Journal of Econometrics 187 (2015) 256–274

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso

Mehmet Caner [a,*], Qingliang Fan [b,**]

[a] North Carolina State University, Department of Economics, 4168 Nelson Hall, Raleigh, NC 27518, United States

[b] Wang Yanan Institute for Studies in Economics (WISE), Department of Statistics and Fujian Key Laboratory of Statistical Science, Xiamen University, 361005, China

ARTICLE INFO

Article history:
Received 17 September 2012
Received in revised form
21 May 2014
Accepted 8 January 2015
Available online 11 March 2015

JEL classification:
C52
C26
C13

Keywords:
Model selection
Near minimax risk bound
Shrinkage estimators

ABSTRACT

In this paper, we use the adaptive lasso estimator to choose the relevant instruments and eliminate the irrelevant instruments. The limit theory of Zou (2006) is extended from univariate iid case to heteroskedastic and non Gaussian data. Then we use the selected instruments in generalized empirical likelihood estimators (GEL). In this sense, these are called hybrid GEL. It is also shown that the lasso estimators are not model selection consistent whereas the adaptive lasso can select the correct model with fixed number of instruments. In simulations we show that hybrid GEL estimators have smaller bias and mean squared error than the other estimators in certain cases.

1. Introduction

One of the important issues in economics is the correct selection of the instruments. A lot of empirical cases involving labor, institutional economics deal with very limited number of instruments. For example Acemoglu et al. (2001), Acemoglu and Johnson (2006), Card (1995) papers use 1 or at most 2 instruments with one endogenous variable at hand. It is critical in those cases to see whether the researchers have strong instruments or not. If there is only one instrument in just identified case, and it is weak, the second stage coefficient is inconsistent which is shown in Staiger and Stock (1997). If there is more than one instrument and only one endogenous variable then the second stage regression will give consistent estimate, but in finite samples there will be bias.

In many instruments setup, there have been several papers analyzing instrument selection recently. Donald and Newey (2001) target mean squared error of the second stage regression coefficients. Their paper do not take into account the weakness of the instruments. Kuersteiner and Okui (2010) use model averaging to pick up instruments. Their approach is similar to Donald and Newey (2001). In a landmark paper, Belloni et al. (2012) introduce a new heteroskedasticity consistent Lasso type estimator to select 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 the 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 (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 uses a penalty term that depends on the residuals from the first stage.

Our paper is interested in selecting the instruments in a fixed number of instruments setup. To achieve that objective, we benefit from the recent statistics literature. In statistics, the model selection and estimation via lasso, bridge estimators are analyzed by Knight and Fu (2000). Caner (2009) analyzes GMM objective function based bridge type penalty. Caner (2009) does not consider instrument selection and only considers model selection for the second stage regression. The smooth penalty functions are introduced by Fan and Li (2001, 2002). However, one of the most recent shrinkage based estimator in statistics is the adaptive lasso

* Corresponding author. Tel.: +1 919 513 0853; fax: +1 919 513 0854.
** Corresponding author.
E-mail addresses: mcaner@ncsu.edu (M. Caner), michaelqfan@gmail.com (Q. Fan).

http://dx.doi.org/10.1016/j.jeconom.2015.01.007
0304-4076/

of Zou (2006). This has optimality properties, and is easier to estimate than the bridge estimator. The lasso, in fixed number of regressor case in least squares, cannot select the correct model with probability one which can be seen in Proposition 1 of Zou (2006). Zou (2006) also shows that the adaptive lasso is model selection consistent, and achieves the near minimax risk bound in iid Gaussian data. Adaptive lasso is oracle efficient which means the nonzero parameters are estimated with standard efficient limit in least squares and the zero parameters are estimated as zero with probability approaching one.

Our paper applies adaptive lasso technique to the instrument selection problem in the reduced form estimates, and then run the generalized empirical likelihood estimators in the second stage regression. The main idea here is to benefit from the very good model selection properties of the adaptive lasso in the first stage. In other words, it can pick the relevant instruments with probability approaching one. Given the first stage results we have a small 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 the adaptive lasso is model selection consistent. Adaptive lasso can differentiate between the irrelevant and strong instruments. In the simulations, the 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 there are two working papers which apply shrinkage based methods to IV regression. The first one, independently written, is by Garcia (2011). He devises the 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 main issue is the structural equation parameter selection. That paper is also an important contribution since it is important to handle high dimensional problems in econometrics. Shrinkage methods will be immensely useful in high dimensional cases for the applied researchers. Other papers that analyze selection of strong instruments among both strong and redundant ones are by Hall et al. (2007), and Cheng and Liao (2012). Hall et al. (2007) shows that information criterion based approach can be used. Cheng and Liao (2012) analyze the many invalid and redundant instrument case. One very important contribution to this literature is by Leeb and Pötscher (2005). They show that if the parameters are varying with 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 the adaptive lasso in the weak instruments context. 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 Pötscher (2005). Leeb and Pötscher (2005) idea applies to the least squares framework. However, we are interested in second stage regression estimates. We show that the 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 the adaptive lasso with all instruments. Adaptive lasso selects the relevant instruments and eliminates the irrelevant ones, and uses this information in the second stage estimation via GEL. Section 3 examines the limit theory for GEL when the estimated instruments in the first stage are used. Section 4 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 5 provides the algorithm that is used. Section 6 introduces an oracle inequality. Section 7 carries out extensive simulations. The Appendix provides all the proofs. Let $\|.\|$, $\|.\|_\infty$ denote the Euclidean norm, and maximal value of a vector respectively.

2. The reduced form estimation via adaptive lasso

In this section, we show that the adaptive lasso can be used in a 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. We will be analyzing Eq. (2), then the next section will take on (1). In matrix form the structural and reduced form equations are, respectively,

$$Y = X\beta_0 + \epsilon, \tag{1}$$

$$X = Z\gamma^0 + \nu, \tag{2}$$

where $Y = (y_1, y_2, \ldots, y_i, \ldots, y_n)'$ is $n \times 1$, and X is $n \times p$ full column rank matrix, X_i represents $p \times 1$ vector for each $i = 1, 2, \ldots, n$, $\beta_0 : p \times 1$ vector which shows the true parameters. Now assume that the instruments are uncorrelated with ϵ_i. Next, the dimensions of the instrument matrix, and the reduced form coefficients are $Z : n \times q$, and $\gamma^0 : q \times p$. We also have $q \geq p$, where q represents all the fitted instruments. We can rewrite (2)

$$X = Z_1\gamma_1^{0'} + Z_2\gamma_2^{0'} + \cdots + Z_q\gamma_q^{0'} + \nu, \tag{3}$$

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

2.1. The objective function and the assumptions

Next we can write the reduced form equation in vector form as:

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

We can write this as:

$$X_v = \tilde{Z}_1\gamma_1^0 + \cdots + \tilde{Z}_q\gamma_q^0 + \nu_v, \tag{4}$$

where $X_v = vec(X)$, $\tilde{Z}_j = (I_p \otimes Z_j)$, $\nu_v = vec(\nu)$. 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, for $j = 1, 2, \ldots, q$. The true number of nonzero vectors are q_0, with $q_0 \geq p$. We can rewrite (4)

$$X_v = \tilde{Z}\gamma_v^0 + \nu_v, \tag{5}$$

where $\tilde{Z} = [\tilde{Z}_1, \ldots \tilde{Z}_q]$ ($np \times pq$ 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:

$$\hat{\gamma}_v = \underset{\gamma_v}{\operatorname{argmin}}[X_v - \tilde{Z}\gamma_v]'[X_v - \tilde{Z}\gamma_v] + \lambda_n \sum_{j=1}^{q}\sum_{k=1}^{p} \hat{w}_{jk}|\gamma_{jk}|. \tag{6}$$

See that the weights are $\hat{w}_{jk} = |\tilde{\gamma}_{jk}|^{-\tau}$, $\tau > 0$, where $\tilde{\gamma}_{jk}$ is the $\sqrt{n}$ consistent LS estimate of γ_{jk}. To understand these better note that γ_v is pq vector, so these are stacked $p \times 1$ vectors, γ_j, for

$j = 1, \ldots, q$. In these q vectors, the nonzero ones will be denoted by

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

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

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

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

Assumption F.1.

$$\frac{v_v' v_v}{n} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{p} v_{ik}^2}{n} \xrightarrow{p} \sigma_v^2 > 0,$$

where $\sigma_v^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p} E v_{ik}^2$, and $\sigma_v^2 < \infty$.

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

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

Assumption F.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 and of full rank, as well as symmetric submatrix of dimensions $q_0 p \times q_0 p$.

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

Assumption F.5. We assume the following Central Limit Theorem

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

Ω is $pq \times pq$ matrix, and it has Ω_{11} as the $q_0 p \times q_0 p$ upper left block, as positive definite, full rank matrix.

Note that Assumptions F.1–F.3 and F.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. Assumption F.4 is used in Zou (2006). This shows the behavior of the penalty. Note that $\frac{\lambda}{n} n^{\tau/2} \to 0$ is not shown in Zou (2006). This is needed for consistency.[1]

Set $\mathcal{A}_n = \{j : \hat{\gamma}_j \neq 0_p\}$. Note that $\hat{\gamma}_{vA}$ represents the adaptive lasso estimator for the first $q_0 p$ elements in $\hat{\gamma}_v$. Let γ_{vA}^0 ($q_0 p \times 1$ vector) represent the corresponding true nonzero elements. The following theorem generalizes the adaptive lasso of Zou (2006) from univariate, iid case to multivariate, heteroskedastic case. Oracle property is also preserved as in Zou (2006).

[1] We thank an anonymous referee for pointing that out.

Theorem 1. *Under Assumptions F.1–F.5,*

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

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

Note that $\mathcal{A}$ definition can be changed to $\mathcal{A}' = \{j, k : \gamma_{j,k} \neq 0, j = 1, \ldots, q, k = 1, \ldots, p\}$. 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 the first q_0 of γ_j^0 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_1^0 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.

3. 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 of (1). Then the first estimator to consider is two step GMM

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

where $Y = (y_1, \ldots, y_n)'$, X is $n \times p$ full column rank matrix, and $\hat{Z} = [Z_1, \ldots, Z_{\hat{q}_0}] : n \times \hat{q}_0$ matrix. So $\hat{Z}$ is the matrix of estimated number of instruments, $\hat{q}_0$. Note this is not the predicted instruments by using the reduced form equations. This is just the result from the adaptive lasso in the first stage. Adaptive lasso picks $\hat{q}_0$ instruments in the first stage, and by Theorem 1(ii), $\hat{q}_0 \xrightarrow{p} q_0$. $\hat{W}$ is the standard efficient weight used in GMM.

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

$$\hat{\beta}_{GEL} = \underset{\beta\in\mathcal{B}}{\text{argmin}} \sup_{\delta\in\Delta_n(\beta)} \sum_{i=1}^{n} \rho(\delta' g_{ie}(\beta)), \tag{7}$$

where $\mathcal{B}$ is a compact subset of R^p, and $\Delta_n(\beta) = \{\delta : \delta' g_{ie}(\beta) \in \mathcal{V}, i = 1, \ldots, n\}$, $\mathcal{V}$ is an open interval containing zero. Since this is a linear model, the sample moments are $g_{ie}(\beta) = \hat{Z}_i(y_i - X_i'\beta)$, and $\hat{Z}_i$ is $\hat{q}_0 \times 1$ vector of instruments selected in the first stage via adaptive lasso, and X_i is the $p \times 1$ vector. When $\rho(\iota) = ln(1-\iota)$ this is the empirical likelihood estimator, when $\rho(\iota) = -exp(\iota)$, this is called the exponential tilting estimator, and with $\rho(\iota) = -(1 + \iota)^2/2$, it is the continuous updating estimator. Also see that the first order partial derivative and the second order partial derivative evaluated at zero are set at $\rho_1(0) = -1$, $\rho_2(0) = -1$ respectively. For standardizations, and further details, see Newey and Smith (2004). If we can obtain q_0, true number of strong instruments in the first stage, in the second stage the standard GEL limits can be derived as shown in Newey and Smith (2004).

