
intercept 0.09528516 0.02296877                NA
X1         0.49358278 0.02357374                NA
X2         0.79970313 0.02371402                NA

Confidence Interval: [0.4391,0.6257], [0.8632,1.0586]

p-value of Sargan: 0.6316876
```

The output is similar to that of ***MVadap.cv***, except that the estimation results only include the post-selection 2SLS estimators and their standard errors, and the median-of-medians estimators.

Parameters specification:

```
library(MASS)
```

```
n <- 2000
L <- 21 # number of candidate instruments
s <- 9  # number of invalid instruments

gamma1 <- runif(L,1.5,2.5);
gamma2 <- runif(L,1.5,2.5);
gamma <- cbind(gamma1, gamma2) # first stage coefficients
cgamma <- 0.4

# direct effects of instruments on the outcome
pi <- c(rep(1,s),rep(0,L-s))
cpi <- 0.6

beta <- c(0.5, 1) # causal effects

# variance-covariance matrix of instruments
Sigmaz <- matrix(rep(0,L*L), nrow=L)
for(i in 1:L){
  for(j in 1:L){
    Sigmaz[i,j] <- 0.5^abs(i-j)
  }
}

# variance and covariance of the error terms
epsvar <- 1
rho1 <- 0.25
rho2 <- 0.3
```

Data are generated by

```
# Generate data
error <- rnorm(n,0,epsvar)
Z <- matrix(mvrnorm(n,mu=rep(0,L),Sigma=Sigmaz), ncol=L)

errorD <- matrix(rnorm(2*n,0,1),nrow=n, ncol=2) + cbind(rho1*
   error, rho2*error)
D <- 0.1 + Z %*% gamma * cgamma + errorD
colnames(D) = NULL

# two exogenous control variables
X <-  matrix(mvrnorm(n,mu=rep(0,2),Sigma=diag(1,2)), ncol=2)
eta <- c(0.5, 0.8)

Y <- 0.1  + D %*% beta + X %*% eta + Z %*% pi * cpi + error
```