

# On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments

Frank Windmeijer<sup>a,b</sup>, Helmut Farbmacher<sup>c</sup>, Neil Davies<sup>b,d</sup>, and George Davey Smith<sup>b,d</sup>

<sup>a</sup>Department of Economics, University of Bristol, Bristol, United Kingdom; <sup>b</sup>MRC Integrative Epidemiology Unit, Bristol, United Kingdom; <sup>c</sup>Center for the Economics of Aging, Max Planck Society Munich, Germany; <sup>d</sup>School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom

## ABSTRACT

We investigate the behavior of the Lasso for selecting invalid instruments in linear instrumental variables models for estimating causal effects of exposures on outcomes, as proposed recently by Kang et al. Invalid instruments are such that they fail the exclusion restriction and enter the model as explanatory variables. We show that for this setup, the Lasso may not consistently select the invalid instruments if these are relatively strong. We propose a median estimator that is consistent when less than 50% of the instruments are invalid, and its consistency does not depend on the relative strength of the instruments, or their correlation structure. We show that this estimator can be used for adaptive Lasso estimation, with the resulting estimator having oracle properties. The methods are applied to a Mendelian randomization study to estimate the causal effect of body mass index (BMI) on diastolic blood pressure, using data on individuals from the UK Biobank, with 96 single nucleotide polymorphisms as potential instruments for BMI. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received June 2016  
Revised July 2018

## KEYWORDS

Causal inference;  
Instrumental variables  
estimation; Invalid  
instruments; Lasso;  
Mendelian randomization.


## 1. Introduction


Instrumental variables estimation is a procedure for the identification and estimation of causal effects of exposures on outcomes where the observed relationships are confounded by non-random selection of exposure. This problem is likely to occur in observational studies, but also in randomized clinical trials if there is selective participant noncompliance. An instrumental variable (IV) can be used to solve the problem of nonignorable selection. To do this, an IV needs to be associated with the exposure, but only associated with the outcome indirectly through its association with the exposure. The former condition is referred to as the “relevance” and the latter as the “exclusion” condition. Examples of instrumental variables are quarter-of-birth for educational achievement to determine its effect on wages, see Angrist and Krueger (1991), randomization of patients to treatment as an instrument for actual treatment when there is non-compliance, see, for example, Greenland (2000), and Mendelian randomization studies use IVs based on genetic information, see, for example, Lawlor et al. (2008). For recent reviews and further examples see, for example, Clarke and Windmeijer (2012), Imbens (2014), Burgess, Small, and Thompson (2017), and Kang et al. (2016).

Whether instruments are relevant can be tested from the observed association between exposure and instruments. The effects on the standard linear IV estimator of “weak instruments,” that is, the case where instruments are only weakly associated with the exposure of interest, have been derived for the linear model using weak instrument asymptotics by Staiger

and Stock (1997). This has led to the derivation of critical values for the simple  $F$ -test statistic for testing the null of weak instruments by Stock and Yogo (2005). Another strand of the literature focuses on instrument selection in potentially high-dimensional settings, see, for example, Belloni et al. (2012), Belloni et al. (2014), Chernozhukov et al. (2015), and Lin et al. (2015), where the focus is on identifying important covariate effects and selecting optimal instruments from a (large) set of a priori valid instruments, where optimality is with respect to the variance of the IV estimator.

In this article, we consider violations of the exclusion condition of the instruments, following closely the setup by Kang et al. (2016) for the linear IV model where some of the available instruments can be invalid in the sense that they can have a direct effect on the outcomes or are associated with unobserved confounders. Kang et al. (2016) proposed a Lasso-type procedure to identify and select the set of invalid instruments. Liao (2013) and Cheng and Liao (2015) also considered shrinkage estimation for identification of invalid instruments, but in their setup there is a subset of instruments that is known to be valid and that contains sufficient information for identification and estimation of the causal effects. In contrast, Kang et al. (2016) did not assume any prior knowledge about which instruments are potentially valid or invalid. This is a similar setup as in Andrews (1999) who proposed a selection procedure using information criteria based on the so-called  $J$ -test of over-identifying restrictions, as developed by Sargan (1958) and Hansen (1982). The Andrews (1999) setup is more general than

**CONTACT** Frank Windmeijer  [f.windmeijer@bristol.ac.uk](mailto:f.windmeijer@bristol.ac.uk)  Department of Economics, University of Bristol, Bristol BS8 1TH, United Kingdom.

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

that of Kang et al. (2016) and requires a large number of model evaluations, which has a negative impact on the performance of the selection procedure.

This article assesses the performance of the Kang et al. (2016) Lasso-type selection and estimation procedure in their setting of a fixed number of potential instruments. If the set of invalid instruments were known, the oracle two-stage least squares (2SLS) estimator would be the estimator of choice in their setting. As the focus is estimation of and inference on the causal effect parameter, denoted by  $\beta$ , and as the standard Lasso approach does not have oracle properties, see, for example, Zou (2006), we show how the adaptive Lasso procedure by Zou (2006) can be used to obtain an estimator with oracle properties. To do so, we propose an initial consistent estimator of the parameters that is consistent also when the irreparable condition for consistent Lasso selection of Zhao and Yu (2006) and Zou (2006) fails. The oracle property in this setup is when an estimator for  $\beta$  has the same limiting distribution as the oracle 2SLS estimator.

Applying the irreparable condition to this IV setup, we derive conditions under which the Lasso method does not consistently select the invalid instruments. As is well known from Zhao and Yu (2006), Zou (2006), Meinshausen and Bühlmann (2006), and Wainwright (2009), certain correlation structures of the variables prevent consistent selection. New in our results are the conditions on the strength of the invalid instruments relative to that of the valid ones that result in violations of the irreparable condition, where the strength of an instrument is its standardized effect on the exposure. From this we can show that consistent selection of the invalid instruments may not be possible if these are relatively strong, even when less than 50% of the instruments are invalid, which is a sufficient condition for the identification of the parameters.

We show that under the condition that less than 50% of the instruments are invalid, a simple median-type estimator is a consistent estimator for the parameters in the model, independent of the strength of the invalid instruments relative to that of the valid instruments, or their correlation structure. It can therefore be considered for use in the adaptive Lasso procedure as proposed by Zou (2006). With  $n$  the sample size, we show that the median estimator converges at the  $\sqrt{n}$  rate, but with an asymptotic bias, as the limiting distribution is that of an order statistic. It does, however, satisfy the conditions for the adaptive Lasso procedure to enjoy oracle properties.

Because of this oracle property, and as in practice instrument strength is very likely to vary by instruments and invalid instruments could be relatively strong, it will be important to consider our adaptive Lasso approach for assessing instrument validity and estimating causal effects. In Mendelian randomization studies it is clear that genetic markers have differential impacts on exposures from examining the results from genome-wide association studies and one cannot rule out *ex ante* that invalid instruments with a direct effect are also stronger predictors for the exposure. (Bowden et al. (2015) and Kolesar et al. (2015) allowed for all instruments to be invalid and showed that the causal effect can be consistently estimated if the number of instruments increases with the sample size under the assumption of uncorrelatedness of the instrument strength and their direct effects on the outcome variable.)

The next section, Section 2, introduces the model and the Lasso estimator as proposed by Kang et al. (2016). In Section 3, we derive the irreparable condition for this particular Lasso selection problem and present the result on the relationship between the relative strengths of the instruments and consistent selection. Section 4 presents the median estimator, establishes its consistency, and shows that its asymptotic properties are such that the adaptive Lasso estimator enjoys oracle properties. Section 5 presents some Monte Carlo simulation results. In Section 5.2, we link the Andrews (1999) method to the Lasso selection problem and show how the test of over-identifying restrictions can be used as a stopping rule. Section 5.3 investigates how close the behavior of the adaptive Lasso estimator is to that of the oracle 2SLS estimator in the Monte Carlo simulations, by comparing the performances of the Wald tests on the causal parameter under the null for different sample sizes. Further analyses and simulation results investigating the effects of varying the information content by varying the strength of the instruments and the size of the direct effects of the invalid instruments on the outcome are presented in Section B in the supplementary materials. In Section 6, the methods are applied to a Mendelian randomization study to estimate the causal effect of body mass index (BMI) on diastolic blood pressure using data on individuals from the UK Biobank, with 96 single nucleotide polymorphisms as potential instruments for BMI. Section 7 concludes.

The following notation is used in the remainder of the article. For a full column rank matrix  $\mathbf{X}$  with  $n$  rows,  $\mathbf{M}_X = \mathbf{I}_n - \mathbf{P}_X$ , where  $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is the projection onto the column space of  $\mathbf{X}$ , and  $\mathbf{I}_n$  is the  $n$ -dimensional identity matrix. A  $k$ -vector of ones is denoted as  $\mathbf{1}_k$ . The  $l_p$ -norm is denoted by  $\|\cdot\|_p$ , and the  $l_0$ -norm,  $\|\cdot\|_0$ , denotes the number of nonzero components of a vector. We use  $\|\cdot\|_\infty$  to denote the maximal element of a vector.

## 2. Model and Lasso Estimator

We follow Kang et al. (2016; KZCS from now on), who considered the following potential outcomes model. For  $i = 1, \dots, n$ , let  $Y_i^{(d,z)}$ , be the potential outcome if the individual  $i$  were to have exposure  $d$  and instrument values  $\mathbf{z}$ . The observed outcome for an individual  $i$  is denoted by the scalar  $Y_i$ , the treatment by the scalar  $D_i$ , and the vector of  $L$  potential instruments by  $\mathbf{Z}_i$ . The instruments may not all be valid and can have a direct or indirect effect. For two possible values of the exposure  $d^*$ ,  $d$  and instruments  $\mathbf{z}^*$ ,  $\mathbf{z}$ , assume the following potential outcomes model

$$Y_i^{(d^*,z^*)} - Y_i^{(d,z)} = (\mathbf{z}^* - \mathbf{z})' \boldsymbol{\phi} + (d^* - d) \beta \quad (1)$$

$$E[Y_i^{(0,0)} | \mathbf{Z}_i] = \mathbf{Z}_i' \boldsymbol{\psi}, \quad (2)$$

where  $\boldsymbol{\phi}$  measures the direct effect of  $\mathbf{z}$  on  $Y$ , and  $\boldsymbol{\psi}$  represents the presence of unmeasured confounders that affect both the instruments and the outcome.

We have a random sample  $\{Y_i, D_i, \mathbf{Z}_i'\}_{i=1}^n$ . Combining (1) and (2), the observed data model for the random sample is given by

$$Y_i = D_i \beta + \mathbf{Z}_i' \boldsymbol{\alpha} + \varepsilon_i, \quad (3)$$

where  $\alpha = \phi + \psi$ ;

$$\varepsilon_i = Y_i^{(0,0)} - E[Y_i^{(0,0)} | Z_i]$$

and hence  $E[\varepsilon_i | Z_i] = 0$ . For ease of exposition, we further assume that  $E[\varepsilon_i^2 | Z_i] = \sigma_\varepsilon^2$ .

The KZCS definition of a valid instrument is then linked to the exclusion restriction and given as follows: Instrument  $j, j \in \{1, \dots, L\}$ , is valid if  $\alpha_j = 0$  and it is invalid if  $\alpha_j \neq 0$ . As in the KZCS setting, we are interested in the identification and estimation of the scalar treatment effect  $\beta$  in large samples with a fixed number  $L$  of potential instruments.