Here in this section we will consider the second stage regression GEL, with estimated number of instruments. The case for two step GMM is the same, and will not be considered therefore. We compare and contrast what may happen in the second stage with different mistakes in the selection step in the first stage. Our analysis is generally asymptotic. In simulations we see finite sample model selection results and their impact on mean squared error, the bias of second stage estimators. The first possibility is that truth is $q_0 < p$. Then via Theorem 1(ii), we get $\hat{q}_0 \xrightarrow{p} q_0$,

and we do not go to a second stage regression at all. But if we do not consider model selection we can continue and estimate a model with weak identification for the second stage. Then the second stage coefficient estimates are inconsistent as can be seen in Theorem 2(i) of Guggenberger and Smith (2005). This is one of the advantages of our model, it can stop in the first step. Also the model selection percentage of several methods is important to compare since we do not want to pick weak/irrelevant instruments in the first stage to cause inconsistent estimates in the second stage in the case described above. This type of analysis has been done in our simulations, and shows the superior model selection property of the adaptive lasso method. The second possibility is $q_0 \geq p$. There, we run the second stage regression using GEL and we show that the first stage selection of the estimated number of instruments $\hat{q}_0$ will not alter the limit of the standard GEL. Under $q_0 \geq p$, there are three cases, without losing any generality in the proofs, assume $\hat{q}_0 \geq q_0 \geq p$, but $\hat{q}_0 \xrightarrow{p} q_0$ as in Theorem 1(ii). Note that two other cases are: $q_0 > \hat{q}_0 \geq p$, and $q_0 \geq p > \hat{q}_0$. We will discuss these cases after the proof of Theorem 2. Also set $Z = [Z_r, Z_w]$, where Z_r represents $n \times q_0$ matrix of strong instruments, Z_{ri} is q_0 column vector, and Z_w are the $n \times (q - q_0)$ matrix of irrelevant instruments. The proofs follow through with other setups of instruments, this one does not burden us with extra notation. Also define $\Sigma = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} EZ_i Z_i' \epsilon_i^2$. The assumptions are set now. These are done for all instruments so that we can compare our estimated number of instruments case with no model selection case. Let rank(M) for a generic matrix M, represent the rank of the matrix M.

Assumption S.1. (ϵ_i, ν_i, Z_i) are independent across $i = 1, \ldots, n$.

Assumption S.2. $EZ_i\epsilon_i = 0$, $EZ_i\nu_i' = 0$, where Z_i is $q \times 1$ vector of all fitted instruments, and $q \geq \hat{q}_0 \geq p$. Also $q_0 \geq p$. $rank(EZ_{ri}X_i') = p$.

Assumption S.3. Set $g_i(\beta) = Z_i(y_i - X_i'\beta)$. Then assume

$$\max_i \sup_\beta E\|g_i(\beta)\|^\xi < \infty, \quad \xi > 2.$$

Assumption S.4.

$$\tilde{\Sigma} = n^{-1} \sum_{i=1}^{n} Z_i Z_i' \tilde{e}_i^2 \xrightarrow{p} \Sigma,$$

where $\tilde{e}_i = y_i - X_i'\tilde{\beta}$, where $\tilde{\beta}$ represents GEL by using all instruments (q instead of $\hat{q}_0$).

$$\Sigma = \begin{pmatrix} \Sigma_r & \Sigma_{rw} \\ \Sigma_{rw}' & \Sigma_w \end{pmatrix},$$

where Σ is $q \times q$ nonsingular matrix. Σ_r is $q_0 \times q_0$ matrix, and Σ_w is $(q - q_0) \times (q - q_0)$ matrix. Σ_r represents the block that corresponds to strong instruments, this structure makes the limit easy to analyze without losing any generality.

Assumption S.5. $\beta \in B$, where B is a compact subset of R^p. $\rho(\iota)$ is twice continuously differentiable in ι, which is in a neighborhood of 0.

Assumption S.6. We assume the following Central Limit Theorem

$$n^{-1/2} \sum_{i=1}^{n} Z_i\epsilon_i \xrightarrow{d} N(0, \Sigma).$$

These are Assumptions 1–2 in Newey and Smith (2004) which are strengthened for the independent data, or we can use (linear sub case) Assumptions Θ, ID, $M\rho$ in Guggenberger and Smith (2005). We do not allow for weak instruments unlike Guggenberger and Smith (2005). Even though we have strong and irrelevant instruments, since $\hat{q}_0 \geq p$, there is identification, and consistency. The proof of consistency is trivial given identification, and is also shown in the proof of Theorem 2 in the Appendix. Assumptions S.1–S.2 are standard and show the validity of the instruments. Assumption S.3 is needed for consistency in GEL. Assumptions S.4–S.6 are standard limit theorems, and description of $\rho(.)$ function. Let $\hat{\beta}$ represent the GEL with estimated number of instruments from now on. This is defined as $\hat{\beta}_{GEL}$ in (7).

Next set $G_r = \Sigma_{zr}\gamma^0$, and $n^{-1}Z_r'Z_r \xrightarrow{p} \Sigma_{zr}$, where $\Sigma_{zr} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} EZ_{ri}Z_{ri}'$. Note that Σ_{zr} is $q_0 \times q_0$ matrix, and G_r is of dimension $q_0 \times p$, and is of full column rank. This can be obtained through Assumption F.3, and the reduced form equation via Assumption F.2. G_w is the cross product of irrelevant and relevant instruments, and it is $(q - q_0) \times p$, full column rank matrix. Namely $G_w = \Sigma_{zw}\gamma^0$, and $n^{-1}Z_w'Z_r \xrightarrow{p} \Sigma_{zw}$, where $\Sigma_{zw} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} EZ_{wi}Z_{ri}'$. Define

$$G_F = \begin{bmatrix} G_r \\ G_w \end{bmatrix},$$

as the limit for $Z'X/n$.

The next theorem shows that even though we use the estimated number of instruments $\hat{q}_0$, the limit of GEL is equivalent to the one if we had used q_0 number of instruments. The main reason for that result is Theorem 1(ii). The proof of Theorem is in the Appendix.

Theorem 2. *Under Assumptions F.2–F.3, S.1–S.6, $\hat{\beta}$ is consistent and*

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{p} N(0, V_r),$$

where $V_r = (G_r'\Sigma_r^{-1}G_r)^{-1}$.

Remarks.

1. The important case to consider is the full set of instruments in GEL. In that case, by using the same set of assumptions we get the following through Theorem 3.2 of Newey and Smith (2004) (also note that in our independent data case, this can be seen from p. 677 of Guggenberger and Smith (2005)). The limit in Guggenberger and Smith (2005) will simplify in the independent case due to $Eg_i(\beta_0) = 0$, and $\Omega(\beta_0) = lim_{n\to\infty} En^{-1} \sum_{i=1}^{n} g_i(\beta_0)g_i(\beta_0)' = \Delta(\beta_0) = \lim_{n\to\infty} E[n^{1/2}(\hat{g})(\beta_0) - E\hat{g}(\beta_0)][n^{1/2}(\hat{g})(\beta_0) - E\hat{g}(\beta_0)]'$. There $\hat{g}(\beta_0) = n^{-1} \sum_{i=1}^{n} g_i(\beta_0)$, and the equality between $\Omega(\beta_0) = \Delta(\beta_0)$ can be seen from $E\hat{g}(\beta_0) = 0$ and independence of moments across for $i = 1, \ldots, n$. The limit for full set of instruments GEL estimator, $\tilde{\beta}$, is

$$n^{1/2}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

where

$$V = (G_F'\Sigma^{-1}G_F)^{-1}.$$

2. One of the important issues is the comparison of the variance terms in both cases. Simple matrix algebra through partitioned inverse (Exercise 5.16a in Abadir and Magnus (2005)) shows

$$G_F'\Sigma^{-1}G_F = G_r'\Sigma_r^{-1}G_r + G_F'\bar{\Sigma}G_F. \tag{8}$$

Given Σ_r is positive definite, and Σ_{rw} is of full rank, by Exercise 8.4.44 of Abadir and Magnus (2005) $\bar{\Sigma}$ is positive semi definite, and is described in detail below. So by (8)

$$(G_r'\Sigma_r^{-1}G_r)^{-1} - (G_F'\Sigma^{-1}G_F)^{-1} \geq 0.$$

This last result clearly shows that as long as the instruments are valid, the model selection by the adaptive lasso does not improve upon the variance of GEL estimator. The intuition is clear here, each orthogonality restriction provides exogenous source of variation,

and hence is useful in reducing variance as long as they are valid. But this is not such a good sign for GEL which uses all instruments, because this may become problematic in testing structural parameters, the low standard errors can cause size distortion. This point is also made in p. 687 of Newey and Windmeijer (2009). Also an important issue is the finite sample bias, and we will discuss this in next remark. For completeness we provide

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_r^{-1} + \Sigma_r^{-1}\Sigma_{rw}\Xi^{-1}\Sigma_{rw}'\Sigma_r^{-1} & -\Sigma_r^{-1}\Sigma_{rw}\Xi^{-1} \\ -\Xi^{-1}\Sigma_{rw}'\Sigma_r^{-1} & \Xi^{-1} \end{bmatrix}.$$

$\Xi = \Sigma_w - \Sigma_{rw}'\Sigma_r^{-1}\Sigma_{rw}$ and

$$\bar{\Sigma} = \begin{bmatrix} \Sigma_r^{-1}\Sigma_{rw}\Xi^{-1}\Sigma_{rw}'\Sigma_r^{-1} & -\Sigma_r^{-1}\Sigma_{rw}\Xi^{-1} \\ -\Xi^{-1}\Sigma_{rw}'\Sigma_r^{-1} & \Xi^{-1} \end{bmatrix}.$$

3. Note that in the simulations we take the estimated number of instruments from the first stage and then use them in two step GMM and GEL in the second stage regression.

4. 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 give us good finite sample results in simulations, so we do not report them to save space.

5. One important point to remember is the finite sample bias of having more than necessary number of instruments. So if we do not perform model selection in the first stage and put all instruments in GEL, the bias is calculated to be $q/2n$ on p. 694 of Newey and Windmeijer (2009). This gives an advantage to adaptive lasso based model selection. We also use other model selection procedures and show that adaptive lasso generally fares better than them in simulations.

6. An important case is when the relevant and irrelevant instruments have zero asymptotic correlation. Then $G_w = 0$, so $G_F = (G_r', 0')'$. With conditional homoskedasticity, $\Sigma_{rw} = 0$. This means that the variance of hybrid GEL and GEL using all instruments is the same in Remark 2.

4. Asymptotic bias and selection inconsistency of lasso

This section will analyze certain lasso estimators that are used in the literature. The first one is the regular lasso, 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 the details. The second one is the heteroskedasticity consistent lasso type estimator of Belloni et al. (2012). Their estimator is a big leap in the literature. This estimator can choose optimal instruments in large number of instruments case, and has the oracle property for the instrumental variable estimation. It also works well with the 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 lasso), and this affects the 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 the 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.

$$y_i = X_i'\beta_0 + \epsilon_i,$$

where X_i is $p \times 1$ endogenous variable vector, and

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

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

Next the reduced form equation is, without losing generality set $p = 1$,

$$X_i = Z_i'\gamma_0 + v_i,$$

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

In Belloni et al. (2012), quite sensibly the number of relevant instruments are approximately "q_0" (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 $q \to \infty$), 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 is equal to q_0. In a simple applied work it is sometimes difficult to find valid instruments (see Acemoglu et al. (2001); Acemoglu and Johnson (2006)) so this is a reasonable assumption in our case.

Next we define the heteroskedasticity consistent lasso:

$$\hat{\gamma}_L = \underset{\gamma}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (X_i - Z_i'\gamma)^2 + \lambda \sum_{j=1}^{q} |\gamma_j \hat{\pi}_j| \right], \tag{9}$$

where Belloni et al. (2012) have a two step process in estimating γ.

First, they set, for each $j = 1, \ldots, q$,

$$\hat{\pi}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2 (X_i - \bar{X})^2},$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$. By using this in the lasso formula above they get Initial Lasso. Then after getting "Initial Lasso estimator $\hat{\gamma}_{Initial-Lasso}$", they setup the following refined loadings

$$\hat{\pi}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2 \hat{v}_i^2}, \quad j = 1, \ldots, q$$

where $\hat{v}_i = X_i - Z_i'\hat{\gamma}_{Initial-Lasso}$. After using the refined loadings in (9) objective function we get $\hat{\gamma}_L$. This is equation (2.4) in Belloni et al. (2012). These are 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 their sequence in this part of the paper as well. We will not analyze the 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 $\tilde{X}_i = X_i - \bar{E}X_i$, $1 \leq i \leq n$. We make the following assumptions for the following Lemmata:

Assumption B.1.

$$\text{(i)} \quad \max_{1\leq j\leq q} \left[\bar{E}(\tilde{X}_i)^2 + \bar{E}(Z_{ij}^2\tilde{X}_i) + \frac{1}{\bar{E}(Z_{ij}^2 v_i^2)} \right] = O_p(1).$$

(ii) $\max_{1\le j\le q} \bar{E}(Z_{ij}^3 v_i^3) = O_p(K_n)$.

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

(iv) $\max_{1\le j\le q} Z_{ij}^2 \dfrac{\log n}{n} = o_p(1)$

and

$$\max_{1\le j\le q} |\bar{E}_n(Z_{ij}^2 v_i^2) - \bar{E}(Z_{ij}^2 v_i^2)| + |\bar{E}_n(Z_{ij}^2 \tilde{X}_i^2) - \bar{E}(Z_{ij}^2 \tilde{X}_i^2)| = o_p(1).$$

(v) $\|Z_i'(\hat{\gamma}_L - \gamma_0)\|_{2,n} = O_p\left(\dfrac{(\log n)^{1/2}}{n^{1/2}}\right)$.

Assumption B.2. (i) $\hat{\gamma}_L$ is consistent.
(ii) $\lambda/n^{1/2} \to \lambda_0 \ge 0$.

