Simulation

Jasmine Ju

June 4th 2016

Outline

- Simulation Results (Thesis Paper Setup)
- Why PostLasso is Better than Lasso?
  Exploration of 1st Stage Estimation
- Simulation Results (JASA Paper Setup)

Simulation Setup (Thesis paper)

- Data generation

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$$

$$\boldsymbol{X} = \mathbf{Z}^T\Pi + \omega$$

$$(\epsilon_i, \omega_i) \sim N\left(0, \begin{bmatrix} 1 & \sigma_{\epsilon\omega} \\ \sigma_{\epsilon\omega} & \sigma_\omega^2 \end{bmatrix}\right)$$

- $\mathbf{z}_i = [z_{i1}, ..., z_{iq}]^T \sim N(0, \Sigma_z)$, $\mathrm{Corr}(z_{ih}, z_{ij}) = 0.5^{|i-h|}$
- Set $\beta = 1$, the strength of the instruments $F = 10, 40, 160$

$$\sigma_\omega^2 = \frac{n\Pi^T\Sigma_z\Pi}{F\Pi^T\Pi}$$

- Consider $\mathrm{Corr}(\epsilon, \omega) = 0.3$ and $\mathrm{Corr}(\epsilon, \omega) = 0.6$

Simulation Setup $n > p$ (Thesis paper)

- Example 1: $\Pi = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $n = 20$, $p = 8$
- Example 2: $\Pi_i = 0.85$, $i = 1, ..., 8$, $n = 20$, $p = 8$
- Example 3: $\Pi = (\text{rep}(3, 15), \text{rep}(0, 15))$, $n = 50$, $p = 40$
  $z_i = W_1 + \nu_i, W_1 \sim N(0, 1), i = 1, ..., 5$
  $z_i = W_2 + \nu_i, W_1 \sim N(0, 1), i = 6, ..., 10$
  $z_i = W_3 + \nu_i, W_1 \sim N(0, 1), i = 11, ..., 15$
  $z_i \sim N(0, 1), i = 16, ..., 40$
  ν_i are i.i.d $N(0, 0, 01)$
- Example 4: $\Pi = (\text{rep}(1, 5), \text{rep}(0, 95))$, $n = 500$, $p = 100$
- For each example, we tried $\text{Corr}(\epsilon, \omega) = 0.3/\text{Corr}(\epsilon, \omega) = 0.6$ and presented the results separately (as in the thesis paper).

Data Generation for Example 3?

- Code in the Thesis:

```
dat.function <- function(n,p,cor,beta,pi1,fstar){
      # create the column of the matrix Z
      x1 <- replicate(5,rnorm(n)) + rnorm(n,sd=sqrt(0.01))
      x2 <- replicate(5,rnorm(n)) + rnorm(n,sd=sqrt(0.01))
      x3 <- replicate(5,rnorm(n)) + rnorm(n,sd=sqrt(0.01))
      x4 <- replicate(25,rnorm(n))
      Z <- cbind(x1,x2,x3,x4)
      # generete e_n and v_n
      cov.matrix <- cov(Z)
      sigmav <- n*(t(pi1)%*%cov.matrix%*%pi1)/(fstar*t(pi1)%*%pi1)
      cov_ve <- cor*sqrt(sigmav)
      covmatr.error <- matrix(c(1,cov_ve,cov_ve,sigmav),ncol=2)
      mat <- rmvnorm(n,sigma=covmatr.error)
      v <- mat[,2]
      e <- mat[,1]
      # generate endogenous variable
      X <- Z%*%pi1 + v
      # Response variable y
      Y <- beta*X + e
      #Output
      out<-list(dat=data.frame(Y=Y,X=X,Z=Z))
      out
}
```

- W part for $Z_1, \ldots, Z_5$ are different?

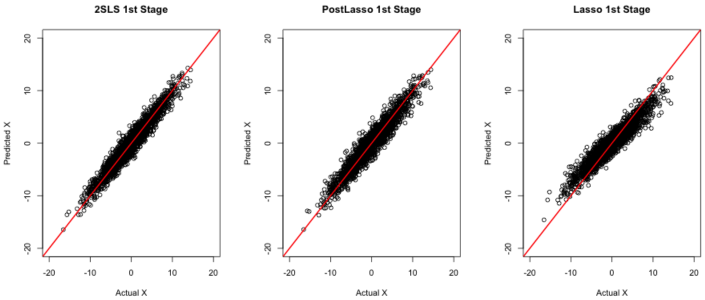
Results: $Cor(\epsilon, v) = 0.3, p < n$, RMSE of $\hat{\beta}$

	2SLS	PostLasso	Lasso	F
Example 1	0.06	0.05	0.11	10
Example 2	0.06	0.06	0.25	10
Example 3	0.01	0.01	0.09	10
Example 4	0.02	0.01	0.75	10
Example 1	0.05	0.05	0.07	40
Example 2	0.07	0.06	0.08	40
Example 3	0.01	0.01	0.04	40
Example 4	0.02	0.01	0.26	40
Example 1	0.05	0.05	0.06	160
Example 2	0.07	0.06	0.07	160
Example 3	0.01	0.01	0.02	160
Example 4	0.02	0.01	0.11	160

