Hybrid Generalized Empirical Likelihood Estimators: Instrument Selection with Adaptive Lasso

Mehmet Caner
North Carolina State University *

Qingliang Fan
Xiamen University

September 17, 2012

Abstract

In this paper we use adaptive lasso estimator to select between relevant and irrelevant instruments in heteroskedastic and non Gaussian data. To do so limit theory of Zou (2006) is extended from univariate iid case. Then we use the selected instruments in generalized empirical likelihood estimators (GEL). In this sense, these are called hybrid GEL. It is also shown in the paper that Lasso estimators are not model selection consistent whereas adaptive lasso can select the correct model in fixed number of instruments case. It is also shown that adaptive lasso estimator can achieve near minimax risk bound even in the case of heteroskedastic Gaussian data. This is a new result and extends the standard normal iid data in Zou (2006). In simulations we show that hybrid GEL estimators have very good bias and mean squared error compared with other estimators.

JEL Codes:C52, C26, C13.

Keywords: Model Selection, Near Minimax Risk Bound, Shrinkage Estimators.

*Corresponding Authors: Mehmet Caner: Department of Economics, 4168 Nelson Hall, Raleigh, NC 27518. email: mcaner@ncsu.edu. Qingliang Fan: Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, China, 361005. Email: qfan@ncsu.edu.

1 Introduction

One of the important issues in economics is selection of the instruments. We think that this is very important since a lot of empirical cases involving labor, institutional economics deal with very limited number of instruments. For example Acemoglu, Johnson and Robinson (2001), Acemoglu and Johnson (2006), Card (1995) papers use 1 or at most 2 instruments with one endogenous variable at hand. It is critical then in those cases to see whether we have strong instruments or not. If we have one instrument in just identified case, and it is weak, the second stage coefficient is inconsistent, see Staiger and Stock (1997). If we have more than one instrument and one endogenous variable then the second stage regression will give consistent estimate, but in finite samples we will have bias. Our simulations in the paper show this.

In many instruments setup, there have been several papers analyzing instrument selection recently in the literature. Donald and Newey (2001) target Mean Squared Error of second stage regression coefficients. Theirs do not take into account the weakness of the instruments. Recently, Kuersteiner and Okui (2010) use model averaging to pick up instruments. Their approach is similar to Donald and Newey (2001), and improve on Mean Squared Error of second stage coefficients. In a landmark paper Belloni, Chen, Chernozhukov, and Hansen (2012) introduce a new heteroskedasticity consistent Lasso type estimator to pick optimal instruments among many of them. This is a very important leap over the statistics literature as well as econometrics literature. They are able to establish performance bounds for Lasso estimator for the first time in heteroskedastic, non Gaussian data. This is very difficult to achieve, since all the results in statistics is for Gaussian and iid data. Belloni, Chen, Chernozhukov, and Hansen (2012)(Belloni et al. (2012) from now on) use moderate deviation theory for self normalized sums to solve this problem. In the meantime, this results in a completely new Lasso based estimator which depend on the residuals as the penalty.

Our paper is interested in selecting instruments in a fixed number of instruments unlike the many instruments case. Toward that end, we benefit from the recent statistics literature. In statistics related to model selection, and estimation, Lasso, Bridge are analyzed by Knight and Fu (2000). Caner (2009) analyze GMM based Bridge estimation. Caner (2009) does not consider instrument selection and only considers model selection in orthogonality restrictions. Recently, in statistics, smooth penalty functions are introduced by Fan and Li (2001, 2002). However, one of the most recent shrinkage based estimator in statistics is adaptive lasso of Zou (2006). This has optimality properties, and easy to estimate compared to Bridge, and it is also shown there that Lasso estimator is not model selection consistent asymptotically. Lasso, in fixed number of regressor case in least squares, cannot select the correct model with probability one, see Proposition 1 of Zou (2006). Zou (2006) also shows that adaptive lasso is model selection consistent, and achieves the near minimax risk bound in iid Gaussian data. Adaptive lasso is also oracle efficient, estimates the nonzero parameters with standard efficient limit in least squares and estimates the zero parameters

as zero with probability one.

Our paper applies adaptive lasso to instrument selection problem in the reduced form estimates, and then run generalized empirical likelihood estimators in the second stage regression. The main idea here is to benefit from the very good model selection properties of adaptive lasso in the first stage. We show that adaptive lasso in the first stage is model selection consistent. In other words, it can pick the relevant instruments with probability approaching one. Given that we have a better bias and MSE for the structural parameters in the second stage. We also extend the near minimax risk bound in Zou (2006) to non iid Gaussian data. In addition to that, we show a variant of Lasso is also subject to model selection problem, whereas adaptive lasso is model selection consistent. Adaptive lasso can differentiate between the irrelevant and strong instruments. Simulations in the paper show that using F test as an indicator may be problematic. In the other simulations, adaptive lasso performs very well in terms of bias/MSE of second stage coefficients compared with another Lasso based estimator, Donald and Newey (2001) procedure, model averaging estimator of Kuersteiner and Okui (2010), LIML, Fuller, and heteroskedasticity consistent version of Fuller estimator of Hausman et al. (2012).

Recently we come across with two working papers which apply shrinkage based methods to IV regression. The first one, independently written, is by Garcia (2011). He devises adaptive lasso for the case of many weak instruments. Note that the asymptotics (of fixed number) of selection of instruments are entirely different from the many weak instruments case. The next paper is by Shi (2011). There the issue is the structural equation parameter selection in increasing number of parameters case with shrinkage. This is also important contribution since it is important to handle high dimensional problems in econometrics. Shrinkage methods will be immensely useful in these cases to applied researchers. One very important contribution to this literature is by Leeb and Pőtscher (2005). They show that if the parameters are varying by the sample size, then the shrinkage methods cannot be uniformly consistent, and hence cannot select the true model with probability approaching one. For this reason, it is impossible to select the correct instruments with adaptive lasso in the weak instruments context when their number is fixed. We analyze this situation also in simulations. Our theories are based on fixed parameter asymptotics, and hence is not subject to the criticism of Leeb and Pőtscher (2005). Leeb and Pőtscher (2005) idea applies to least squares framework. However, we are interested in second stage regression estimates. We show that finite sample distribution of second stage coefficients is not bi-modal, and normally distributed. This is shown in the simulation section in a simple overidentified case.

Section 2 provides the limit theory for adaptive lasso when relevant and irrelevant instruments exist. Adaptive lasso picks the relevant ones and eliminates the irrelevant ones, and uses this information in second stage GEL. Section 3 provides the result that even a new version of Lasso is model selection inconsistent whereas adaptive lasso can select the model correctly in the case of fixed number of instruments. Section 4 provides the algorithm that is used. Section 5 introduces

an oracle inequality. Section 6 carries out extensive simulations. Appendix provides all the proofs. $\|.\|, \|.\|_\infty$ represent the Euclidean norm, and maximal value of a vector respectively.

2 The Reduced Form Estimation Via Adaptive Lasso

In this section we show that we can use adaptive lasso in multivariate setting to select and estimate the reduced form coefficients simultaneously. In this sense, this section extends the univariate adaptive lasso of Zou (2006). So we will be able to eliminate the irrelevant instruments and keep the relevant ones. In matrix form the reduced form equations are:

$$X = Z\gamma^0 + \nu,$$

where $X : n \times p$, $Z : n \times q$, and $\gamma^0 : q \times p$ matrix, we also have $q \geq p$, q represents all the fitted instruments. We can rewrite that

$$X = Z_1\gamma_1^{0'} + Z_2\gamma_2^{0'} + \cdots + Z_q\gamma_q^{0'} + \nu, \tag{1}$$

where $\gamma_j^0 : p \times 1$ vector $j = 1, 2 \cdots q$. Each Z_j is $n \times 1$, $j = 1, \cdots q$.

Next we can write these in vector form as:

$$vecX = (I_p \otimes Z_1)vec(\gamma_1^{0'}) + (I_p \otimes Z_2)vec(\gamma_2^{0'}) + \cdots + (I_p \otimes Z_q)vec(\gamma_q^{0'}) + vec(\nu).$$

We can write this in new notation as:

$$X_v = \tilde{Z}_1\gamma_1^0 + \cdots + \tilde{Z}_q\gamma_q^0 + \nu_v, \tag{2}$$

where $X_v = vecX$, $\tilde{Z}_j = (I_p \otimes Z_j)$, $\nu_v = vec(\nu)$. So $\tilde{Z}_j$ is $np \times p$ matrix for $j = 1, 2 \cdots q$, and X_v is $np \times 1$ vector.

Note that γ_j^0 are the true population coefficient vectors, $j = 1, 2, \cdots q$, and the true number of nonzero vectors are q_0, with $q_0 \geq p$. We can further rewrite (2)

$$X_v = \tilde{Z}\gamma_v^0 + \nu_v, \tag{3}$$

where $\tilde{Z} = [\tilde{Z}_1, \cdots \tilde{Z}_q]$ ($np \times pq$ matrix), and

$$\gamma_v^0 = \begin{bmatrix} \gamma_1^0 \\ \vdots \\ \gamma_q^0 \end{bmatrix},$$

where $\gamma_v^0 : pq \times 1$ vector. The objective function is:

$$\hat{\gamma}_v = argmin_{\gamma_v}[X_v - \tilde{Z}\gamma_v]'[X_v - \tilde{Z}\gamma_v] + \lambda_n \sum_{j=1}^{q}\sum_{k=1}^{p} \hat{w}_{jk}|\gamma_{jk}|. \tag{4}$$

See that the weights are $\hat{w}_{jk} = |\tilde{\gamma}_{jk}|^{-\tau}$, where $\tilde{\gamma}_{jk}$ is the $\sqrt{n}$ consistent LS estimator, and $0 < \tau \leq 1$. To understand these better note that γ_v is pq vector, so these are stacked $p \times 1$ vectors, γ_j, $j = 1, \cdots, q$.

In these q vectors, the nonzero ones will be denoted by

$$\mathcal{A} = \{j : \gamma_j^0 \neq 0_p\},$$

where 0_p represents a $p \times 1$ vector of zeros. Without losing any generality, we can designate the last $\gamma_{q_0+1}^0, \gamma_{q_0+2}^0, \cdots \gamma_q^0$ as zero vectors (all $p \times 1$ cells are zero). This can be written as

$$\mathcal{A}^c = \{j : \gamma_j^0 = 0_p\},$$

$j = q_0 + 1, q_0 + 2, \cdots q$. So we can represent $\mathcal{A} = \{1, 2, \cdots, q_0\}$. Now we assume that for all relevant instruments ($j = 1, 2, \cdots, q_0$) all p cells in vectors $\gamma_j^0$are nonzero. This is just done for the simplicity. We will also talk about the possibility of zero cells in γ_j^0, $j = 1, 2 \cdots q_0$ after Theorem 1.

Assumptions.

1.

$$\frac{\nu_v'\nu_v}{n} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{p}\nu_{ik}^2}{n} \xrightarrow{p} \sigma_\nu^2 > 0,$$

where $\sigma_v^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p} E\nu_{ik}^2$. Also this variance is finite.

2. We have the following Law of Large Numbers result:

$$\frac{\tilde{Z}'\nu_v}{n} \xrightarrow{p} 0.$$

3.

$$\frac{\tilde{Z}'\tilde{Z}}{n} \xrightarrow{p} C < \infty.$$

Also matrix C ($pq \times pq$ matrix) is of full rank. We can write C as

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix},$$

where C_{11} is positive definite full rank, and symmetric submatrix of dimensions $q_0 p \times q_0 p$.

4. For the penalty term $\lambda_n/\sqrt{n} \to 0$, $\frac{\lambda_n}{n^{1/2}} n^{\tau/2} \to \infty$ and $0 < \tau \leq 1$.

5. We assume the following Central Limit Theorem

$$\frac{\tilde{Z}'\nu_v}{n^{1/2}} \xrightarrow{d} N(0,\Omega) \equiv W.$$

Ω is $pq \times pq$ matrix, and it has Ω_{11} as the $q_0p \times q_0p$ upper left block, as positive definite, full rank matrix.

Note that Assumptions 1, 2 , 3 , 5 are high level assumptions. These can be proved via suitable moment conditions on the errors and the instruments as in Davidson (1994). Note that we can use independent data for Theorem 1 below. Assumption 4 is used in Zou (2006). This shows the behavior of the penalty. The main difference with Zou (2006) is $0 < \tau \leq 1$. This is needed for the consistency proof. This is not shown in Zou (2006).

Set $\mathcal{A}_n = \{j : \hat{\gamma}_j \neq 0_p\}$. Note that $\hat{\gamma}_{vA}$ represents the adaptive lasso estimator for the first q_0p elements in $\hat{\gamma}_v$. Let γ^0_{vA} ($q_0p \times 1$ vector) represent the corresponding true nonzero elements. The following theorem generalizes adaptive lasso of Zou (2006) from univariate, iid case to multivariate, heteroskedastic case. Oracle property is also preserved as in Zou (2006).

Theorem 1.*Under Assumptions 1-5,*

(i).

$$n^{1/2}(\hat{\gamma}_{vA} - \gamma^0_{vA}) \xrightarrow{d} N(0, C_{11}^{-1}\Omega_{11}C_{11}^{-1}).$$

(ii).

$$\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1.$$

Note that $\mathcal{A}$ definition can be changed to $\mathcal{A}' = \{j, k : \gamma_{j,k} \neq 0, j = 1, \cdots q, k = 1, \cdots p\}$. Our definition of $\mathcal{A}$ is just for simplification. One issue is how the results may be affected when we have a combination of zero and nonzero cells in first q_0 γ^0_j vectors. From the proof of Theorem 1, we see that $\hat{\gamma}_v$ is $pq \times 1$ vector, but each cell in that vector is penalized separately. So we will find zero and nonzero cells in $pq \times 1$ cells. So as an example, if $p = 2, q = 3, q_0 = 2$, we may estimate $\gamma_1 = (1, 0)', \gamma_2 = (2, 3)'$ and $\gamma_3 = (0, 0)$. So our method will find that $\gamma_{12} = 0$,(first coefficient vector, 2nd cell) but still clearly that γ^0_1 is relevant since $\gamma_{11} = 1$, and the instrument corresponding to that will be put in the second stage regression. But since $\gamma_{3,1} = \gamma_{3,2} = 0$, the third instrument will not be used in second stage regression. An irrelevant instrument means that all p cells of γ_j are zero.

2.1 Second Stage Regression

Since we find the relevant instruments in the first stage via adaptive lasso, we can use them in the second stage regression . In other words,

$$y_i = x_i'\beta + \epsilon_i,$$

for $i = 1, \cdots n$, and $\beta : p \times 1$ vector. Now assume that the instruments are uncorrelated with ϵ_i. Then the first estimator to consider is two step GMM

$$\hat{\beta}_{GMM} = (X'Z_r\hat{W}Z_r'X)^{-1}(X'Z_r\hat{W}Z_r'Y),$$

where $Y = (y_1, \cdots, y_n)'$ X is $n \times p$ matrix and $Z_r = [Z_1, \cdots, Z_{q_0}] : n \times q_0$ matrix. $\hat{W}$ is the standard efficient weight used in GMM. Application of the standard asymptotic theory to $\hat{\beta}_{GMM}$ will provide the efficient GMM limit.

Define the function $\rho(\iota)$ where ι is a scalar, and the function is concave on its domain. Next, we consider generalized empirical likelihood (GEL from now on) estimators for the second stage. These are defined in Newey and Smith (2004).

$$\hat{\beta}_{GEL} = argmin_{\beta\in\mathcal{B}} \sup_{\delta\in\Delta_n(\beta)} \sum_{i=1}^{n} \rho(\delta' g_i(\beta)),$$

where $\mathcal{B}$ is a compact subset of R^p, and $\Delta_n(\beta) = \{\delta : \delta' g_i(\beta) \in \mathcal{V}, i = 1, \cdots, n\}$, $\mathcal{V}$ is an open interval containing zero. Since this is a linear model $g_i(\beta) = Z_i\epsilon_i = Z_i(y_i - x_i'\beta)$, and Z_i is $q_0 \times 1$ vector. When $\rho(\iota) = ln(1-\iota)$ this is empirical likelihood estimator, when $\rho(\iota) = -exp(\iota)$, this is called exponential tilting, and when $\rho(\iota) = -(1+\iota)^2/2$, this is called continuous updating estimator. For standardizations, and further detail, see Newey and Smith (2004). Using the correct number of relevant instruments q_0 from the first stage will result in standard GEL limits as shown in Newey and Smith (2004).

An important point is why we use immediately Z_r matrix of dimensions $n \times q_0$ (i.e. the correct number of instruments). The reality is we have $\hat{q}$ number of instruments from the first stage. But Theorem 1(ii) clearly shows that $\hat{q} - q_0 \xrightarrow{p} 0$. Selection consistency provides the reason to use the correct number of instruments in the second stage. We could have shown this in extensive proofs, but this is not difficult to show, and could have unnecessarily lengthened the proofs. But to show the main argument, take the linear GMM and analyze $\|\hat{Z}'\epsilon/n^{1/2}\|$ where $\hat{Z} = (Z_1, \cdots, Z_{|\hat{q}-q_0|})$. This is one of the limit terms. Note that $\hat{Z}$ is a $n \times |\hat{q} - q_0|$ dimensional matrix, where $\hat{q}$ is the estimated number of relevant instruments in the first stage (reduced form). See also that $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$.

$$\|\hat{Z}'\epsilon/n^{1/2}\| \leq \sqrt{|\hat{q} - q_0|}\|Z'\epsilon/n^{1/2}\|_\infty = o_p(1),$$

given stochastically bounded $\|Z'\epsilon/n^{1/2}\|_\infty$. The other terms can be treated similarly, so both in linear GMM, linear GEL, selection consistency provides us with starting the second stage with correct number of relevant instruments. This does not make any difference in the limit. Note that in the simulations we use estimated number of instruments from the first stage and then put them in two step GMM and GEL in the second stage.