Note that Assumption B.1(i)–(iv) are Assumption RF(i)–(iv) in Belloni et al. (2012). Assumption B.1(v) is Theorem 1 of Belloni et al. (2012), we use this theorem to shorten the proofs. We briefly discuss Assumption B.2(ii). If this had been $\lambda/n^{1/2} \to 0$ as in adaptive lasso, we see that right hand side of (44) would have been zero. So there will be no model selection issues for the nonzero parameters. But also the right hand side of (43) will be zero, so the zeros will be selected as nonzeros wrongly. In other words changing the assumption leads to no model selection. Given (13) (14), same results will be obtained for Lasso of Belloni et al. (2012). Then another point is the comparison of this assumption with the one in Belloni et al. (2012). Since we deal with fixed number of instruments, Belloni et al. (2012), assumption is $\lambda = O(n^{-1/2})$. But ours is different, and the reason is that we did not scale the objective function by dividing with n. If we do that then our tuning parameter $\lambda/n = O(n^{-1/2})$ has the same rate as well. So this is the same assumption as theirs if we had scaled the objective function. We did not scale it here to get the limits with the methodology of Knight and Fu (2000).

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*

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

where

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

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

This result will provide the selection inconsistency of the variables 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).

Note that $\mathcal{A}_{Ln} = \{j : \hat{\gamma}_j \neq 0\}$, and $\mathcal{A}_L = \{j : \gamma_{j0} \neq 0\}$.

Lemma 2. *Under Assumptions B.1–B.2,*

$$\limsup_n P(\mathcal{A}_{Ln} = \mathcal{A}_L) \le c < 1,$$

where c is a constant depending on the true model.

Our result extends Proposition 1 of Zou (2006) from the regular lasso to the lasso that is used by Belloni et al. (2012). We show that usage of this type of lasso estimators in fixed instruments context may lead to inconsistent instrument selection, which may affect the second stage regressions as discussed above. We show that the adaptive lasso is immune to the instrument inconsistency problem. Bühlmann 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.

5. Algorithm and the choice of the tuning parameter

Adaptive lasso estimates can be computed efficiently by a modification of LARS (LAR: 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 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 $i = 1, \ldots, n$

$$X_i = Z_i'\gamma + \nu_i, \tag{10}$$

where X_i is the univariate endogenous variable, $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ is the associated q-dimensional instruments, $\gamma = (\gamma_1, \ldots, \gamma_q)'$ is the coefficient vector and ν_i is the random error with mean 0 and variance σ_ν^2. For the multivariate model we vectorize X, Z, ν, etc., the algorithm works as in the univariate case. Assume that we fit q predictors and the true model has q_0 variables ($1 \le q_0 \le q$).

Adaptive Lasso Algorithm

1. Create new covariates $Z_j^* = Z_j/\hat{w}_j, j = 1, 2, \ldots, q$, where $\hat{w}_j$ is the adaptive weight as defined in Section 2. Note that both Z_j, Z_j^* are of dimension $n \times 1$, for each j.
2. Solve the LASSO via LARS algorithm for given λ.
$$\tilde{\gamma} = \text{argmin}_\gamma \left\| X - \sum_{j=1}^q Z_j^{*'}\gamma_j \right\|^2 + \lambda \sum_{j=1}^q |\gamma_j| \tag{11}$$
where $X = (X_1, \ldots, X_n)'$.
3. Output is $\hat{\gamma}_j = \tilde{\gamma}_j/\hat{w}_j, j = 1, 2, \ldots, q$. This is the adaptive lasso estimate.
4. Then we substitute $\hat{\gamma}_j, j = 1, 2, \ldots, q$, adaptive lasso estimate from Step 3 in Eq. (12). This provides us a BIC value for a given λ. Note that tuning parameter optimization is explained in detail 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 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 the 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

$$BIC_\lambda = \hat{\sigma}_\lambda^2 + DF_\lambda \log(n)/n, \tag{12}$$

where $\hat{\sigma}_\lambda^2 = n^{-1}\left\|X - Z\hat{\gamma}\right\|^2$, DF_λ is the number of nonzero coefficients in $\hat{\gamma}$ which is described in Step 3 of adaptive lasso Algorithm. Z is $n \times q$ matrix where the vector form is described in (10). The reason we use this method is that BIC method provides selection consistency when the sample size n is approaching ∞. But other methods such as AIC or GCV (generalized cross validation) are not selection consistent (see Wang et al., 2009).

6. Oracle inequality

In this section we extend an important result due to Zou (2006). Zou (2006) provides the proof that the adaptive lasso achieves near minimax risk in iid standard normal data. Here we show that this can be extended to non iid-Gaussian, and 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 the adaptive lasso in the first step helps in selecting instruments, and then 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 the adaptive lasso estimator $\hat{\mu}_i$. The risk is defined as in Zou (2006)

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

The ideal risk is defined in Eq. (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 Eq. (5) of Zou (2006) as

$$\hat{\mu}_i = \underset{u}{\operatorname{argmin}}\left[\frac{1}{2}(X_i - u)^2 + \frac{\lambda_i}{|X_i|^\tau}|u|\right], \tag{13}$$

for $i = 1, 2 \cdots, n$, and $\lambda_i = (\sqrt{2\sigma_i^2 \log n})^{1+\tau}$. 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 $|X_i|^{-\tau}$. So oracle inequality here in Theorem 3 extends Zou (2006) from iid Gaussian case to heteroskedastic-Gaussian data.

Minimization of (13) provides us the following as in Zou (2006)

$$\hat{\mu}_i = \left[|X_i| - \frac{\lambda_i}{|X_i|^\tau}\right]_+ sgn(X_i), \tag{14}$$

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

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

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, the adaptive lasso still attains the near minimax risk as shown in Zou (2006) in iid-Gaussian case. Theorem 3 here gives us a basic oracle inequality, extending Zou (2006). In Theorem 3 of Zou (2006), Gaussian-iid case, he finds

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

Compared to Zou (2006), because of non iid Gaussian nature here, our constants b, c, d depend on σ_i^2. Zou (2006) takes $\sigma_i^2 = 1$. Theorem 3 is a new result, and the proof technique is not the same as in Zou (2006).

7. Simulation

In this section we want to answer four questions. The first one is whether our test can do better in selecting the irrelevant instruments compared with a sequential t-test. The second question is to compare the adaptive lasso with a technique that uses all available instruments. The third question is whether the adaptive lasso selection of instruments will deliver better second stage finite sample results than several competitors. The fourth question concerns the bi-modality of shrinkage estimators in least squares context that is raised by Leeb and Pötscher (2005). We want to see whether this bi-modality in the first stage can affect the second stage structural coefficients.

7.1. 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) estimators, specifically, the continuous updating (CUE), exponential tilting (ET) and empirical likelihood (EL). We also analyze TSLS for the second stage regression (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). To compare and contrast with other estimators, we include F-TSLS, F-GMM, F-CUE. These are the estimators when we use all available instruments in TSLS, GMM, CUE respectively. So there is no model selection involved in the first stage. We do not report Full Empirical Likelihood and Full Exponential Tilting since F-CUE did better than them in the finite samples. Next, to see a benchmark case, we use only strong instruments in the second stage. These are called Oracle estimators, O-TSLS, O-CUE for TSLS and CUE respectively. To understand the effects of weak and irrelevant instruments we also consider Weak TSLS (W-TSLS), Weak CUE (W-CUE) as well. W-TSLS, W-CUE only use Z_3 below, which is either a weak or irrelevant instrument depending on the model and then we run GEL with only one instrument Z_3.

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 Pötscher (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

$$\hat{\beta} = (X'P^K X)^{-1} X'P^K Y \tag{15}$$

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 $\hat{u}^K = (I - P^K)X$, $\hat{u}^K_\lambda = \hat{u}^K\tilde{H}\hat{\lambda}$. So Mallow's criteria are $\hat{R}^m_\lambda(K) = \frac{\hat{u}^{K\prime}_\lambda\hat{u}^K_\lambda}{n} + \hat{\sigma}^2_\lambda(2K/n)$. Finally, the approximate MSE of the 2SLS estimator is $\hat{S}_\lambda(K) = \hat{\sigma}^2_{\lambda\epsilon}\frac{K^2}{n} + \hat{\sigma}^2_\epsilon\left(\hat{R}^m_\lambda(K) - \sigma^2_\lambda\frac{K}{n}\right)$.

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, where Z_m be the matrix $(Z_{m,1}, \ldots, Z_{m,N})$ and let $P_m = Z_m(Z'_mZ_m)^{-1}Z'_m$. 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, there is the Post Lasso estimator by Belloni et al. (2012) which estimates the optimal instruments set. The Post Lasso is running OLS with heteroskedasticity consistent lasso selected variables. This type of lasso estimator is shown in Section 4.

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

$$\hat{\beta} = \left(X'PX - \sum_{i=1}^n P_{ii}X_iX'_i - \hat{\alpha}X'X\right)^{-1} \times \left(X'Py - \sum_{i=1}^n P_{ii}X_iy_i - \hat{\alpha}X'y\right). \tag{16}$$

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

7.1.1. Simulation results for conditional homoskedasticity

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

$$y_i = X_i\beta_0 + \epsilon_i \tag{17}$$

$$X_i = \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \gamma_3 Z_{3i} + \nu_i \tag{18}$$

where $i = 1, 2, \ldots, n$. The true $\beta_0 = 1$. Assume the IV matrix $Z = [Z_1, Z_2, Z_3]$ has full rank and satisfies

$$Z'Z/n \xrightarrow{p} \Sigma_{zz} = \begin{bmatrix} \sigma^2_{\gamma_1} & \sigma_{\gamma_1\gamma_2} & \sigma_{\gamma_1\gamma_3} \\ \sigma_{\gamma_1\gamma_2} & \sigma^2_{\gamma_2} & \sigma_{\gamma_2\gamma_3} \\ \sigma_{\gamma_3\gamma_1} & \sigma_{\gamma_3\gamma_2} & \sigma^2_{\gamma_3} \end{bmatrix}$$

as $n \to \infty$. We further assume $\sigma^2_{\gamma_1} = \sigma^2_{\gamma_2} = \sigma^2_{\gamma_3} = 1$ and the correlation between z_1 and z_2 is $\rho_1 = \sigma_{\gamma_1\gamma_2}/(\sigma_{\gamma_1}\sigma_{\gamma_2})$. The errors $[\epsilon_i, \nu_i]'$ $(i = 1, 2, \ldots, n)$ are assumed to be i.i.d. $N(0, \Omega)$, where 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, 0.5 and 0.99. However, we do not report the results with $\rho_2 = 0.99$ since this is similar to $\rho_2 = 0.5$ case. The variances of each error are normalized at four. Each model is replicated 500 times. Note that all Tables run with the same correlation structure which is described above. Now we use four setup of γ's:

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

Model 2: two nonzero (strong) and one local to zero (weak) coefficients $\gamma = (1, 1, \frac{t}{\sqrt{n}})'$, where t is a scalar, 2.5, 3.54 (for sample size $n = 100$, 200 respectively), so we have $\gamma_3/\sigma_{\gamma_3} = 0.25$ as in Leeb and Pötscher (2005).

In the reduced form equations, in Models 1–2, there are three instruments, two of which are strong. Thus the selection procedure would have to select either one, two or three instruments. Also note that the correlation between Z_1, Z_2: ρ_1 is 0.7 for Models 1–2, the same correlation of 0.7 is also assumed between the first and third instruments, as well as between the second and third instruments. We show in the following tables the model selection results and see how it affects the finite sample properties of the hybrid GEL estimator. We also use a sequential t-test under 5% significance level. The reduced form equation model settings correspond to the potential bi-modal density of LS estimator in Fig. 2 of Leeb and Pötscher (2005). We analyze the critique of Leeb and Pötscher (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 Pötscher (2005).

Now we introduce Models 3–4. These are intended to understand the finite sample properties of these estimators when the full model is slightly far away from the true one. Also we want to see the effects of weak instruments compared with Model 2, where we can consider Model 2 setup as nearly-weak.

Model 3: one nonzero (strong) and two exact zero (irrelevant) coefficients $\gamma = (1, 0, 0)'$.

Model 4: one nonzero (strong) and one local to zero (weak) coefficients, and one irrelevant instrument $\gamma = (1, \frac{1}{\sqrt{n}}, 0)'$.

For Models 3–4, the correlation between Z_1, Z_2: $\rho_1 = 0$. Also we impose zero correlation between first and third, and second and third instruments. Before discussing the results, denote the method of Donald and Newey (2001) by DN, and the shrinkage method of Belloni et al. (2012) by PL, and Kuersteiner and Okui (2010) by MA. First we show detailed results for model selection in Tables 1–2. Tables 1a–1b analyze the cases for the Models 1 and 2. We see that the adaptive lasso does very well, and DN, PL can overshoot the true model. In Table 1a, with $n = 100$, the adaptive lasso picks the true model 70%–77% of time whereas DN, PL pick the true model 47%–49%, 47%–71% of the time respectively. Sequential t also does a good job, but very rarely picks the only weak instrument. In Tables 2a–2b, we see that PL is very good and adaptive lasso and sequential t test came very close second in selecting the correct model.