Let  $\mathbf{y}$  and  $\mathbf{d}$  be the  $n$ -vectors of  $n$  observations on  $\{Y_i\}$  and  $\{D_i\}$ , respectively, and let  $\mathbf{Z}$  be the  $n \times L$  matrix of potential instruments. As an intercept is implicitly present in the model,  $\mathbf{y}, \mathbf{d}$ , and the columns of  $\mathbf{Z}$  have all been taken in deviation from their sample means. Following the notation of Zou (2006), let  $\mathbf{Z}_A$  be the set of invalid instruments,  $A = \{j : \alpha_j \neq 0\}$  and  $\alpha_A$  the associated coefficient vector. The oracle instrumental variables or two-stage least square (2SLS) estimator is obtained when the set  $\mathbf{Z}_A$  is known. Let  $\mathbf{R}_A = [\mathbf{d} \ \mathbf{Z}_A]$ , the oracle 2SLS estimator is then given by

$$\hat{\theta}_{or} = \begin{pmatrix} \hat{\beta}_{or} \\ \hat{\alpha}_A \end{pmatrix} = (\mathbf{R}'_A \mathbf{P}_Z \mathbf{R}_A)^{-1} \mathbf{R}'_A \mathbf{P}_Z \mathbf{y}. \tag{4}$$

Let  $\hat{\mathbf{d}} = \mathbf{P}_Z \mathbf{d}$ , with individual elements  $\hat{D}_i$ , then  $\hat{\theta}_{or}$  is the OLS estimator in the model

$$Y_i = \hat{D}_i \beta + \mathbf{Z}'_{A,i} \alpha_A + \xi_i,$$

where  $\xi_i$  is defined implicitly, and hence

$$\hat{\alpha}_A = (\mathbf{Z}'_A \mathbf{M}_{\hat{\mathbf{d}}} \mathbf{Z}_A)^{-1} \mathbf{Z}'_A \mathbf{M}_{\hat{\mathbf{d}}} \mathbf{y} = (\mathbf{Z}'_A \mathbf{M}_{\hat{\mathbf{d}}} \mathbf{Z}_A)^{-1} \mathbf{Z}'_A \mathbf{M}_{\hat{\mathbf{d}}} \mathbf{P}_Z \mathbf{y}. \tag{5}$$

The oracle 2SLS estimator for  $\beta$  is given by

$$\hat{\beta}_{or} = (\hat{\mathbf{d}}' \mathbf{M}_{\mathbf{Z}_A} \hat{\mathbf{d}})^{-1} \hat{\mathbf{d}}' \mathbf{M}_{\mathbf{Z}_A} \mathbf{y}.$$

Under standard assumptions, as defined below,

$$\sqrt{n} (\hat{\beta}_{or} - \beta) \xrightarrow{d} N(0, \sigma_{\hat{\beta}_{or}}^2), \tag{6}$$

where

$$\sigma_{\hat{\beta}_{or}}^2 = \sigma_\varepsilon^2 \{ E[\mathbf{Z}_i D_i]' E[\mathbf{Z}_i \mathbf{Z}'_i]^{-1} E[\mathbf{Z}_i D_i] - E[\mathbf{Z}_{A,i} D_i]' E[\mathbf{Z}_{A,i} \mathbf{Z}'_{A,i}]^{-1} E[\mathbf{Z}_{A,i} D_i] \}^{-1}. \tag{7}$$

The vector  $\hat{\mathbf{d}}$  is the linear projection of  $\mathbf{d}$  on  $\mathbf{Z}$ . If we define  $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{d}$ , then  $\hat{\mathbf{d}} = \mathbf{Z} \hat{\boldsymbol{\gamma}}$ , or  $\hat{D}_i = \mathbf{Z}'_i \hat{\boldsymbol{\gamma}}$ . We specify

$$D_i = \mathbf{Z}'_i \boldsymbol{\gamma} + v_i, \tag{8}$$

where  $\boldsymbol{\gamma} = E[\mathbf{Z}_i \mathbf{Z}'_i]^{-1} E[\mathbf{Z}_i D_i]$ , and hence  $E[\mathbf{Z}_i v_i] = 0$ . Further, as in KZCS, let  $\boldsymbol{\Gamma} = E[\mathbf{Z}_i \mathbf{Z}'_i]^{-1} E[\mathbf{Z}_i Y_i] = \boldsymbol{\gamma} \beta + \boldsymbol{\alpha}$ . Then define  $\pi_j$  as

$$\pi_j \equiv \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j}, \tag{9}$$

for  $j = 1, \dots, L$ . **Theorem 1** in KZCS states the conditions under which, given knowledge of  $\boldsymbol{\gamma}$  and  $\boldsymbol{\Gamma}$ , a unique solution exists for values of  $\beta$  and  $\alpha_j$ . A necessary and sufficient condition to identify  $\beta$  and the  $\alpha_j$  is that the valid instruments form the largest group, where instruments form a group if they have the same value of  $\pi$ . **Corollary 1** in KZCS then states a sufficient

condition for identification. Let  $s = \|\boldsymbol{\alpha}\|_0$  be the number of invalid instruments. A sufficient condition is that  $s < L/2$ , as then clearly the largest group is formed by the valid instruments.

In model (3), some elements of  $\boldsymbol{\alpha}$  are assumed to be zero, but it is not known ex ante which ones they are and the selection problem therefore consists of correctly identifying those instruments with nonzero  $\alpha$ . KZCS proposed to estimate the parameters  $\boldsymbol{\alpha}$  and  $\beta$  by using  $l_1$  penalization on  $\boldsymbol{\alpha}$  and to minimize

$$(\hat{\boldsymbol{\alpha}}^{(n)}, \hat{\beta}^{(n)}) = \arg \min_{\boldsymbol{\alpha}, \beta} \frac{1}{2} \|\mathbf{P}_Z (\mathbf{y} - \mathbf{d}\beta - \mathbf{Z}\boldsymbol{\alpha})\|_2^2 + \lambda_n \|\boldsymbol{\alpha}\|_1, \tag{10}$$

where  $\|\boldsymbol{\alpha}\|_1 = \sum_j |\alpha_j|$ . This method is closely related to the Lasso, and the regularization parameter  $\lambda_n$  determines the sparsity of the vector  $\hat{\boldsymbol{\alpha}}^{(n)}$ . From (5), a fast two-step algorithm is proposed as follows. For a given  $\lambda_n$  solve

$$\hat{\boldsymbol{\alpha}}^{(n)} = \arg \min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{M}_{\hat{\mathbf{d}}} \mathbf{P}_Z \mathbf{y} - \mathbf{M}_{\hat{\mathbf{d}}} \mathbf{Z} \boldsymbol{\alpha}\|_2^2 + \lambda_n \|\boldsymbol{\alpha}\|_1 \tag{11}$$

and obtain  $\hat{\beta}^{(n)}$  by

$$\hat{\beta}^{(n)} = \frac{\hat{\mathbf{d}}' (\mathbf{y} - \mathbf{Z} \hat{\boldsymbol{\alpha}}^{(n)})}{\hat{\mathbf{d}}' \hat{\mathbf{d}}}. \tag{12}$$

To find  $\hat{\boldsymbol{\alpha}}^{(n)}$  in (11), the Lasso modification of the LARS algorithm of Efron et al. (2004) can be used and KZCS had developed an R-routine for this purpose, called *sisVIVE* (some invalid and some valid IV estimator), where the regularization parameter  $\lambda_n$  is obtained by cross-validation.

For the random variables and iid sample  $\{Y_i, D_i, \mathbf{Z}'_i\}_{i=1}^n$ , and model (3) and (8), we assume throughout that the following conditions hold:

*Assumption 1.*  $E[\mathbf{Z}_i \mathbf{Z}'_i] = \mathbf{Q}$ , with  $\mathbf{Q}$  a finite and full-rank matrix.

*Assumption 2.* Let  $\mathbf{u}_i = (\varepsilon_i v_i)'$ . Then  $E[\mathbf{u}_i] = 0$ ;  $E[\mathbf{u}_i \mathbf{u}'_i] = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{bmatrix} = \Sigma$ . The elements of  $\Sigma$  are finite.

*Assumption 3.*  $\text{plim}(n^{-1} \mathbf{Z}' \mathbf{Z}) = E[\mathbf{Z}_i \mathbf{Z}'_i]$ ;  $\text{plim}(n^{-1} \mathbf{Z}' \mathbf{d}) = E[\mathbf{Z}_i D_i]$ ;  $\text{plim}(n^{-1} \mathbf{Z}' \boldsymbol{\varepsilon}) = E[\mathbf{Z}_i \varepsilon_i] = 0$ ;  $\text{plim}(n^{-1} \mathbf{Z}' \mathbf{v}) = E[\mathbf{Z}_i v_i] = 0$ ;  $\text{plim}(n^{-1} \sum_{i=1}^n \mathbf{u}_i) = 0$ ;  $\text{plim}(n^{-1} \sum_{i=1}^n \mathbf{u}_i \mathbf{u}'_i) = \Sigma$ .

*Assumption 4.*  $\gamma = (E[\mathbf{Z}_i \mathbf{Z}'_i])^{-1} E[\mathbf{Z}_i D_i]$ ,  $\gamma_j \neq 0, j = 1, \dots, L$ .

The setting is thus a relatively straightforward one with fixed parameters  $\beta, \boldsymbol{\alpha}$ , and  $\boldsymbol{\gamma}$ , and fixed number  $L \ll n$  of potential instruments. This is the setting under which the oracle 2SLS estimator has the limiting distribution (6), and is a setting of interest in many applications. To identify in this simple setting an ex ante unknown subset of invalid instruments using the Lasso is challenging, as highlighted in the next section where we investigate the irrerepresentable condition for this setting.

For the case of many weak instruments, even the oracle 2SLS estimator would not be the estimator of choice, due to its poor asymptotic performance, and the median estimator may not be consistent. Oracle estimators with better asymptotic properties in this setting are the limited information maximum likelihood (LIML) estimator, see Bekker (1994) and Hansen, Hausman and Newey (2008), or the continuous updating estimator (CUE), see Newey and Windmeijer (2009). Selection of invalid instruments in this setting is outside the scope of this article.