We could have done model selection in the second stage given the number of instruments, but we believe that they have to be analyzed jointly. There may be identification problems with this. But we believe this is an open question to be analyzed. We also develop theory for the case of using the predictors from the first stage to be used in second stage. But this did not get us good finite sample results in simulations, so we do not report them to save space.

3 Asymptotic Bias and Selection Inconsistency of Lasso

This section will analyze certain Lasso estimators that are used in the literature. The first one is regular lasso, and it has asymptotic bias and is model selection inconsistent as shown in Theorem 2 of Knight and Fu (2000). So we refer the reader to Knight and Fu (2000) for details. The second one is heteroskedasicity consistent Lasso type estimator of Belloni et al. (2012). Their estimator is a big leap in the literature. This estimator, in large number of instruments case, can choose optimal instruments, and has the oracle property for the instrumental variable estimation. This estimator also works well with heteroskedastic and non-Gaussian cases. Here we show that with fixed number of instruments, there is an asymptotic bias in estimating the relevant instruments with their method (a variant of lasso as well as post-lasso which is least squares after running least squares), and this affects selection consistency in return. Note that the setup of Belloni et al. (2012) involves many instruments, and hence we are not analyzing that case. We ask ourselves the question that what if their method could have applied to fixed number of instruments, can we have selection consistency? Since it has also data dependent weights as adaptive lasso here, this will be a good estimator to compare.

But we want to make one point crystal clear, their paper is path breaking in this literature. Our paper is not attempting to take away from large contributions that they made. So we setup the model in Belloni et al. (2012) with fixed number of instruments.

$$y_i = d_i'\alpha_0 + \epsilon_i,$$

where d_i is $k_d \times 1$ endogenous variable vector, and

$$E[\epsilon_i|z_i] = 0,$$

for each $i = 1, \cdots, n$. z_i is $p \times 1$ vector of instruments, and p is fixed, unlike Belloni et al. (2012).

Next the reduced form equation is

$$d_i = z_i'\gamma_0 + v_i,$$

where v_i, ϵ_i are correlated, and γ_0 is $p \times 1$ vector, $E(v_i|z_i) = 0$, $p \geq k_d$. With no loss of generality, we abstract away from including control variables in both reduced form and structural equations. Also we set $k_d = 1$ to simplify the notation. With a vector of endogenous variables, $k_d > 1$, we could have followed the methodology in section 2, and sum the penalty terms over all k_d times p parameters (reduced form matrix elements). Then our Assumption B.1 will be true with the added $\max_{1 \leq l \leq k_d}$ condition. To simplify the proofs/assumptions we set $k_d = 1$.

In Belloni et al. (2012), quite sensibly the number of relevant instruments are approximately "s" (Condition AS in Belloni et al. (2012)). This is a very good idea in their case, since with increasing number of instruments (in their case $p \to \infty$), it makes sense to describe the relevant instruments as approximate number. However, in our case we setup the fixed number of instruments as possible instrument candidates, and then the true number of instruments is fixed and equal to s. In a simple applied work it is sometimes difficult to find valid instruments (see Acemoglu et al. (2001, 2006)) so this is a reasonable assumption in our case.

Next we define the heteroskedasticity consistent Lasso:

$$\hat{\gamma}_L = argmin_\gamma [\sum_{i=1}^{n} (d_i - z_i'\gamma)^2 + \lambda \sum_{j=1}^{p} |\gamma_j \hat{\pi}_j|], \tag{5}$$

where they have two step process in estimating γ.

First, they set, for each $j = 1, \cdots p$,

$$\hat{\pi}_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} z_{ij}^2 (d_i - \bar{d})^2},$$

where $\bar{d} = n^{-1} \sum_{i=1}^{n} d_i$. By using this in the Lasso formula above they get Initial Lasso. Then after getting "Initial Lasso", they setup the following refined loadings

$$\hat{\pi}_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} z_{ij}^2 \hat{v}_i^2}, \quad j = 1, \cdots p$$

where $\hat{v}_i = d_i - z_i'\hat{\gamma}_{Initial-Lasso}$. After using the refined loadings in (5) objective function we get $\hat{\gamma}_L$. This is equation (2.4) in Belloni et al. (2012). The steps to get lasso estimator is well described in Algorithm 1 in Appendix of Belloni et al. (2012). They only select the instruments in the reduced form, and there is no model selection in the second stage. We will follow this sequence in this part of the paper as well. We will not analyze post lasso estimator in their paper since this is just a

regular-unpenalized least squares estimator after running lasso estimator with refined loadings.

Before the assumptions we introduce some of the notation that is used in Belloni et al (2012). Let $\|f_i\|_{2,n} = \sqrt{\frac{1}{n}\sum_{i=1}^n f_i^2}$, for a generic random variable f_i. Then let $E_n(f_i) = \frac{1}{n}\sum_{i=1}^n f_i$, and $\bar{E}(f_i) = \lim_{n\to\infty} E[\frac{1}{n}\sum_{i=1}^n f_i]$, and $\tilde{d}_i = d_i - \bar{E}d_i, 1 \leq i \leq n$. We make the following assumptions for the following Lemmata:

Assumption B.1

(i).

$$\max_{1\leq j\leq p}[\bar{E}(\tilde{d}_i)^2 + \bar{E}(z_{ij}^2\tilde{d}_i) + \frac{1}{\bar{E}(z_{ij}^2 v_i^2)}] = O_p(1).$$

(ii).

$$\max_{1\leq j\leq p} \bar{E}(z_{ij}^3 v_i^3) = O_p(K_n).$$

(iii).

$$K_n^2 log^3 n = o(n).$$

(iv).

$$\max_{1\leq j\leq p} z_{ij}^2 \frac{logn}{n} = o_p(1).$$

and

$$\max_{1\leq j\leq p} |\bar{E}_n(z_{ij}^2 v_i^2) - \bar{E}(z_{ij}^2 v_i^2)| + |\bar{E}_n(z_{ij}^2 \tilde{d}_i^2) - \bar{E}(z_{ij}^2 \tilde{d}_i^2)| = o_p(1).$$

(v).

$$\|z_i'(\hat{\gamma}_{Lasso} - \gamma_0)\|_{2,n} = O_p(\frac{(logn)^{1/2}}{n^{1/2}}).$$

Assumption B.2

(i). $\hat{\gamma}_{Lasso}$ is consistent.

(ii). $\lambda/n^{1/2} \to \lambda_0 \geq 0$.

Note that Assumption B1(i)-(iv) are Assumption RF(i)-(iv) in Belloni et al. (2012). Assumption B1(v) is Theorem 1 of Belloni et al. (2012), we use this to shorten the proofs. Before the limit of Lasso in Lemma 1, set $\hat{u} = n^{1/2}(\hat{\gamma}_L - \gamma)$. Define $\pi_j^0 = \sqrt{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n Ez_{ij}^2 v_i^2}$ for all $i = 1, 2\cdots n$.

Lemma 1. *Under Assumptions B.1-B.2 we have the following limit*

$$n^{1/2}(\hat{\gamma}_L - \gamma_0) \xrightarrow{d} argmin_u V(u),$$

where

$$V(u) = -2u'W + u'\Sigma u + \lambda_0 \sum_{j=1}^p [\pi_j^0 u_j sgn(\gamma_{j0}\pi_j^0) 1_{\{\gamma_{j0}\neq 0\}} + |u_j \pi_j^0| 1_{\{\gamma_{j0}=0\}}],$$

and $W \equiv N(0, \Sigma_{Zv})$, *where* $\Sigma_{Zv} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E z_i z_i' v_i^2$ *with* $n^{-1} \sum_{i=1}^{n} z_i z_i' \xrightarrow{p} \Sigma = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E z_i z_i'$.

This clearly shows the asymptotic bias of Lasso with respect to nonzero coefficients of Belloni et al. (2012) in the special case of fixed number of instruments. This is similar to the limit of Lasso in Knight and Fu (2000). Before going over our second result in this section, we introduce the following notation from Zou (2006). See that

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where Σ_{11} is a square matrix, corresponds to limit of second moments of relevant instruments, it is also invertible and positive definite, Σ_{22} corresponds to limit of second moments of irrelevant instruments, and Σ_{12} is the limit of sample cross product of the relevant with irrelevant instruments, $\Sigma_{21} = \Sigma_{12}'$. Note that $\mathcal{A}_n = \{j : \hat{\gamma}_j \neq 0\}$, and $\mathcal{A} = \{j : \gamma_{j0} \neq 0\}$.

Next we show that the asymptotic bias results in selection inconsistency of the Lasso type estimator considered in this section. Namely, if there is a weak or irrelevant instrument, then this Lasso may not eliminate that and put the weak instrument in the second stage. In a simple model with one endogenous regressor and one instrument, this results in inconsistent estimation of the second stage coefficients through Staiger and Stock (1997) asymptotics.

Lemma 2. *Under Assumptions B.1-B.2,*

$$\limsup_{n} P(\mathcal{A}_n = \mathcal{A}) \leq c < 1,$$

where c is a constant depending on the true model.

Our result extends Proposition 1 of Zou (2006) from regular Lasso to the Lasso that is used by Belloni et al. (2012). We show that usage of this Lasso estimators in fixed instruments context may lead to inconsistent instrument selection, which may affect the second stage regressions as discussed above. However, we should note that the Lasso used in Belloni et al. (2012) is for large number of instruments with an approximate number of relevant instruments, we do not analyze their estimators behavior in a large instruments context. The Lasso that they introduce is a huge leap in the literature. They show clearly that with heteroskedastic and non Gaussian data, Lasso satisfies an oracle inequality. Theirs is very innovative work both on theoretical terms as well as its practical use. Ours is an attempt to analyze what may happen to Lasso type estimators in a fixed number of instruments framework. We show that adaptive lasso is immune to the instrument inconsistency problem. Bűhlmann and van de Geer (2010) also show/discuss better selection consistency properties of adaptive lasso over lasso in sections 2.8.3, and 7.8.3 of their

book.

4 Algorithm and The Choice of the Tuning Parameter

Adaptive lasso estimates can be computed efficiently by a modification of LARS (Least Angle Regression, and S suggesting 'LASSO' and 'Stagewise') algorithm (Efron et al. (2004)). The computational efficiency is an advantage of adaptive lasso in practice compared to other Oracle methods such as SCAD (Fan and Li, 2001) and bridge estimator. In this section, we briefly discuss the implementation of LARS in adaptive lasso.

In Zou (2006), a simple modified version of LARS can be adopted for the adaptive lasso estimation. It works as follows.

To illustrate the method, we set up a basic linear model, for $i = 1, \ldots, n$

$$x_i = Z_i'\gamma + \nu_i. \tag{6}$$

where x_i is the univariate endogenous variable, $Z_i = (Z_{i1}, \ldots, Z_{id})'$ is the associated d-dimensional instruments, $\gamma = (\gamma_1, \ldots, \gamma_d)'$ is the coefficient vector and ν_i is the random error with mean 0 and variance σ_ν^2. For multivariate model we vectorize X, Z, ν etc., the algorithm works as in the univariate case. Assume that we fit d predictors and the true model has d_0 variables ($1 \leq d_0 \leq d$).

Adaptive Lasso Algorithm

1. Create new covariates $Z_j^* = Z_j/\hat{w}_j$, $j = 1, 2, \ldots, d$, where $\hat{w}_j$ is the adaptive weight as defined in Section 2. Note that each Z_j, Z_j^* is $n \times 1$ vector.

2. Solve the LASSO via LARS algorithm for given λ.

$$\tilde{\gamma} = arg\min_{\gamma} \left\| X - \sum_{j=1}^{d} Z_j^{*'}\gamma_j \right\|^2 + \lambda \sum_{j=1}^{d} |\gamma_j| \tag{7}$$

   where $X = (X_1, \ldots, X_n)'$.

3. Output is $\hat{\gamma}_j = \tilde{\gamma}_j/\hat{w}_j$, $j = 1, 2, \ldots, d$. This is the adaptive lasso estimate.

4. Then we put $\hat{\gamma}_j$, $j = 1, 2, \ldots, d$, adaptive lasso estimate from Step 3 in equation (8) below. This provides us a BIC value for a given λ. Note that tuning is explained in details after the Algorithm.

5. Repeat Steps 2-4 for each remaining λ in a set of Λ (e.g., $\Lambda = \{\lambda_1, \ldots, \lambda_{100}\}$) and record each BIC_λ for given $\hat{\gamma}(\lambda)$.

6. Choose the pair of $\hat{\gamma}(\lambda)$, which minimizes BIC over λ.

In Step 2 we use LARS algorithm to compute LASSO. Then in Step 3, we convert this to adaptive lasso solution. Now we explain the simple intuition behind LARS. LARS procedure works as follows (see Efron et al. (2004) for more details). Assume for simplicity that we have standardized our explanatory variables to have zero mean and unit variance, and that our response variable also has zero mean. We start with all coefficients being equal to zero (no variables in the model), and find the predictor most correlated with endogenous variable x, say it's the first covariate z_1 (as we can always switch the 'position' of the covariates). The reasoning here is the largest correlation possibly shows significance of the variable, so we take out that variable and include in our model. Since the covariate z most correlated with the residual is equivalently the one that makes the least angle with the residual, the name of the method is called 'least angle regression'. We take the largest step possible in the direction of this predictor until some other predictor, say z_2, has as much correlation with the current residual. Different from classic Forward Selection which takes a 'full step' with z_1, LARS now proceeds in a direction equiangular between the two predictors z_1 and z_2 (so that the residual makes equal angles with both covariates) until a third variable z_3 becomes equally correlated with the current residual. LARS then proceeds equiangularly between z_1, z_2 and z_3, that is, along the 'least angle direction', until a fourth variable enters, and so on. The result of $\hat{\gamma}$ in Step 2 depends on λ choice. We explain this in detail next in choice of λ. As shown in Theorem 1 of Efron et al. (2004), a slight modification to LARS can get us the full solution path of LASSO.

In this part we explain the method to select tuning parameter λ which we use in the adaptive lasso simulations. Recall that the tuning parameter λ controls the penalty level and therefore the model complexity. We follow the tuning parameter selector by Wang and Leng (2007). In their paper, it has been shown that with the BIC method, tuning parameter can achieve the oracle properties of adaptive lasso. BIC method is selection consistent for adaptive lasso under fixed predictor dimension and a slight modification of the BIC method is also consistent under diverging number of parameters (Wang et al. (2009)). In Wang and Leng (2007), they show that the tuning parameter by BIC method can select the correct model with probability approaching one.

The BIC method for λ selector is to minimize the following

$$BIC_\lambda = \hat{\sigma}_\lambda^2 + DF_\lambda \log(n)/n \tag{8}$$

where $\hat{\sigma}_\lambda^2 = n^{-1} \|X - Z\hat{\gamma}\|^2$, DF_λ is the number of nonzero coefficients in $\hat{\gamma}$ which is described in Step 3 of adaptive lasso Algorithm. Z is $n \times d$ matrix where the vector form is described in (6). The reason we use this method is that BIC method can get us the model selection consistency when the sample size n is approaching ∞. But other methods such as AIC or GCV (generalized cross validation) can not get us there (see Wang et al. (2009)).

5 Oracle Inequality

In this section we extend an important result due to Zou (2006). Zou (2006) provides the proof that adaptive lasso achieves near minimax risk in iid standard normal data. Here we show that this can be extended to non iid, but Gaussian heteroskedastic data. The proof is generally different than Zou (2006). A more general regression setup can be achieved with heteroskedastic non Gaussian data. This may be possible by using moderate deviation theorems as in lasso of Belloni et al. (2012). This is a different scope than the current paper and will extend the paper enormously in volume. Oracle inequalities on general non Gaussian settings are difficult to establish, and the main aim of our paper is to show that adaptive lasso first step helps in selecting instruments, and improves on the second stage GEL finite sample properties. The model that we use is from Zou (2006).

$$x_i = \mu_i + \nu_i,$$

where ν_i is the error term, and $E\nu_i^2 = \sigma_i^2 > 0$, for each $i = 1, 2 \cdots n$. The aim is to estimate μ_i with adaptive lasso estimator $\hat{\mu}_i$. The risk is defined as in Zou (2006)

$$R(\hat{\mu}) = E[\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2].$$

The ideal risk is defined in equation (16) of Donoho and Johnstone (1994) as

$$R(ideal) = \sum_{i=1}^{n} min(\mu_i^2, \sigma_i^2),$$

where $\min(a, b)$ represents the minimum of the scalars a, b. Adaptive lasso estimate in this case is derived in equation (5) of Zou (2006) as

$$\hat{\mu}_i = argmin_u[\frac{1}{2}(x_i - u)^2 + \frac{\lambda_i}{|x_i|^\tau}|u|], \tag{9}$$

for $i = 1, 2 \cdots, n$, and $\lambda_i = (\sqrt{2\sigma_i^2 logn})^{1+\tau}$. This type of λ_i is used in p.113 of Averkamp and Houdre (2003). Note that the λ_i is compatible with the results in section 2. We abstract away from estimation of σ_i^2 for the main purpose of showing the oracle inequality. As stated in Zou (2006), since we have one observation for each μ_i, the weight is defined as $|x_i|^{-\tau}$, where $0 < \tau \leq 1$. So oracle inequality here in Theorem 3 extends Zou (2006) from iid Gaussian case to heteroskedastic-Gaussian data.