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, Z_2 being selected but not Z_3 (Model Selection %) in Models 1–2, and selecting Z_1 in models 3–4. The detailed model selection results are in Tables 1–2. We also show in Figs. 1 and 2 the finite sample densities of the hybrid estimators $\hat{\beta}$ with $n = 100$, $n = 200$ and with Model 2. This corresponds to Leeb and Pötscher (2005) model.

Table 1a
Model selection results: $n = 100$, $\rho_2 = 0.5$, $\rho_1 = 0.7$.

		(Z_1, Z_2, Z_3)	(Z_1, Z_2)	(Z_1, Z_3)	(Z_2, Z_3)	(Z_1)	(Z_2)	(Z_3)	None
Model 1	AL	0.082	0.770	0.004	0.020	0.082	0.042	0.000	0.000
	DN	0.480	0.492	0.014	0.012	0.002	0.000	0.000	0.000
	PL	0.244	0.712	0.012	0.006	0.020	0.006	0.000	0.000
	ST	0.052	0.702	0.008	0.006	0.124	0.102	0.002	0.004
Model 2	AL	0.178	0.696	0.022	0.030	0.052	0.022	0.000	0.000
	DN	0.470	0.470	0.028	0.032	0.000	0.000	0.000	0.000
	PL	0.494	0.474	0.024	0.008	0.000	0.000	0.000	0.000
	ST	0.074	0.680	0.030	0.038	0.102	0.070	0.004	0.002

AL is the adaptive lasso method. DN is the method that is proposed by Donald and Newey (2001). PL is the post lasso estimator proposed by Belloni et al. (2012). ST is the sequential t-test under 5% level. Each column shows the fraction of that specific choice given the total number of iterations. The true model is (Z_1, Z_2) column, and for example AL picks true model 77% in Model 1.

Table 1b
Model selection results: $n = 200$, $\rho_2 = 0.5$, $\rho_1 = 0.7$.

		(Z_1, Z_2, Z_3)	(Z_1, Z_2)	(Z_1, Z_3)	(Z_2, Z_3)	(Z_1)	(Z_2)	(Z_3)	None
Model 1	AL	0.050	0.936	0.002	0.000	0.008	0.004	0.000	0.000
	DN	0.398	0.600	0.000	0.002	0.000	0.000	0.000	0.000
	PL	0.252	0.746	0.000	0.000	0.002	0.000	0.000	0.000
	ST	0.052	0.934	0.002	0.000	0.008	0.004	0.000	0.000
Model 2	AL	0.220	0.770	0.004	0.002	0.002	0.002	0.000	0.000
	DN	0.470	0.526	0.002	0.002	0.000	0.000	0.000	0.000
	PL	0.672	0.328	0.000	0.000	0.000	0.000	0.000	0.000
	ST	0.220	0.766	0.004	0.002	0.006	0.002	0.002	0.002

AL is the adaptive lasso method. DN is the method that is proposed by Donald and Newey (2001). PL is the post lasso estimator proposed by Belloni et al. (2012). ST is the sequential t-test under 5% level. Each column shows the fraction of that specific choice given the total number of iterations. The true model is (Z_1, Z_2) column, and for example AL picks true model 93.6% in Model 1.

Table 2a
Model selection results: $n = 100$, $\rho_2 = 0.5$, $\rho_1 = 0$.

		(Z_1, Z_2, Z_3)	(Z_1, Z_2)	(Z_1, Z_3)	(Z_2, Z_3)	(Z_1)	(Z_2)	(Z_3)	None
Model 3	AL	0.004	0.038	0.040	0.000	0.912	0.000	0.000	0.000
	DN	0.430	0.252	0.226	0.000	0.092	0.000	0.000	0.000
	PL	0.000	0.006	0.002	0.000	0.950	0.000	0.000	0.000
	ST	0.004	0.040	0.042	0.000	0.910	0.000	0.000	0.000
Model 4	AL	0.006	0.070	0.036	0.000	0.882	0.000	0.000	0.000
	DN	0.466	0.260	0.198	0.000	0.076	0.000	0.000	0.000
	PL	0.000	0.018	0.002	0.000	0.938	0.000	0.000	0.042
	ST	0.002	0.070	0.044	0.000	0.880	0.002	0.000	0.002

AL is the adaptive lasso method. DN is the method that is proposed by Donald and Newey (2001). PL is the post lasso estimator proposed by Belloni et al. (2012). ST is the sequential t-test under 5% level. Each column shows the fraction of that specific choice given the total number of iterations. The true model is (Z_1) column, and for example AL picks true model 91% in Model 3.

Table 2b
Model selection results: $n = 200$, $\rho_2 = 0.5$, $\rho_1 = 0$.

		(Z_1, Z_2, Z_3)	(Z_1, Z_2)	(Z_1, Z_3)	(Z_2, Z_3)	(Z_1)	(Z_2)	(Z_3)	None
Model 3	AL	0.006	0.032	0.034	0.000	0.928	0.000	0.000	0.000
	DN	0.404	0.246	0.262	0.000	0.088	0.000	0.000	0.000
	PL	0.000	0.006	0.006	0.000	0.988	0.000	0.000	0.000
	ST	0.006	0.056	0.048	0.000	0.890	0.004	0.000	0.000
Model 4	AL	0.008	0.060	0.030	0.000	0.902	0.000	0.000	0.000
	DN	0.412	0.272	0.244	0.000	0.072	0.000	0.000	0.000
	PL	0.000	0.008	0.006	0.000	0.986	0.000	0.000	0.000
	ST	0.008	0.106	0.046	0.000	0.840	0.000	0.000	0.000

AL is the adaptive lasso method. DN is the method that is proposed by Donald and Newey (2001). PL is the post lasso estimator proposed by Belloni et al. (2012). ST is the sequential t-test under 5% level. Each column shows the fraction of that specific choice given the total number of iterations. The true model is (Z_1) column, and for example AL picks true model 92.8% in Model 4.

First we consider the bias. For Models 1–2, we consider Tables 3–4. We see clearly that the biases of H-CUE, H-EL, H-ET, PL, F-CUE, LIML are small and close to each other, the biases of DN, MA, Fuller methods are higher. For example, from Table 3, Model 1, we see that the PL has the lowest bias at −0.001, H-CUE has a bias of 0.006 and DN has a bias of 0.012. In Tables 7–8, we analyze Models 3–4. This is the model that has only one strong instrument among three and there is no correlation between the weak and strong instruments. There we see that H-CUE and LIML have the lowest bias. Specifically, H-CUE and LIML, in Model 3, Table 7, have the bias of 0.034, DN and PL have the bias of 0.062, 0.052 respectively. We also see that among the models with all instruments, F-CUE does better in bias compared with F-TSLS. In Model 3, Table 7, the bias of F-CUE is 0.037 whereas F-TSLS has the bias of 0.071. MA also does not do well in terms of bias, in Table 7, Model 3, model averaging estimator has bias of 0.071. As expected with only an irrelevant instrument in the second stage, the bias of W-TSLS is 0.486 in Table 7, Model 3, and it is 0.480 in Table 7, Model 4.

Second we consider MSE. All the estimators have similar MSE in Tables 3–4 and 8. However, in Table 7, for Model 3, we see that

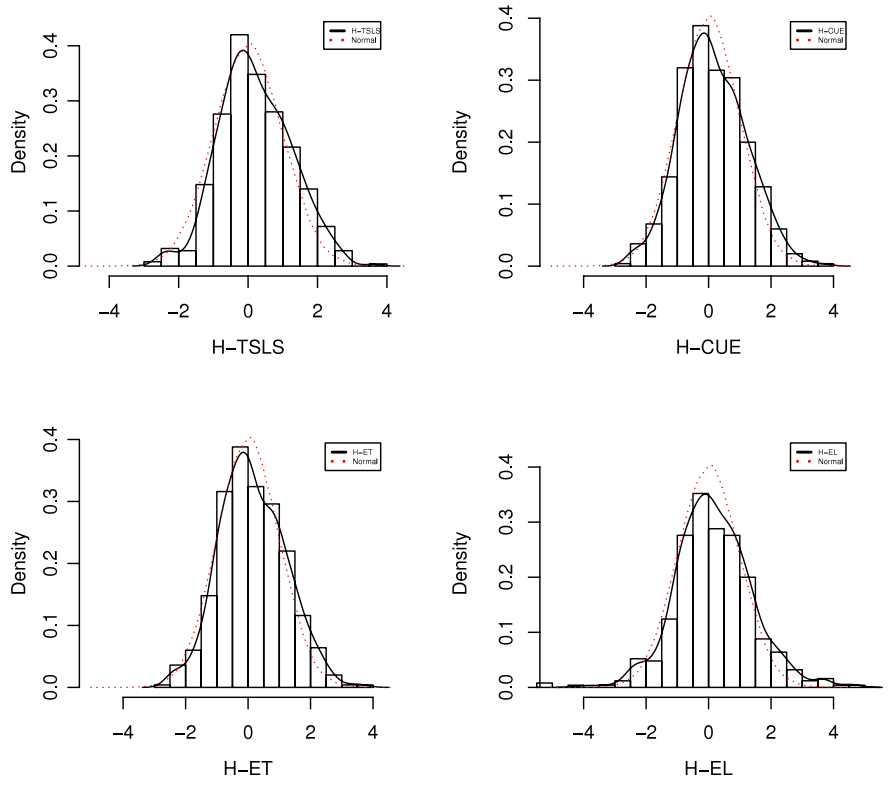

Fig. 1. Finite sample densities of hybrid estimators: Model 2, $n = 100$, $\gamma_3 = 2.5/\sqrt{n}$, $\rho_2 = 0.5$, $\rho_1 = 0.7$, $\sigma_\epsilon = 2$.

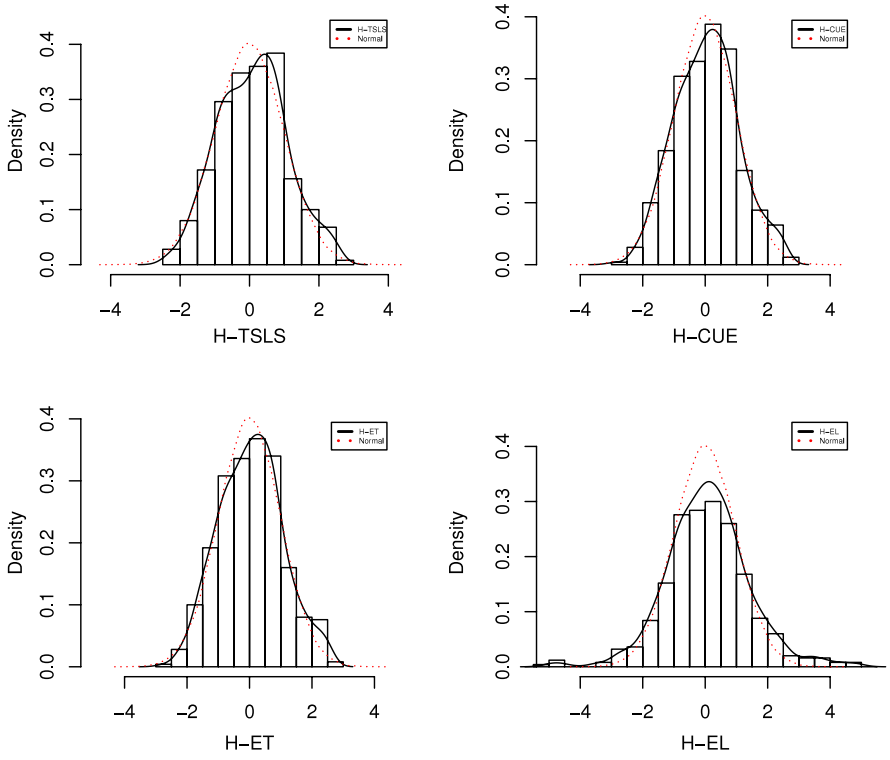

Fig. 2. Finite sample densities of hybrid estimators: Model 2, $n = 200$, $\gamma_3 = 3.54/\sqrt{n}$, $\rho_2 = 0.5$, $\rho_1 = 0.7$, $\sigma_\epsilon = 2$.

H-CUE, and F-CUE has the best MSE with 0.040, compared to 0.044 of PL, and 0.046 of DN. Again in Table 7, Model 4, we see that F-CUE, H-CUE have the minimum MSE of 0.039, 0.040 respectively, whereas the PL has MSE of 0.043, and DN has MSE of 0.041. In Table 7, Model 3 we see that LIML, Fuller have large MSE of 0.056, 0.048 respectively. We also observe that F-TSLS has higher MSE of 0.046 compared to 0.040 of F-CUE in Table 7, Model 3. Weak instruments setup in Table 7 has very high MSE compared with Table 3 since in Table 3, the instrument is not as weak as in Table 7.

When we look at the coverage rates for a 95% confidence interval, we see that all methods are slightly under the nominal rate generally. But H-TSLS comes closer to 95% among other

Table 3
Second stage results: $n = 100$, $\rho_2 = 0.5$, $\rho_1 = 0.7$.