### 3. Irrepresentable Condition

As  $\mathbf{Z}'\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{P}_{\mathbf{Z}}\mathbf{y} = \mathbf{Z}'\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{P}_{\mathbf{Z}}\mathbf{y} = \mathbf{Z}'\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{y}$ , it follows that

$$\begin{aligned} \|\mathbf{M}_{\tilde{\mathbf{d}}}(\mathbf{P}_{\mathbf{Z}}\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha})\|_2^2 &= \mathbf{y}'\mathbf{P}_{\mathbf{Z}}\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{P}_{\mathbf{Z}}\mathbf{y} - 2\mathbf{y}'\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\alpha}'\mathbf{Z}'\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{Z}\boldsymbol{\alpha} \\ &= \mathbf{y}'\mathbf{P}_{\mathbf{Z}}\mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{P}_{\mathbf{Z}}\mathbf{y} - 2\mathbf{y}'\tilde{\mathbf{Z}}\boldsymbol{\alpha} + \boldsymbol{\alpha}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\boldsymbol{\alpha}, \end{aligned}$$

where  $\tilde{\mathbf{Z}} = \mathbf{M}_{\tilde{\mathbf{d}}}\mathbf{Z}$ . As

$$\|\mathbf{y} - \tilde{\mathbf{Z}}\boldsymbol{\alpha}\|_2^2 = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\tilde{\mathbf{Z}}\boldsymbol{\alpha} + \boldsymbol{\alpha}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\boldsymbol{\alpha},$$

it follows that the Lasso estimator  $\hat{\boldsymbol{\alpha}}^{(n)}$  as defined in (11) can equivalently be obtained as

$$\hat{\boldsymbol{\alpha}}^{(n)} = \arg \min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{y} - \tilde{\mathbf{Z}}\boldsymbol{\alpha}\|_2^2 + \lambda_n \|\boldsymbol{\alpha}\|_1. \quad (13)$$

This minimization problem looks very much like a standard Lasso approach with  $\tilde{\mathbf{Z}}$  as explanatory variables. However, an important difference is that  $\tilde{\mathbf{Z}}$  does not have full rank, but its rank is equal to  $L - 1$ . This is related to the standard Lasso case where we have an overcomplete dictionary implying that the OLS solution is not feasible. Intuitively, we cannot set  $\lambda_n = 0$  in (13) as we have to shrink at least one element of  $\boldsymbol{\alpha}$  to zero to identify the parameter  $\beta$ . All just-identified models with  $L - 1$  instruments included as invalid result in a residual correlation of 0, and hence setting  $\lambda_n = 0$  does not lead to a unique 2SLS estimator.

We assume throughout that  $E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i']$  is finite. Let  $\mathbf{C} = \text{plim}(n^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})$ , then it follows from Assumptions 1, 3, and 4 that  $\mathbf{C} = \mathbf{Q} - \mathbf{Q}\boldsymbol{\gamma}(\boldsymbol{\gamma}'\mathbf{Q}\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}'\mathbf{Q}$  is finite.

We follow Zhao and Yu (2006) and Zou (2006), who developed the irrepresentable conditions for consistent Lasso variable selection. As before, let  $A = \{j : \alpha_j \neq 0\}$  and assume wlog that  $A = \{1, 2, \dots, s\}$ ,  $s < L$ . (We will use subscripts  $A$  and 1 interchangeably from here onward, and subscript 2 for associations with the set  $A^c = \{j : \alpha_j = 0\}$ .) Let

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}'_{21} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}, \quad (14)$$

where  $\mathbf{C}_{11}$  is an  $s \times s$  matrix. Further, define  $\hat{A}_n = \{j : \hat{\alpha}_j^{(n)} \neq 0\}$ . Let  $\mathbf{s}(\boldsymbol{\alpha}_1)$  denote the vector  $\text{sgn}(\boldsymbol{\alpha}_1)$ , where  $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_A = (\alpha_1, \dots, \alpha_s)'$ ,  $\text{sgn}(a) = 1$  if  $a > 0$ ,  $\text{sgn}(a) = -1$  if  $a < 0$ , and  $\text{sgn}(a) = 0$  if  $a = 0$ . The irrepresentable condition

$$\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)\|_{\infty} < 1, \quad (15)$$

is an (almost) necessary and sufficient condition for consistent Lasso variable selection. While (15) refers to the formulation of the weak irrepresentable condition of Zhao and Yu (2006), they showed that in this setting of a random design with fixed  $L$  and constant parameters  $\boldsymbol{\alpha}$ , their strong and weak irrepresentable conditions are equivalent to (15) almost surely (Zhao and Yu 2006, p. 2544).

If (15) is satisfied, and if  $\lambda_n$  satisfies  $\lambda_n/n \rightarrow 0$  and  $\lambda_n/\sqrt{n} \rightarrow \infty$ , then  $\lim_{n \rightarrow \infty} P(\hat{A}_n = A) = 1$ , see Theorem 1 in Zhao and Yu (2006). Necessity means that consistent model selection implies the irrepresentable condition. As Zou (2006) showed, if  $\lim_{n \rightarrow \infty} P(\hat{A}_n = A) = 1$  and under the same conditions  $\lambda_n/n \rightarrow 0$  and  $\lambda_n/\sqrt{n} \rightarrow \infty$ , then the following condition must hold

$$\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)\|_{\infty} \leq 1. \quad (16)$$

While in the standard linear model setup  $\lambda_n/n \rightarrow 0$  guarantees estimation consistency, see Lemma 1 in Zou (2006), this is not the case in the IV setup here because of the rank deficiency of  $\tilde{\mathbf{Z}}$ . Choosing  $\lambda_n = 0$  in the standard setup would simply result in consistent OLS estimation of a model that includes all variables, which is not possible here as discussed above. Therefore, if the necessary irrepresentable condition (16) does not hold, consistent Lasso selection is not possible and even  $\lambda_n/n \rightarrow 0$  does not guarantee estimation consistency in this rank deficient IV case.

We now analyze under what conditions the irrepresentable condition does or does not hold in the IV setup, focusing particularly on the relative strengths  $\boldsymbol{\gamma}_1$  and  $\boldsymbol{\gamma}_2$  of the invalid and valid instruments.

Partition  $\mathbf{Q} = \text{plim}(n^{-1}\mathbf{Z}'\mathbf{Z})$  and  $\boldsymbol{\gamma}$  commensurate with the partitioning of  $\mathbf{C}$  as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}'_{21} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix}, \quad (17)$$

where the instruments have been standardized such the diagonal elements of  $\mathbf{Q}$  are equal to 1. In contrast to  $\mathbf{C}$ ,  $\mathbf{Q}$  is not rank deficient. Then for the Lasso specification (13), we have the following result.

*Proposition 1.* Consider the observational models (3) and (8) under Assumptions 1, 3, and 4. Let  $\mathbf{C} = \text{plim}(n^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})$ ;  $\mathbf{Q} = \text{plim}(n^{-1}\mathbf{Z}'\mathbf{Z})$ ; and  $\mathbf{C}_{11}$ ,  $\mathbf{C}_{21}$ ,  $\mathbf{Q}_{11}$ ,  $\mathbf{Q}_{21}$ ,  $\mathbf{Q}_{22}$ ,  $\boldsymbol{\gamma}_1$ , and  $\boldsymbol{\gamma}_2$  as specified in (14) and (17). Then  $\mathbf{C}_{21}\mathbf{C}_{11}^{-1}$  is given by

$$\mathbf{C}_{21}\mathbf{C}_{11}^{-1} = \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1} - \tilde{\mathbf{Q}}_{22}\boldsymbol{\gamma}_2 \frac{\boldsymbol{\gamma}'_1 + \boldsymbol{\gamma}'_2\mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}}{\boldsymbol{\gamma}'_2\tilde{\mathbf{Q}}_{22}\boldsymbol{\gamma}_2}, \quad (18)$$

where

$$\tilde{\mathbf{Q}}_{22} = \mathbf{Q}_{22} - \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\mathbf{Q}'_{21} = \text{plim}(n^{-1}\mathbf{Z}'_2\mathbf{M}_{\mathbf{Z}_1}\mathbf{Z}_2).$$

*Proof.* See Section A.1 in the supplementary materials.  $\square$

Proposition 1 shows that consistent selection of the instruments is not only affected by the correlation structure of the instruments, but also by the values of  $\boldsymbol{\gamma}_1$  and  $\boldsymbol{\gamma}_2$ . The next Proposition derives conditions on  $\boldsymbol{\gamma}_1$  and  $\boldsymbol{\gamma}_2$  under which the necessary condition for consistent variable selection (16) does not hold.

*Proposition 2.* Under the assumptions of Proposition 1, if  $|\boldsymbol{\gamma}'_1\mathbf{s}(\boldsymbol{\alpha}_1)| > \|\boldsymbol{\gamma}_2\|_1$ , then  $\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)\|_{\infty} > 1$ .

*Proof.* It follows from (18) that

$$|\boldsymbol{\gamma}'_2\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)| = |\boldsymbol{\gamma}'_1\mathbf{s}(\boldsymbol{\alpha}_1)|.$$

Therefore,

$$\begin{aligned} \|\boldsymbol{\gamma}_2\|_1 \|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)\|_{\infty} &\geq |\boldsymbol{\gamma}'_1\mathbf{s}(\boldsymbol{\alpha}_1)| \\ \|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)\|_{\infty} &\geq \frac{|\boldsymbol{\gamma}'_1\mathbf{s}(\boldsymbol{\alpha}_1)|}{\|\boldsymbol{\gamma}_2\|_1}. \end{aligned}$$

Hence,  $\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)\|_{\infty} > 1$  if  $|\boldsymbol{\gamma}'_1\mathbf{s}(\boldsymbol{\alpha}_1)| > \|\boldsymbol{\gamma}_2\|_1$ .  $\square$

*Remark 1.* If  $\mathbf{s}(\boldsymbol{\alpha}_1) = \mathbf{s}(\boldsymbol{\gamma}_1)$ , then  $|\boldsymbol{\gamma}'_1\mathbf{s}(\boldsymbol{\alpha}_1)| = \|\boldsymbol{\gamma}_1\|_1$ , its maximum. Regardless of the correlation structure of the instruments,  $\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}(\boldsymbol{\alpha}_1)\|_{\infty} > 1$  and hence the necessary condition for consistent Lasso variable selection does not hold in that case



if  $\|\boldsymbol{\gamma}_1\|_1 > \|\boldsymbol{\gamma}_2\|_1$ , that is, when the invalid instruments are stronger (in  $l_1$ -norm) than the valid ones.

From Proposition 1, we can investigate consistent selection for various cases of interest. Related to the Monte Carlo simulations in KZCS and in Section 5, Corollary 1 considers the case with  $\boldsymbol{\gamma}_1 = \tilde{\gamma}_1 \boldsymbol{t}_s$  and  $\boldsymbol{\gamma}_2 = \tilde{\gamma}_2 \boldsymbol{t}_{L-s}$ .

*Corollary 1.* If  $\boldsymbol{\gamma}_1 = \tilde{\gamma}_1 \boldsymbol{t}_s$  and  $\boldsymbol{\gamma}_2 = \tilde{\gamma}_2 \boldsymbol{t}_{L-s}$ , then  $|\boldsymbol{\gamma}'_1 \boldsymbol{s}(\boldsymbol{\alpha}_1)| > \|\boldsymbol{\gamma}_2\|_1$  if  $|\frac{\tilde{\gamma}_1}{\tilde{\gamma}_2} \boldsymbol{t}'_s \boldsymbol{s}(\boldsymbol{\alpha}_1)| > L - s$ . Let  $g = |\boldsymbol{t}'_s \boldsymbol{s}(\boldsymbol{\alpha}_1)|$ , then it follows that  $\|C_{21} C_{11}^{-1} \boldsymbol{s}(\boldsymbol{\alpha}_1)\|_\infty > 1$  if  $|\frac{\tilde{\gamma}_1}{\tilde{\gamma}_2} g| > L - s$ . Hence, if  $g = s$ ,  $\|C_{21} C_{11}^{-1} \boldsymbol{s}(\boldsymbol{\alpha}_1)\|_\infty > 1$  if  $s > L / (1 + |\frac{\tilde{\gamma}_1}{\tilde{\gamma}_2}|)$ .

