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

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

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

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

- $\mathbf{z}_i = [z_{i1}, ..., z_{iq}]^T \sim N(0, \Sigma_z)$, $\text{Corr}(z_{ih}, z_{ij}) = 0.5^{|i-h|}$
- 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 $\text{Corr}(\epsilon, \omega) = 0.3$ and $\text{Corr}(\epsilon, \omega) = 0.6$

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 $Z_1, \dots, Z_5$ are different?

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

	OLS	2SLS	PostLasso	Lasso	F
Example 1	0.053	0.051	0.051	0.298	10
Example 2	0.065	0.061	0.061	0.700	10
Example 3	0.006	0.006	0.006	0.079	10
Example 4	0.026	0.020	0.014	1.922	10
Example 1	0.050	0.051	0.051	0.121	40
Example 2	0.066	0.065	0.065	0.203	40
Example 3	0.006	0.006	0.006	0.037	40
Example 4	0.041	0.022	0.013	0.716	40
Example 1	0.050	0.050	0.050	0.070	160
Example 2	0.065	0.065	0.065	0.087	160
Example 3	0.006	0.006	0.006	0.025	160
Example 4	0.045	0.018	0.013	0.266	160

- λ is chosen such that the five-folds cross validation error is within one standard error from the minimum
- PostLasso has the best performance

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

	F=10	F=40	F=160	Actual s
Example 1	4.05	4.24	4.21	3
Example 2	4.72	7.13	7.95	8
Example 3	15.57	16.95	15.71	15
Example 4	1.69	4.83	5.14	5

- For strong instruments, Lasso selects more variables (close to the actual number of instruments)

Results: $Cor(\epsilon, v) = 0.6$, $n < p$, RMSE of $\hat{\beta}$

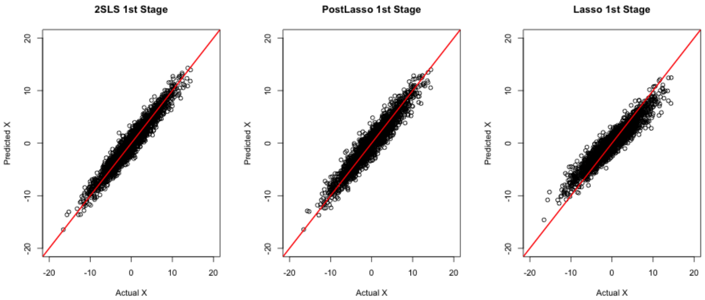
	OLS	2SLS	PostLasso	Lasso	F
Example 1	0.065	0.052	0.052	0.304	10
Example 2	0.090	0.067	0.067	0.728	10
Example 3	0.007	0.006	0.006	0.080	10
Example 4	0.052	0.038	0.015	1.936	10
Example 1	0.053	0.051	0.050	0.122	40
Example 2	0.077	0.067	0.067	0.213	40
Example 3	0.006	0.006	0.006	0.037	40
Example 4	0.082	0.039	0.014	0.717	40
Example 1	0.050	0.050	0.050	0.069	160
Example 2	0.068	0.065	0.065	0.090	160
Example 3	0.006	0.006	0.006	0.026	160
Example 4	0.088	0.028	0.013	0.267	160

Results: $Cor(\epsilon, v) = 0.6$, $n > p$, Number of Selected Vars

	F=10	F=40	F=160	Actual s
Example 1	4.05	4.22	4.24	3
Example 2	4.72	7.14	7.95	8
Example 3	15.59	16.92	15.70	15
Example 4	1.69	4.83	5.15	5