		O-TSLS	W-TSLS	F-TSLS	F-CUE	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 1	Bias	0.009	−0.007	0.012	−0.003	0.014	0.006	0.006	0.007	0.014	0.012	−0.001	0.004	0.009
	MAD	0.112	0.152	0.108	0.114	0.108	0.114	0.114	0.115	0.108	0.108	0.110	0.110	0.111
	95% Coverage Rate	0.932	0.966	0.930	0.900	0.928	0.904	0.912	0.918	0.928	0.926	0.926	0.928	0.926
	MSE	0.012	0.021	0.012	0.010	0.012	0.011	0.011	0.011	0.012	0.011	0.012	0.012	0.012
	Model Selection %	–	–	–	–	0.770	0.770	0.770	0.770	0.492	–	0.712	–	–
Model 2	Bias	0.008	−0.006	0.010	−0.002	0.011	0.004	0.005	0.005	0.012	0.010	0.000	0.006	0.008
	MAD	0.101	0.128	0.100	0.103	0.100	0.103	0.104	0.105	0.099	0.100	0.101	0.102	0.103
	95% Coverage Rate	0.932	0.962	0.928	0.904	0.928	0.908	0.920	0.924	0.928	0.924	0.932	0.926	0.928
	MSE	0.010	0.015	0.011	0.009	0.010	0.009	0.009	0.009	0.010	0.009	0.010	0.010	0.010
	Model Selection %	–	–	–	–	0.696	0.696	0.696	0.696	0.470	–	0.474	–	

O-TSLS is the TSLS estimator based on the strong instruments only. W-TSLS is the TSLS estimator based on Z_3, which is weak in Model 2, irrelevant instrument in Models 1, 3,4. F-TSLS is the TSLS estimator that we use all available instruments in TSLS, F-CUE is the CUE estimator with all 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.

Table 4
Second stage results: $n = 200$, $\rho_2 = 0.5$, $\rho_1 = 0.7$.

		O-TSLS	W-TSLS	F-TSLS	F-CUE	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 1	Bias	0.005	0.002	0.010	0.002	0.008	0.004	0.004	0.003	0.010	0.010	0.004	0.006	0.010
	MAD	0.078	0.110	0.079	0.077	0.079	0.078	0.078	0.078	0.080	0.079	0.077	0.077	0.080
	95% Coverage Rate	0.942	0.954	0.942	0.934	0.942	0.928	0.928	0.930	0.940	0.942	0.940	0.940	0.942
	MSE	0.006	0.010	0.006	0.005	0.006	0.005	0.006	0.006	0.006	0.006	0.006	0.006	0.006
	Model Selection %	–	–	–	–	0.936	0.936	0.936	0.936	0.600	–	0.746	–	–
Model 2	Bias	0.004	0.001	0.009	0.005	0.009	0.005	0.004	0.004	0.009	0.009	0.003	0.005	0.009
	MAD	0.070	0.093	0.072	0.072	0.073	0.070	0.071	0.071	0.072	0.072	0.070	0.070	0.071
	95% Coverage Rate	0.942	0.952	0.940	0.930	0.940	0.930	0.930	0.930	0.940	0.938	0.942	0.940	0.940
	MSE	0.005	0.007	0.005	0.004	0.005	0.004	0.005	0.005	0.005	0.005	0.005	0.005	0.005
	Model Selection %	–	–	–	–	0.770	0.770	0.770	0.770	0.526	–	0.328	–	–

O-TSLS is the TSLS estimator based on the strong instruments only. W-TSLS is the TSLS estimator based on Z_3, which is weak in Model 2, irrelevant instrument in Models 1, 3,4. F-TSLS is the TSLS estimator that we use all available instruments in TSLS, F-CUE is the CUE estimator. 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.

Table 5
Second stage with heteroskedasticity results: $n = 100$, $\rho_2 = 0.5$, $\rho_1 = 0.7$.

		O-CUE	W-CUE	F-GMM	F-CUE	H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	H-FULL
Model 1	Bias	−0.006	−0.021	−0.008	0.010	0.020	0.018	0.013	0.012	0.059	0.061	0.185	0.029
	MAD	0.333	0.431	0.316	0.308	0.334	0.335	0.323	0.325	0.277	0.272	0.363	0.280
	95% Coverage Rate	0.838	0.878	0.886	0.870	0.866	0.836	0.850	0.858	0.800	0.922	0.568	0.922
	MSE	0.096	0.308	0.108	0.088	0.105	0.084	0.086	0.095	0.045	0.092	0.056	0.107
	Model Selection %	–	–	–	–	0.352	0.352	0.352	0.352	0.252	–	0.242	–
Model 2	Bias	−0.008	−0.013	−0.005	−0.007	0.009	0.004	0.004	0.003	0.029	0.034	0.106	0.018
	MAD	0.302	0.352	0.287	0.272	0.299	0.293	0.286	0.287	0.255	0.255	0.296	0.263
	95% Coverage Rate	0.836	0.890	0.890	0.866	0.886	0.852	0.880	0.880	0.792	0.924	0.664	0.918
	MSE	0.072	0.171	0.085	0.071	0.083	0.069	0.071	0.076	0.034	0.074	0.034	0.083
	Model Selection %	–	–	–	–	0.360	0.360	0.360	0.360	0.236	–	0.254	–

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. (2012). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the GMM estimator based on all instruments. FCUE is the CUE estimator that uses full set of instruments. O-CUE, W-CUE are the oracle (only strong instruments) and the only weak/irrelevant instrument Z_3 used in CUE respectively.

methods generally in Tables 3–4, 7–8. Mean absolute deviations are similar for all models.

From Figs. 1–2, we see that the bi-modality of the reduced form equation such as the ones shown in Fig. 2 of Leeb and Pötscher (2005) in the reduced form equations does not affect the empirical distribution of hybrid GEL estimators of the second stage in overidentified case. We only show Table 3-Model 2, and Table 4-Model 2, the other cases in homoskedastic setups display the same shape so to save space they are not included. The figures for heteroskedastic cases are not shown since the main idea is to show that structural parameter estimates in the overidentified framework are not affected by bi-modality of reduced form coefficients.

To summarize, in the homoskedastic case, H-CUE has very good bias, MSE combination among all estimators. Also we should note that F-CUE is a close competitor.

7.1.2. 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 5–6, 9–10. First we consider the bias. For Models 1–2, we consider Tables 5–6. We see clearly that the bias of F-GMM, F-CUE are small, the bias of DN, MA, H-Fuller methods have large bias. In Model 1,

Table 6
Second stage with heteroskedasticity results: $n = 200$, $\rho_2 = 0.5$, $\rho_1 = 0.7$.

		O-CUE	W-CUE	F-GMM	F-CUE	H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	H-FULL
Model 1	Bias	0.004	0.008	−0.002	−0.002	0.006	0.008	0.008	0.009	0.031	0.033	0.041	0.016
	MAD	0.211	0.281	0.206	0.202	0.206	0.209	0.212	0.211	0.203	0.207	0.214	0.202
	95% Coverage Rate	0.896	0.934	0.916	0.902	0.918	0.892	0.906	0.914	0.818	0.942	0.806	0.944
	MSE	0.039	0.097	0.041	0.037	0.040	0.037	0.038	0.040	0.019	0.043	0.019	0.046
	Model Selection %	–	–	–	–	0.602	0.602	0.602	0.602	0.364	–	0.540	–
Model 2	Bias	0.005	0.008	0.000	0.004	−0.001	0.005	0.006	0.008	0.026	0.026	0.029	0.019
	MAD	0.195	0.234	0.187	0.188	0.196	0.199	0.197	0.199	0.188	0.188	0.196	0.187
	95% Coverage Rate	0.904	0.930	0.918	0.902	0.918	0.892	0.910	0.916	0.802	0.946	0.804	0.944
	MSE	0.031	0.055	0.034	0.030	0.033	0.030	0.031	0.032	0.016	0.036	0.016	0.037
	Model Selection %	–	–	–	–	0.540	0.540	0.540	0.540	0.382	–	0.448	–

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. (2012). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the GMM estimator based on all instruments. FCUE is the CUE estimator that uses full set of instruments. O-CUE, W-CUE are the oracle (only strong instruments) and the only weak/irrelevant instrument Z_3 used in CUE respectively.

Table 7
Second stage results: $n = 100$, $\rho_2 = 0.5$, $\rho_1 = 0$.

		O-TSLS	W-TSLS	F-TSLS	F-CUE	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 3	Bias	0.032	0.486	0.071	0.037	0.041	0.034	0.036	0.035	0.062	0.071	0.052	0.034	0.052
	MAD	0.216	1.398	0.222	0.246	0.222	0.223	0.223	0.223	0.220	0.222	0.224	0.228	0.227
	95% Coverage Rate	0.956	0.990	0.932	0.896	0.944	0.910	0.918	0.922	0.936	0.924	0.914	0.956	0.950
	MSE	0.050	3.496	0.046	0.040	0.046	0.040	0.041	0.041	0.046	0.044	0.044	0.056	0.048
	Model Selection %	–	–	–	–	0.912	0.912	0.912	0.912	0.092	–	0.950	–	–
Model 4	Bias	0.032	0.361	0.069	0.038	0.048	0.037	0.037	0.037	0.058	0.069	0.056	0.028	0.049
	MAD	0.218	1.197	0.222	0.237	0.218	0.222	0.221	0.223	0.223	0.222	0.222	0.227	0.226
	95% Coverage Rate	0.956	0.990	0.930	0.894	0.938	0.906	0.914	0.916	0.932	0.924	0.912	0.958	0.948
	MSE	0.049	2.685	0.045	0.039	0.046	0.040	0.040	0.040	0.041	0.045	0.043	0.044	0.053
	Model Selection %	–	–	–	–	0.882	0.882	0.882	0.882	0.076	–	0.938	–	–

O-TSLS is the TSLS estimator based on the strong instruments only. W-TSLS is the TSLS estimator based on Z_3, which is weak in Model 2, irrelevant instrument in Models 1, 3,4. F-TSLS is the TSLS estimator that we use all available instruments in TSLS, F-CUE is the CUE estimator. 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.

Table 8
Second stage results: $n = 200$, $\rho_2 = 0.5$, $\rho_1 = 0$.

		O-TSLS	W-TSLS	F-TSLS	F-CUE	H-TSLS	H-CUE	H-ET	H-EL	DN	MA	PL	LIML	Fuller
Model 3	Bias	−0.001	0.476	0.021	−0.003	0.004	0.000	−0.001	−0.002	0.015	0.021	0.001	−0.003	0.011
	MAD	0.137	1.526	0.137	0.140	0.136	0.136	0.137	0.136	0.137	0.137	0.137	0.133	0.132
	95% Coverage Rate	0.954	0.990	0.938	0.918	0.950	0.930	0.932	0.936	0.940	0.926	0.952	0.942	0.940
	MSE	0.022	3.282	0.021	0.019	0.021	0.020	0.020	0.021	0.021	0.020	0.022	0.022	0.021
	Model Selection %	–	–	–	–	0.928	0.928	0.928	0.928	0.088	–	0.988	–	–
Model 4	Bias	−0.001	0.342	0.021	−0.001	0.004	0.000	0.000	0.002	0.017	0.021	0.001	0.002	0.012
	MAD	0.137	1.235	0.132	0.138	0.131	0.136	0.135	0.135	0.134	0.132	0.137	0.136	0.130
	95% Coverage Rate	0.954	0.982	0.936	0.918	0.950	0.932	0.932	0.934	0.940	0.926	0.952	0.938	0.934
	MSE	0.022	2.645	0.021	0.019	0.021	0.020	0.020	0.020	0.021	0.020	0.022	0.022	0.021
	Model Selection %	–	–	–	–	0.902	0.902	0.902	0.902	0.072	–	0.986	–	–

O-TSLS is the TSLS estimator based on the strong instruments only. W-TSLS is the TSLS estimator based on Z_3, which is weak in Model 2, irrelevant instrument in Models 1, 3,4. F-TSLS is the TSLS estimator that we use all available instruments in TSLS, F-CUE is the CUE estimator. 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.

in Table 5, we see that FCUE, F-GMM, H-EL have small median bias. H-EL, H-CUE have the lowest bias in Model 2 on Table 5 among all estimators. For example, from Table 5, Model 2, we see that the PL has the largest bias with 0.106, H-EL has the smallest bias of 0.003. In between, DN has a bias of 0.029, whereas F-GMM has a bias of −0.005 in the same setup. In Table 6, we see that with $n = 200$, the bias of F-GMM, F-CUE have the lowest bias closely followed by our hybrid methods. DN, MA, PL display large bias compared to other methods. In Tables 9–10, we analyze Models 3–4. This is the model that has only one strong instrument among three and there is no correlation between the weak and strong instruments. There we see that F-CUE and F-GMM have the lowest bias, followed by hybrid estimators. Specifically, in Table 9, F-GMM has the lowest bias with −0.021, and H-GMM has a bias of 0.114, and DN has 0.211 bias.