When instruments are uncorrelated, such that  $\mathbf{Q} = \mathbf{I}_L$ , it follows that  $\|C_{21} C_{11}^{-1} \boldsymbol{s}(\boldsymbol{\alpha}_1)\|_\infty < 1$  if  $s < L - |\frac{\tilde{\gamma}_1}{\tilde{\gamma}_2}| g$ . Hence, if  $g = s$ ,  $\|C_{21} C_{11}^{-1} \boldsymbol{s}(\boldsymbol{\alpha}_1)\|_\infty < 1$  if  $s < L / (1 + |\frac{\tilde{\gamma}_1}{\tilde{\gamma}_2}|)$ .

*Remark 2.* For equal strength instruments,  $\tilde{\gamma}_1 = \tilde{\gamma}_2$ , the result of Corollary 1 shows that the necessary condition (16) does not hold for all possible configurations of  $\boldsymbol{\alpha}_1$  if  $s > L/2$ . For uncorrelated equal strength instruments, the irrepresentable condition (15) holds for all possible configurations of  $\boldsymbol{\alpha}_1$  if  $s < L/2$ .

#### 4. A Consistent Estimator when $s < L/2$ and Adaptive Lasso

As the results above highlight, the Lasso path may not include the correct model, leading to an inconsistent estimator of  $\beta$ . This is the case even if less than 50% of the instruments are invalid because of differential instrument strength and/or correlation patterns of the instruments. Indeed, we find in the simulation exercise of Section 5.1 that the Lasso selects the valid instruments as invalid if these are relatively weak,  $\|\boldsymbol{\gamma}_2\|_1 < \|\boldsymbol{\gamma}_1\|_1$ , for a design with  $\boldsymbol{s}(\boldsymbol{\alpha}_1) = \boldsymbol{s}(\boldsymbol{\gamma}_1)$ . In this section, we present an estimation method that consistently selects the invalid instruments when less than 50% of the potential instruments are invalid. This is the same condition as that for the Lasso selection problem to satisfy the irrepresentable condition for equal strength uncorrelated instruments, but the proposed estimator below is consistent when the instruments have differential strength and/or have a general correlation structure.

We consider the adaptive Lasso approach of Zou (2006) using an initial consistent estimator of the parameters. In the standard linear case, the OLS estimator in the model with all explanatory variables included is consistent. As explained in Section 3, in the instrumental variables model this option is not available. We build on the result of Han (2008), who shows that the median of the  $L$  IV estimates of  $\beta$  using one instrument at the time is a consistent estimator of  $\beta$  in a model with invalid instruments, but where the instruments cannot have direct effects on the outcome, unless the instruments are uncorrelated.

Let  $\hat{\Gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ ;  $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{d}$  and let  $\hat{\boldsymbol{\pi}}$  be the  $L$ -vector with  $j$ th element

$$\hat{\pi}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}. \tag{19}$$

Under the standard assumptions, Theorem 1 shows that the median of the  $\hat{\pi}_j$ , denoted  $\hat{\beta}_m$ , is a consistent estimator for  $\beta$  when  $s < L/2$ , without any further restrictions on the relative strengths or correlations of the instruments. Theorem 1 also

shows that  $\sqrt{n}(\hat{\beta}_m - \beta)$  converges in distribution to that of an order statistic. From these results it follows that the consistent estimator  $\hat{\boldsymbol{\alpha}}_m = \hat{\Gamma} - \hat{\boldsymbol{\gamma}}\hat{\beta}_m$  can be used for the adaptive Lasso approach of Zou (2006), resulting in oracle properties of the resulting estimator of  $\beta$ .

*Theorem 1.* Under model specifications (3) and (8) with Assumptions 1–4, let  $\hat{\boldsymbol{\pi}}$  be the  $L$ -vector with elements as defined in (19). If  $s < L/2$ , then the estimator  $\hat{\beta}_m$  defined as

$$\hat{\beta}_m = \text{median}(\hat{\boldsymbol{\pi}})$$

is a consistent estimator for  $\beta$ ,

$$\text{plim}(\hat{\beta}_m) = \beta.$$

Let  $\hat{\boldsymbol{\pi}}_2$  be the  $L - s$  vector with elements  $\hat{\pi}_j$ ,  $j = s + 1, \dots, L$ . The limiting distribution of  $\hat{\beta}_m$  is given by

$$\sqrt{n}(\hat{\beta}_m - \beta) \xrightarrow{d} q_{[l],L-s},$$

where for  $L$  odd,  $q_{[l],L-s}$  is the  $l$ th-order statistic of the limiting normal distribution of  $\sqrt{n}(\hat{\boldsymbol{\pi}}_2 - \beta \boldsymbol{t}_{L-s})$ , where  $l$  is determined by  $L, s$ , and the signs of  $\delta_j = \frac{\alpha_j}{\gamma_j}$ ,  $j = 1, \dots, s$ . For  $L$  even,  $q_{[l],L-s}$  is defined as the average of either the  $[l]$  and  $[l - 1]$ -order statistics, or the  $[l]$  and  $[l + 1]$ -order statistics.

*Proof.* See Section A.2 in the supplementary materials. □

Given the consistent estimator  $\hat{\beta}_m$ , we obtain a consistent estimator for  $\boldsymbol{\alpha}$  as

$$\hat{\boldsymbol{\alpha}}_m = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{d}\hat{\beta}_m) = \hat{\Gamma} - \hat{\boldsymbol{\gamma}}\hat{\beta}_m,$$

which can then be used for the adaptive Lasso specification of (13) as proposed by Zou (2006). The adaptive Lasso estimator for  $\boldsymbol{\alpha}$  is defined as

$$\hat{\boldsymbol{\alpha}}_{ad}^{(n)} = \arg \min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{y} - \tilde{\mathbf{Z}}\boldsymbol{\alpha}\|_2^2 + \lambda_n \sum_{l=1}^L \frac{|\alpha_l|}{|\hat{\alpha}_{m,l}|^v}, \tag{20}$$

and, for given values of  $v$  can be estimated straightforwardly using the LARS algorithm, see Zou (2006). The resulting adaptive Lasso estimator for  $\beta$  is obtained as

$$\hat{\beta}_{ad}^{(n)} = \frac{\hat{\mathbf{d}}'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_{ad}^{(n)})}{\hat{\mathbf{d}}'\hat{\mathbf{d}}}.$$

As the result for the limiting distribution of the median estimator shows,  $\hat{\beta}_m$ , although converging at the  $\sqrt{n}$  rate, has an asymptotic bias. This clearly also results in an asymptotic bias of  $\hat{\boldsymbol{\alpha}}_m$ . As  $\sqrt{n}(\hat{\boldsymbol{\alpha}}_m - \boldsymbol{\alpha}) = O_p(1)$ , Theorem 2 together with Remark 1 in Zou (2006) states the following properties of the adaptive Lasso estimator  $\hat{\boldsymbol{\alpha}}_{ad}^{(n)}$ , where  $\hat{A}_{ad,n} = \{j : \hat{\alpha}_{ad,j}^{(n)} \neq 0\}$ .

*Proposition 3.* Suppose that  $\lambda_n = o(\sqrt{n})$  and  $(\sqrt{n})^{v-1}\lambda_n \rightarrow \infty$ , then the adaptive Lasso estimator  $\hat{\boldsymbol{\alpha}}_{ad}^{(n)}$  satisfies

1. Consistency in variable selection:  $\lim_{n \rightarrow \infty} P(\hat{A}_{ad,n} = A) = 1$ .
2. Asymptotic normality:  $\sqrt{n}(\hat{\boldsymbol{\alpha}}_{ad,A}^{(n)} - \boldsymbol{\alpha}_A) \xrightarrow{d} N(0, \sigma^2 C_{11}^{-1})$ .

*Proof.* See Zou (2006), Theorem 2 and Remark 1. □

From the results of Proposition 3, it follows that the limiting distribution of  $\widehat{\beta}_{ad}^{(n)}$  is that of the oracle 2SLS estimator of  $\beta$ , as stated in the next Corollary.

Corollary 2. Under the conditions of Proposition 3, the limiting distribution of the adaptive Lasso estimator  $\widehat{\beta}_{ad}^{(n)}$  is given by

$$\sqrt{n}(\widehat{\beta}_{ad}^{(n)} - \beta) \xrightarrow{d} N(0, \sigma_{\beta_{or}}^2), \tag{21}$$

with  $\sigma_{\beta_{or}}^2$  as defined in (7).

### 5. Simulation Results

#### 5.1. Relative Strength of Instruments

We start with presenting some estimation results from a Monte Carlo exercise which is similar to that in KZCS. The data are generated from

$$Y_i = D_i\beta + Z_i'\alpha + \varepsilon_i$$

$$D_i = Z_i'\gamma + v_i,$$

where

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right);$$

$$Z_i \sim N(0, \mathbf{I}_L);$$

and we set  $\beta = 0$ ;  $L = 10$ ;  $\rho = 0.25$ ;  $s = 3$ , and the first  $s$  elements of  $\alpha$  are equal to  $a = 0.2$ . Further,  $\gamma_1 = \tilde{\gamma}_1 t_s$  and  $\gamma_2 = \tilde{\gamma}_2 t_{L-s}$ . Note that none of the estimation results presented here and below depend on the value of  $\beta$ . Table 1 presents estimation results for estimators of  $\beta$  in terms of bias, standard deviation, root mean squared error (rmse), and median absolute deviation (mad) for 1000 replications for sample sizes of  $n = 500$ ,  $n = 2000$ , and  $n = 10,000$  for an equal strength design, with  $\tilde{\gamma}_1 = \tilde{\gamma}_2 = 0.2$ .

The information content for IV estimation can be summarized by the concentration parameter, see Rothenberg (1984).

For the oracle estimation of  $\beta$  by 2SLS, the concentration parameter is given by  $\mu_n^2 = \gamma_2' Z_2' M_{Z_1} Z_2 \gamma_2 / \sigma_v^2$ . For this data-generating process with independent instruments, the concentration parameter is therefore approximately  $n(L - s)(0.2^2)$  and hence equal to 140, 560, and 2800 for the three sample sizes.  $\mu_n^2$  can be seen as a population Wald statistic for testing  $H_0 : \gamma_2 = 0$ . The corresponding population  $F$ -statistics are equal to  $n(0.2^2)$ , or 20, 80, and 400 for the sample sizes 500, 2000, and 10,000, respectively.

A summary measure of the information content for Lasso selection is the (squared) signal-to-noise ratio (SNR), denoted by  $\eta^2$ . It is defined as

$$\eta^2 = \frac{\alpha_1' C_{11} \alpha_1}{\sigma_\varepsilon^2},$$

see, for example, Bühlmann and van der Geer (2011, p. 25). Analogously to the concentration parameter,  $n\eta^2$  can be interpreted as a population Wald statistic for testing  $H_0 : \alpha_1 = 0$ . We analyze the effects of varying  $\mu_n^2$  and  $\eta^2$  more extensively in Section B.2 in the supplementary materials, where we derive that, for this design,

$$\eta^2 = \frac{(L - s) a^2}{\left(\frac{\tilde{\gamma}_1}{\tilde{\gamma}_2}\right)^2 + \frac{L-s}{s}}, \tag{22}$$

resulting in  $\eta^2 = 0.084$  for the parameter values considered in Table 1.