So the minimization of (9) above provides us the following as in Zou (2006)

$$\hat{\mu}_i = \left[|x_i| - \frac{\lambda_i}{|x_i|^\tau}\right]_+ sgn(x_i), \tag{10}$$

where $[.]_+$ denotes the positive part of the expression inside, otherwise this is set as zero, i.e. $[k]_+ = k, if k > 0$, otherwise $[k]_+ = 0$. The following is one of the main results of the paper. We take σ_i to be the positive root of variance σ_i^2. The proof has to be modified slightly otherwise.

Theorem 2.*(Oracle Inequality). Let* $\lambda_i = (2\sigma_i^2 logn)^{(1+\tau)/2}$, *then*

$$R(\hat{\mu}) \leq (2logn + b + \frac{b}{\tau})[R(ideal) + \frac{c^{1/2}}{d^{1/2}} \frac{1}{2\pi^{1/2}} (logn)^{-1/2}],$$

where $d > \max_i \sigma_i$, $0 < d < \infty$, $c > d/min_i \sigma_i^2$, $b = max(2c, 4d + 1)$. *Note that* b, c, d, σ_i *are all positive constants, and do not depend on* n.

This shows that even though we have heteroskedastic-Gaussian data, adaptive lasso still attains the near minimax risk as shown in Zou (2006) in iid-Gaussian case. Theorem 2 here gives us a basic oracle inequality, extending Zou (2006). In Theorem 3 of Zou (2006), Gaussian-iid case, he finds

$$R(\hat{\mu}) \leq (2logn + 5 + \frac{4}{\tau})[R(ideal) + \frac{1}{\sqrt{2\pi}} (logn)^{-1/2}].$$

Compared to Zou (2006), because of non iid Gaussian nature here, our constants b, c, d depend on σ_i^2. In Zou (2006), he takes $\sigma_i^2 = 1$. Theorem 2 is a new result, and the proof technique is not the same as in Zou (2006).

6 Simulation

In this section we want to answer three questions. The first one is whether our test can do better in selecting the irrelevant instruments compared with F test? The second question is to compare the adaptive lasso with a full model (i.e. no model selection) for the second stage regression? The third question is whether adaptive lasso selection of instruments will deliver better second stage finite sample results compared with several competitors. The fourth question concerns the bi-modality of shrinkage estimators in least squares context that is raised by Leeb and Pőtscher (2005). We want to see whether this bi-modality in the first stage can affect the second stage structural coefficients. To answer the first question we use a basic setup. To answer the rest of the questions, and especially the fourth one, we use a model which includes Leeb and Pőtscher(2005) setup.

6.1 Performance of Adaptive Lasso in First Stage Selection

First, we look at the performance of adaptive lasso in the reduced form equation. We report the percentage of correct model selection, when the nonzero (strong) instruments are estimated as nonzero and zero (or local to zero) instruments are estimated as zero. The LARS algorithm that is used to get adaptive lasso estimates as well as tuning parameters are described in section 4.

TABLE 1: Comparison of Adaptive Lasso with F test

	Simulation results for Model 1		Simulation Results for Model 2, $t = 3$	
	MSP	F-MSP	MSP	F-MSP
$n = 60, \sigma = 6$	.65	.04	.41	.03
$n = 60, \sigma = 3$	.98	.74	.81	.79
$n = 120, \sigma = 6$	.87	.24	.77	.19
$n = 120, \sigma = 3$	.98	1.00	.87	1.00
$n = 300, \sigma = 6$	.97	.94	.95	.91
$n = 300, \sigma = 3$	.99	1.00	.92	1.00

MSP is the rate of correct model selection, 1 being perfect model selection, 0 being lowest. This is for adaptive lasso. F-MSP is the rate of F-statistics greater than 10.

We use the following design for the reduced form equation.

$$X = Z\gamma + \nu \tag{11}$$

where X is $n \times 1$ endogenous variables, Z is $n \times 2$ matrix of instruments. $Z_i \sim N(0, I_2)$ i.i.d. and $E(\nu_i|Z_i) = 0$ for $i = 1, 2, \ldots, n$. $\nu_i \sim N(0, \sigma^2)$ i.i.d. Let $\gamma = (\gamma_1, \gamma_2)$ is 2×1 true parameter vector. The IV model has two settings of parameter values:

Model 1 : one strong and one irrelevant instrument. $\gamma = (2, 0)'$

Model 2: one strong and one weak (local to zero) instrument. $\gamma = (2, \frac{t}{\sqrt{n}})'$, where t is a constant real number

In both models we simulated 100 datasets for each combination of (σ, n). We use three sample sizes, n =60, 120 and 300 and σ takes on values 6, 3 in corresponding model setup. We set $t = 3$. This is a small scale simulation, since LARS is computationally intensive. Still, this provides us with the information that F test is not working well to separate the weak instrument from the strong one.

In Table 1, we use F test (joint test on both instruments) on the reduced form equation and report the percentage of when F-statistic is greater than 10. It is common in applied studies that diagnose instruments to be weak if F-statistic is less than 10 (Staiger and Stock, 1997). F-statistic is also approximately increasing with concentration parameter which is a unitless measure of instrument strength (Stock et al. (2002)). Note that we are not using 11.39 as the critical value as advocated by Stock et al. (2002)). This could have made the results much worse for them. We will compare the model selection performance of adaptive lasso and first stage F-test 'rule of thumb' (use all instruments whenever F-statistic is greater than 10).

In our simulations, F-test 'rule of thumb' tends to miss the mark of model selection since it does not reject the H_0 that both coefficients are zero when $\sigma = 6, n = 60, 120$. Also, it is known that rejection of the null hypothesis by no means implies there is no weak instrument (Staiger and

Stock, 1997). On the other hand, adaptive lasso not only shows there are weak instruments in the model, it specifically tells which ones are (by shrinking them to 0). This simulation setup here is just used to illustrate the simple problems that may arise with ad hoc F-test with a large noise and a mixed quality of the instruments. A setup that favors F-test may provide good results for F-test.

6.2 Comparison of Hybrid Estimators with Other Estimators

We present here several 'hybrid' estimators. We call it hybrid since in the first stage we use adaptive lasso to select instruments. In the second stage we use generalized empirical likelihood (GEL) estimator, specifically, the continuous updating (CUE), exponential tilting (ET) and empirical likelihood (EL), as well as TSLS (or GMM in heteroskedasticity case). We therefore name these hybrid estimators, respectively, H-CUE, H-ET, H-EL and H-TSLS (H-GMM in heteroskedasticity case). In simulations we also include the Donald and Newey (2001) estimator, Kuersteiner and Okui (2010) model averaging TSLS estimator, Belloni et al. (2012)'s Post-Lasso estimator and the traditional limited information maximum likelihood (LIML), Fuller's estimator and the heteroskedasticity robust Fuller (Hausman et al., (2012)) . We compare the results of these structural equation parameter estimators in terms of finite sample properties. We also adopt the model setup in Leeb and Pőtscher (2005) in the reduced form equation.

We now briefly explain other estimators which we compare our hybrid estimators. First, we show Donald and Newey (2001) estimator which chooses the number of instruments to minimize the leading term of Nagar (1959) type MSE. The 2SLS estimator is

$$\hat{\beta} = (X'P^K X)^{-1} X'P^K Y \tag{12}$$

where $X = (x_1, \ldots, x_n)'$, $Y = (y_1, \ldots, y_n)'$, $P^K = Z^K(Z^{K\prime}Z^K)^{-1}Z^K$, and K is the index for the number of instruments which are included in the regression. Now we define the necessary variables to minimize MSE with respect to K as described in Donald and Newey (2001). Let $\tilde{\beta}$ be some preliminary estimator of β, e.g., it can be the regular 2SLS estimator. Let $\tilde{\epsilon} = Y - X\tilde{\beta}$, $\tilde{H} = X'P^K X/n$, and $\tilde{u} = (I - P^K)X$. Let $\tilde{u}_\lambda = \tilde{u}\tilde{H}^{-1}\tilde{\lambda}$, where $\tilde{\lambda} = 1$. We have the following variables: $\hat{\sigma}^2_\epsilon = \tilde{\epsilon}'\tilde{\epsilon}/n$, $\hat{\sigma}^2_\lambda = \tilde{u}'_\lambda\tilde{u}_\lambda/n$, $\hat{\sigma}_{\lambda\epsilon} = \tilde{u}'_\lambda\tilde{\epsilon}/n$. These preliminary estimators do not depend on K, they remain as constants as the approximate MSE are calculated. We can use cross validation or Mallow's in the calculation. Taking Mallow's criterion as an example, first, let $\hat{u}^K = (I - P^K)X$, $\hat{u}^K_\lambda = \hat{u}^K\tilde{H}\tilde{\lambda}$. So the Mallow's criteria is $\hat{R}^m_\lambda(K) = \dfrac{\hat{u}^{K\prime}_\lambda \hat{u}^K_\lambda}{n} + \hat{\sigma}^2_\lambda(2K/n)$. Finally, the approximate MSE of the 2SLS estimator is $\hat{S}_\lambda(K) = \hat{\sigma}^2_{\lambda\epsilon}\dfrac{K^2}{n} + \hat{\sigma}^2_\epsilon\left(\hat{R}^m_\lambda(K) - \sigma^2_\lambda\dfrac{K}{n}\right)$.

Second, the model averaging estimator by Kuersteiner and Okui (2010) is considered. Set a weighting vector W, where $W = w_1, \ldots, w_M$, and $\sum_{m=1}^M w_m = 1$ for some M which is the number of all possible instruments. Let $Z_{m,i}$ be the vector of the first m elements of $Z_{M,i}$ which is an

$M \times 1$ vector of instruments, let Z_m be the matrix $(Z_{m,1}, \ldots Z_{m,N})$ and let $P_m = Z_m(Z_m'Z_m)^{-1}Z_m'$. Define $P(W) = \sum_{i=1}^{M} w_m P_m$. The model averaging two stage least squares estimator (MA2SLS) is defined as $\hat{\beta} = (X'P(W)X)^{-1}X'P(W)y$.

Third, the Post Lasso estimator by Belloni et al. (2012) which estimates the optimal instruments set. The Post Lasso is essentially OLS with Lasso selected variables. The Lasso estimator of the reduced form β is

$$\begin{aligned} \hat{\beta} &= arg \min_{\beta \in R^p} \hat{Q}(\beta) + \frac{\lambda}{n} \|\hat{\Upsilon}\beta\|_1 \\ &= arg \min_{\beta \in R^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} (x_i - z_i'\beta)^2 + \frac{\lambda}{n} \sum_{j=1}^{p} |\hat{\Upsilon}_j \beta_j| \right\} \end{aligned}$$

where $\hat{Q}(\beta)$ is the sum of squared residuals (least squares) from running x_i (endogenous variable) on z_i (the instruments) and $\|\hat{\Upsilon}\beta\|_1$ is the sum of absolute values, $\hat{\Upsilon}$ is penalty loadings defined as follows:

Initial (or basic option) penalty loadings. Each (j,j) element of the $p \times p$ diagonal $\hat{\Upsilon}$ matrix is

$$\hat{\Upsilon}_j = \sqrt{\sum_{i=1}^{n} \frac{z_{ij}^2 (x_i - \bar{x})^2}{n}} \quad j = 1, 2, \ldots, p, \tag{13}$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. Refined penalty loadings are formulated in this way. Each element of the $p \times p$ diagonal $\hat{\Upsilon}$ matrix is

$$\hat{\Upsilon}_j = \sqrt{\sum_{i=1}^{n} \frac{z_{ij}^2 \hat{\nu}_i^2}{n}} \quad j = 1, 2, \ldots, p \tag{14}$$

where $\hat{\nu}_i = x_i - z_i'\hat{\beta}$, where $\hat{\beta}$ can be either initial LASSO or Post-Lasso (LASSO using the initial penalty loadings) after a finite number of iterations.

Fourth, the heteroskedasticity robust Fuller's estimator (Hausman et al. 2012) is given as follows. Let $P = Z(Z'Z)^{-1}Z'$, P_{ij} denote the ij^{th} element of P, and $\bar{X} = [y, X]$. Let $\tilde{\alpha}$ be the smallest eigenvalues of $(\bar{X}'\bar{X})^{-1}(\bar{X}'P\bar{X} - \sum_{i=1}^{n} P_{ii}\bar{X}_i\bar{X}_i')$. For a constant C let $\hat{\alpha} = [\tilde{\alpha} - (1 - \alpha)C/T]/[1 - (1 - \tilde{\alpha})C/T]$. The heteroskedasticity robust Fuller's estimator (HFUL) is given by

$$\hat{\beta} = (X'PX - \sum_{i=1}^{n} P_{ii}X_iX_i' - \hat{\alpha}X'X)^{-1}(X'Py - \sum_{i=1}^{n} P_{ii}X_iy_i - \hat{\alpha}X'y) \tag{15}$$

The asymptotic variance estimator is shown in p. 215 of Hausman et al. (2012), which we use in the calculation of HFUL variance.

6.2.1 Simulation Results for Conditional Homoskedasticity

The linear IV regression model with a single endogenous regressor and no included exogenous variable is:

$$y_i = \beta_0 x_i + \epsilon_i \tag{16}$$

$$x_i = \gamma_1 z_{1i} + \gamma_2 z_{2i} + \nu_i \tag{17}$$

where $i = 1, 2, \cdots, n$. The true $\beta_0 = 1$. Assume the IV matrix $Z = [z_1, z_2]$ has full rank and satisfies

$$Z'Z/n \to \Sigma_z = \begin{bmatrix} \sigma^2_{\gamma_1} & \sigma_{\gamma_1\gamma_2} \\ \sigma_{\gamma_1\gamma_2} & \sigma^2_{\gamma_2} \end{bmatrix}$$

as $n \to \infty$. We further assume $\sigma^2_{\gamma_1} = \sigma^2_{\gamma_2} = 1$ and the correlation between z_1 and z_2 is $\rho_1 = \sigma_{\gamma_1\gamma_2}/(\sigma_{\gamma 1}\sigma_{\gamma_2}) = .7$. The errors $[\epsilon_i, \nu_i]'$ ($i = 1, 2, \cdots, n$) are assumed to be i.i.d. $N(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \sigma^2_\epsilon & \sigma_{\epsilon\nu} \\ \sigma_{\epsilon\nu} & \sigma^2_\nu \end{bmatrix}.$$

Let $\sigma_\epsilon = \sigma_\nu = 2$ and ρ_2 is correlation between the two error terms ϵ and ν. The closer ρ_2 is to 1, the stronger the endogeneity of x. We use two values for ρ_2 in simulations, .5 and .99. Each model is replicated 500 times. Now we use two setup of γ's:

Model 1 : one nonzero (strong) and one exact zero (irrelevant) coefficients $\gamma = (1, 0)'$

Model 2: one nonzero (strong) and one local to zero (weak) coefficients $\gamma = (1, \frac{t}{\sqrt{n}})'$, where t is a scalar, 2.5, 3.54 (for sample size n= 100, 200 respectively), so we have $\gamma_2/\sigma_{\gamma_2} = .25$ as in Leeb and Pőtscher (2005).

The reduced form equation model settings corresponds to the potential bi-modal density of LS estimator in Figure 2 of Leeb and Pőtscher (2005). We analyze the critique of Leeb and Pőtscher (2005) and we show in simulations that the our second stage coefficients are immune to bi-modality. In the case of irrelevant instruments, we do not expect bi-modality since all parameters are constants. See Proposition A.9 of Leeb and Pőtscher (2005).

In the following tables we report the median bias of the estimates (Bias), median absolute deviation (MAD), coverage rate of a nominal 95% confidence interval (95% Coverage Rate), mean squared error (MSE) and the percentage of z_1 being selected but not z_2 (Model Selection %). We also show in the following figures the finite sample densities of the hybrid estimators $\hat{\beta}$. We also did simulations for $n = 1000$, and the case that $\gamma_2/\sigma_{\gamma_2} = .21$ as in Leeb and Pőtscher (2005). The results of $\gamma_2/\sigma_{\gamma_2} = .21$ are very similar to the results of $\gamma_2/\sigma_{\gamma_2} = .25$. For $n = 1000$, the results are similar to $n = 200$. Therefore these figures and tables are not shown here for the sake of space.