Second we consider MSE. In Models 1–2, Tables 5–6, clearly we see that DN is the best and PL comes very close. Specifically, in Table 7, Model 1, DN has MSE of 0.045, and PL has 0.056, wheres H-CUE has 0.084. Full models F-GMM, F-CUE have higher MSE compared to other estimators. In Models 3–4, Tables 9–10, DN generally has the lowest MSE, and PL, H-CUE are very close in MSE. In Table 9, Model 3, H-CUE has the lowest MSE of 0.177, DN has 0.179, and PL has MSE of 0.201. In Model 4, Table 9, DN has the best MSE of 0.173, followed by H-CUE (0.180) and PL (0.200). The other methods, except H-ET, H-EL, do fare well in MSE. (See Table 9.)

Table 9
Second stage with heteroskedasticity results: $n = 100$, $\rho_2 = 0.5$, $\rho_1 = 0$.

		O-CUE	W-CUE	F-GMM	F-CUE	H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	H-FULL
Model 3	Bias	−0.002	−1.047	−0.021	−0.043	0.114	0.135	0.139	0.140	0.211	0.206	0.430	0.156
	MAD	0.473	1.553	0.524	0.518	0.517	0.508	0.507	0.507	0.452	0.451	0.640	0.438
	95% Coverage Rate	0.894	0.944	0.866	0.842	0.794	0.748	0.754	0.764	0.802	0.876	0.238	0.902
	MSE	0.679	11.702	0.478	0.401	0.212	0.177	0.186	0.205	0.179	0.249	0.201	0.327
	Model Selection %	–	–	–	–	0.366	0.366	0.366	0.366	0.028	–	0.274	–
Model 4	Bias	0.001	−0.833	−0.022	−0.038	0.133	0.137	0.148	0.139	0.208	0.208	0.430	0.144
	MAD	0.472	1.483	0.530	0.524	0.516	0.506	0.509	0.509	0.449	0.452	0.639	0.429
	95% Coverage Rate	0.896	0.938	0.866	0.834	0.790	0.756	0.764	0.782	0.794	0.880	0.234	0.896
	MSE	0.692	9.905	0.450	0.370	0.222	0.180	0.184	0.205	0.173	0.240	0.200	0.318
	Model Selection %	–	–	–	–	0.360	0.360	0.360	0.360	0.030	–	0.270	–

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. (2012). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the GMM estimator based on all instruments. FCUE is the CUE estimator that uses full set of instruments. O-CUE, W-CUE are the oracle (only strong instruments) and the only weak/irrelevant instrument Z_3 used in CUE respectively.

Table 10
Second stage with heteroskedasticity results: $n = 200$, $\rho_2 = 0.5$, $\rho_1 = 0$.

		O-CUE	W-CUE	F-GMM	F-CUE	H-GMM	H-CUE	H-ET	H-EL	DN	MA	PL	H-FULL
Model 3	Bias	−0.018	−0.712	−0.027	−0.026	0.027	0.022	0.033	0.034	0.080	0.080	0.258	0.021
	MAD	0.302	1.476	0.325	0.323	0.323	0.319	0.312	0.325	0.288	0.292	0.466	0.296
	95% Coverage Rate	0.932	0.968	0.890	0.886	0.878	0.874	0.876	0.878	0.864	0.914	0.552	0.938
	MSE	0.146	8.545	0.136	0.126	0.104	0.098	0.101	0.104	0.074	0.113	0.092	0.164
	Model Selection %	–	–	–	–	0.452	0.452	0.452	0.452	0.028	–	0.624	–
Model 4	Bias	−0.017	−0.818	−0.025	−0.022	0.027	0.029	0.036	0.037	0.077	0.078	0.258	0.017
	MAD	0.302	1.483	0.323	0.323	0.311	0.309	0.313	0.320	0.299	0.299	0.467	0.295
	95% Coverage Rate	0.932	0.964	0.892	0.888	0.882	0.874	0.878	0.882	0.862	0.914	0.552	0.932
	MSE	0.145	9.831	0.136	0.124	0.101	0.097	0.099	0.101	0.073	0.112	0.092	0.165
	Model Selection %	–	–	–	–	0.436	0.436	0.436	0.436	0.026	–	0.624	–

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. (2012). HFUL is the heteroskedasticity robust Fuller's estimator by Hausman et al. (2012). FGMM is the GMM estimator based on all instruments. FCUE is the CUE estimator that uses full set of instruments. O-CUE, W-CUE are the oracle (only strong instruments) and the only weak/irrelevant instrument Z_3 used in CUE respectively.

When we look at the coverage rates for a 95% confidence interval, we see that all methods are under the nominal coverage rate generally. We can say that MA is the best among the methods in terms of coverage, and PL fares the worst. Mean absolute deviations are similar for all models except from PL, which has high mean absolute deviation. To summarize, in the heteroskedastic case, Hybrid methods with Full methods (no model selection) has very good bias, but in MSE methods of DN, H-CUE, PL do very well. We also see that except from Table 9, H-TSLS, H-CUE are very close to O-TSLS, O-CUE in bias and MSE in all Tables.

8. 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 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.

Acknowledgments

Authors thank co-editor Han Hong for advice and help. They also thank two anonymous referees and associate editor whose comments changed the paper substantially. Qingliang Fan's research was supported by National Natural Science Foundation of China (NNSFC) grant 71301134.

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:

$$L_n(\gamma_v) = [X_v - \tilde{Z}\gamma_v]'[X_v - \tilde{Z}\gamma_v] + \lambda_n \sum_{j=1}^{q}\sum_{k=1}^{p} \hat{w}_{jk}|\gamma_{jk}|. \tag{19}$$

Using (4) see that sum of squared errors part in that equation can be written as

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

First, by Assumption F.1

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

Then by Assumption F.2

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

Next via Assumption F.3

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

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

$$\frac{1}{n}(X_v - \tilde{Z}\gamma_v)'(X_v - \tilde{Z}\gamma_v) \xrightarrow{p} \sigma_v^2 + (\gamma_v - \gamma_v^0)'C(\gamma_v - \gamma_v^0). \tag{20}$$

Next we consider the penalty term in our objective function. First since $\tilde{\gamma}_{jk} = O_p(n^{-1/2})$, by Assumption F.4

$$\frac{\lambda_n}{n}\hat{w}_{jk} = \frac{\lambda}{n}\frac{1}{|\tilde{\gamma}_{jk}|^\tau} = \frac{\lambda}{n}\frac{n^{\tau/2}}{|n^{1/2}\tilde{\gamma}_{jk}|^\tau} \xrightarrow{p} 0.$$

So

$$\frac{\lambda_n}{n}\sum_{j=1}^{q}\sum_{k=1}^{p}\hat{w}_{jk}|\gamma_{jk}| \xrightarrow{p} 0. \tag{21}$$

Noting that $L_n(\gamma_v)$ is convex by (20) (21)

$$L_n(\gamma_v) \xrightarrow{p} \sigma_v^2 + (\gamma_v - \gamma_v^0)'C(\gamma_v - \gamma_v^0) = L(\gamma_v). \tag{22}$$

$$\hat{\gamma}_v = O_p(1), \tag{23}$$

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

$$\operatorname{argmin} L_n(\gamma_v) \xrightarrow{p} \operatorname{argmin} L(\gamma_v).$$

Given that C is full rank, unique minimum is at γ_v^0 for the limit term in (23). Consistency is proved and

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

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

$$\hat{\gamma}_v = \begin{bmatrix} \gamma_1^0 + \dfrac{\hat{u}_1}{\sqrt{n}} \\ \vdots \\ \gamma_q^0 + \dfrac{\hat{u}_q}{\sqrt{n}} \end{bmatrix}. \tag{24}$$

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

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

Note that

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

where

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

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

$$u = \begin{bmatrix} u_1 \\ \vdots \\ u_q \end{bmatrix} \tag{25}$$

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

$$\begin{aligned} V_n(u) &= \Psi_n(u) - \Psi_n(0) \\ &= u'\left(\frac{\tilde{Z}'\tilde{Z}}{n}\right)u - 2u'\left(\frac{\tilde{Z}'v_v}{\sqrt{n}}\right) \\ &\quad + \frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^{q}\sum_{k=1}^{p}\hat{w}_{jk}\sqrt{n}(|\gamma_{jk}^0 + u_{jk}/\sqrt{n}| - |\gamma_{jk}^0|). \end{aligned} \tag{26}$$

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

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

Next by Assumption F.5 (via Central Limit Theorem)

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

The limit for the penalty in (26) will be discussed next. Depending on γ_{jk}^0 there are two possibilities. First if $\gamma_{jk}^0 \neq 0$ ($j = 1, 2, \ldots, q_0, k = 1, 2, \ldots, p$) 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 F.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

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

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

$$\frac{\lambda_n}{n^{1/2}}\hat{w}_{jk} = \frac{\lambda_n}{n^{1/2}}n^{\tau/2}(n^{1/2}\tilde{\gamma}_{jk})^{-\tau} \xrightarrow{p} \infty, \tag{27}$$

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

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

Define $u_{\mathcal{A}}$ as the first pq_0 elements of u vector which is of dimension pq. Set C_{11} as 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 $\mathcal{A} = \{1, \ldots, q_0\}$ without losing any generality.)

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

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 $\mathcal{A}^c = \{q_0 + 1, \ldots, q\}$ 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

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

Also see that by Assumption F.4, for $k = 1, \ldots, p$, as in (27)

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

Rewrite left term of the first order condition above as

$$\frac{2\tilde{Z}_{j'}'[\nu_v - \tilde{Z}(\hat{\gamma}_v - \gamma_v^0)]}{n^{1/2}} = \frac{2\tilde{Z}_{j'}'\nu_v}{n^{1/2}} - \frac{2\tilde{Z}_{j'}'\tilde{Z}}{n} n^{1/2}(\hat{\gamma}_v - \gamma_v^0). \tag{28}$$

By the arguments in the proof of the asymptotic normality, Assumptions F.3, F.5, Theorem 1(i), (28) converges to a normal distribution, so

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

Proof of Theorem 2. Here we start with the scenario of estimated number of instruments being $\hat{q}_0$, instead of q_0, their true number. Also we assume here $q \geq \hat{q}_0 \geq q_0 \geq p$. Instead of putting q_0 instruments in the second stage, we use $\hat{q}_0$. But of course Theorem 1 shows us that $\hat{q}_0 \xrightarrow{p} q_0$. The analysis here looks at what happens when we fit instruments that is larger than or equal to number of endogenous variables, but these are not equal to true number (only in large samples, they are equal). Note that we will discuss the other two possibilities $q \geq q_0 > \hat{q}_0 \geq p$, and $q \geq q_0 \geq p > \hat{q}_0$ at the end of the proof. The first issue is consistency. This will be preserved since even though there may be a mistake in numbers of instruments, we still have $\hat{q}_0 \geq p$. To see this point clearly, we just need full column rank of our γ^0 matrix. This is maintained here, also we see this in p. 674 of Guggenberger and Smith (2005) which includes weak, irrelevant and strong instruments in GEL (their Π_B matrix is of full column rank). Note the key is to have more or equal number of strong instruments compared with endogenous variables to identify the parameters on them. Any parameter that has only weak instruments for identification will be inconsistent as in Theorem 2(i) of Guggenberger and Smith (2005). But any parameter which is identified through a combination of strong and irrelevant instruments (at least one strong instrument per parameter) is consistent, as also can be seen from p. 677 of Guggenberger and Smith (2005).

Just to summarize the main point in consistency,

$$\max_i \sup_\beta E\|g_{ie}(\beta)\|^\xi \leq \max_i \sup_\beta E\|g_i(\beta)\|^\xi < \infty,$$

where $\xi > 2$, and $g_{ie}(.)$ represents $\hat{q}_0$ number of orthogonality restrictions, and $g_i(.)$ represent the maximum fit of q of them. After that inequality, all the proofs in Lemmas A.7–A.9 will follow in Guggenberger and Smith (2005) in the case of strong and irrelevant instrument case. Then the consistency is achieved. Note that the same analysis can be achieved through Newey and Smith (2004), by extending their iid assumptions to independent case.

Next we will develop the asymptotic normality of our estimators. Before that we need three results that will show making a mistake in the number of instruments ($\hat{q}_0$ instead of q_0) in the first stage, (the mistake converges in probability to zero) will lead to the same limit that we can find if we had used the true number of instruments q_0.

Denote the extra estimated instruments in matrix form as Z_M, which is an $n \times (\hat{q}_0 - q_0)$ dimensional matrix, without losing any generality assume that $\hat{q}_0 \geq q_0$. See that Z_r was the matrix of strong instruments, which is an $n \times q_0$ matrix. We have the following result

$$\left\| \frac{Z_M' Z_r}{n} \right\| \leq \sqrt{\hat{q}_0 - q_0} \left\| \frac{Z' Z_r}{n} \right\|_\infty = o_p(1), \tag{29}$$

since by Theorem 1(ii), $\hat{q}_0 - q_0 \xrightarrow{p} 0$, and by Assumption F.3 $\|\frac{Z'Z_r}{n}\|_\infty = O_p(1)$.