The “2SLS” results are for the naive 2SLS estimator of  $\beta$  that treats all instruments as valid. The probability limit of this estimator is given by

$$\begin{aligned} \text{plim}(\widehat{\beta}_{naive}) &= \beta + \frac{\gamma' Q \alpha}{\gamma' Q \gamma} \\ &= \beta + \frac{\gamma_1' Q_{11} \alpha_1 + \gamma_2' Q_{21} \alpha_1}{\gamma_1' Q_{11} \gamma_1 + 2\gamma_2' Q_{21} \gamma_1 + \gamma_2' Q_{22} \gamma_2}. \end{aligned} \tag{23}$$

Table 1. Estimation results for 2SLS and Lasso estimators for  $\beta$ ;  $L = 10, s = 3, \tilde{\gamma}_1 = \tilde{\gamma}_2$ .

$\beta$	bias	std dev	rmse	mad	av. # instr selected as invalid [min, max]	freq. all invalid instr selected
<i>n</i> = 500						
2SLS	0.2966	0.0808	0.3074	0.2944	0	0
2SLS or	0.0063	0.0843	0.0845	0.0570	3	1
Lasso <sub>cv</sub>	0.1384	0.0965	0.1687	0.1352	6.41 [2,9]	0.990
Post-Lasso <sub>cv</sub>	0.1169	0.1136	0.1630	0.1143		
Lasso <sub>cvse</sub>	0.2206	0.0847	0.2363	0.2174	3.16 [0,8]	0.664
Post-Lasso <sub>cvse</sub>	0.0905	0.1243	0.1537	0.0994		
<i>n</i> = 2000						
2SLS	0.3019	0.0387	0.3044	0.3007	0	0
2SLS or	0.0047	0.0422	0.0424	0.0285	3	1
Lasso <sub>cv</sub>	0.0721	0.0509	0.0882	0.0705	6.64 [3,9]	1
Post-Lasso <sub>cv</sub>	0.0617	0.0577	0.0845	0.0644		
Lasso <sub>cvse</sub>	0.1140	0.0430	0.1218	0.1165	3.76 [3,8]	1
Post-Lasso <sub>cvse</sub>	0.0277	0.0521	0.0590	0.0387		
<i>n</i> = 10,000						
2SLS	0.2996	0.0177	0.3002	0.2992	0	0
2SLS or	0.0006	0.0182	0.0182	0.0126	3	1
Lasso <sub>cv</sub>	0.0317	0.0236	0.0395	0.0311	6.44 [3,9]	1
Post-Lasso <sub>cv</sub>	0.0272	0.0267	0.0380	0.0282		
Lasso <sub>cvse</sub>	0.0479	0.0187	0.0514	0.0489	3.81 [3,9]	1
Post-Lasso <sub>cvse</sub>	0.0118	0.0238	0.0265	0.0176		

NOTE: Results from 1000 MC replications;  $\beta = 0$ ;  $\rho = 0.25$ ;  $a = 0.2$ ;  $\tilde{\gamma}_2 = 0.2$ .

Therefore, in the design specified here, we have  $\text{plim}(\hat{\beta}_{\text{naive}}) = s/L = 0.3$ .

The “2SLS or” is the oracle 2SLS estimator that correctly includes the three invalid instruments in the model as explanatory variables. For the Lasso estimates, the value for  $\lambda_n$  has been obtained by 10-fold cross-validation, using the one-standard error rule, as in KZCS. This estimator is denoted “Lasso<sub>cvse</sub>” and is the one produced by the *sisVIVE* routine. We also present results for the cross-validated estimator that does not use the one-standard error rule, denoted “Lasso<sub>cv</sub>.” For the Lasso estimation procedure, we standardize throughout such that the diagonal elements of  $\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}/n$  are equal to 1.

We further present results for the so-called post-Lasso estimator, see, for example, Belloni et al. (2012), which is called the LARS-OLS hybrid by Efron et al. (2004). This is here simply the 2SLS estimator in the model that includes  $\mathbf{Z}_{\hat{A}_n}$ , the set of instruments with nonzero estimated Lasso coefficients. Clearly, when  $\hat{A}_n = A$ , the post-Lasso 2SLS estimator is equal to the oracle 2SLS estimator. The post-Lasso 2SLS estimator is expected to have a smaller bias as it avoids the bias in the Lasso estimate of  $\beta$  due to the shrinkage of the Lasso estimate of  $\alpha$  toward  $\mathbf{0}$ , see also Hastie, Tibshirani, and Friedman (2009, p. 91). This shrinkage bias effect on  $\hat{\beta}^{(n)}$  for models where  $A \subseteq \hat{A}_n$  is in the direction of the bias of  $\hat{\beta}_{\text{naive}}$ , where  $\alpha$  is assumed to be  $\mathbf{0}$ . (In an OLS setting, Belloni and Chernozhukov (2013) showed that the post-Lasso estimator can perform at least as well as Lasso in terms of rate of convergence, but is less biased even if the Lasso-based model selection misses some components of the true model.)

Further entries in Table 1 are the average number of instruments selected as invalid, that is, the average number of instruments in  $\hat{A}_n = \{j : \hat{\alpha}_j^{(n)} \neq 0\}$ , together with the minimum and maximum number of selected instruments, and the proportion of times the instruments selected as invalid include all three invalid instruments.

The results in Table 1 reveal some interesting patterns. First of all, the Lasso<sub>cv</sub> estimator outperforms the Lasso<sub>cvse</sub> estimator in terms of bias, rmse, and mad for all sample sizes, but this is reversed for the post-Lasso estimators, that is, the post-Lasso<sub>cvse</sub> outperforms the post-Lasso<sub>cv</sub>. The Lasso<sub>cv</sub> estimator selects on average around 6.5 instruments as invalid, which is virtually independent of the sample size. The Lasso<sub>cvse</sub> estimator selects on average around 3.8 instruments as invalid for  $n = 2000$  and  $n = 10,000$ , but fewer, 3.16 for  $n = 500$ . Although the three invalid instruments are always jointly selected as invalid for the larger sample sizes, the Lasso<sub>cvse</sub> is substantially biased, the biases being larger than twice the standard deviations. The post-Lasso<sub>cvse</sub> estimator performs best, but is still outperformed by the oracle 2SLS estimator at  $n = 10,000$ . Although the post-Lasso<sub>cvse</sub> estimator has a larger standard deviation than the Lasso<sub>cvse</sub> estimator, it has a smaller bias, rmse, and mad for all sample sizes.

We focus below on the performance of the median and adaptive Lasso estimators for a design with invalid instruments that are stronger than the valid ones, but for comparison we present results for these estimators for this equal strength instruments design in Section B.1 in the supplementary materials, which also includes a more detailed analysis of the differences in performances of the Lasso and post-Lasso estimators in this design.

Table 2 presents estimation results for the same Monte Carlo design as in Table 1, but now with stronger invalid than valid instruments, with  $\tilde{\gamma}_2 = 0.2$  and  $\tilde{\gamma}_1 = 3\tilde{\gamma}_2$ . At these relative values, the necessary condition (16) is not satisfied and the Lasso selection will here select the valid instruments as invalid. Note that the behavior of the oracle 2SLS estimator is the same as in Table 1. In this case,  $\beta + a/\tilde{\gamma}_2 = 0 + 0.2/0.6 = 0.33$ , which is the parameter value estimated by the invalid instruments. From (22), it follows that the SNR is smaller here, with  $\eta^2 = 0.0247$ . The estimation results for the adaptive Lasso are based on

**Table 2.** Estimation results for estimators of  $\beta$ ;  $L = 10, s = 3, \tilde{\gamma}_1 = 3\tilde{\gamma}_2$ .

$\beta$	bias	std dev	rmse	mad	av. # instr selected as invalid [min, max]	freq. all invalid instr selected
<i>n</i> = 500						
Post-Lasso <sub>cv</sub>	0.2696	0.0583	0.2759	0.2718	5.06 [0,9]	0.03
Post-Lasso <sub>cvse</sub>	0.2658	0.0429	0.2692	0.2651	0.45 [0,8]	0
$\hat{\beta}_m$	0.1128	0.0936	0.1466	0.1129		
ALasso <sub>cv</sub>	0.1735	0.0952	0.1979	0.1830	3.73 [0,9]	0.48
Post-ALasso <sub>cv</sub>	0.1324	0.1321	0.1870	0.1591		
ALasso <sub>cvse</sub>	0.2586	0.0420	0.2620	0.2568	0.46 [0,6]	0.04
Post-ALasso <sub>cvse</sub>	0.2428	0.0787	0.2552	0.2568		
<i>n</i> = 2000						
Post-Lasso <sub>cv</sub>	0.3004	0.0308	0.3020	0.3023	8.89 [3,9]	0.01
Post-Lasso <sub>cvse</sub>	0.2910	0.0352	0.2931	0.2932	6.58 [0,9]	0.00
$\hat{\beta}_m$	0.0634	0.0500	0.0808	0.0649		
ALasso <sub>cv</sub>	0.0600	0.0527	0.0798	0.0596	4.42 [3,9]	0.998
Post-ALasso <sub>cv</sub>	0.0360	0.0626	0.0722	0.0442		
ALasso <sub>cvse</sub>	0.1656	0.0489	0.1726	0.1668	3.07 [0,6]	0.89
Post-ALasso <sub>cvse</sub>	0.0281	0.0774	0.0823	0.0348		
<i>n</i> = 10,000						
Post-Lasso <sub>cv</sub>	0.3197	0.0120	0.3199	0.3202	8.97 [8,9]	0
Post-Lasso <sub>cvse</sub>	0.3202	0.0122	0.3204	0.3204	8.70 [7,9]	0
$\hat{\beta}_m$	0.0278	0.0226	0.0358	0.0284		
ALasso <sub>cv</sub>	0.0153	0.0222	0.0270	0.0190	3.92 [3,9]	1
Post-ALasso <sub>cv</sub>	0.0092	0.0253	0.0269	0.0177		
ALasso <sub>cvse</sub>	0.0661	0.0212	0.0694	0.0668	3.02 [3,6]	1
Post-ALasso <sub>cvse</sub>	0.0010	0.0186	0.0187	0.0129		

NOTE: Results from 1000 MC replications;  $a = 0.2$ ;  $\beta = 0$ ;  $\tilde{\gamma}_2 = 0.2, \rho = 0.25$ .

setting  $\nu = 1$ . The resulting estimators are denoted as ‘‘ALasso.’’ As  $L$  is even here, the median is defined as  $\widehat{\beta}_m = (\widehat{\pi}_{[5]} + \widehat{\pi}_{[6]})/2$ , where  $\widehat{\pi}_{[j]}$  is the  $j$ th-order statistic.

The results in Table 2 confirm that, for large sample sizes, the Lasso selects the valid instruments as invalid because of the relative strength of the invalid instruments. The post-ALasso<sub>cvse</sub> estimator does not perform well for  $n = 500$ , but does for the sample sizes of  $n = 2000$ , and  $n = 10,000$ , with results for the latter very similar to the oracle 2SLS results. The Post-ALasso<sub>cv</sub> estimator performs better at  $n = 500$ , as it selects more instruments as invalid with a larger proportion correctly selecting all invalid instruments, although it is outperformed there by the simple median estimator  $\widehat{\beta}_m$ .