First we consider the bias. From Table 2, we see that full model TSLS is worse compared to all hybrid methods, Post-Lasso of Belloni et al. (2012) and full model LIML. For Model 1, the median

bias of the full model is 0.032, while the bias of hybrid TSLS, CUE, ET, EL are 0.019, 0.019, 0.020 and 0.018 respectively. Therefore we do not recommend using the full model TSLS. LIML has the best bias term (which is 0.012) given our linear homoskedastic model. If we use TSLS with only the strong instrument (SO), the bias is 0.015. Post-Lasso has bias 0.024 which is worse than the hybrid estimators. Donald and Newey (2001) and Kuersteiner and Okui (2010) model averaging TSLS estimators (with bias 0.035 and 0.036 respectively) have higher bias than the full model. Fuller's estimator has bias 0.031 which is very close to the full model TSLS. Model 2 in Table 1 has also very similar results. When we analyze Tables 2-4, we see that LIML has the best bias followed by H-CUE estimator.

Second we consider MSE. Overall, hybrid CUE-ET estimators have the best MSE in Tables 2-5 followed by post Lasso and Fuller estimators. In Table 2, Model 1, the MSE for hybrid CUE-ET are 0.040, 0.040 respectively. Full model TSLS has 0.051. LIML's MSE is 0.057 which is worse than all except from the weak only (WO) TSLS case (MSE is 0.079). TSLS with strong instrument only (SO) has MSE of 0.055. Donald and Newey (2001), Kuersteiner and Okui (2010) model averaging TSLS estimators have MSE of 0.051 and 0.050 respectively. Post-Lasso has MSE of 0.043 whereas Fuller has MSE of 0.044. In Model 2, hybrid GEL estimators are the best (0.028, 0.028 and 0.029 for H-CUE, H-ET, H-EL respectively). Kuersteiner and Okui (2010) model averaging TSLS and Fuller's estimator are (closely) second best (0.030) and post Lasso has MSE of 0.031. Full model TSLS, hybrid TSLS and Donald and Newey (2001) estimator and Post-Lasso are closely behind (0.031). MSE of TSLS with strong IV only (SO) and LIML are both 0.034. In Table 3, we see the same trends as in Table 1. But in Tables 4-5, all models are pretty close to each other in MSE when $n = 200$.

When we look at the coverage rates for a 95% confidence interval, we see that all methods are slightly under the nominal rate generally. In terms of model selection rates, Hybrid estimators perform very well, Donald and Newey (2001) method does not do that well, and post Lasso is in between the two cases. To give an example in Tables 2-3, H-CUE has correct model selection 77-92% of time, whereas Donald and Newey (2001) method has 22-51% correct model rate, and post Lasso of Belloni et al. (2012) has a rate of 53-82%. From the following Figures 1-8, we see that the bi-modality of the reduced form equation such as the ones shown in Fig. 2 of Leeb and Pőtscher (2005) in the reduced form equations does not affect the empirical distribution of hybrid GEL estimators of the second stage in overidentified case. The figures for heteroskedastic case are not shown since the main idea is to show that structural parameter estimates in overidentified framework are not affected by bi-modality of reduced form coefficients.

To summarize, in the homoskedastic case, LIML has the best bias, but hybrid CUE-ET have the best MSE terms, also we see that hybrid estimators do a very good job on selecting the model.

Table 2: Second Stage Results: $n = 100$, $\rho_2 = .5$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		SO	WO	Full model	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 1	Bias	.015	.021	.032	.019	.019	.020	.018	.035	.036	.024	.012	.031
	MAD	.194	.259	.197	.194	.199	.200	.200	.194	.196	.194	.196	.192
	95% Coverage Rate	.944	.956	.936	.934	.924	.926	.936	.936	.932	.920	.940	.934
	MSE	.055	.079	.051	.045	.040	.040	.041	.051	.050	.043	.057	.044
	Model Selection %	-	-	-	.924	.924	.924	.924	.274	-	.812	-	-
Model 2	Bias	.013	.014	.026	.023	.019	.016	.017	.026	.025	.022	.013	.027
	MAD	.166	.194	.162	.162	.165	.164	.167	.162	.162	.164	.166	.165
	95% Coverage Rate	.948	.956	.940	.938	.918	.920	.940	.940	.934	.938	.942	.934
	MSE	.034	.044	.031	.031	.028	.028	.029	.031	.030	.031	.034	.030
	Model Selection %	-	-	-	.770	.770	.770	.770	.216	-	.544	-	-

SO is the model which we use the strong instrument only. WO is the model which we use the weak instrument only. Full model is the one that we use all available instruments. H-TSLS is the hybrid method that we use adaptive lasso selection in first stage and TSLS in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-TSLS except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post-lasso estimator proposed by Belloni et al. (2012). LIML and Fuller's estimator are the conventional methods without any instruments selection.

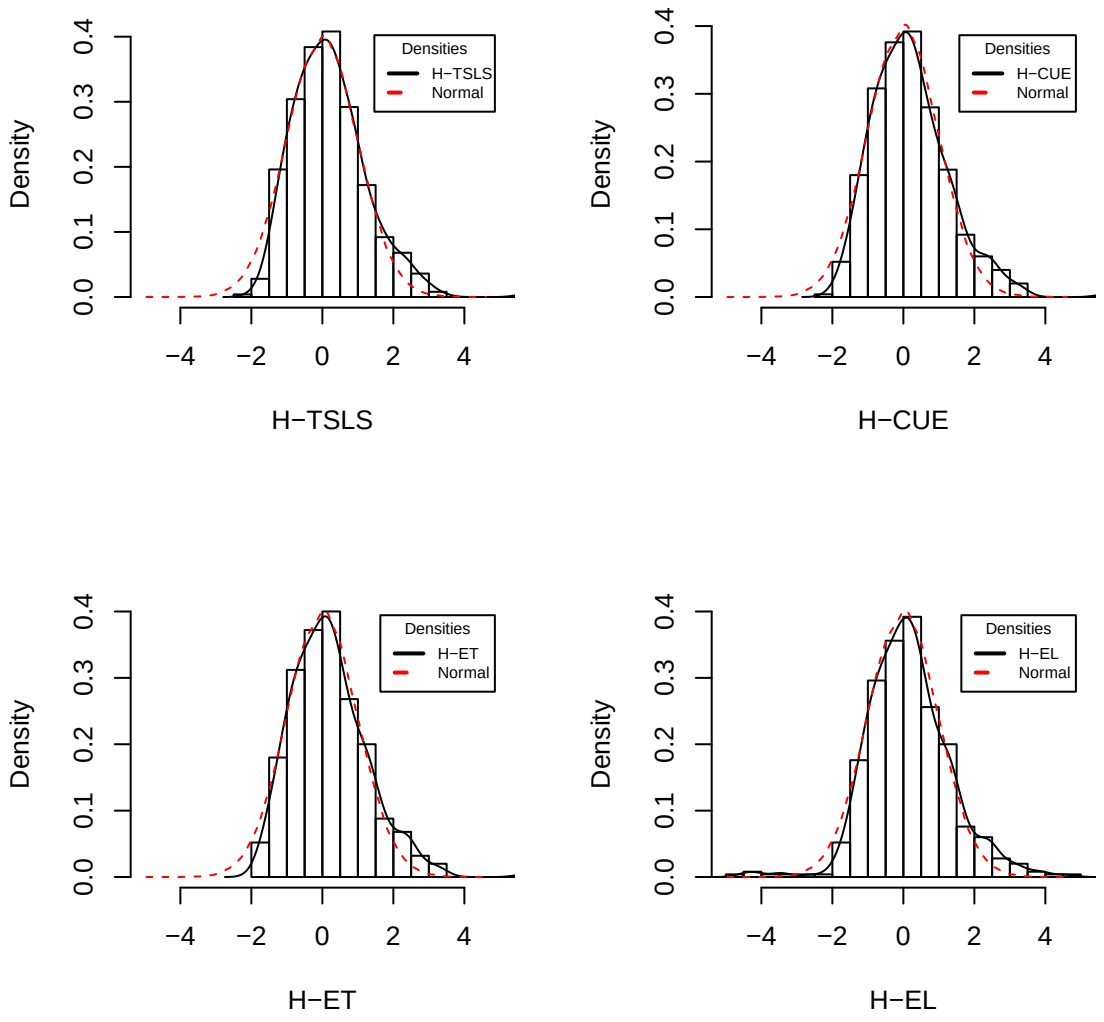

Figure 1: Finite sample densities of hybrid estimators: $n = 100$, $\gamma_2 = 0$, $\rho_2 = .5$, $\sigma_\epsilon = 2$

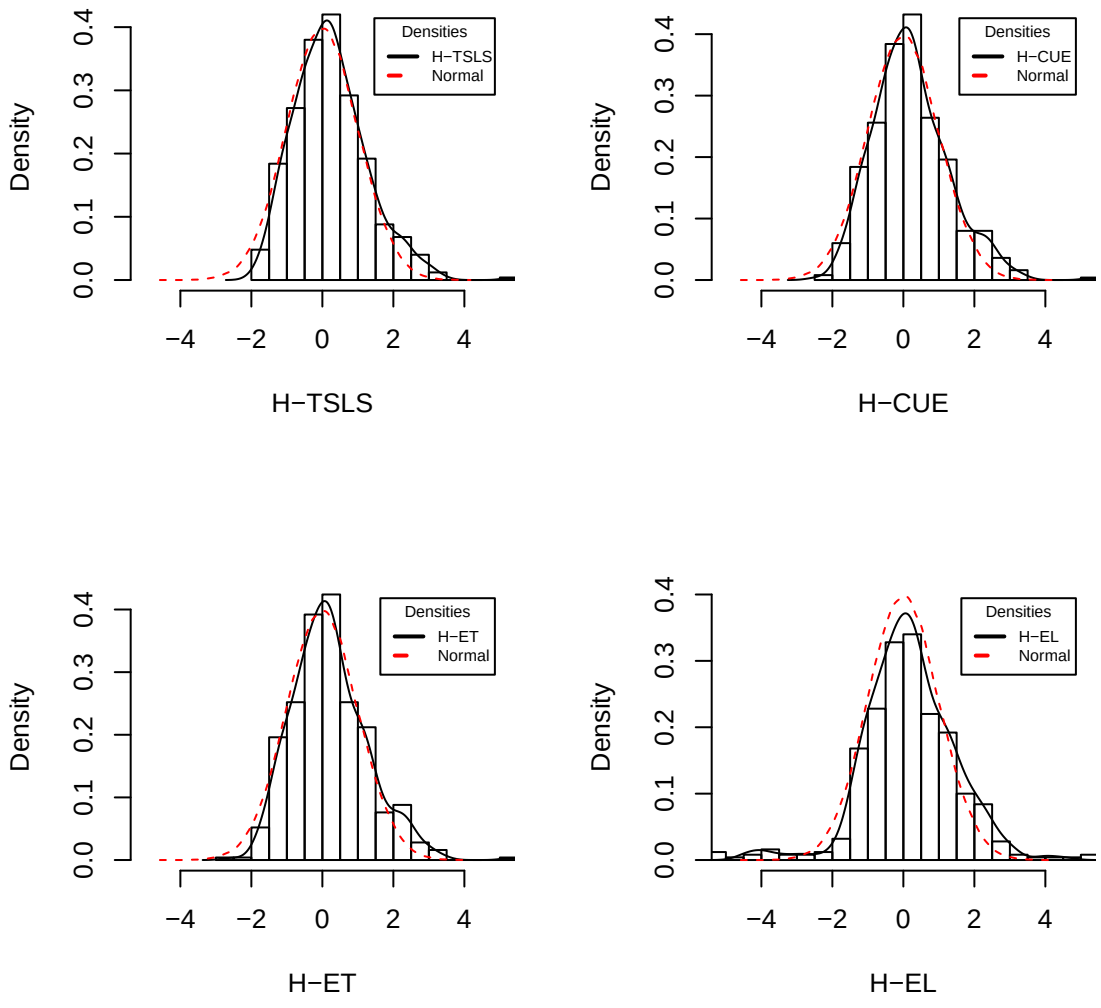

Figure 2: Finite sample densities of hybrid estimators: $n = 100$, $\gamma_2 = 2.5/\sqrt{n}$, $\rho_2 = .5$, $\sigma_\epsilon = 2$

Table 3:Second Stage Results: $n = 100$, $\rho_2 = .99$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		SO	WO	Full model	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 1	Bias	.013	.009	.049	.030	.011	.011	.013	.043	.055	.030	.013	.051
	MAD	.185	.277	.194	.191	.192	.192	.192	.193	.198	.191	.182	.185
	95% Coverage Rate	.922	.926	.898	.904	.910	.914	.898	.900	.884	.904	.928	.898
	MSE	.068	.074	.059	.047	.042	.043	.043	.059	.060	.047	.074	.045
	Model Selection %	-	-	-	.912	.912	.912	.912	.512	-	.816	-	-
Model 2	Bias	.011	.007	.035	.031	.017	.018	.017	.035	.037	.035	.009	.034
	MAD	.159	.205	.165	.162	.164	.161	.160	.163	.167	.163	.159	.158
	95% Coverage Rate	.930	.926	.906	.912	.910	.910	.906	.912	.898	.910	.930	.906
	MSE	.037	.042	.032	.032	.028	.029	.029	.032	.031	.032	.036	.031
	Model Selection %	-	-	-	.772	.772	.772	.772	.470	-	.528	-	-

SO is the model which we use the strong instrument only. WO is the model which we use the weak instrument only. Full model is the one that we use all available instruments. H-TSLS is the hybrid method that we use adaptive lasso selection in first stage and TSLS in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-TSLS except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post-lasso estimator proposed by Belloni et al. (2012). LIML and Fuller's estimator are the conventional methods without any instruments selection.

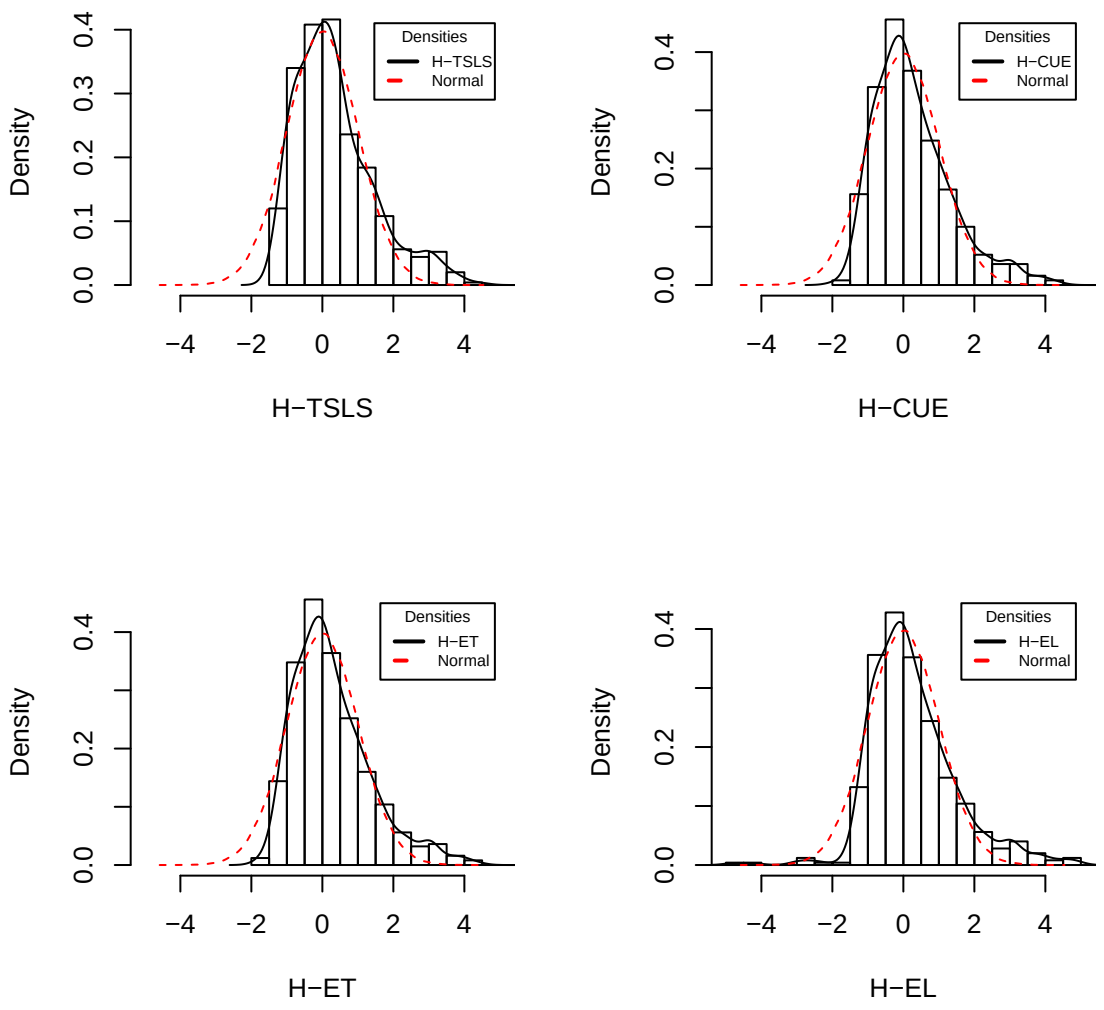

Figure 3: Finite sample densities of hybrid estimators: $n = 100$, $\gamma_2 = 0$, $\rho_2 = .99$, $\sigma_\epsilon = 2$