Next we prove

$$\left\| \frac{\sum_{i=1}^n Z_{Mi} Z_{Mi}' \hat{e}_i^2}{n} \right\| \leq \sqrt{\hat{q}_0 - q_0} \left\| \frac{\sum_{i=1}^n Z_i Z_i' \hat{e}_i^2}{n} \right\|_\infty = o_p(1), \tag{30}$$

where we use Theorem 1(ii), $\hat{q}_0 - q_0 = o_p(1)$ and Assumption S.4, $\|\frac{\sum_{i=1}^n Z_i Z_i' \hat{e}_i^2}{n}\|_\infty = O_p(1)$, where $\hat{e}_i = y_i - X_i'\hat{\beta}$. Here any consistent estimator in the residual term will give the same asymptotic results in this proof, including $\tilde{\beta}$ (GEL estimator using all instruments from the first stage). Note that Z_{Mi} is $(\hat{q} - q_0) \times 1$ vector of the matrix Z_M, Z_i is the $q \times 1$ vector of the matrix Z.

Next

$$\left\| \frac{Z_M' \epsilon}{n^{1/2}} \right\| \leq \sqrt{\hat{q}_0 - q_0} \left\| \frac{Z' \epsilon}{n^{1/2}} \right\|_\infty = o_p(1), \tag{31}$$

since by Theorem 1(ii), $\hat{q}_0 - q_0 \xrightarrow{p} 0$, and by Assumption S.6 $\|\frac{Z'\epsilon}{n^{1/2}}\|_\infty = O_p(1)$. We will use the results (29)–(31) in the asymptotic normality proof below.

Next, see that by Assumptions F.2–F.3, since by definition $G_r = \Sigma_{zr}\gamma^0$

$$\begin{aligned} \frac{Z_r' X}{n} &= \frac{Z_r' Z_r}{n}\gamma^0 + \frac{Z_r' v}{n} \\ &\xrightarrow{p} \Sigma_{zr}\gamma^0 = G_r + o_p(1), \end{aligned}$$

where we use $Z_r' v/n \xrightarrow{p} 0$. By Assumption S.6, we have

$$\frac{Z_r' \epsilon}{n^{1/2}} \xrightarrow{d} N(0, \Sigma_r).$$

Note that $\hat{e}_i = y_i - X_i'\hat{\beta}$, since $\hat{\beta}$ is also consistent as $\tilde{\beta}$, Assumption S.4 can be used to have

$$\frac{1}{n}\sum_{i=1}^n Z_{ri} Z_{ri}' \hat{e}_i^2 \xrightarrow{p} \Sigma_r.$$

Using (29)–(31) with $\hat{Z} = [Z_m, Z_r]$, and the equations immediately above, with $Z_M' v/n \xrightarrow{p} 0$

$$\left\| \frac{\hat{Z}' X}{n} - G_r \right\| \xrightarrow{p} 0. \tag{32}$$

$$\left\| \frac{\hat{Z}' \epsilon}{n^{1/2}} - l_1 \right\| \xrightarrow{p} 0 \quad \text{where } l_1 \equiv N(0, \Sigma_r) \tag{33}$$

$$\left\| \frac{\sum_{i=1}^{n} \hat{Z}_i \hat{Z}_i' \hat{e}_i^2}{n} - \Sigma_r \right\| \xrightarrow{p} 0. \tag{34}$$

By Assumptions F.3 and S.1, following the first order condition expansions for GEL estimators, without losing any generality, either by using independent version of (A.7) in Newey and Smith (2004), or by excluding weak identification version in (A.15) of Guggenberger and Smith (2005),

$$0 = \begin{pmatrix} 0 \\ -\hat{g}(\beta_0) \end{pmatrix} + \bar{M} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\delta} \end{pmatrix}, \tag{35}$$

where $\hat{\delta}$ represents the estimate of the Lagrange multiplier in GEL, and $\hat{g}(\beta_0) = n^{-1} \sum_{i=1}^{n} g_i(\beta_0)$. Note that $g_{ie}(\beta_0) = \hat{Z}_i \epsilon_i$, which is of $\hat{q}_0 \times 1$ vector. Note that in our linear case, $\hat{g}_{ie} = \hat{Z}_i(y_i - X_i'\hat{\beta}) = \hat{Z}_i \hat{e}_i$, which is of $\hat{q}_0 \times 1$ dimension. Then $G_i(\hat{\beta}) = \hat{Z}_i X_i'$ which is of $\hat{q}_0 \times p$ dimension. For the mean value expansion below, $\bar{\beta} \in (\beta_0, \hat{\beta})$.

$$\bar{M} = \begin{pmatrix} 0 & \sum_{i=1}^{n} \rho_1(\hat{\delta}'\hat{g}_i) G_i(\hat{\beta})'/n \\ \sum_{i=1}^{n} \rho_1(\bar{\delta}'\hat{g}_i) G_i(\bar{\beta})/n & \sum_{i=1}^{n} \rho_2(\bar{\delta}'\hat{g}_i) g_i(\bar{\beta}) \hat{g}_i'/n \end{pmatrix}. \tag{36}$$

Note that since $\bar{\delta} = O_p(n^{-1/2})$, as well as $\hat{\delta} = O_p(n^{-1/2})$ as in Lemmas 7–8 in Guggenberger and Smith (2005), using Assumption S.3 and Markov's inequality

$$\max_i |\bar{\delta}'\hat{g}_i| \xrightarrow{p} 0.$$

So, by $\rho_1(0) = -1$, $\rho_2(0) = -1$ as in Guggenberger and Smith (2005) or Newey and Smith (2004)

$$\max_i |\rho_1(\bar{\delta}'\hat{g}_i) + 1| \xrightarrow{p} 0, \tag{37}$$

$$\max_i |\rho_2(\bar{\delta}'\hat{g}_i) + 1| \xrightarrow{p} 0. \tag{38}$$

Note that also Eq. (37) is true with $\hat{\delta}$ in $\rho_1(.)$ function. Next use (32)–(38)

$$\|\bar{M} - M\| \xrightarrow{p} 0,$$

where

$$M = -\begin{pmatrix} 0 & G_r' \\ G_r & \Sigma_r \end{pmatrix}.$$

See that

$$M^{-1} = \begin{pmatrix} -V_r & H_r \\ H_r' & P_r \end{pmatrix},$$

$V_r = (G_r' \Sigma_r G_r)^{-1}$, $H_r = V_r G_r' \Sigma_r^{-1}$, $P_r = \Sigma_r^{-1} - \Sigma_r^{-1} G_r V_r G_r' \Sigma_r^{-1}$. Then note that $\hat{g}(\beta_0) = n^{-1} \sum_{i=1}^{n} \hat{Z}_i \epsilon_i$, using (33) with (35), and some simple algebra

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_r).$$

Now we analyze two other possible mistakes in the first stage. The case for $q_0 > \hat{q}_0 \geq p$ is the same in consistency as $\hat{q}_0 \geq p$, and the asymptotic analysis carries over with $|q_0 - \hat{q}_0|$ in all equations in this proof instead of $\hat{q}_0 - q_0$. The case for $q_0 \geq p > \hat{q}_0$ tells us that we should stop in the first stage and do not continue with second stage since $p > \hat{q}_0$, but the probability of this event goes to zero by Theorem 1. □

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

$$\hat{\pi}_j \xrightarrow{p} \pi_j^0, \tag{39}$$

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

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

and $\hat{v}_i = X_i - Z_i \hat{\gamma}_{InitialLasso}$. Next denote

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

We want to prove specifically

$$\max_{1\leq j\leq q} |\hat{\pi}_j^2 - \tilde{\pi}_j^2| \xrightarrow{p} 0. \tag{40}$$

$$\max_{1\leq j\leq q} |\tilde{\pi}_j^2 - (\pi_j^0)^2| \xrightarrow{p} 0. \tag{41}$$

Step 3 in p. 37 (via Assumption B.1) of the proof of Theorem 1 of Belloni et al. (2012) provides the proof of (41). Then proof of Lemma 11 of online appendix in Belloni et al. (2012) shows (40) via Assumption B.1. □

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

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

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

$$V_n(u) = \sum_{i=1}^{n} \{[v_i - u'Z_i/n^{1/2}]^2 - v_i^2\} + \lambda \left[\sum_{j=1}^{q} |\hat{\pi}_j(\gamma_{j0} + u_j/n^{1/2})| - |\hat{\pi}_j \gamma_{j0}| \right].$$

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

$$\sum_{i=1}^{n} \{[v_i - u'Z_i/n^{1/2}]^2 - v_i^2\} \xrightarrow{d} -2u'W + u'\Sigma_{zz}u, \tag{42}$$

where $W \equiv N(0, \Sigma_{Zv})$, and $n^{-1} \sum_{i=1}^{n} Z_i Z_i' \xrightarrow{p} \Sigma_{zz}$. 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 zeros then the penalty term is:

$$\lambda \left[\sum_{j=1}^{q} |\hat{\pi}_j(\gamma_{j0} + u_j/n^{1/2})| - |\hat{\pi}_j \gamma_{j0}| \right] \xrightarrow{p} \lambda_0 \sum_{j=1}^{q} |\pi_j^0 u_j|, \tag{43}$$

by (39), and the Assumption of $\lambda/n^{1/2} \to \lambda_0 \geq 0$, and γ_{j0} is the jth element of γ_0 vector, $j = 1, \ldots, q$. If the γ_0 coefficients are nonzero then the limit of the penalty is

$$\lambda \left[\sum_{j=1}^{q} |\hat{\pi}_0(\gamma_{j0} + u_j/n^{1/2})| - |\hat{\pi}_j \gamma_{0j}| \right]$$

$$\xrightarrow{p} \lambda_0 \sum_{j=1}^{q} \pi_j^0 u_j sgn(\gamma_{j0}\pi_j^0), \tag{44}$$

where we use again the proof of (39) and consistency of $\hat{\gamma}_L$ in Assumption B.2. Now combine (42) (43) (44) to have

$$V_n(u) \xrightarrow{d} V(u) = -2u'W + u'\Sigma_{zz}u$$

$$+ \lambda_0 \sum_{j=1}^{q} [\pi_j^0 u_j sgn(\gamma_{j0}\pi_j^0) 1_{\{\gamma_{j0} \neq 0\}} + |u_j \pi_j^0| 1_{\{\gamma_{j0}=0\}}]. \quad \square \tag{45}$$

Proof of Lemma 2. Note that $\hat{\gamma}_{Lj}$ for all $j = 1, \ldots, q$ represents the estimator in (9). 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. (2012) Lasso estimator.

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

$$\mathcal{A}_{Ln} = \{j : \hat{\gamma}_{Lj} \neq 0\},$$
$$\mathcal{A}_L = \{j : \gamma_{Lj0} \neq 0\}.$$

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

Let $u^* = \text{argmin}\, V(u)$ in (45), then

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

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

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

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

We need to show

$$c = P(u_j^* = 0, \forall j \notin \mathcal{A}_L) < 1. \tag{46}$$

The second part of the proof has some modification compared with 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 (46) (the same in Zou (2006)) hence it is omitted. By Kuhn–Tucker optimality condition, and Σ_{zz} being defined in Lemma 1,

$$-2W_j + 2(\Sigma_{zz}u^*)_j + \lambda_0 \pi_j^0 sgn(\pi_j^0 \gamma_{j0}) = 0, \quad \forall j \in \mathcal{A}_L.$$
$$|-2W_j + 2(\Sigma_{zz}u^*)_j| \leq \lambda_0 \pi_j^0, \quad \forall j \notin \mathcal{A}_L.$$

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

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

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

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

$$-2W_{\mathcal{A}_L} + 2\Sigma_{11}u^*_{\mathcal{A}_L} + \lambda_0 \pi^0_{\mathcal{A}_L} sgn(\pi^0_{\mathcal{A}_L}\gamma_{\mathcal{A}_L}) = 0. \tag{47}$$

$$|-2W_{\mathcal{A}_L^c} + 2\Sigma_{21}u^*_{\mathcal{A}_L}| \leq \lambda_0 \pi^0_{\mathcal{A}_L^c}. \tag{48}$$

Note that this is the difference with Zou (2006) proof, we have π^0 terms in (47) (48). Next combine (47) (48) componentwise

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

This means that

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

Note that in the above equation if the truth is zero coefficient, the weight in adaptive lasso takes positive infinite value unlike $\pi^0_{\mathcal{A}_L^c}$ 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 the heteroskedasticity consistent lasso of Belloni et al. (2012). $\square$

Proof of Theorem 3. The first part of the proof (Eqs. (49)–(51)) 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

$$E\left[\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2\right] = E\left[\sum_{i=1}^{n}(\hat{\mu}_i - X_i) + (X_i - \mu_i)\right]^2$$
$$= E\left[\sum_{i=1}^{n}(\hat{\mu}_i - X_i)^2\right] + E\left[\sum_{i=1}^{n}(X_i - \mu_i)^2\right]$$
$$+ 2E\left[\sum_{i=1}^{n}\hat{\mu}_i(X_i - \mu_i)\right]$$
$$- 2E\left[\sum_{i=1}^{n}X_i(X_i - \mu_i)\right]. \tag{49}$$