### 5.2. Alternative Stopping Rule

The results for the Lasso estimator in Table 1 show that the 10-fold cross-validation method tends to select too many valid instruments as invalid over and above the invalid ones, and that the ad hoc one-standard error rule does improve the selection. The fact that the cross-validation method selects too many variables is well known, see, for example, Buhlmann and van der Geer (2011), who argued that use of the cross-validation method is appropriate for prediction purposes, but that the penalty parameter needs to be larger for variable selection, as achieved by the one-standard error rule. Selecting valid instruments as invalid in addition to correctly selecting the invalid instruments clearly does not lead to an asymptotic bias, but results in a less efficient estimator as compared to the oracle estimator.

We propose a stopping rule for the LARS/Lasso algorithm based on the approach of Andrews (1999) for moment selection, which is particularly well-suited for the IV selection problem. We can use this approach because the number of instruments  $L \ll n$ . This stopping rule is computationally less expensive than cross-validation.

Consider again the oracle model

$$y = \mathbf{d}\beta + \mathbf{Z}_A\alpha_A + \varepsilon = \mathbf{R}_A\theta_A + \varepsilon. \tag{24}$$

Let  $\mathbf{g}_n(\theta_A) = n^{-1}\mathbf{Z}'(y - \mathbf{R}_A\theta_A)$ , and  $\mathbf{W}_n$  a  $k_z \times k_z$  weight matrix, then the oracle generalized method of moments (GMM) estimator is defined as

$$\widehat{\theta}_{A,\text{gmm}} = \arg \min_{\theta_A} \mathbf{g}_n(\theta_A)' \mathbf{W}_n^{-1} \mathbf{g}_n(\theta_A),$$

see Hansen (1982). 2SLS is a one-step GMM estimator, setting  $\mathbf{W}_n = n^{-1}\mathbf{Z}'\mathbf{Z}$ . Given the moment conditions  $E[\mathbf{Z}_i\varepsilon_i] = 0$ , 2SLS is efficient under conditional homoscedasticity,  $E(\varepsilon_i^2|\mathbf{Z}_i) = \sigma_\varepsilon^2$ . Under general forms of conditional heteroscedasticity, an efficient two-step oracle GMM estimator is obtained by setting

$$\mathbf{W}_n = \mathbf{W}_n(\widehat{\theta}_{A,1}) = n^{-1} \sum_{i=1}^n ((y_i - \mathbf{R}'_{A,i}\widehat{\theta}_{A,1})^2 \mathbf{Z}_i \mathbf{Z}_i'),$$

where  $\widehat{\theta}_{A,1}$  is an initial consistent estimator, with a natural choice the 2SLS estimator. Then, under the null that the moment conditions are correct,  $E[\mathbf{Z}_i\varepsilon_i] = 0$ , the Hansen (1982)  $J$ -test statistic and its limiting distribution are given by

$$J_n(\widehat{\theta}_{A,\text{gmm}}) = n \mathbf{g}_n(\widehat{\theta}_{A,\text{gmm}})' \mathbf{W}_n^{-1}(\widehat{\theta}_{A,1}) \mathbf{g}_n(\widehat{\theta}_{A,\text{gmm}}) \xrightarrow{d} \chi^2_{(L-\dim(\mathbf{R}_A))}.$$

For any set  $A^+$ , such that  $A \subset A^+$ , we have that

$$J_n(\widehat{\theta}_{A^+,\text{gmm}}) \xrightarrow{d} \chi^2_{(L-\dim(\mathbf{R}_{A^+}))},$$

whereas for any set  $A^-$ , such that  $A \not\subset A^-$ ,  $J_n(\widehat{\theta}_{A^-\text{,gmm}}) = O_p(n)$ .

Note that the  $J$ -test is a robust score, or Lagrange multiplier, test for testing  $H_0 : \alpha_C = 0$  in the just identified specification

$$y = \mathbf{d}\beta + \mathbf{Z}_B\alpha_B + \mathbf{Z}_C\alpha_C + \varepsilon,$$

where  $\mathbf{Z}_B$  is a  $k_B$  set of instruments included in the model and  $\mathbf{Z}_C$  is any selection of  $L - k_B - 1$  instruments from the  $L - k_B$  set of instruments not in  $\mathbf{Z}_B$ , see, for example, Davidson and MacKinnon (1993, p. 235). This makes clear the link between the  $J$ -test and testing for additional invalid instruments of the form as specified in model (3).

We can now combine the LARS/Lasso algorithm with the Hansen  $J$ -test, which is a directed downward testing procedure in the terminology of Andrews (1999). Compute  $J_n(\widehat{\theta}_{\widehat{A}_n^{[j]}})$  at every LARS/Lasso step  $j = 0, 1, 2, \dots$ , where  $\widehat{A}_n^{[0]} = \emptyset$  and  $\|\widehat{A}_n^{[1]}\|_0 = 1$ , compare it to a corresponding critical value  $\zeta_{n,L-k}$  of the  $\chi^2_{(L-k)}$  distribution, where  $k = \dim(\mathbf{R}_{\widehat{A}_n^{[j]}})$ . We then select the model with the largest degrees of freedom  $L - k$ , for which  $J_n(\widehat{\theta}_{\widehat{A}_n^{[j]}})$  is smaller than the critical value. If two models of the same dimension pass the test, which can happen with a Lasso step, the model with the smallest value of the  $J$ -test gets selected. (If there is no empirical evidence at all for any invalid instruments, that is, if  $J_n(\widehat{\theta}_{\widehat{A}_n^{[0]}})$  is smaller than its corresponding critical value, then the model with all instruments as valid gets selected.) Clearly, this approach is a post-Lasso approach, where the LARS/Lasso algorithm is used purely for selection of the invalid instruments. For consistent model selection, the critical values  $\zeta_{n,L-k}$  need to satisfy

$$\zeta_{n,L-k} \rightarrow \infty \text{ for } n \rightarrow \infty, \text{ and } \zeta_{n,L-k} = o(n), \tag{25}$$

see Andrews (1999).

As the oracle model is on the adaptive LARS/Lasso path in large samples, this approach leads to consistent selection,  $\lim_{n \rightarrow \infty} P(\widehat{A}_{n,\text{ah}}^{\text{ad}} = A) = 1$ , the subscript ah standing for Andrews/Hansen. As Guo et al. (2018, Theorem 2) showed, consistent selection implies that the limiting distribution of the 2SLS estimator  $\widehat{\beta}_{A_{n,\text{ah}}^{\text{ad}}}$  is the same as that of the oracle 2SLS estimator, that is,  $\sqrt{n}(\widehat{\beta}_{A_{n,\text{ah}}^{\text{ad}}} - \beta) \xrightarrow{d} N(0, \sigma_{\beta_{\text{or}}}^2)$ . We call  $\widehat{\beta}_{A_{n,\text{ah}}^{\text{ad}}}$  the post-ALasso<sub>ah</sub> estimator. This approach also leads to consistent selection along the Lasso path when the irrepresentable condition (15) holds, resulting in oracle properties of the resulting post-Lasso<sub>ah</sub> estimator.

Let  $\zeta_{n,L-k} = \chi^2_{L-k}(p_n)$  be the  $1 - p_n$  quantile of the  $\chi^2_{L-k}$  distribution. Here,  $p_n$  is the  $p$ -value of the test. This combination of the Andrews/Hansen method with the LARS/Lasso steps therefore results in having to choose a  $p$ -value  $p_n$  instead of a penalty parameter  $\lambda_n$ . Keeping  $n$  fixed, choosing a large value for  $p_n$  leads to selecting a larger set as invalid instruments as compared to choosing a smaller value for  $p_n$ . Finite sample inference will not be straightforward, as this method is essentially a sequential approach where the model at step  $j$  is only considered when the model at step  $j - 1$  is rejected. Using the consistent selection properties, we will investigate the behavior of the Wald test in the



**Table 3.** Results for post-(A)Lasso<sub>ah</sub> 2SLS estimators for  $\beta$ ;  $L = 10, s = 3$ .

	$n$	bias	std dev	rmse	mad	av. # instr selected as invalid [min, max]	freq. all invalid instr selected
$\tilde{\gamma}_1 = \tilde{\gamma}_2$	500	0.0896	0.1252	0.1539	0.1007	2.56 [0,5]	0.391
	2000	0.0055	0.0430	0.0434	0.0286	3.02 [3,5]	1
	10,000	0.0009	0.0186	0.0186	0.0129	3.02 [3,5]	1
$\tilde{\gamma}_1 = 3\tilde{\gamma}_2$	500	0.2172	0.1091	0.2431	0.2471	0.86 [0,5]	0.07
	2000	0.0173	0.0677	0.0699	0.0303	3.05 [1,5]	0.93
	10,000	0.0008	0.0186	0.0186	0.0129	3.01 [3,5]	1

NOTE: Results from 1000 MC replications;  $\beta = 0$ ;  $a = 0.2$ ;  $\tilde{\gamma}_2 = 0.2$ ;  $\rho = 0.25$ .

next section and find in our simulation designs that this method performs quite well and similar to the ALasso<sub>cvse</sub> method in the unequal instrument strength design, and also performs well using the post-Lasso<sub>ah</sub> estimator for the equal strength design.

Table 3 presents the estimation results using this stopping rule as a selection device for the Lasso estimator for the design with equal strength instruments and the adaptive Lasso estimator for the unequal instrument strength design, as in Tables 1 and 2. We denote the resulting 2SLS estimators as "post-(A)Lasso<sub>ah</sub>." The  $p$ -values here are chosen as  $p_n = 0.1/\ln(n)$ , following Belloni et al. (2012), and are equal to 0.0161, 0.0132, and 0.0109 for  $n$  equal to 500, 2000, and 10,000, respectively. For the equal strength design, the ah approach selects too few invalid instruments for  $n = 500$ , resulting in an upward bias, with bias, std dev, rmse, and mad very similar to those of the post-Lasso<sub>cvse</sub> estimator in Table 1. For  $n = 2000$  and  $n = 10,000$ , this post-Lasso procedure performs well with properties very similar to that of the oracle 2SLS estimator, and with smaller bias, rmse, and mad than the post-Lasso<sub>cvse</sub> method. For the unequal strength design, for  $n = 10,000$  the results are virtually identical to those of the oracle and post-ALasso<sub>cvse</sub> estimators, whereas the post-ALasso<sub>ah</sub> estimator performs better in terms of bias, std dev, rmse, and mad than the post-ALasso<sub>cvse</sub> estimator when  $n = 2000$ . Again, when  $n = 500$ , the method does not select the invalid instruments.

