Selecting (In)Valid Instruments for Instrumental Variables Estimation

Frank Windmeijer[a,e], Helmut Farbmacher[b], Neil Davies[c,e]
George Davey Smith[c,e], Ian White[d]

[a]Department of Economics
University of Bristol, UK
[b]Center for the Economics of Aging
Max Planck Institute Munich, Germany
[c]School of Social and Community Medicine
University of Bristol, UK
[d]MRC Biostatistics Unit, Cambridge, UK
[e]MRC Integrative Epidemiology Unit, Bristol, UK

October 2015, Preliminary, please do not quote.

Abstract

We investigate the behaviour of the Lasso for identifying invalid instruments in linear models, as proposed recently by Kang, Zhang, Cai and Small (2015, Journal of the American Statistical Association, in press), where invalid instruments are such that they enter the model as explanatory variables. We show that for this setup, the Lasso may not select all invalid instruments in large samples if they are relatively strong. Consistent selection also depends on the correlation structure of the instruments. We propose an initial estimator that is consistent when less than 50% of the instruments are invalid, but its consistency does not depend on the relative strength of the instruments or their correlation structure. This estimator can therefore be used for adaptive Lasso estimation. We develop an alternative selection method based on Hansen's J-test of overidentifying restrictions, but limiting the number of models to be evaluated. This latter selection procedure is consistent under weaker conditions on the number of invalid instruments and is aligned with the general identification result for this particular model.

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 randomised clinical trials if there is selective participant non-compliance. An instrumental variable (IV) can be used to solve the problem of non-ignorable selection. In order 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'. Examples of instrumental variables are quarter-of-birth for educational achievement to determine its effect on wages, see Angrist and Krueger (1991), randomisation of patients to treatment as an instrument for actual treatment when there is non-compliance, see e.g. Greenland (2000), and Mendelian randomisation studies use IVs based on genetic information, see e.g. Lawlor et al. (2008). For recent reviews and further examples see e.g. Clarke and Windmeijer (2012), Imbens (2014), Burgess et al. (2015) and Kang et al. (2015).

Whether instruments are relevant can be tested from the observed association between exposure and instruments. The effects of 'weak instruments', i.e. 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), leading to critical values for the simple F-test statistic for testing the null of weak instruments, derived by Stock and Yogo (2005).

In this paper we consider violations of the exclusion condition of the instruments, following closely the setup of Kang et al. (2015) 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. (2015) propose 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. (2015) do not assume any prior knowledge of

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 the Kang et al. (2015) setup and requires a large number of model evaluations, which has a negative impact on the performance of the selection procedure.

In this paper we assess the performance of the Kang et al. (2015) Lasso type selection and estimation procedure. By evaluating the LARS algorithm of Efron et al. (2004), using large sample asymptotics, we show that the Lasso method may not consistently select the correct invalid instruments. Consistent selection depends on the relative strength of the instruments and/or the instrument correlation structure, even when less than 50% of the instruments are invalid. We show that under the latter condition a simple median type estimator is a consistent estimator for the parameters in the model, independent of the relative strength of the instruments or their correlation structure. This estimator can therefore be used for an adaptive Lasso procedure as proposed by Zou (2006). We also show that the specific model structure of Kang et al. (2015) makes upward and downward testing procedures as proposed by Andrews (1999) feasible, as the number of models to be evaluated is limited by the model structure. These latter procedures result in consistent model selection under weaker conditions on the number of invalid instruments, with the necessary condition being that the valid instruments form the largest group, where a group of instruments is formed by instruments giving the same value of the estimate of the causal effect.[1]

Instrument strength is very likely to vary by instruments, so it will be important to use our proposed estimators and selection methods for assessing instrument validity. In Mendelian randomization studies it is clear that genetic markers have differential impacts on exposures from examining the results from genome wide association studies. The issue of correlation between genetic markers seems less of an issue as genes are generally independently distributed by Mendel's second law.

Finally, another strand of the literature focuses on instrument selection in potentially

[1] Bowden et al. (2015) and Kolesar et al. (2015) allow for all instruments to be invalid and show that the causal effect can be consistently estimated if the number of instruments increases with the sample size under the relatively strong assumption of uncorrelatedness of the instrument strength and their direct effects on the outcome variable.

high-dimensional settings, see e.g. Belloni et al. (2012) and Lin et al. (2015). Here 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. Belloni et al. (2012) propose a new method for selecting the Lasso penalty parameter. We analyse its behaviour for the Lasso selection method in cases where this method consistently selects the instruments. As our setting is that of a fixed number of potential instruments, we show that simply using Hansen's J-test performs as well, and that the so-called post-Lasso selection IV estimator has better finite sample properties than the Lasso estimator.