Note that

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

since μ_i is constant and v_i has zero mean. Substituting these in (49) we obtain

$$E\left[\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2\right] = E\left[\sum_{i=1}^{n}(\hat{\mu}_i - X_i)^2\right] - \sum_{i=1}^{n}\sigma_i^2$$
$$+ 2E\sum_{i=1}^{n}\hat{\mu}_i(X_i - \mu_i). \tag{50}$$

Now we consider the first term on the right hand side of (50). Using (14) for each $i = 1, 2, \ldots, n$ we have

$$(\hat{\mu}_i - X_i)^2 = \begin{array}{ll} X_i^2 & \text{if } |X_i| \leq \lambda_i^{1/1+\tau} \\ \dfrac{\lambda_i^2}{|X_i|^{2\tau}} & \text{if } |X_i| > \lambda_i^{1/1+\tau}. \end{array} \tag{51}$$

We will benefit from (51) in evaluating the first term on the right hand side of (50). Next, we consider the third term on the right hand side of (50). First by Stein's Lemma (Lemma 5.1 in de la Peña et al. (2009))

$$\begin{aligned} E\left[\sum_{i=1}^n \hat{\mu}_i(X_i - \mu_i)\right] &\leq E\left[\sum_{i=1}^n \hat{\mu}_i\left(\frac{X_i - \mu_i}{\sigma_i}\right)\right] \max_i \sigma_i \\ &\leq E\left[\sum_{i=1}^n \frac{\partial \hat{\mu}_i}{\partial X_i}\right] d, \end{aligned} \tag{52}$$

where $\max_i \sigma_i \leq d$ and $0 < d < \infty$, $\sigma_i > 0$. Since by (14), for each $i = 1, 2 \cdots, n$

$$\frac{\partial \hat{\mu}_i}{\partial X_i} = \begin{array}{ll} 0 & \text{if } |X_i| \leq \lambda_i^{1/1+\tau} \\ 1 + \dfrac{\lambda_i}{\tau |X_i|^{1+\tau}} & \text{if } |X_i| > \lambda_i^{1/1+\tau}. \end{array} \tag{53}$$

Combine (51) (52) (53) in (50) we can rewrite

$$\begin{aligned} E\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 &\leq E\sum_{i=1}^n (X_i^2 1_{\{|X_i| \leq \lambda_i^{1/(1+\tau)}\}}) \\ &\quad + E\sum_{i=1}^n \left(\frac{\lambda_i^2}{|X_i|^{2\tau}} 1_{\{|X_i| > \lambda_i^{1/(1+\tau)}\}}\right) - \sum_{i=1}^n \sigma_i^2 \\ &\quad + E\left[\sum_{i=1}^n \left(2 + \frac{2\lambda_i}{\tau |X_i|^{1+\tau}}\right) 1_{\{|X_i| > \lambda_i^{1/(1+\tau)}\}}\right] d \\ &= E\left[\sum_{i=1}^n X_i^2 1_{\{|X_i| \leq \lambda_i^{1/(1+\tau)}\}}\right] \\ &\quad + E\left[\sum_{i=1}^n \left(\frac{\lambda_i^2}{|X_i|^{2\tau}} + 2d + \frac{2\lambda_i d}{\tau |X_i|^{1+\tau}}\right) 1_{\{|X_i| > \lambda_i^{1/(1+\tau)}\}}\right] \\ &\quad - \sum_{i=1}^n \sigma_i^2. \end{aligned} \tag{54}$$

By using $|X_i| \leq \lambda_i^{1/1+\tau}$ for the first right hand side term in (54), 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 (54) and $\frac{\lambda_i}{|X_i|^{1+\tau}} \leq 1$

$$\begin{aligned} E\sum_{i=1}^n [\hat{\mu}_i - \mu_i]^2 &\leq \sum_{i=1}^n [\lambda_i^{2/1+\tau} P(|X_i| \leq \lambda_i^{1/1+\tau})] \\ &\quad + \sum_{i=1}^n [(2d + 2d/\tau + \lambda_i^{2/1+\tau}) P(|X_i| > \lambda_i^{1/1+\tau})] - \sum_{i=1}^n \sigma_i^2 \\ &\leq \sum_{i=1}^n \lambda_i^{2/1+\tau} + 2d + 2d/\tau. \end{aligned} \tag{55}$$

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

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

where $c > \frac{d}{\min_i \sigma_i^2}$. So the bound in (55) can be written as

$$E\sum_{i=1}^n [(\hat{\mu}_i - \mu_i)^2] \leq \left[2 \log n + 2c + \frac{2}{\tau} c\right] \sum_{i=1}^n \sigma_i^2. \tag{56}$$

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 (56). Use (54)

$$\begin{aligned} E\left[\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2\right] &\leq E\sum_{i=1}^n X_i^2 \\ &\quad + E\left[\sum_{i=1}^n \left(\frac{\lambda_i^2}{|X_i|^{2\tau}} + 2d + \frac{2\lambda_i d}{\tau |X_i|^{1+\tau}} - X_i^2\right) 1_{\{|X_i| > \lambda_i^{1/1+\tau}\}}\right] \\ &\quad - \sum_{i=1}^n \sigma_i^2 \\ &= E\left[\sum_{i=1}^n \left(\frac{\lambda_i^2}{|X_i|^{2\tau}} + 2d + \frac{2\lambda_i d}{\tau |X_i|^{1+\tau}} - X_i^2\right) 1_{\{|X_i| > \lambda_i^{1/1+\tau}\}}\right] \\ &\quad + \sum_{i=1}^n \mu_i^2. \end{aligned} \tag{57}$$

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

$$\frac{\lambda_i^2}{|X_i|^{2\tau}} - X_i^2 \leq \frac{\lambda_i^2}{|X_i|^{2\tau}} - \lambda_i^{2/1+\tau}, \tag{58}$$

and

$$\frac{1}{|X_i|^{2\tau}} \leq \frac{1}{\lambda_i^{2\tau/(1+\tau)}} \tag{59}$$

so

$$\begin{aligned} \frac{\lambda_i^2}{|X_i|^{2\tau}} - \lambda_i^{2/1+\tau} &\leq \frac{\lambda_i^2}{\lambda_i^{2\tau/(1+\tau)}} - \lambda_i^{2/(2+\tau)} \\ &= \lambda_i^{2/(1+\tau)} - \lambda_i^{2/(1+\tau)} = 0. \end{aligned} \tag{60}$$

By (58)–(60), if $|X_i| > \lambda_i^{1/(1+\tau)}$

$$\frac{\lambda_i^2}{|X_i|^{2\tau}} - X_i^2 \leq 0. \tag{61}$$

So use (61) in (57) to have

$$\begin{aligned} &E\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \\ &\leq E\left[\sum_{i=1}^n \left(\frac{2\lambda_i d}{\tau |X_i|^{1+\tau}} + 2d\right) 1_{\{|X_i| > \lambda_i^{1/(1+\tau)}\}}\right] + \sum_{i=1}^n \mu_i^2. \end{aligned} \tag{62}$$

When $|X_i| > \lambda_i^{1/(1+\tau)}$ we can rewrite (62) as

$$\begin{aligned} &E\left[\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2\right] \\ &\leq \left(\frac{2d}{\tau} + 2d\right) \sum_{i=1}^n P(|X_i| > \lambda_i^{1/(1+\tau)}) + \sum_{i=1}^n \mu_i^2. \end{aligned} \tag{63}$$

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

$$P(|X_i| > t_i) \leq \frac{2}{\sqrt{2\pi\sigma_i^2 t_i}} e^{-t_i^2/2\sigma_i^2} + 2\mu_i^2$$
$$\leq \frac{1}{n\sqrt{\pi\sigma_i^2}} (\log n)^{-1/2} + 2\mu_i^2, \tag{64}$$

where we use t_i definition and $\lambda_i = (2\sigma_i^2 \log n)^{(1+\tau)/2}$ in the last step. See also the equations after (A.12) in Zou (2006). Use (64) in (63)

$$E\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \leq \left(\frac{4d}{\tau} + 4d\right) \max_i \left(\frac{1}{2\sqrt{\pi\sigma_i^2}} (\log n)^{-1/2}\right)$$
$$+ \left(\frac{4d}{\tau} + 4d + 1\right) \sum_{i=1}^n \mu_i^2$$
$$\leq \left(\frac{4d}{\tau} + 4d\right) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (\log n)^{-1/2}$$
$$+ \left(\frac{4d}{\tau} + 4d + 1\right) \sum_{i=1}^n \mu_i^2, \tag{65}$$

by c, d definitions. Add $2\log n \sum_{i=1}^n \mu_i^2$ and $(2\log n + 1)(c^{1/2}/d^{1/2}) \frac{1}{2\sqrt{\pi}} (\log n)^{-1/2}$ to (65) so that it is compatible with the bound in (56)

$$E\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \leq \left(2\log n + \frac{4d}{\tau} + 4d + 1\right) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (\log n)^{-1/2}$$
$$+ \left(2\log n + \frac{4d}{\tau} + 4d + 1\right) \sum_{i=1}^n \mu_i^2. \tag{66}$$

Then set $b = max(2c, 4d + 1)$. Use b definition to rewrite (66) as

$$E\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \leq \left(2\log n + \frac{b}{\tau} + b\right) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (\log n)^{-1/2}$$
$$+ \left(2\log n + \frac{b}{\tau} + b\right) \sum_{i=1}^n \mu_i^2. \tag{67}$$

Next add $(2\log n + b + \frac{b}{\tau}) c^{1/2}/d^{1/2} \frac{1}{2\sqrt{\pi}} (\log n)^{-1/2}$ to (56) and use b definition as well to have

$$E\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \leq \left(2\log n + \frac{b}{\tau} + b\right) \frac{1}{2\sqrt{\pi}} \frac{c^{1/2}}{d^{1/2}} (\log n)^{-1/2}$$
$$+ \left(2\log n + \frac{b}{\tau} + b\right) \sum_{i=1}^n \sigma_i^2. \tag{68}$$

The result can be deducted from (67) (68). □

References

Abadir, K., Magnus, J., 2005. Matrix Algebra. Cambridge University Press.
Acemoglu, D., Johnson, S., 2006. Unbundling institutions. J. Political Economy 113, 949–995.
Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The Colonial origins of a comparative development: An empirical investigation. Amer. Econ. Rev. 91, 1369–1401.
Anderson, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. Ann. Statist. 10, 1100–1120.
Averkamp, R., Houdre, C., 2003. Wavelet thresholding for non-necessarily gaussian noise: idealism. Ann. Statist. 31, 110–151.
Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse method and models for optimal instruments with an application to eminent domain. Econometrica 44, 115–130.
Bűhlmann, P., van de Geer, S., 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: Christofedes, L.N., et al. (Eds.), Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp. University of Toronto Press, Toronto, pp. 201–221.
Cheng, X., Liao, Z., 2012. Select the valid and relevant moments: A one step procedure for GMM with many moments, PIER Working Paper, 12-045.
Davidson, J., 1994. Stochastic Limit Theory. Oxford University Press.
de la Peña, V., Lai, T.L., Shao, Q.M., 2009. Self-normalized processes. In: Probability and Applications. Springer Verlag.
Donald, S., Newey, W., 2001. Choosing the number of instruments. Econometrica 69, 1161–1191.
Donoho, D., Johnstone, I., 1994. Ideal spatial adaptation via wavelet shrinkages. Biometrika 81, 425–455.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–499.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Fan, J., Li, R., 2002. Variable election for Cox's proportional hazards model and frailty model. Ann. Statist. 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.
Guggenberger, P., Smith, R., 2005. Generalized empirical likelihood estimators and tests under partial, weak and strong identification. Econometric Theory 21, 667–709.
Hall, A.R., Inoue, A., Jana, K., Shin, C., 2007. Information in generalized method of moments estimation and entropy based moment selection. J. Econometrics 138 (2), 488–512.
Hausman, J.A., Newey, W.K., Woutersen, T., Chao, J.C., Swanson, N.R., 2012. IV Estimation with heteroscedasticity and many instruments. Quant. Econom. 3, 211–255.
Knight, K., Fu, W., 2000. Asymptotics for lasso type estimators. Ann. Statist. 28, 1356–1378.
Kuersteiner, G., Okui, R., 2010. Constructing optimal instruments by first stage prediction averaging. Econometrica 78, 698–718.
Leeb, H., Pőtscher, B., 2005. Model selection and inference: facts and fiction. Econometric Theory 21, 21–59.
Nagar, A.L., 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. Econometrica 27, 575–595.
Newey, W., Smith, R., 2004. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica 72, 219–257.
Newey, W., Windmeijer, F., 2009. Many weak moment asymptotics for generalized empirical likelihood estimators. Econometrica 77, 687–721.
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., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65, 557–586.
van der Vaart, A., Wellner, J., 1996. Weak Convergence and Rmpirical Processes. Springer Verlag.
Wang, H., Leng, C., 2007. Unified LASSO estimation by least squares approximation. J. Amer. Statist. Assoc. 102, 1039–1049.
Wang, H., Li, B., Leng, C., 2009. Shrinkage tuning parameter selection with a diverging number of parameters. J. Roy. Statist. Soc. Ser. B 71, 671–683.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.