### 5.3. Inference

From the limiting distribution result (21), a simple approach to estimating the asymptotic variance of the post-ALasso 2SLS estimator for  $\beta$  is by calculating the standard 2SLS variance estimator. The post-ALasso 2SLS estimator is given by

$$\hat{\beta}_{ad,post}^{(n)} = (\hat{\mathbf{d}}' \mathbf{M}_{Z_{\hat{A}_{ad,n}}} \hat{\mathbf{d}})^{-1} \hat{\mathbf{d}}' \mathbf{M}_{Z_{\hat{A}_{ad,n}}} \mathbf{y}$$

and its estimated variance given by

$$\widehat{\text{var}} \left( \hat{\beta}_{ad,post}^{(n)} \right) = \hat{\sigma}_\varepsilon^2 \left( \hat{\mathbf{d}}' \mathbf{M}_{Z_{\hat{A}_{ad,n}}} \hat{\mathbf{d}} \right)^{-1}, \tag{26}$$

where  $\hat{\sigma}_\varepsilon^2 = \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}} / n$ ,  $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{d}} \hat{\beta}_{ad,post}^{(n)} - \mathbf{Z}_{\hat{A}_{ad,n}} \hat{\boldsymbol{\alpha}}_{\hat{A}_{ad,n},post}^{(n)}$ . Under the conditions of Proposition 3, the standard assumptions and conditional homoscedasticity,  $n \widehat{\text{var}} \left( \hat{\beta}_{ad,post}^{(n)} \right) \xrightarrow{p} \sigma_{\beta_{or}}^2$ . A standard robust version, robust to general forms of heteroscedasticity, is

given by

$$\begin{aligned} \widehat{\text{var}}_r \left( \hat{\beta}_{ad,post}^{(n)} \right) &= \left( \hat{\mathbf{d}}' \mathbf{M}_{Z_{\hat{A}_{ad,n}}} \hat{\mathbf{d}} \right)^{-1} \hat{\mathbf{d}}' \mathbf{M}_{Z_{\hat{A}_{ad,n}}} \hat{\mathbf{H}} \mathbf{M}_{Z_{\hat{A}_{ad,n}}} \hat{\mathbf{d}} \left( \hat{\mathbf{d}}' \mathbf{M}_{Z_{\hat{A}_{ad,n}}} \hat{\mathbf{d}} \right)^{-1}, \end{aligned}$$

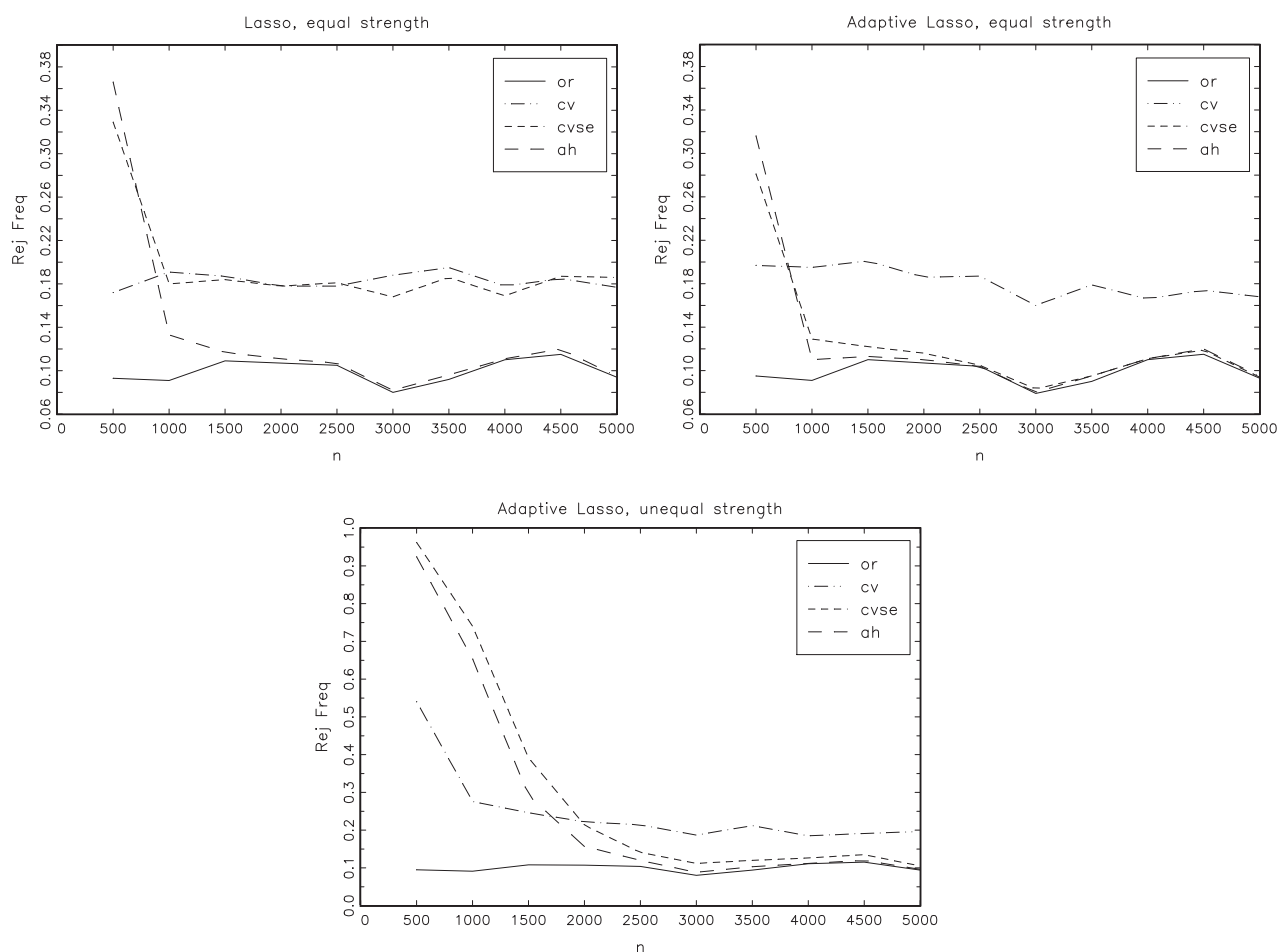
where  $\hat{\mathbf{H}}$  is an  $n \times n$  diagonal matrix with diagonal elements  $\hat{H}_{ii} = \hat{\varepsilon}_i^2$ , for  $i = 1, \dots, n$ . The robust Wald test for the null  $H_0 : \beta = \beta_0$  is then given by

$$W_{\beta,r} = \frac{\left( \hat{\beta}_{ad,post}^{(n)} - \beta_0 \right)^2}{\widehat{\text{var}}_r \left( \hat{\beta}_{ad,post}^{(n)} \right)}.$$

From the results for the post-ALasso<sub>cvse</sub> and post-ALasso<sub>ah</sub> estimators for the unequal strength instruments design as presented in Tables 2 and 3, respectively, one would expect this approach to work well for the large sample case,  $n = 10,000$ , as there the estimation results are very close to those of the oracle 2SLS estimator. The robust Wald test for the null  $H_0 : \beta = 0$ , the true value of  $\beta$ , at the 10% level for  $n = 10,000$  has a rejection frequency of 9.3% and 9.2% for the post-ALasso<sub>cvse</sub> and post-ALasso<sub>ah</sub> estimators, respectively, very close to that of the robust Wald test based on the oracle 2SLS estimator, which has a rejection frequency of 9.0%.

For the equal strength instruments design, we perform the same analysis for the post-Lasso estimators. Figures 1(a)–1(c) shows the performance of the robust Wald test  $W_{\beta,r}$ , its rejection frequency at the 10% level, as a function of the sample size in steps of 500,  $n = 500, 1000, \dots, 5000$ . Figures 1(a) and 1(b) shows the results for the post-Lasso and post-ALasso estimators for the equal strength instruments design. Figure 1(c) shows the results for the post-ALasso estimators for the unequal strength instruments design.

Figure 1(a) clearly shows that the Lasso<sub>cv</sub> and Lasso<sub>cvse</sub> procedures do not result in consistent selection and the resulting post-Lasso estimators do not have oracle properties. The Wald test rejection frequencies remain constant for increasing sample size and larger than those of the oracle estimator. In contrast, the post-Lasso<sub>ah</sub> estimator behaves very similar to the oracle estimator in this design from  $n = 1500$  onward. Figure 1(b) shows that both the post-ALasso<sub>cvse</sub> and post-ALasso<sub>ah</sub> behave like the oracle estimator, again from  $n = 1500$  onward in this design. The results in Figure 1(c) show that for the unequal instruments



**Figure 1.** (a–c) Rejection frequencies of robust Wald tests for  $H_0: \beta = 0$  at 10% level as a function of sample size, in steps of 500. Equal strength instruments design, Post-Lasso in (a), Post-ALasso in (b). Unequal strength instruments design, Post-ALasso in (c). Based on 1000 MC replications for each sample size.

strength design considered here, the performances of the post-adaptive Lasso estimators are far from that of the oracle estimator in small samples, as expected from the results in Tables 2 and 3. The post-ALasso<sub>ah</sub> behaves like the oracle estimator here from  $n = 4000$  onward, with the post-ALasso<sub>cvse</sub> estimator behaving similarly, but having a larger rejection frequency for all sample sizes considered here that are less than  $n = 5000$ .

The results in Tables 1–3 and Figures 1(a)–1(c) show clearly that the information content in the data, given the parameter values chosen here, is insufficient at  $n = 500$  for the (adaptive) Lasso procedures to correctly select the invalid instruments and hence the resulting estimators have poor properties, far removed from those of the oracle estimator. At these levels of information, the ALasso<sub>cv</sub> estimator is actually the preferred estimator as it counteracts the selection of too few invalid instruments of the ALasso<sub>cvse</sub> and ALasso<sub>ah</sub> estimators. We further explore how the performances of the estimators depend on the information content of the data-generating process in Section B.2 in the supplementary materials.

## 6. The Effect of BMI on Diastolic Blood Pressure Using Genetic Markers as Instruments

We use data on 105,276 individuals from the UK Biobank and investigate the effect of BMI on diastolic blood pressure (DBP). See Sudlow et al. (2015) for further information on the UK

Biobank. We use 96 single nucleotide polymorphisms (SNPs) as instruments for BMI as identified in independent GWAS studies, see Locke et al. (2015).

With Mendelian randomization studies, the SNPs used as potential instruments can be invalid for various reasons, such as linkage disequilibrium, population stratification, and horizontal pleiotropy, see, for example, von Hinke et al. (2016) or Davey Smith and Hemani (2014). For example, an SNP has pleiotropic effects if it not only affects the exposure but also has a direct effect on the outcome. While we guard against population stratification by considering only white European origin individuals in our data, the use of the Lasso methods can be extremely useful here to identify the SNPs with direct effects on the outcome and to estimate the causal effect of BMI on diastolic blood pressure taking account of this.

Because of skewness, we log-transformed both BMI and DBP. The linear model specification includes age, age<sup>2</sup>, and sex, together with 15 principal components of the genetic relatedness matrix as additional explanatory variables. Table 4 presents the estimation results for the causal effect parameter, which is here the percentage change in DBP due to a 1% change in BMI. As  $p$ -value for the Hansen test-based procedures we take again  $0.1/\ln(n) = 0.0086$ .