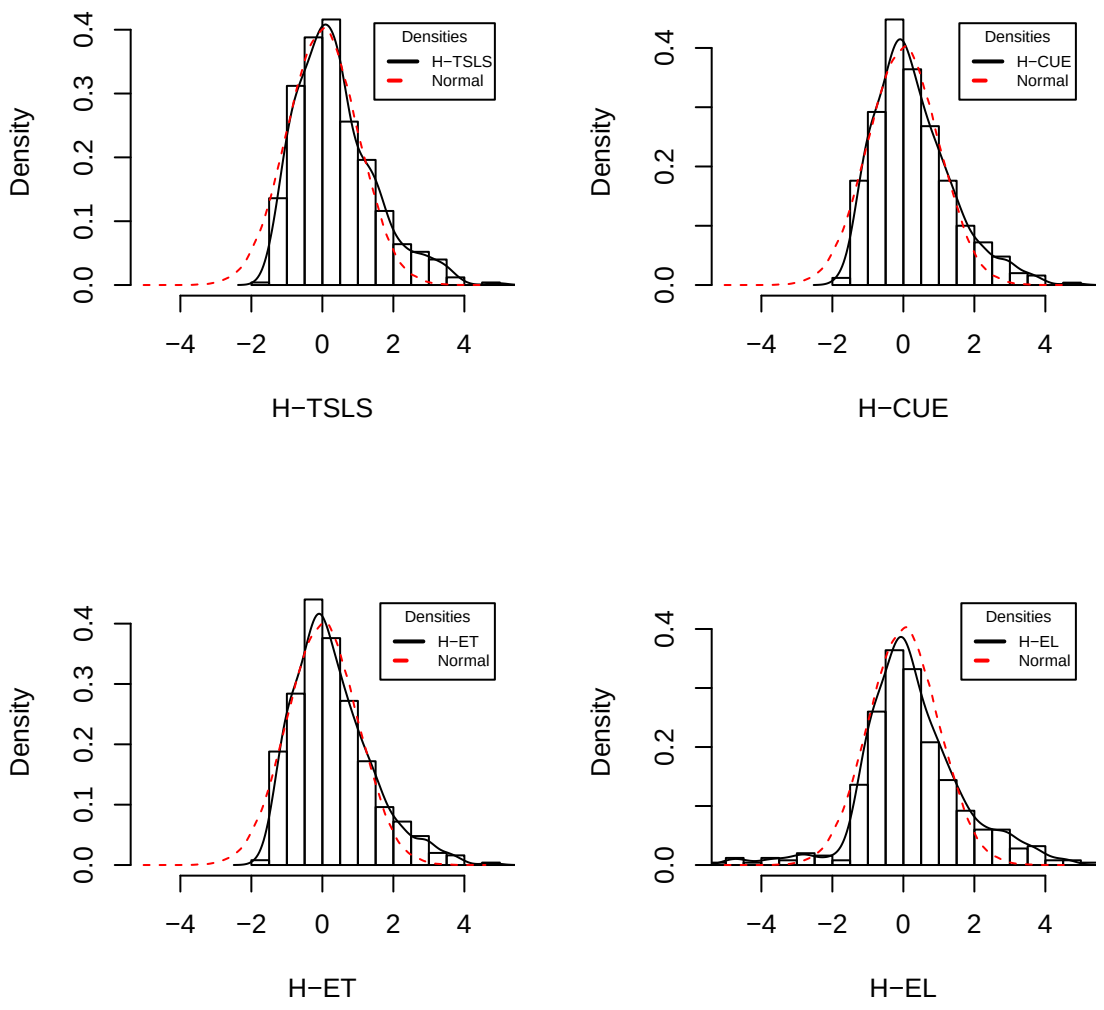

Figure 4: Finite sample densities of hybrid estimators: $n = 100$, $\gamma_2 = 2.5/\sqrt{n}$, $\rho_2 = .99$, $\sigma_\epsilon = 2$

Table 4: Second Stage Results: $n = 200$, $\rho_2 = .5$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		SO	WO	Full model	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 1	Bias	.026	.028	.030	.028	.025	.026	.025	.031	.031	.027	.023	.032
	MAD	.138	.203	.135	.137	.141	.140	.140	.135	.136	.138	.139	.137
	95% Coverage Rate	.948	.942	.950	.948	.932	.932	.950	.950	.936	.950	.944	.938
	MSE	.022	.042	.022	.022	.021	.021	.021	.022	.022	.022	.023	.022
	Model Selection %	-	-	-	.954	.954	.954	.954	.332	-	.856	-	-
Model 2	Bias	.022	.020	.026	.025	.025	.025	.025	.026	.026	.026	.019	.026
	MAD	.118	.151	.117	.117	.119	.119	.119	.117	.119	.119	.116	.116
	95% Coverage Rate	.944	.942	.942	.946	.930	.930	.942	.942	.934	.942	.932	.934
	MSE	.017	.024	.016	.016	.016	.016	.016	.016	.016	.015	.016	.016
	Model Selection %	-	-	-	.804	.804	.804	.804	.140	-	.412	-	-

SO is the model which we use the strong instrument only. WO is the model which we use the weak instrument only. Full model is the one that we use all available instruments. H-TSLS is the hybrid method that we use adaptive lasso selection in first stage and TSLS in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-TSLS except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post-lasso estimator proposed by Belloni et al. (2012). LIML and Fuller's estimator are the conventional methods without any instruments selection.

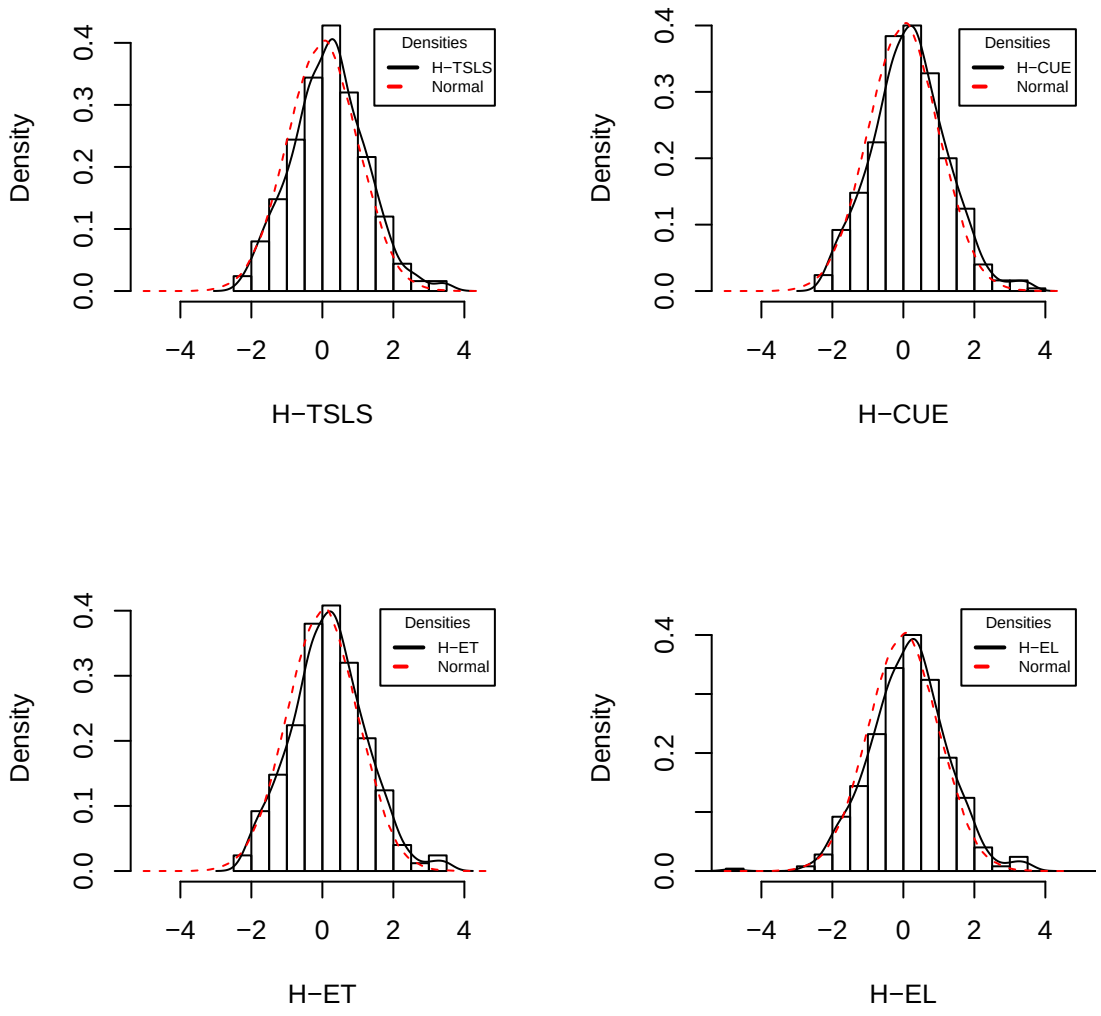

Figure 5: Finite sample densities of hybrid estimators: $n = 200$, $\gamma_2 = 0$, $\rho_2 = .5$, $\sigma_\epsilon = 2$

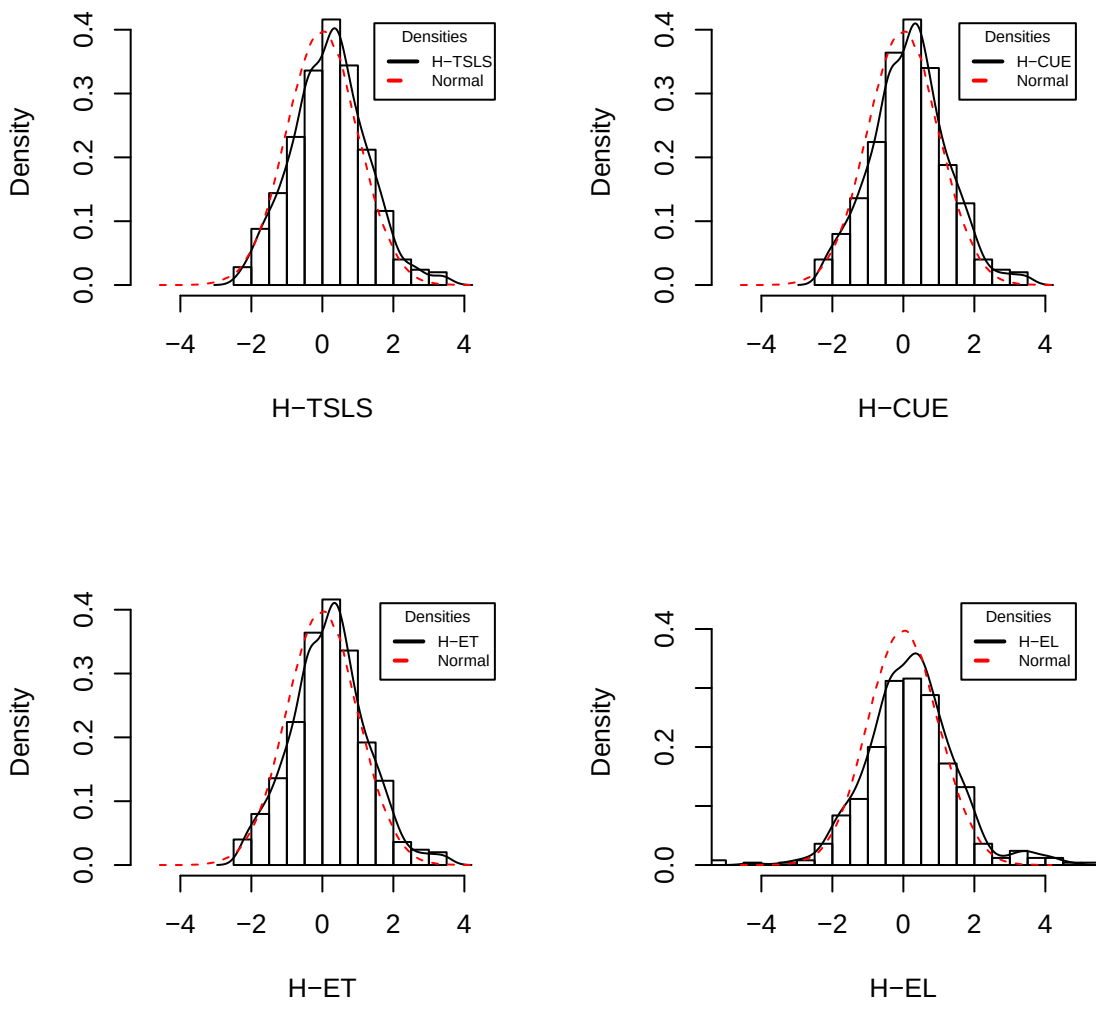

Figure 6: Finite sample densities of hybrid estimators: $n = 200$, $\gamma_2 = 3.54/\sqrt{n}$, $\rho_2 = .5$, $\sigma_\epsilon = 2$

Table 5: Second Stage Results: $n = 200$, $\rho_2 = .99$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		SO	WO	Full model	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 1	Bias	.015	.016	.037	.020	.017	.017	.017	.035	.044	.021	.013	.032
	MAD	.142	.198	.144	.144	.142	.143	.143	.144	.147	.146	.142	.137
	95% Coverage Rate	.930	.928	.910	.926	.924	.926	.910	.910	.906	.922	.938	.916
	MSE	.023	.039	.022	.023	.021	.021	.022	.022	.022	.022	.023	.022
	Model Selection %	-	-	-	.962	.962	.962	.962	.480	-	.840	-	-
Model 2	Bias	.013	.012	.023	.022	.017	.017	.017	.024	.025	.022	.009	.023
	MAD	.121	.149	.122	.123	.121	.121	.120	.122	.124	.123	.126	.122
	95% Coverage Rate	.934	.942	.922	.922	.916	.914	.922	.918	.922	.924	.940	.922
	MSE	.016	.022	.015	.015	.015	.015	.015	.015	.015	.015	.015	.015
	Model Selection %	-	-	-	.786	.786	.786	.786	.358	-	.396	-	-

SO is the model which we use the strong instrument only. WO is the model which we use the weak instrument only. Full model is the one that we use all available instruments. H-TSLS is the hybrid method that we use adaptive lasso selection in first stage and TSLS in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-TSLS except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post-lasso estimator proposed by Belloni et al. (2012). LIML and Fuller's estimator are the conventional methods without any instruments selection.

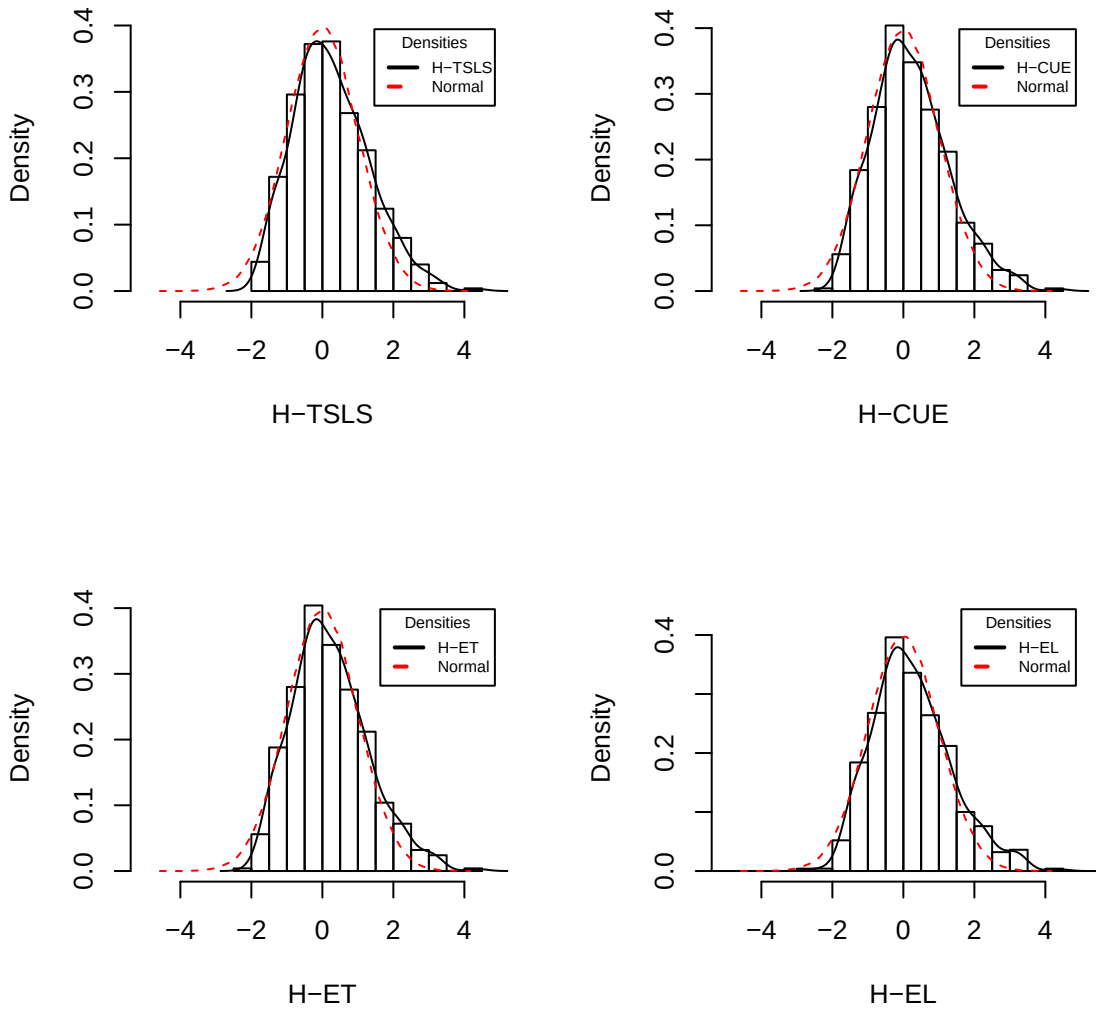

Figure 7: Finite sample densities of hybrid estimators: $n = 200$, $\gamma_2 = 0$, $\rho_2 = .99$, $\sigma_\epsilon = 2$