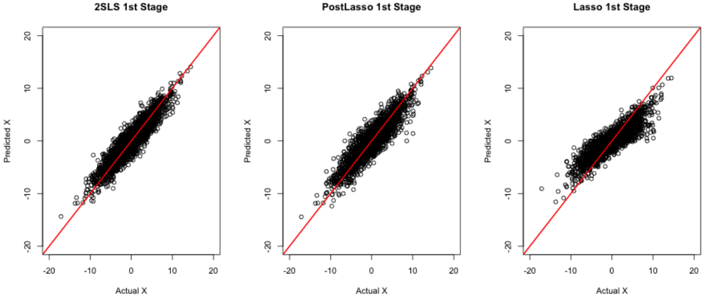
- For strong instruments, Lasso selects more variables (close to the actual number of instruments)

Explore 1st Stage (Example 1, 100 iterations)

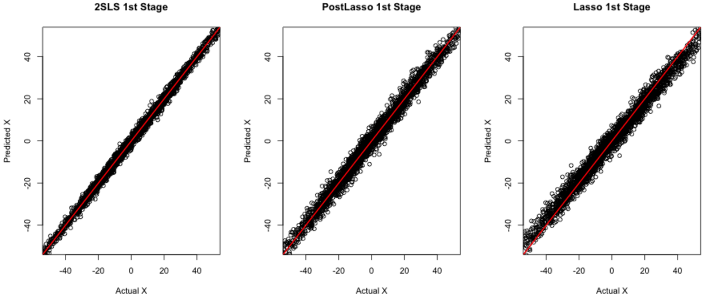


- Lasso underestimate positive X, overestimate negative X
- Overestimate β in the second stage ($\hat{\beta} > 1$)
- Similar discovery for other examples

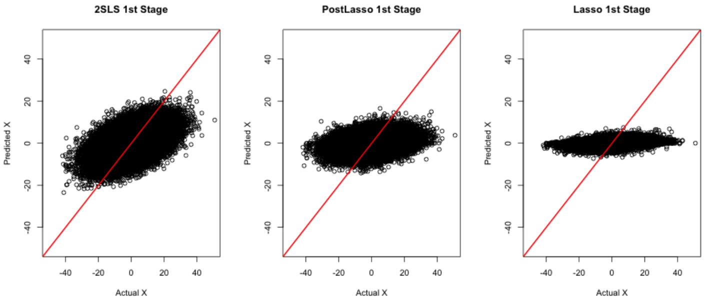
Explore 1st Stage (Example 2, 100 iterations)



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 1st Stage is biased (RMSE of the 1st Stage prediction)

	2SLS	PostLasso	Lasso	F
Example 1	1.20	1.36	1.70	10
Example 2	1.61	1.83	2.25	10
Example 3	1.52	2.71	3.46	10
Example 4	9.42	10.66	10.89	10
Example 1	0.60	0.67	0.81	40
Example 2	0.81	0.84	1.02	40
Example 3	0.78	1.37	1.74	40
Example 4	4.72	5.25	5.44	40
Example 1	0.30	0.34	0.41	160
Example 2	0.40	0.40	0.46	160
Example 3	0.39	0.73	1.04	160
Example 4	2.36	2.62	2.72	160

- Explain the overestimation of β in the second stage
- The performance of high dimension scenarios? $(n < p)$

Simulation Setup $n < p$ (Thesis paper)

- Example 5:
  $\Pi = (\text{rep}(2, 25), \text{rep}(0, 20), \text{rep}(2, 25), \text{rep}(0, 10))$,
  $n = 40, p = 80, s = 50$

- Example 6:
  $\Pi = (\text{rep}(2, 15), \text{rep}(0, 30), \text{rep}(2, 5), \text{rep}(0, 30))$,
  $n = 40, p = 80, s = 20$

- Example 7:
  $\Pi_i = 0.85, o = 1, ..., 80, n = 40, p = 80$

- Example 8:
  $\Pi = (\text{rep}(1, 100), \text{rep}(0, 900))$,$n = 50, p = 1000, s = 100$

- 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).