2 Model and sisVIVE Estimator

We follow Kang, Zhang, Cai and Small (2015) (KZCS from now on), who considered the following potential outcomes model. The 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 resulting in the following potential outcomes model

$$\begin{aligned} Y_i^{(d',\mathbf{z}')} - y_i^{(d,\mathbf{z})} &= (\mathbf{z}' - \mathbf{z})\,\phi + (d' - d)\,\beta \quad (1) \\ E\left[Y_i^{(0,0)} | \mathbf{Z}_{i.}\right] &= \mathbf{Z}_{i.}'\psi \quad (2) \end{aligned}$$

where ϕ measures the direct effect of $\mathbf{z}$ on Y, and ψ represents the effect of confounders in the relationship between $\mathbf{Z}_{i.}$ and $Y_i^{(0,0)}$.

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

$$Y_i = D_i\beta + \mathbf{Z}_{i.}'\alpha + \varepsilon_i \quad (3)$$

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

$$\varepsilon_i = Y_i^{(0,0)} - E\left[Y_i^{(0,0)} | \mathbf{Z}_{i.}\right]$$

and hence $E\left[\varepsilon_i | \mathbf{Z}_{i.}\right] = 0$. The KZCS definition of a valid instrument is then given as follows: Instrument j, $j \in \{1, ..., L\}$, 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 β 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. Let $\mathbf{Z}_{sel}$ be a subset of instruments included in the equation, and let $\mathbf{R} = \left[\begin{array}{cc} \mathbf{d} & \mathbf{Z}_{sel} \end{array} \right]$. The standard Instrumental Variables, or Two-Stage Least Squares (2SLS), estimator is then given by

$$\widehat{\theta} = \left(\begin{array}{c} \widehat{\beta} \\ \widehat{\alpha}_{sel} \end{array} \right) = \left(\mathbf{R}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{R} \right)^{-1} \mathbf{R}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{y}. \tag{4}$$

Let $\widehat{\mathbf{d}} = \mathbf{P}_Z \mathbf{d}$, $\mathbf{P}_Z = \mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'$, then $\widehat{\theta}$ is equivalent to the OLS estimator in the model

$$Y_i = \widehat{D}_i \beta + \mathbf{Z}'_{sel,i.} \alpha_{sel} + \xi_i$$

and hence

$$\begin{aligned} \widehat{\alpha}_{sel} &= \left(\mathbf{Z}'_{sel} \mathbf{M}_{\widehat{d}} \mathbf{Z}_{sel} \right)^{-1} \mathbf{Z}'_{sel} \mathbf{M}_{\widehat{d}} \mathbf{y} \\ &= \left(\mathbf{Z}'_{sel} \mathbf{M}_{\widehat{d}} \mathbf{Z}_{sel} \right)^{-1} \mathbf{Z}'_{sel} \mathbf{M}_{\widehat{d}} \mathbf{P}_Z \mathbf{y}. \end{aligned} \tag{5}$$

where $\mathbf{M}_{\widehat{d}} = \mathbf{I}_n - \mathbf{P}_{\widehat{d}}$, where $\mathbf{I}_n$ is the identity matrix of order n.

$\widehat{\mathbf{d}}$ is the linear projection of $\mathbf{d}$ on $\mathbf{Z}$. If we define $\widehat{\gamma} = \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{d}$, then $\widehat{\mathbf{d}} = \mathbf{Z}\widehat{\gamma}$, or $\widehat{D}_i = \mathbf{Z}'_{i.} \widehat{\gamma}$, and we specify

$$D_i = \mathbf{Z}'_{i.} \gamma + v_i, \tag{6}$$

where $\gamma = E\left[\mathbf{Z}_{i.} \mathbf{Z}'_{i.} \right]^{-1} E\left[\mathbf{Z}_{i.} D_i \right]$, and hence $E\left[\mathbf{Z}_{i.} v_i \right] = 0$. Further, as in KZCS, let $\Gamma = E\left[\mathbf{Z}_{i.} \mathbf{Z}'_{i.} \right]^{-1} E\left[\mathbf{Z}_{i.} Y_i \right] = \gamma\beta + \alpha$. Clearly, both γ and Γ can be consistently estimated under standard assumptions. Assuming that $\gamma_j \neq 0 \ \forall j$, then define π_j as

$$\pi_j \equiv \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j}. \tag{7}$$