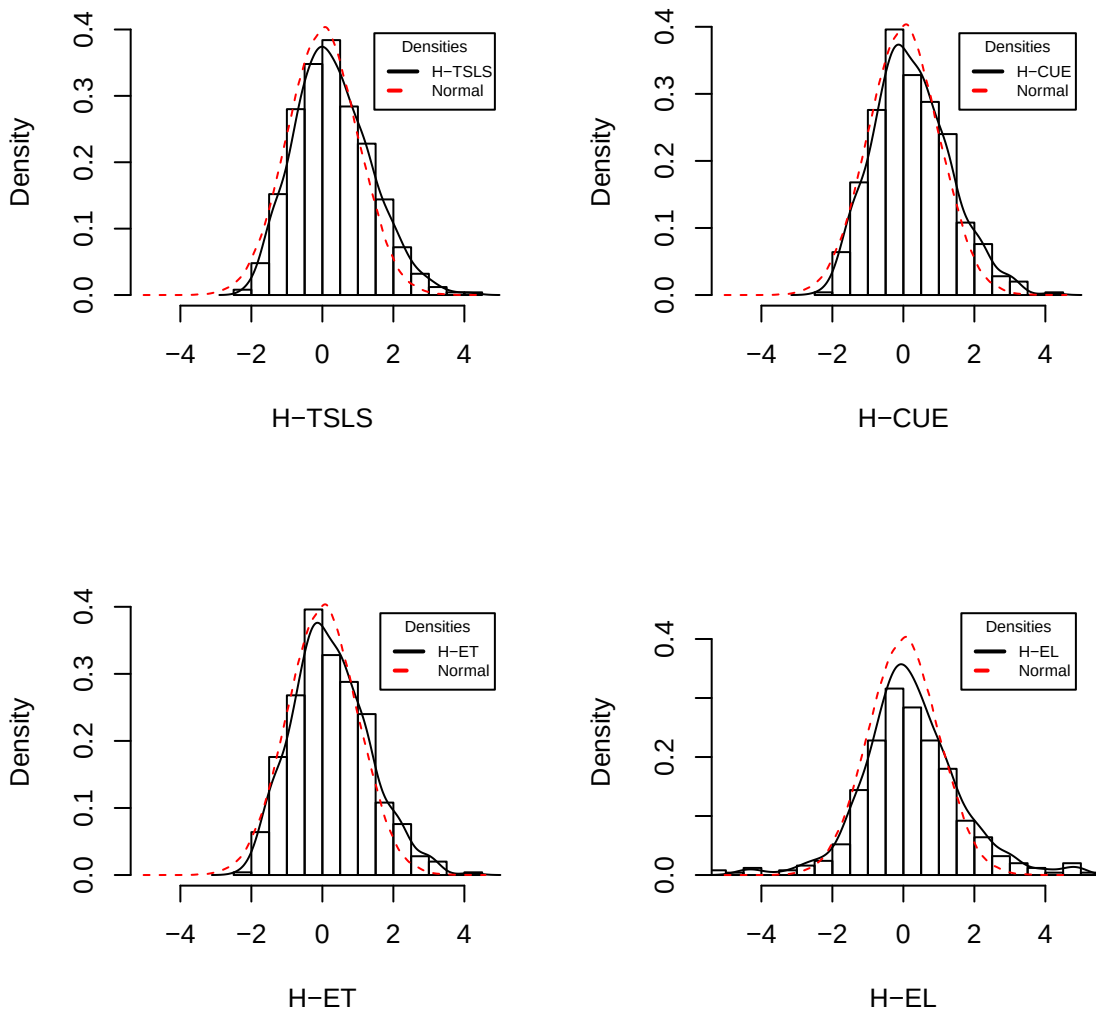

Figure 8: Finite sample densities of hybrid estimators: $n = 200$, $\gamma_2 = 3.54/\sqrt{n}$, $\rho_2 = .99$, $\sigma_\epsilon = 2$

6.2.2 Simulation Results for Conditional Heteroskedasticity

The IV regression model we use here is similar to the one used in conditional homoskedasticity simulations, but with modification of the error terms, replacing ϵ_i by $\epsilon_i = \|Z_i\|\epsilon_i$ ($\|\|$ is the Euclidean norm) and ν_i by $\nu_i = \|Z_i\|\nu_i$ to have the desired heteroskedasticity.

We now describe the simulation results as shown in Tables 6-9. In the tables we present finite sample results for the new hybrid estimators, respectively, hybrid GMM (H-GMM), hybrid CUE (H-CUE), hybrid ET (H-ET) and hybrid EL (H-EL) estimators. We also include Donald and Newey (2001) estimator, Kuersteiner and Okui (2010) model averaging estimator, Post Lasso by Belloni et al. (2012), heteroskedasticity robust Fuller's estimator by Hausman et al. (2012), the conventional full model (use all instruments) GMM, and full model CUE estimators. We report median bias, median absolute deviation, nominal 95% coverage rate, MSE and the percentage of only the strong IV being selected by the model selection methods.

First, we describe the results of MSE. In Table 6, Post Lasso is the best in terms of MSE. The hybrid estimators are the second best in MSE category. In Model 1 of Table 6, Post Lasso has MSE of 0.164. The hybrid GMM, CUE, ET, EL have MSE of 0.196, 0.189, 0.187 and 0.195 respectively. Heteroskedasticity robust Fuller's estimator has 0.288. Donald and Newey (2001) estimator has 0.359. Full model GMM and CUE have MSE of 0.513 and 0.517 respectively. Since Kuersteiner and Okui (2010) estimator is designed for homoskedastic data, its MSE is the highest, which is 0.600. Model 2 has similar results. In Table 7, with small sample size and higher endogeneity, we find that Post Lasso has the best MSE. In Model 1, Donald and Newey (2001) estimator has the second best MSE with 0.135. The hybrid estimators (H-GMM, H-CUE, H-ET, H-EL) have 0.188,0.187,0.186 and 0.186 respectively. In Model 2, Post Lasso has the best MSE. Hybrid estimators, H-GMM, H-CUE, H-ET, H-EL, are the second best in terms of MSE, which take the values of 0.157, 0.156, 0.154 and 0.157 respectively. Donald and Newey (2001) estimator is the third in MSE category with 0.165 in Table 7, Model 2. Tables 8-9 show results when $n = 200$. In terms of median bias, and mean absolute deviation, hybrid estimators perform well compared with post Lasso. Tables 6-9 show that post Lasso has higher bias compared with hybrid estimators. In terms of coverage rates, all methods undercover. Hybrid estimators also have very good model selection percentage in reduced form equation compared with post Lasso and Donald-Newey (2001) estimators as can be seen in Tables 6-9. As expected, in the case of weak instruments, the correct model selection percentages of all methods suffer.

7 Conclusion

This paper proposes hybrid estimators. The first stage is adaptive lasso estimation/model selection. This method penalizes irrelevant instruments and do not use them in the second stage. In the second

Table 6: Second Stage with Heteroskedasticity Results: $n = 100$, $\rho_2 = .5$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	HFUL	FGMM	FCUE
Model 1	Bias	0.075	0.077	0.082	0.084	0.091	0.101	0.379	0.076	0.002	-0.007
	MAD	0.444	0.447	0.447	0.449	0.410	0.401	0.582	0.390	0.471	0.448
	95% Coverage Rate	0.810	0.818	0.822	0.824	0.824	0.926	0.348	0.912	0.904	0.914
	MSE	0.196	0.189	0.187	0.195	0.359	0.600	0.164	0.288	0.513	0.517
	Model Selection %	0.634	0.634	0.634	0.634	0.162	-	0.368	-	-	-
Model 2	Bias	0.028	0.028	0.031	0.032	0.069	0.068	0.298	0.058	0.008	0.008
	MAD	0.359	0.349	0.349	0.352	0.348	0.340	0.495	0.331	0.404	0.389
	95% Coverage Rate	0.872	0.866	0.866	0.866	0.826	0.932	0.472	0.930	0.900	0.912
	MSE	0.161	0.164	0.158	0.161	0.182	0.354	0.112	0.222	0.334	0.333
	Model Selection %	0.588	0.588	0.588	0.588	0.148	-	0.384	-	-	-

H-GMM is the hybrid method that we use adaptive lasso selection in first stage and GMM in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-GMM except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post lasso estimator proposed by Belloni et al. (2010). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the full model GMM. FCUE is the full model CUE.

Table 7: Second Stage with Heteroskedasticity Results: $n = 100$, $\rho_2 = .99$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	HFUL	FGMM	FCUE
Model 1	Bias	-0.008	-0.010	0.001	0.007	0.060	0.073	0.028	0.071	-0.002	0.008
	MAD	0.448	0.452	0.430	0.426	0.423	0.428	0.441	0.385	0.456	0.457
	95% Coverage Rate	0.880	0.880	0.880	0.878	0.800	0.890	0.802	0.870	0.906	0.904
	MSE	0.188	0.187	0.186	0.186	0.135	0.264	0.097	0.180	0.205	0.205
	Model Selection %	0.912	0.912	0.912	0.912	0.426	-	0.816	-	-	-
Model 2	Bias	0.063	0.062	0.059	0.065	0.109	0.120	0.353	0.062	0.009	0.004
	MAD	0.368	0.359	0.363	0.364	0.355	0.353	0.533	0.334	0.368	0.374
	95% Coverage Rate	0.814	0.808	0.808	0.812	0.756	0.818	0.392	0.850	0.866	0.858
	MSE	0.157	0.156	0.154	0.157	0.165	0.316	0.139	0.281	0.363	0.364
	Model Selection %	0.580	0.580	0.580	0.580	0.420	-	0.404	-	-	-

H-GMM is the hybrid method that we use adaptive lasso selection in first stage and GMM in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-GMM except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post lasso estimator proposed by Belloni et al. (2010). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the full model GMM. FCUE is the full model CUE.

Table 8: Second Stage with Heteroskedasticity Results: $n = 200$, $\rho_2 = .5$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	HFUL	FGMM	FCUE
Model 1	Bias	0.082	0.088	0.079	0.080	0.070	0.070	0.189	0.066	0.053	0.054
	MAD	0.316	0.314	0.314	0.318	0.304	0.305	0.387	0.294	0.325	0.328
	95% Coverage Rate	0.912	0.898	0.898	0.894	0.806	0.928	0.616	0.934	0.926	0.920
	MSE	0.108	0.108	0.106	0.107	0.072	0.150	0.061	0.155	0.172	0.170
	Model Selection %	0.768	0.768	0.768	0.768	0.222	-	0.690	-	-	-
Model 2	Bias	0.053	0.054	0.053	0.052	0.054	0.055	0.091	0.049	0.038	0.039
	MAD	0.265	0.266	0.264	0.265	0.260	0.248	0.284	0.260	0.264	0.263
	95% Coverage Rate	0.922	0.920	0.918	0.916	0.792	0.938	0.722	0.938	0.934	0.930
	MSE	0.088	0.088	0.087	0.088	0.039	0.089	0.034	0.098	0.100	0.099
	Model Selection %	0.652	0.652	0.652	0.652	0.200	-	0.646	-	-	-

H-GMM is the hybrid method that we use adaptive lasso selection in first stage and GMM in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-GMM except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post lasso estimator proposed by Belloni et al. (2010). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the full model GMM. FCUE is the full model CUE.

Table 9: Second Stage with Heteroskedasticity Results: $n = 200$, $\rho_2 = .99$, $\rho_1 = .7$, $\sigma_\epsilon = 2$

		H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	HFUL	FGMM	FCUE
Model 1	Bias	0.056	0.054	0.052	0.048	0.088	0.095	0.185	0.067	0.034	0.032
	MAD	0.297	0.297	0.299	0.300	0.311	0.310	0.335	0.293	0.289	0.287
	95% Coverage Rate	0.868	0.864	0.868	0.866	0.740	0.842	0.566	0.880	0.894	0.888
	MSE	0.119	0.118	0.118	0.118	0.214	0.424	0.055	0.192	0.213	0.217
	Model Selection %	0.760	0.760	0.760	0.760	0.398	-	0.716	-	-	-
Model 2	Bias	0.032	0.033	0.031	0.032	0.061	0.069	0.097	0.052	0.030	0.028
	MAD	0.262	0.260	0.261	0.266	0.269	0.266	0.279	0.261	0.265	0.268
	95% Coverage Rate	0.896	0.890	0.886	0.890	0.750	0.878	0.670	0.896	0.902	0.896
	MSE	0.087	0.086	0.087	0.087	0.048	0.109	0.033	0.117	0.115	0.114
	Model Selection %	0.666	0.666	0.666	0.666	0.406	-	0.646	-	-	-

H-GMM is the hybrid method that we use adaptive lasso selection in first stage and GMM in second stage using the selected instruments. H-CUE, H-ET, H-EL are similar to H-GMM except that we use the GEL estimators after selection. DN is the method proposed by Donald and Newey (2001). MA is the model averaging method proposed by Kuersteiner and Okui (2010). PL is the post lasso estimator proposed by Belloni et al. (2010). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the full model GMM. FCUE is the full model CUE.

stage we try two step GMM, as well as Continuous Updating (CUE), Exponential Tilting, Empirical Likelihood estimators. We show that hybrid estimators have good finite sample properties compared with existing methods. We think that a useful extension is to find a way of jointly analyzing reduced and structural form equations in adaptive lasso framework. But this poses identification issues. To overcome them will be a major step.

REFERENCES

Acemoglu, D., S. Johnson, and J. A. Robinson (2001): "The Colonial Origins of a Comparative Development: An Empirical Investigation," *American Economic Review* 91, 1369-1401.

Acemoglu, D., S. Johnson, (2006): "Unbundling Institutions," *Journal of Political Economy* 113, 949-995.

Anderson, P. K. and R.D. Gill (1982): "Cox's Regression Model for Counting Processes: a Large Sample Study," *Annals of Statistics* 10, 1100-1120.

Averkamp, R. and C. Houdre (2003): "Wavelet Thresholding for Non-Necessarily Gaussian Noise: Idealism" *Annals of Statistics* 31, 110-151.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Method and Models for Optimal Instruments with an Application to Eminent Domain," Forthcoming, *Econometrica*.

Bűhlmann, P. and S. van de Geer (2010): "Statistics for High-Dimensional Data," Springer Verlag.

Caner, M. (2009): "Lasso Type GMM estimator," *Econometric Theory*, 25, 270-291.

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate Returns to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp* eds. L.N. Christofedes et. al Toronto: University of Toronto Press, 201-221.

Davidson, J. (1994): "Stochastic Limit Theory," Oxford University Press.

de la Peña, V., T.L. Lai, Q.M. Shao (2009): *Self-normalized processes*. Probability and Applications, Springer Verlag.

Donald, S. and W. Newey (2001): "Choosing The Number of Instruments," *Econometrica*, 69, 1161-1191.

Donoho, D., and I. Johnstone (1994): "Ideal spatial adaptation via wavelet shrinkages," *Biometrika*, 81, 425-455.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004): "Least Angle Regression" *Annals of Statistics*, 32, 407-499.

Fan, J. and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association* 96, 1348-1360.

Fan, J. and R. Li (2002): "Variable Selection for Cox's Proportional Hazards Model and Frailty Model," *Annals of Statistics* 30, 74-99.

Garcia, P. E. (2011): "Linear Regression with a Large Number of Weak Instruments using a Post l_1 Penalized Estimator," . Working Paper. Department of Economics. University of Wisconsin-

Madison.

Hausman, J.A., W.K. Newey, T. Woutersen, J.C. Chao, and N.R. Swanson, (2012): "IV Estimation with Heteroscedasticity and Many Instruments," *Quantitative Economics*, 3, 211-255.

Knight, K. and W. Fu (2000): "Asymptotics for Lasso Type Estimators" *Annals of Statistics*, 28, 1356-1378.

Kuersteiner, G. and R. Okui (2010): "Constructing Optimal Instruments by First Stage Prediction Averaging," *Econometrica* 78, 698-718.

Leeb, H., and B. Pőtscher (2005): "Model selection and inference: facts and fiction. *Econometric Theory*, 21, 21-59.

Newey, W. and R. Smith (2004): "Higher order properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica* 72, 219-257.

Pollard, D. (1991): "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186-199.

Shi, Z. (2011): "Estimation of High Dimensional Linear Structural Model," Working Paper. Department of Economics. Yale University.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica* 65, 557-586.

Stock, J.H. and J. Wright, M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business Economics and Statistics*, 20, 518-528.

van der Vaart, A., and J. Wellner (1996): "Weak Convergence and Empirical Processes" Springer Verlag.