Results: $Cor(\epsilon, v) = 0.6$, $n < p$, RMSE of $\hat{\beta}$

	OLS	PostLasso	Lasso	F
Example 5	0.008	0.007	0.279	10
Example 6	0.015	0.013	0.272	10
Example 7	0.017	0.016	0.699	10
Example 8	0.011	0.014	0.781	10
Example 5	0.007	0.007	0.253	40
Example 6	0.013	0.012	0.150	40
Example 7	0.014	0.014	0.595	40
Example 8	0.009	0.014	0.789	40
Example 5	0.007	0.007	0.260	160
Example 6	0.013	0.012	0.121	160
Example 7	0.013	0.013	0.517	160
Example 8	0.008	0.013	0.789	160

Results: $Cor(\epsilon, v) = 0.6$, $n < p$, Number of Selected Vars

	F=10	F=40	F=160	Actual s
Example 5	24.32	24.74	25.26	50
Example 6	20.38	22.75	23.22	20
Example 7	21.55	22.88	23.25	80
Example 8	23.56	22.99	22.76	100

Simulation Setup (JASA paper)

- Setup:

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

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}_0 + \mathbf{E}$$

- $\boldsymbol{\beta}_0 = 1$; $\boldsymbol{\Gamma}_0$ is a $1 \times p$ vector
- $(\boldsymbol{\epsilon}_i, \eta_i) \sim N(\mathbf{0}, \Sigma)$
- For Σ, set $\sigma_{i,j} = (0.2)^{|i-j|}$, for $i, j = 1, 2$
- Assume that each variable is centered

Simulation Setup (JASA paper, $p < n$)

The nonzero entries in the columns of $\mathbf{\Gamma}_0$ are sampled from the uniform distribution $U([-b, -a] \cup [a, b])$

$\mathbf{Z} \sim Bernoulli(p_0)$, where $p_0 = 0.5$ for Model 1-3, $p_0 \sim U([0, 0.5])$ for Model 4

When $p < n$,

- Model 1: $(n, p, s) = (200, 100, 5)$, $(a, b) = (0.75, 1)$
- Model 2: $(n, p, s) = (400, 200, 5)$, $(a, b) = (0.75, 1)$
- Model 3: $(n, p, s) = (400, 200, 5)$, $(a, b) = (0.5, 0.75)$
- Model 4 (realistic): $(n, p, s) = (400, 200, 50)$
  Five of $(a, b) = (0.5, 1)$ and forty-five of $(a, b) = (0.05, 0.1)$

Simulation Results (JASA paper, $p < n$)

	OLS	2SLS	PostLasso	Lasso
Example 1	0.112	0.088	0.072	1.466
Example 2	0.102	0.074	0.052	1.139
Example 3	0.139	0.112	0.078	1.868
Example 4	0.114	0.088	0.062	1.345

Table: RMSE for $\hat{\beta}$

	Lasso	Actual s
Example 1	17.44	5
Example 2	33.09	5
Example 3	18.28	5
Example 4	27.20	50

Table: Number of Selected Variables

Simulation Setup (JASA paper, $p > n$)

When $p > n$,

- Model 5: $(n, p, s) = (300, 600, 5)$, $(a, b) = (0.75, 1)$
- Model 6: $(n, p, s) = (500, 1000, 5)$, $(a, b) = (0.75, 1)$
- Model 7: $(n, p, s) = (500, 1000, 5)$, $(a, b) = (0.5, 0.75)$
- Model 8 (realistic): $(n, p, s) = (500, 1000, 50)$
  Five of $(a, b) = (0.5, 1)$ and forty-five of $(a, b) = (0.05, 0.1)$

Simulation Results (JASA paper, $p > n$)

	OLS	PostLasso	Lasso
Example 5	0.104	0.061	1.603
Example 6	0.105	0.052	1.296
Example 7	0.144	0.091	2.017
Example 8	0.123	0.074	1.756

Table: RMSE for $\hat{\beta}$

	Lasso	Actual s
Example 1	26.31	5
Example 2	41.86	5
Example 3	21.15	5
Example 4	29.16	50

Table: Number of Selected Variables