The OLS estimate of the causal parameter is equal to 0.206 (s.e. 0.003), whereas the 2SLS estimate treating all 96 instruments as valid is much smaller at 0.087 (s.e. 0.016), with a 95%

**Table 4.** Estimation results, the effect of ln(BMI) on ln(DBP).

	estimate	rob st err	# instr selected as invalid	p-value J-test
OLS	0.206	0.003		
2SLS	0.087	0.016	0	0.0000
Lasso <sub>cv</sub>	0.126		56	
Post-Lasso <sub>cv</sub>	0.145	0.033		1.0000
Lasso <sub>cvse</sub>	0.111		20	
Post-Lasso <sub>cvse</sub>	0.142	0.020		0.6435
Post-Lasso <sub>ah</sub>	0.122	0.018	12	0.0123
median, $\hat{\beta}_m$	0.148			
ALasso <sub>cv</sub>	0.158		54	
Post-ALasso <sub>cv</sub>	0.161	0.029		1.0000
ALasso <sub>cvse</sub>	0.131		17	
Post-ALasso <sub>cvse</sub>	0.151	0.019		0.4091
Post-ALasso <sub>ah</sub>	0.163	0.018	11	0.0102

NOTE: Sample size  $n = 105,276$ ;  $L = 96$ .

confidence interval of [0.056, 0.118]. The  $J$ -test, however, rejects the null that all the instruments are valid. The Lasso<sub>cv</sub> estimator identifies a large number of 56 instruments as invalid and the Lasso<sub>cv</sub> estimate is equal to 0.126, the post-Lasso<sub>cv</sub> estimate is equal to 0.145. The Lasso<sub>cvse</sub> procedure identifies 20 instruments as invalid and the Lasso<sub>cvse</sub> estimate is equal to 0.111. The post-Lasso<sub>cvse</sub> estimate is larger and equal to 0.142, which is in line with our findings above that the Lasso estimator is biased toward the 2SLS estimator that treats all instruments as valid due to shrinkage. The post-Lasso<sub>ah</sub> procedure selects a subset of 12 instruments as invalid, and the post-Lasso<sub>ah</sub> parameter estimate is equal to 0.122.

The median estimate  $\hat{\beta}_m$  is equal to 0.148. Using this estimate for the adaptive Lasso results in the cv method selecting 54 instruments as invalid and the cvse method selecting 17 instruments as invalid. The adaptive Lasso<sub>ah</sub> method selects a subset of 11 instruments as invalid. The post-ALasso<sub>cv</sub>, post-ALasso<sub>cvse</sub>, and post-ALasso<sub>ah</sub> estimates are equal to 0.161, 0.151, and 0.163, respectively, with the 95% confidence intervals of the post-ALasso<sub>cvse</sub> and post-ALasso<sub>ah</sub> estimators given by [0.113, 0.189] and [0.127, 0.198], respectively. These results indicate that the OLS estimator is less confounded than suggested by the 2SLS estimation results using all 96 instruments as valid instruments.

The strongest potential instrument is the FTO SNP. For all Lasso estimators in Table 4, it is selected as an invalid instrument. The value for  $\hat{\pi}_{\text{FTO}} = -0.009$ , that is, negative, which is contrary to the direction of the found causal effect.

The  $F$ -test statistic for  $H_0 : \gamma_2 = 0$  for the model resulting from the ALasso<sub>ah</sub> procedure is equal to 18.21 with the associated estimate of the concentration parameter equal to 1547.81. The  $F$ -test result indicates that the 2SLS estimator may have some many weak instruments bias, see Stock and Yogo (2005). However, the LIML (limited information maximum likelihood) estimator in this model is very similar to the 2SLS estimator and is equal to 0.159 (s.e. 0.019), indicating that there is not a many weak instruments problem here, see Davies et al. (2015).

## 7. Conclusions

Instrumental variables estimation is a well-established procedure for the identification and estimation of causal effects of exposures on outcomes where the observed relationships

are confounded by nonrandom selection of exposure. The main identifying assumption is that the instruments satisfy the exclusion restriction, that is, they only affect the outcomes through their relationship with the exposure. In an important contribution, Kang et al. (2016) showed that the Lasso method for variable selection can be used to select invalid instruments in linear IV models, even though there is no prior knowledge about which instruments are valid.

We have shown here that, even under the sufficient condition for identification that less than 50% of the instruments are invalid, the Lasso selection may select the valid instruments as invalid if the invalid instruments are relatively strong, that is, the case where an invalid instrument explains more of the exposure variance than a valid instrument. Consistent selection of invalid instruments also depends on the correlation structure of the instruments.

We show that a median estimator is consistent when less than 50% of the instruments are invalid, and its consistency does not depend on the relative strength of the instruments or their correlation structure. This initial consistent estimator can be used for the adaptive Lasso estimator of Zou (2006) and we show that it performs well for larger sample sizes/information settings in our simulations. This adaptive Lasso estimator has the same limiting distribution as the oracle 2SLS estimator, and solves the inconsistency problem of the Lasso method when the relative strength of the invalid instruments is such that the Lasso method selects the valid instruments as invalid.

## Supplementary Materials

The document contains the proofs of Proposition 1 and Theorem 1 in Section A, and further simulation results and discussions in Section B.

The Stata module “SIVREG” implements the post-ALasso<sub>ah</sub> method. Further details and documentation are provided in Farbmacher (2017).

## Acknowledgments

Helpful comments were provided by Kirill Evdokimov, Chirok Han, Whitney Newey, Hyunseung Kang, Chris Skeels, Martin Spindler, Jonathan Temple, Ian White, and seminar participants at Amsterdam, Bristol, Lausanne, Monash, Oxford, Princeton, Seoul, Sydney, the RES Conference Brighton, the Info-Metrics Conference Cambridge, and the UK Causal Inference Meeting London.

## Funding

This research was partly funded by the Medical Research Council, MC\_UU\_12013/1, MC\_UU\_12013/9, and MC\_UU\_00011/1. Neil Davies further acknowledges support from the Economics and Social Research Council via a Future Leaders Grant, ES/N000757/1. Helmut Farbmacher acknowledges funding from the Fritz Thyssen Stiftung.

## References

- Andrews, D. W. K. (1999), “Consistent Moment Selection Procedures for Generalized Method of Moments Estimation,” *Econometrica*, 67, 543–564. [1339,1340,1346]
- Angrist, J. D., and Krueger, A. B. (1991), “Does Compulsory School Attendance Affect Schooling and Earnings?” *Quarterly Journal of Economics*, 106, 979–1014. [1339]
- Bekker, P. A. (1994), “Alternative Approximations to the Distributions of Instrumental Variable Estimators,” *Econometrica*, 62, 657–681. [1341]

- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369–2429. [1339,1345,1347]
- Belloni, A., and Chernozhukov, V. (2013), "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19, 521–547. [1345]
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650. [1339]
- Bowden, J., Smith, G. D., Burgess, S. (2015), "Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression," *International Journal of Epidemiology*, 44, 512–525. [1340]
- Bühlmann, P., and van der Geer, S. (2011), *Statistics for High-Dimensional Data. Methods Theory and Applications (Springer Series in Statistics)*, Heidelberg: Springer. [1344,1346]
- Burgess, S., Small, D. S., and Thompson, S. G. (2017), "A Review of Instrumental Variable Estimators for Mendelian Randomization," *Statistical Methods in Medical Research*, 26, 2333–2355. [1339]
- Cheng, X., and Liao, Z. (2015), "Select the Valid and Relevant Moments: An Information-based LASSO for GMM with Many Moments," *Journal of Econometrics*, 186, 443–464. [1339]
- Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," *American Economic Review*, 105, 486–490. [1339]
- Clarke, P. S., and Windmeijer, F. (2012), "Instrumental Variable Estimators for Binary Outcomes," *Journal of the American Statistical Association*, 107, 1638–1652. [1339]
- Davey Smith, G., and Hemani, G. (2014), "Mendelian randomization: Genetic Anchors for Causal Inference in Epidemiological Studies," *Human Molecular Genetics*, 23, R89–R98. [1348]
- Davidson, R., and MacKinnon, J. G. (1993), *Estimation and Inference in Econometrics*, Oxford: Oxford University Press. [1346]
- Davies, N. M., von Hinke Kessler Scholder, S., Farbmacher, H., Burgess, S., Windmeijer, F., and Smith, G. D. (2015), "The Many Weak Instruments Problem and Mendelian Randomization," *Statistics in Medicine*, 34, 454–468. [1349]
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–451. [1341,1345]
- Farbmacher, H. (2017), *SIVREG: Stata Module to Perform Adaptive Lasso with Some Invalid Instruments*, Statistical Software Components S458394, Boston College Department of Economics. [1349]
- Greenland, S. (2000), "An Introduction to Instrumental Variables for Epidemiologists," *International Journal of Epidemiology*, 29, 722–729. [1339]
- Guo, Z., Kang, H., Cai, T., and Small, D. (2018), "Confidence Intervals for Causal Effects with Invalid Instruments using Two-Stage Hard Thresholding with Voting," *Journal of the Royal Statistical Society, Series B*, 80, 793–815. [1346]
- Han, C., (2008), "Detecting Invalid Instruments using  $L_1$ -GMM," *Economics Letters*, 101, 285–287. [1343]
- Hansen, C., Hausman, J., and Newey, W. K. (2008), "Estimation with Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26, 398–422. [1341]
- Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [1339,1346]
- Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction (Springer Series in Statistics, 2nd ed.)*, New York: Springer Science and Business Media. [1345]
- Imbens, G. W. (2014), "Instrumental Variables: An Econometrician's Perspective," *Statistical Science*, 29, 323–358. [1339]
- Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016), "Instrumental Variables Estimation with some Invalid Instruments and its Application to Mendelian Randomization," *Journal of the American Statistical Association*, 111, 132–144. [1339,1340,1349]
- Kolesar, M., Chetty, R., Friedman, J., Glaeser, E., Imbens, G. W. (2015), "Identification and Inference with Many Invalid Instruments," *Journal of Business and Economic Statistics*, 33, 474–484. [1340]
- Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008), "Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology," *Statistics in Medicine*, 27, 1133–1163. [1339]
- Liao, Z. (2013), "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection," *Econometric Theory*, 29, 857–904. [1339]
- Lin, W., Feng, R., and Li, H. (2015), "Regularization Methods for High-Dimensional Instrumental Variables Regression With an Application to Genetical Genomics," *Journal of the American Statistical Association*, 110, 270–288. [1339]
- Locke, A. E., (2015), "Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology," *Nature*, 518, 197–206. [1348]
- Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *Annals of Statistics*, 34, 1436–1462. [1340]
- Newey, W. K., and Windmeijer, F. (2009), "Generalized Methods of Moments with Many Weak Moment Conditions," *Econometrica*, 77, 687–719. [1341]
- Rothenberg, T. J. (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics," in *Handbook of Econometrics (Vol. 2)*, eds. Z. Griliches, and M. D. Intriligator, Amsterdam: North Holland, pp. 881–935. [1344]
- Sargan, J. D. (1958), "The Estimation of Economic Relationships Using Instrumental Variables," *Econometrica*, 26, 393–415. [1339]
- Staiger, D., and Stock, J. H. (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586. [1339]
- Stock, J. H., and Yogo, M. (2005), "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, eds. D. W. K. Andrews, and J. H. Stock, New York: Cambridge University Press, pp. 80–108. [1339,1349]
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015), "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," *PLoS Medicine*, 12, e1001779. [1348]
- von Hinke, S., Smith, G. D., Lawlor, D. A., Propper, C., and Windmeijer, F. (2016), "Genetic Markers as Instrumental Variables," *Journal of Health Economics*, 45, 131–148. [1348]
- Wainwright, M. J. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using  $\ell_1$ -Constrained Quadratic Programming (Lasso)," *IEEE Transactions on Information Theory*, 55, 2183–2202. [1340]
- Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [1340,1342]
- Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1340,1341,1342,1343,1349]