Wang, H., and C. Leng (2007): "Unified LASSO Estimation by Least Squares Approximation," *Journal of The American Statistical Association*, 102, 1039-1049.

Wang, H., B. Li, and C. Leng (2009): "Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters," *Journal of the Royal Statistical Society Series B*, 71, 671-683.

Zou, H. (2006): "The adaptive lasso and its Oracle Properties,", *Journal of The American Statistical Association,"*101, p.1418-1429.

APPENDIX

Proof of Theorem 1.

Consistency is analyzed first, then in part (i) we consider asymptotic normality, then in part (ii) selection consistency is proved. Denote the loss function as:

$$L_n(\gamma_v) = [X_v - \tilde{Z}\gamma_v]'[X_v - \tilde{Z}\gamma_v] + \lambda_n \sum_{j=1}^{q}\sum_{k=1}^{p} \hat{w}_{jk}|\gamma_{jk}|. \tag{18}$$

Using (2) see that sum of squared errors part in that equation can be written as

$$\begin{aligned}\frac{1}{n}(X_v - \tilde{Z}\gamma_v)'(X_v - \tilde{Z}\gamma_v) &= \frac{1}{n}[\nu_v - \tilde{Z}(\gamma_v - \gamma_v^0)]'[\nu_v - \tilde{Z}(\gamma_v - \gamma_v^0)] \\ &= \frac{\nu_v'\nu_v}{n} - 2\frac{\nu_v'\tilde{Z}(\gamma_v - \gamma_v^0)}{n} \\ &+ (\gamma_v - \gamma_v^0)'\left(\frac{\tilde{Z}'\tilde{Z}}{n}\right)(\gamma_v - \gamma_v^0).\end{aligned}$$

First, by Assumption 1

$$\frac{\nu_v'\nu_v}{n} = \frac{\sum_{i=1}^n \sum_{k=1}^p \nu_{ik}^2}{n} \xrightarrow{p} \sigma_\nu^2 > 0.$$

Then by Assumption 2

$$\frac{\tilde{Z}'\nu_v}{n} \xrightarrow{p} 0.$$

Next via Assumption 3

$$\frac{\tilde{Z}'\tilde{Z}}{n} \xrightarrow{p} C < \infty.$$

Combining those in the sum of squared errors part of our objective function

$$\frac{1}{n}(X_v - \tilde{Z}\gamma_v)'(X_v - \tilde{Z}\gamma_v) \xrightarrow{p} \sigma_\nu^2 + (\gamma_v - \gamma_v^0)'C(\gamma_v - \gamma_v^0). \tag{19}$$

Next we consider the penalty term in our objective function. First since $\tilde{\gamma}_{jk} = O_p(n^{-1/2})$,

$$\hat{w}_{jk} = O_p(n^{\tau/2}).$$

Then by Assumption 4

$$\frac{\lambda_n}{n}\hat{w}_{jk} \xrightarrow{p} 0,$$

So

$$\frac{\lambda_n}{n}\sum_{j=1}^q \sum_{k=1}^p \hat{w}_{jk}|\gamma_{jk}| \xrightarrow{p} 0. \tag{20}$$

So since $L_n(\gamma_v)$ is convex by (19)(20)

$$L_n(\gamma_v) \xrightarrow{p} \sigma_\nu^2 + (\gamma_v - \gamma_v^0)'C(\gamma_v - \gamma_v^0) = L(\gamma_v). \tag{21}$$

$$\hat{\gamma}_v = O_p(1), \tag{22}$$

by applying the standard results in Anderson and Gill (1982), Pollard (1991) as in the proof of Theorem 1 in Knight and Fu (2000). So given the last two results we have the consistency of our

estimator, using

$$argminL_n(\gamma_v) \xrightarrow{p} argminL(\gamma_v).$$

See that unique minimum is at γ_v^0 for the limit term in (22) given that C is full rank. So the consistency is proved and

$$\hat{\gamma}_v \xrightarrow{p} \gamma_v^0.$$

(i). We start the asymptotic normality proof now. Set $\hat{u} = \sqrt{n}(\hat{\gamma}_v - \gamma_v^0)$. Specifically we can write $\hat{\gamma}_v$ as

$$\hat{\gamma}_v = \begin{bmatrix} \gamma_1^0 + \frac{\hat{u}_1}{\sqrt{n}} \\ \vdots \\ \gamma_q^0 + \frac{\hat{u}_q}{\sqrt{n}} \end{bmatrix}. \tag{23}$$

and define the following $p \times 1$ vector for each $j = 1, \cdots, q$

$$\gamma_j^0 + \frac{\hat{u}_j}{\sqrt{n}} = \begin{pmatrix} \gamma_{j1}^0 + \frac{\hat{u}_{j1}}{\sqrt{n}} \\ \vdots \\ \gamma_{jp}^0 + \frac{\hat{u}_{jp}}{\sqrt{n}} \end{pmatrix}.$$

Note that

$$\hat{u} = argmin\Psi_n(u),$$

where

$$\Psi_n(u) = [X_v - \tilde{Z}(\gamma_v^0 + \frac{u}{\sqrt{n}})]'[X_v - \tilde{Z}(\gamma_v^0 + \frac{u}{\sqrt{n}})] + \lambda_n \sum_{j=1}^{q}\sum_{k=1}^{p} \hat{w}_{jk}|\gamma_{jk}^0 + \frac{u_{jk}}{\sqrt{n}}|,$$

where $u : pq \times 1$ vector, and u is stacked in the same way as γ_v:

$$u = \begin{bmatrix} u_1 \\ \vdots \\ u_q \end{bmatrix}. \tag{24}$$

each u_j, $j = 1, 2 \cdots q$, is $p \times 1$ vector. Now we can consider the following function

$$\begin{aligned} V_n(u) &= \Psi_n(u) - \Psi_n(0) \\ &= u' \left(\frac{\tilde{Z}'\tilde{Z}}{n} \right) u - 2u' \left(\frac{\tilde{Z}'\nu_v}{\sqrt{n}} \right) \\ &+ \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{q}\sum_{k=1}^{p} \hat{w}_{jk}\sqrt{n}(|\gamma_{jk}^0 + u_{jk}/\sqrt{n}| - |\gamma_{jk}^0|). \end{aligned} \tag{25}$$

See that $\hat{u} = argmin V_n(u)$. Then by Assumption 3

$$\frac{\tilde{Z}'\tilde{Z}}{n} \xrightarrow{p} C < \infty.$$

Next by Assumption 5 (via Central Limit Theorem)

$$\frac{\tilde{Z}'\nu_v}{n^{1/2}} \xrightarrow{d} N(0,\Omega) \equiv W.$$

The limit for the penalty in (25) will be discussed next. Depending on γ_{jk}^0 there are two possibilities. First if $\gamma_{jk}^0 \neq 0$ $(j = 1, 2, \cdots q_0, k = 1, 2, \cdots p)$ we have

$$\hat{w}_{jk} \xrightarrow{p} \frac{1}{|\gamma_{jk}^0|^\tau}.$$

So in that case

$$\sqrt{n}(|\gamma_{jk}^0 + u_{jk}/n^{1/2}| - |\gamma_{jk}^0|) \to u_{jk} sgn(\gamma_{jk}^0),$$

and with Assumption 4 $(\lambda_n/n^{1/2} \to 0)$

$$\frac{\lambda_n}{n^{1/2}}\hat{w}_{jk}[n^{1/2}(|\gamma_{jk}^0 + u_{jk}/n^{1/2}| - |\gamma_{jk}^0|)] \xrightarrow{p} 0.$$

The second case is when $\gamma_{jk}^0 = 0$, we have

$$\sqrt{n}(|\gamma_{jk}^0 + u_{jk}/n^{1/2}| - |\gamma_{jk}^0|) = |u_{jk}|,$$

and with $\hat{w}_{jk}$ definition and in the case of zero parameters $(\gamma_{jk}^0 = 0)$ since the first stage estimator $n^{1/2}\tilde{\gamma}_{jk} = O_p(1)$,

$$\frac{\lambda_n}{n^{1/2}}\hat{w}_{jk} = \frac{\lambda_n}{n^{1/2}} n^{\tau/2}(n^{1/2}\tilde{\gamma}_{jk})^{-\tau} \xrightarrow{p} \infty. \tag{26}$$

by Assumption 4. So unless $u_{jk} = 0$

$$\frac{\lambda_n}{n^{1/2}}\hat{w}_{jk} n^{1/2}(|\gamma_{jk}^0 + u_{jk}/n^{1/2}| - |\gamma_{jk}^0|) \xrightarrow{p} \infty.$$

Clearly given the above results, and defining $u_{\mathcal{A}}$ as the first pq_0 elements of u vector which is of dimension pq, and by C_{11} being the $pq_0 \times pq_0$ upper left block in C matrix, and $W_{\mathcal{A}}$ being the first pq_0 elements of pq vector W, (These designations are done since $\mathcal{A} = \{1, \cdots q_0\}$ without losing any generality)

$$\begin{aligned} V_n(u) \xrightarrow{d} V(u) &= u_{\mathcal{A}}' C_{11} u_{\mathcal{A}} - 2u_{\mathcal{A}}' W_{\mathcal{A}} \quad \text{if} \quad u_{jk} = 0, j = q_0 + 1, \cdots, q, k = 1, \cdots p \\ &= \infty \quad otherwise. \end{aligned}$$

Since V_n is convex and the unique minimum of V is $C_{11}^{-1}W_{\mathcal{A}}$, then by epiconvergence result of Knight and Fu (2000) we get

$$\hat{u}_{\mathcal{A}} \xrightarrow{d} N(0, C_{11}^{-1}\Omega_{11}C_{11}^{-1}),$$

since $W_{\mathcal{A}} = N(0, \Omega_{11})$ where Ω_{11} is the full rank, $pq_0 \times pq_0$ upper left block in Ω ($pq \times pq$ matrix). Also

$$\hat{u}_{\mathcal{A}^c} \xrightarrow{d} 0,$$

where $\mathcal{A}^c = \{q_0 + 1, \cdots, q\}$ by $\hat{u}$ definition. So the limit theory is done.

(ii). Now we prove selection consistency. First $\forall j \in \mathcal{A}$, the consistency shows that

$$P(j \in \mathcal{A}_n) \to 1.$$

We have to show also, $\forall j' \notin \mathcal{A}$,

$$P(j' \in \mathcal{A}_n) \to 0.$$

So for all $j' \notin \mathcal{A}$, take an event $j' \in \mathcal{A}_n$. By Karush-Kuhn-Tucker optimality condition

$$2\tilde{Z}'_{j'}(X_v - \tilde{Z}\hat{\gamma}_v) = \lambda_n(\hat{w}_{j'1}, \cdots, \hat{w}_{j'p})'.$$

Also see that by Assumption 4, for $k = 1, \cdots p$, as in (26)

$$\frac{\lambda_n \hat{w}_{j'k}}{n^{1/2}} = \frac{\lambda_n}{n^{1/2}} n^{\tau/2} \frac{1}{|n^{1/2}\tilde{\gamma}_{j'k}|^{\tau}} \xrightarrow{p} \infty.$$

Rewrite left term of the first order condition above as

$$\frac{2\tilde{Z}'_{j'}[\nu_v - \tilde{Z}(\hat{\gamma}_v - \gamma_v^0)]}{n^{1/2}} = \frac{2\tilde{Z}'_{j'}\nu_v}{n^{1/2}} - \frac{2\tilde{Z}_{j'}\tilde{Z}}{n} n^{1/2}(\hat{\gamma}_v - \gamma_v^0). \tag{27}$$

By the arguments in the proof of the asymptotic normality, Assumptions 3,5, Theorem 1(i), (27) converges to a normal distribution, so

$$P(j' \in \mathcal{A}_n) \leq P(2\tilde{Z}'_{j'}(X_v - \tilde{Z}\hat{\gamma}_v) = \lambda_n(\hat{w}_{j'1}, \cdots, \hat{w}_{j'p})') \to 0.$$

Q.E.D.

Proof of Lemma 1. The proof consists of two parts. First we prove a result regarding refined loadings, then the asymptotic bias result is presented.

Proof of Refined Loadings. First, we need to prove

$$\hat{\pi}_j \xrightarrow{p} \pi_j^0, \tag{28}$$

for $j = 1, \cdots, p$, where $\pi_j^0 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n E z_{ij}^2 v_i^2$ for the refined loadings. We show the proof for refined loadings. The proof for initial loadings are very similar, and hence it is skipped. Define, for each $j = 1, 2 \cdots, p$,

$$\hat{\pi}_j^2 = \frac{1}{n} \sum_{i=1}^n z_{ij}^2 \hat{v}_i^2,$$

and $\hat{v}_i = d_i - z_i' \hat{\gamma}_{InitialLasso}$. Next denote

$$\tilde{\pi}_j^2 = \frac{1}{n} \sum_{i=1}^n z_{ij}^2 v_i^2,$$

We want to prove specifically

$$\max_{1\le j\le p} |\hat{\pi}_j^2 - \tilde{\pi}_j^2| \xrightarrow{p} 0, \tag{29}$$

$$\max_{1\le j\le p} |\tilde{\pi}_j^2 - (\pi_j^0)^2| \xrightarrow{p} 0, \tag{30}$$

Step 3 in p.37 (via Assumption B.1) of the proof of Theorem 1 of Belloni et al (2012) provides the proof of (30). Then proof of Lemma 11 of online appendix in Belloni et al. (2012) shows (29) via Assumption B.1.

Q.E.D.

Proof of Asymptotic Bias of Lasso. Now, we assume consistency of Lasso type estimator that is already proved in Theorem 1 of Belloni et al. (2012). Next, we provide the important step in proving the asymptotic bias of Lasso type estimator of Belloni et al. (2012). Denote the objective function of Belloni et al. (2012) as

$$Q_n(u) = \sum_{i=1}^n (v_i - u' z_i / n^{1/2})^2 + \lambda \sum_{j=1}^p |\hat{\pi}_j (\gamma_{j0} + u_j / n^{1/2})|,$$

where $\hat{u}$ minimizes $Q_n(u)$. Now see that $\hat{u}$ also minimizes the following

$$V_n(u) = Q_n(u) - Q_n(0),$$

where

$$\begin{aligned} V_n(u) &= \sum_{i=1}^n \{[v_i - u' z_i / n^{1/2}]^2 - v_i^2\} \\ &+ \lambda [\sum_{j=1}^p |\hat{\pi}_j (\gamma_{j0} + u_j / n^{1/2})| - |\hat{\pi}_j \gamma_{j0}|]. \end{aligned}$$

Then the first part of proof follows much from Theorem 2 of Knight and Fu (2000) and

$$\sum_{i=1}^{n}\{[v_i - u'z_i/n^{1/2}]^2 - v_i^2\} \xrightarrow{d} -2u'W + u'\Sigma u, \tag{31}$$

where $W \equiv N(0, \Sigma_{Zv})$, and $n^{-1}\sum_{i=1}^{n} z_i z_i' \xrightarrow{p} \Sigma$. This is true through Law of Large Numbers and the Central Limit Theorem given Assumption B.1, and Lemma 3, Condition 1 in Belloni et al. (2012). Next if the true γ are zeroes then the penalty term is:

$$\lambda[\sum_{j=1}^{p}|\hat{\pi}_j(\gamma_{j0} + u_j/n^{1/2})| - |\hat{\pi}_j\gamma_{j0}|] \xrightarrow{p} \lambda_0\sum_{j=1}^{p}|\pi_j^0 u_j|, \tag{32}$$

by (28), and the Assumption of $\lambda/n^{1/2} \to \lambda_0 \geq 0$, and γ_{j0} is the jth element of γ_0 vector, $j = 1, \cdots, p$. If the γ_0 coefficients are nonzero then the limit of the penalty is

$$\lambda[\sum_{j=1}^{p}|\hat{\pi}_0(\gamma_{j0} + u_j/n^{1/2})| - |\hat{\pi}_j\gamma_{0j}|] \xrightarrow{p} \lambda_0\sum_{j=1}^{p}\pi_j^0 u_j sgn(\gamma_{j0}\pi_j^0), \tag{33}$$

where we use again the proof of (28) and consistency of $\hat{\gamma}_{Lasso}$ in Assumption B.2. Now combine (31)(32)(33) to have

$$V_n(u) \xrightarrow{d} V(u) = -2u'W + u'\Sigma u + \lambda_0\sum_{j=1}^{p}[\pi_j^0 u_j sgn(\gamma_{j0}\pi_j^0)1_{\{\gamma_{j0}\neq 0\}} + |u_j\pi_j^0|1_{\{\gamma_{j0}=0\}}]. \tag{34}$$

Q.E.D.

Proof of Lemma 2. Note that $\hat{\gamma}_{Lj}$ for all $j = 1, \cdots p$ represents the estimator in (5). The proof is similar to the proof of the Proposition 1 of Zou (2006). It consists of two parts. The first part is a repeat of Zou (2006) with no change. In the second part of the proof, there is a change due to usage of different penalty factor in Belloni et.al (2010) Lasso estimator.