- Table shows the RMSE of the estimated coefficient $\hat{\beta}$
- λ is chosen by five-folds cross validation
- Lasso has better performance compared with the thesis paper

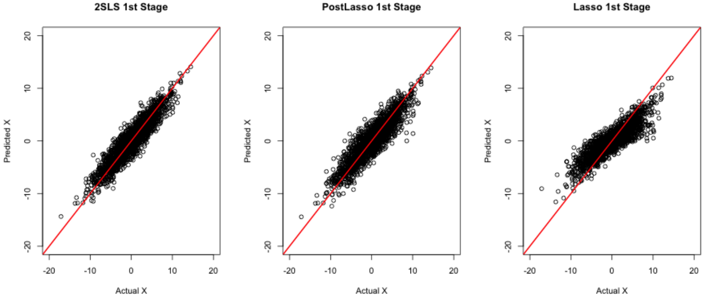
Results: $Cor(\epsilon, v) = 0.3, p < n$, Number of Selected Vars

Explore 1st Stage (Example 1, 100 iterations)

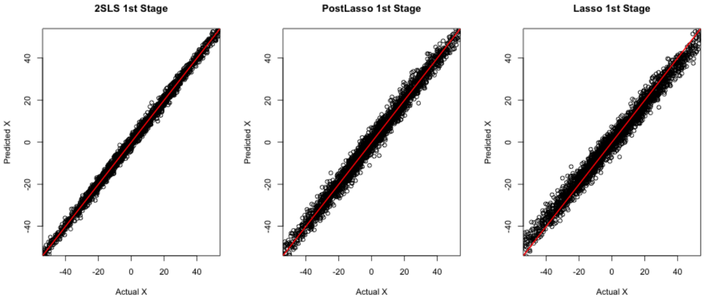


- Lasso underestimate positive X, overestimate negative X
- Overestimate β in the second stage ($\hat{\beta} > 1$)
- Similar discovery for other examples

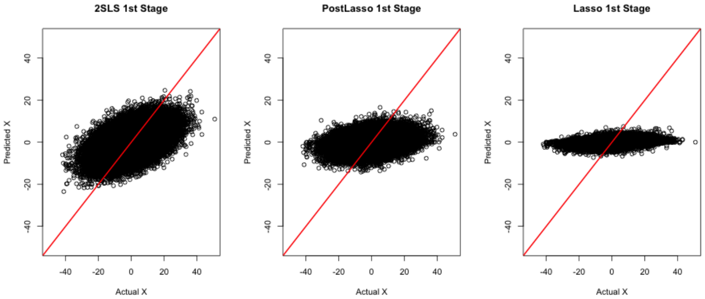
Explore 1st Stage (Example 2, 100 iterations)



Explore 1st Stage (Example 3, 50 iterations)



Explore 1st Stage (Example 4, 100 iterations)



Explore 1st Stage (Summary)

- Lasso is useful for selecting variables, but the predicted value in the first stage is always biased
- Explain the overestimation of β in the second stage
- The performance of high dimension scenarios? ($n < p$)

	2SLS	PostLasso	Lasso	F
Example 1	0.06	0.05	0.12	10
Example 2	0.07	0.06	0.26	10
Example 3	0.02	0.02	0.10	10
Example 4	0.04	0.02	0.76	10
Example 1	0.05	0.05	0.07	40
Example 2	0.07	0.06	0.09	40
Example 3	0.01	0.01	0.04	40
Example 4	0.04	0.02	0.27	40
Example 1	0.05	0.05	0.06	160
Example 2	0.07	0.06	0.07	160
Example 3	0.01	0.01	0.02	160
Example 4	0.03	0.02	0.11	160

My Simulation Results: $Cor(\epsilon, v) = 0.3, p > n$

	PostLasso	Lasso	F
Example 5	0.01	0.09	10
Example 6	0.01	0.08	10
Example 7	0.01	0.31	10
Example 8	0.01	0.67	10
Example 5	0.01	0.07	40
Example 6	0.01	0.04	40
Example 7	0.01	0.19	40
Example 8	0.01	0.67	40
Example 5	0.01	0.06	160
Example 6	0.01	0.03	160
Example 7	0.01	0.17	160
Example 8	0.01	0.66	160

My Simulation Results: $Cor(\epsilon, v) = 0.6, p > n$

	PostLasso	Lasso	F
Example 5	0.01	0.09	10
Example 6	0.01	0.09	10
Example 7	0.02	0.31	10
Example 8	0.01	0.67	10
Example 5	0.01	0.07	40
Example 6	0.01	0.04	40
Example 7	0.01	0.19	40
Example 8	0.01	0.67	40
Example 5	0.01	0.06	160
Example 6	0.01	0.03	160
Example 7	0.01	0.15	160
Example 8	0.01	0.65	160