The first part shows us the main idea behind the proof, hence it is repeated from Zou (2006). We set

$$\mathcal{A}_n = \{j : \hat{\gamma}_{Lj} \neq 0\},$$

$$\mathcal{A} = \{j : \gamma_{Lj0} \neq 0\}.$$

For ease of use set also $\gamma_0 = (\gamma_{\mathcal{A}}, 0_{\mathcal{A}^c})$, where $\gamma_{\mathcal{A}}$ are coefficients that corresponds to the set of nonzero instruments (relevant ones), and $0_{\mathcal{A}^c}$ represents the zero coefficients.

Let $u^* = argminV(u)$ in (34), then

$$P(\mathcal{A}_n = \mathcal{A}) \leq P(\sqrt{n}\hat{\gamma}_{Lj} = 0, \forall j \notin \mathcal{A}).$$

By Lemma 1 $\sqrt{n}\hat{\gamma}_{\mathcal{A}^c} \xrightarrow{d} u^*_{\mathcal{A}^c} = 0$, where $\hat{\gamma}_{\mathcal{A}^c}$ represents the estimators that correspond to "zero population coefficients", and $\mathcal{A}^c = \{j : \gamma_j = 0\}$. Note the typo in the proof of Proposition 1 in Zou (2006) where $\mathcal{A}$ is used instead of $\mathcal{A}^c$ in the previous argument.

But by Portmentaeu Theorem 1.3.4.iii in van der Vaart and Wellner (1996)

$$\limsup P(\sqrt{n}\hat{\gamma}_j = 0, \forall j \notin \mathcal{A}) \leq P(u^*_j = 0, \forall j \notin \mathcal{A}).$$

We need to show

$$c = P(u^*_j = 0, \forall j \notin \mathcal{A}) < 1. \tag{35}$$

The second part of the proof has some modification to Zou (2006) since Lasso of Belloni et al. (2012) is different, in penalty terms, compared to regular Lasso. We only analyze the case of $\lambda_0 > 0$, the case of $\lambda_0 = 0$ is trivial since $c = 0$ in (35) (the same in Zou (2006)) hence it is omitted. By Kuhn-Tucker optimality condition, and Σ being defined in Lemma 1,

$$-2W_j + 2(\Sigma u^*)_j + \lambda_0 \pi^0_j sgn(\pi^0_j \gamma_{j0}) = 0, \forall j \in \mathcal{A}.$$

$$|-2W_j + 2(\Sigma u^*)_j| \leq \lambda_0 \pi^0_j, \forall j \notin \mathcal{A}.$$

We introduce notation that will be useful for the proof. See that

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where Σ_{11} is a square matrix, corresponds to limit of second moments of relevant instruments, it is also invertible and positive definite, Σ_{22} corresponds to limit of second moments of irrelevant instruments, and Σ_{12} is the limit of sample cross product of the relevant with irrelevant instruments, $\Sigma_{21} = \Sigma'_{12}$. $W_{\mathcal{A}}$ are W_j's where $j \in \mathcal{A}$, $W_{\mathcal{A}^c}$ represents W_j's where $j \in \mathcal{A}^c$. Also observe that $u^*_{\mathcal{A}^c} = 0$, $u^*_{\mathcal{A}}$ represents the optimal u with respect to nonzero coefficients. Similarly $\pi^0_{\mathcal{A}}$ is the vector of π^0_j where $j \in \mathcal{A}$, and $\pi^0_{\mathcal{A}^c}$ is the vector of π_j, where $j \in \mathcal{A}^c$.

If $u^*_j = 0$, for all $j \notin \mathcal{A}$, then the optimality condition can be written as componentwise

$$-2W_{\mathcal{A}} + 2\Sigma_{11} u^*_{\mathcal{A}} + \lambda_0 \pi^0_{\mathcal{A}} sgn(\pi^0_{\mathcal{A}} \gamma_{\mathcal{A}}) = 0. \tag{36}$$

$$|-2W_{\mathcal{A}^c} + 2\Sigma_{21} u^*_{\mathcal{A}}| \leq \lambda_0 \pi^0_{\mathcal{A}^c}. \tag{37}$$

Note that this is the difference with Zou (2006) proof, we have π^0 terms in (36)(37). Next combine (36)(37) componentwise

$$|-2W_{\mathcal{A}^c} + \Sigma_{21}\Sigma^{-1}_{11}(2W_{\mathcal{A}} - \lambda_0 \pi^0_{\mathcal{A}} sgn(\pi^0_{\mathcal{A}} \gamma_{\mathcal{A}}))| \leq \lambda_0 \pi^0_{\mathcal{A}^c}.$$

This means that

$$c \leq P[|-2W_{\mathcal{A}^c} + \Sigma_{21}\Sigma_{11}^{-1}(2W_{\mathcal{A}} - \lambda_0 \pi^0_{\mathcal{A}} sgn(\pi^0_{\mathcal{A}}\gamma_{\mathcal{A}}))| \leq \lambda_0 \pi^0_{\mathcal{A}^c}] < 1.$$

Note that in the above equation if the truth is zero coefficient, the weight in adaptive lasso takes positive infinite value unlike $\pi^0_{\mathcal{A}^c}$ of heteroskedastic lasso and makes the right hand side probability equal to one, in the case of adaptive lasso. So just from this equation, also it is possible to compare the adaptive lasso and heteroskedasticity consistent lasso of Belloni et al. (2012). **Q.E.D.**

Proof of Theorem 2.

The first part of the proof (equations (38)-(40)) follows from the proof of Theorem 3 in Zou (2006). Zou (2006) specifically uses iid standard normal random variables in the proof. Since we allow for Gaussian and heteroskedastic data, our proof is different from his. First, we add and subtract from the risk formula

$$\begin{aligned} E[\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2] &= E[\sum_{i=1}^{n}(\hat{\mu}_i - x_i) + (x_i - \mu_i)]^2 \\ &= E[\sum_{i=1}^{n}(\hat{\mu}_i - x_i)^2] + E[\sum_{i=1}^{n}(x_i - \mu_i)^2] \\ &+ 2E[\sum_{i=1}^{n}\hat{\mu}_i(x_i - \mu_i)] - 2E[\sum_{i=1}^{n} x_i(x_i - \mu_i)]. \end{aligned} \tag{38}$$

Note that

$$E[\sum_{i=1}^{n}(x_i - \mu_i)^2] = E[\sum_{i=1}^{n} v_i^2] = \sum_{i=1}^{n}\sigma_i^2,$$

$$E\sum_{i=1}^{n} x_i(x_i - \mu_i) = E[\sum_{i=1}^{n}(\mu_i + v_i)v_i] = E[\sum_{i=1}^{n} v_i^2] = \sum_{i=1}^{n}\sigma_i^2,$$

since μ_i is constant and v_i has zero mean. Substituting these in (38) we obtain

$$E[\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2] = E[\sum_{i=1}^{n}(\hat{\mu}_i - x_i)^2] - \sum_{i=1}^{n}\sigma_i^2 + 2E\sum_{i=1}^{n}\hat{\mu}_i(x_i - \mu_i). \tag{39}$$

Now we consider the first term on the right hand side of (39). Using (10) for each $i = 1, 2 \cdots, n$ we have

$$(\hat{\mu}_i - x_i)^2 = \begin{array}{cc} x_i^2 & \text{if} \quad |x_i| \leq \lambda_i^{1/1+\tau} \\ \frac{\lambda_i^2}{|x_i|^{2\tau}} & \text{if} \quad |x_i| > \lambda_i^{1/1+\tau} \end{array}. \tag{40}$$

We will benefit from (40) in evaluating the first term on the right hand side of (39). Next, we consider the third term on the right hand side of (39). First by Stein's Lemma (Lemma 5.1 in de

la Peña et al. (2009))

$$E[\sum_{i=1}^{n}\hat{\mu}_i(x_i-\mu_i)] \leq E[\sum_{i=1}^{n}\hat{\mu}_i(\frac{x_i-\mu_i}{\sigma_i})]\max_i \sigma_i \leq E\left[\sum_{i=1}^{n}\frac{\partial\hat{\mu}_i}{\partial x_i}\right]d, \tag{41}$$

where $\max_i \sigma_i \leq d$ and $0 < d < \infty$, $\sigma_i > 0$. Since by (10), for each $i = 1, 2\cdots, n$

$$\frac{\partial\hat{\mu}_i}{\partial x_i} = \begin{array}{ll} 0 & \text{if} \quad |x_i| \leq \lambda_i^{1/1+\tau} \\ 1+\frac{\lambda_i}{\tau|x_i|^{1+\tau}} & \text{if} \quad |x_i| > \lambda_i^{1/1+\tau} \end{array}. \tag{42}$$

Combine (40)(41)(42) in (39) we can rewrite

$$\begin{aligned} E\sum_{i=1}^{n}(\hat{\mu}_i-\mu_i)^2 &\leq E\sum_{i=1}^{n}(x_i^2 1_{\{|x_i|\leq\lambda_i^{1/(1+\tau)}\}}) + E\sum_{i=1}^{n}\left(\frac{\lambda_i^2}{|x_i|^{2\tau}}1_{\{|x_i|>\lambda_i^{1/(1+\tau)}\}}\right) - \sum_{i=1}^{n}\sigma_i^2 \\ &+ E\left[\sum_{i=1}^{n}\left(2+\frac{2\lambda_i}{\tau|x_i|^{1+\tau}}\right)1_{\{|x_i|>\lambda_i^{1/(1+\tau)}\}}\right]d \\ &= E\left[\sum_{i=1}^{n}x_i^2 1_{\{|x_i|\leq\lambda_i^{1/(1+\tau)}\}}\right] + E\left[\sum_{i=1}^{n}\left(\frac{\lambda_i^2}{|x_i|^{2\tau}}+2d+\frac{2\lambda_i d}{\tau|x_i|^{1+\tau}}\right)1_{\{|x_i|>\lambda_i^{1/(1+\tau)}\}}\right] \\ &- \sum_{i=1}^{n}\sigma_i^2. \end{aligned} \tag{43}$$

By using $|x_i| \leq \lambda_i^{1/1+\tau}$ for the first right hand side term in (43), then using $|x_i| > \lambda_i^{1/1+\tau}$ to get $\frac{1}{|x_i|^{2\tau}} \leq \frac{1}{|\lambda_i|^{2\tau/1+\tau}}$ in the second term on the right hand side of (43) and $\frac{\lambda_i}{|x_i|^{1+\tau}} \leq 1$

$$\begin{aligned} E\sum_{i=1}^{n}[\hat{\mu}_i-\mu_i]^2 &\leq \sum_{i=1}^{n}[\lambda_i^{2/1+\tau}P(|x_i|\leq\lambda_i^{1/1+\tau})] \\ &+ \sum_{i=1}^{n}[(2d+2d/\tau+\lambda_i^{2/1+\tau})P(|x_i|>\lambda_i^{1/1+\tau})] - \sum_{i=1}^{n}\sigma_i^2 \\ &\leq \sum_{i=1}^{n}\lambda_i^{2/1+\tau}+2d+2d/\tau. \end{aligned} \tag{44}$$

Now we will simplify this expression for further use. We can rewrite, using $\lambda_i = (2\sigma_i^2 logn)^{(1+\tau)/2}$

$$\begin{aligned} \lambda_i^{2/1+\tau}+2d+2d/\tau &= \sigma_i^2(2logn+\frac{2d}{\sigma_i^2}+\frac{2d}{\tau}\frac{1}{\sigma_i^2}) \leq \sigma_i^2[2logn+\frac{2d}{\min_i\sigma_i^2}+\frac{2d}{\tau}\frac{1}{\min_i\sigma_i^2}] \\ &\leq \sigma_i^2[2logn+2c+\frac{2}{\tau}c], \end{aligned}$$

where $c > \frac{d}{min_i \sigma_i^2}$. So the bound in (44) can be written as

$$E\sum_{i=1}^{n}[(\hat{\mu}_i - \mu_i)^2] \leq [2logn + 2c + \frac{2}{\tau}c]\sum_{i=1}^{n}\sigma_i^2. \tag{45}$$

In the next part of the proof we will get a new bound for the estimated risk, and then we compare with the one that we found in (45). Use (43)

$$\begin{aligned} E[\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2] &\leq E\sum_{i=1}^{n}x_i^2 + E[\sum_{i=1}^{n}\left(\frac{\lambda_i^2}{|x_i|^{2\tau}} + 2d + \frac{2\lambda_i d}{\tau|x_i|^{1+\tau}} - x_i^2\right)1_{\{|x_i|>\lambda_i^{1/1+\tau}\}}] - \sum_{i=1}^{n}\sigma_i^2 \\ &= E[\sum_{i=1}^{n}\left(\frac{\lambda_i^2}{|x_i|^{2\tau}} + 2d + \frac{2\lambda_i d}{\tau|x_i|^{1+\tau}} - x_i^2\right)1_{\{|x_i|>\lambda_i^{1/1+\tau}\}}] + \sum_{i=1}^{n}\mu_i^2. \end{aligned} \tag{46}$$

When $|x_i| > \lambda_i^{1/(1+\tau)}$,

$$\frac{\lambda_i^2}{|x_i|^{2\tau}} - x_i^2 \leq \frac{\lambda_i^2}{|x_i|^{2\tau}} - \lambda_i^{2/1+\tau}. \tag{47}$$

and

$$\frac{1}{|x_i|^{2\tau}} \leq \frac{1}{\lambda_i^{2\tau/(1+\tau)}}. \tag{48}$$

so

$$\begin{aligned} \frac{\lambda_i^2}{|x_i|^{2\tau}} &- \lambda_i^{2/1+\tau} \leq \frac{\lambda_i^2}{\lambda_i^{2\tau/(1+\tau)}} - \lambda_i^{2/(2+\tau)} \\ &= \lambda_i^{2/(1+\tau)} - \lambda_i^{2/(1+\tau)} = 0. \end{aligned} \tag{49}$$

By (47)-(49), if $|x_i| > \lambda_i^{1/(1+\tau)}$

$$\frac{\lambda_i^2}{|x_i|^{2\tau}} - x_i^2 \leq 0. \tag{50}$$

So use (50) in (46) to have

$$E\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2 \leq E\left[\sum_{i=1}^{n}\left(\frac{2\lambda_i d}{\tau|x_i|^{1+\tau}} + 2d\right)\ 1_{\{|x_i|>\lambda_i^{1/(1+\tau)}\}}\right] + \sum_{i=1}^{n}\mu_i^2. \tag{51}$$

When $|x_i| > \lambda_i^{1/(1+\tau)}$ we can rewrite (51) as

$$E[\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2] \leq (\frac{2d}{\tau} + 2d)\sum_{i=1}^{n}P(|x_i| > \lambda_i^{1/(1+\tau)}) + \sum_{i=1}^{n}\mu_i^2. \tag{52}$$

Now we try to evaluate the $P(|x_i| > \lambda_i^{1/(1+\tau)})$. Set $t_i = \lambda_i^{1/1+\tau}$, and proceed as in p.1427-1428 of

Zou (2006) to have

$$\begin{aligned} P(|x_i| > t_i) &\leq \frac{2}{\sqrt{2\pi\sigma_i^2}t_i} e^{-t_i^2/2\sigma_i^2} + 2\mu_i^2. \\ &\leq \frac{1}{n\sqrt{\pi\sigma_i^2}} (logn)^{-1/2} + 2\mu_i^2, \end{aligned} \tag{53}$$

where we use t_i definition and $\lambda_i = (2\sigma_i^2 logn)^{(1+\tau)/2}$ in the last step. See also the equations after (A.12) in Zou (2006). Use (53) in (52)

$$\begin{aligned} E\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2 &\leq (\frac{4d}{\tau} + 4d) \max_i (\frac{1}{2\sqrt{\pi\sigma_i^2}} (logn)^{-1/2}) + (\frac{4d}{\tau} + 4d + 1) \sum_{i=1}^{n} \mu_i^2 \\ &\leq (\frac{4d}{\tau} + 4d) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (logn)^{-1/2} + (\frac{4d}{\tau} + 4d + 1) \sum_{i=1}^{n} \mu_i^2, \end{aligned} \tag{54}$$

by c, d definitions. Add $2logn \sum_{i=1}^{n} \mu_i^2$ and $(2logn + 1)(c^{1/2}/d^{1/2}) \frac{1}{2\sqrt{\pi}} (logn)^{-1/2}$ to (54) so that it is compatible with the bound in (45)

$$E\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2 \leq (2logn + \frac{4d}{\tau} + 4d + 1) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (logn)^{-1/2} + (2logn + \frac{4d}{\tau} + 4d + 1) \sum_{i=1}^{n} \mu_i^2. \tag{55}$$

Then set $b = max(2c, 4d + 1)$. Use b definition to rewrite (55) as

$$E\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2 \leq (2logn + \frac{b}{\tau} + b) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (logn)^{-1/2} + (2logn + \frac{b}{\tau} + b) \sum_{i=1}^{n} \mu_i^2. \tag{56}$$

Next add $(2logn + b + \frac{b}{\tau}) c^{1/2}/d^{1/2} \frac{1}{2\sqrt{\pi}} (logn)^{-1/2}$ to (45) and use b definition as well to have

$$E\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2 \leq (2logn + \frac{b}{\tau} + b) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (logn)^{-1/2} + (2logn + \frac{b}{\tau} + b) \sum_{i=1}^{n} \sigma_i^2. \tag{57}$$

The result can be deducted from (56)(57).

Q.E.D.