Theorem 1 in KZCS states the conditions under which, given knowledge of γ and Γ, a unique solution exists for values of β and α_j. A necessary and sufficient condition to identify β and the α_j is then that the valid instruments form the largest group, where instruments form a group if they have the same value of π. Corollary 1 in KZCS then states a sufficient condition for identification. Let s be the number of invalid instruments, then if $s < L/2$, the parameters are identified as then clearly the largest group is formed by the valid instruments.

In model (3), some elements of α 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 non-zero α. KZCS propose to estimate the parameters α and β by using ℓ_1 penalisation on α and to minimise

$$\left(\widehat{\alpha}^{\lambda}, \widehat{\beta}^{\lambda}\right) = \arg\min_{\alpha,\beta} \frac{1}{2}\|\mathbf{P}_Z\left(\mathbf{y} - \mathbf{d}\beta - \mathbf{Z}\alpha\right)\|_2^2 + \lambda\|\alpha\|_1, \tag{8}$$

where the ℓ_1 norm $\|\alpha\|_1 = \sum_j |\alpha_j|$ and the squared ℓ_2 norm is $(\mathbf{y} - \mathbf{d}\beta - \mathbf{Z}\alpha)' \mathbf{P}_Z (\mathbf{y} - \mathbf{d}\beta - \mathbf{Z}\alpha)$. This method is closely related to the Lasso, and the regularization parameter λ determines the sparsity of the vector $\widehat{\alpha}^{\lambda}$. From (5), a fast two-step algorithm is proposed that runs as follows. For a given λ solve

$$\widehat{\alpha}^{\lambda} = \arg\min_{\alpha} \frac{1}{2}\|\mathbf{M}_{\widehat{d}}\mathbf{P}_Z\mathbf{y} - \mathbf{M}_{\widehat{d}}\mathbf{Z}\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{9}$$

and estimate $\widehat{\beta}^{\lambda}$ by

$$\widehat{\beta}^{\lambda} = \frac{\widehat{\mathbf{d}}'\left(\mathbf{y} - \mathbf{Z}\widehat{\alpha}^{\lambda}\right)}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}}. \tag{10}$$

In order to find $\widehat{\alpha}^{\lambda}$ in (9), the Lasso modification of the LARS algorithm of Efron, Hastie, Johnstone and Tibshirani (2004) can be used and KZCS have developed an R-routine for this purpose and call the resulting estimator *sisVIVE* (some invalid and some valid IV estimator), where the regularisation parameter λ is obtained by cross-validation. Following the notation of Zou (2006), let A be the set of invalid instruments, $A = \{j : \alpha_j \neq 0\}$. Let $A_n = \left\{j : \widehat{\alpha}_j^{\lambda} \neq 0\right\}$. We will first investigate under what conditions the Lasso method consistently selects invalid instruments such that $\lim_{n\to\infty} P\left(A_n = A\right) = 1$, or for a weaker version, such that $\lim_{n\to\infty} P\left(A_n \supseteq A\right) = 1$.

An important difference with the standard Lasso approach for linear models is that the matrix of explanatory variables $\mathbf{M}_{\widehat{d}}\mathbf{Z}$ in (9) is not full rank, but its rank is equal to $L-1$. Whereas $\lambda = 0$ would simply include all regressors in the standard linear model and the resulting OLS estimator is consistent, setting $\lambda = 0$ in (9) does not lead to a unique 2SLS estimator, as all models with $L-1$ instruments included as invalid would result in a residual correlation of 0 and hence $\lambda = 0$. Therefore the LARS algorithm has to start from a model without any instruments included as invalid, and at the last LARS/Lasso step one instrument is excluded from the model, i.e. treated as valid. When $L-1$ instruments have been selected as invalid and included in the model, the resulting

Lasso estimator is the (just identified) 2SLS estimator and this final model is the model for which $\lambda = 0$. Clearly, it can then be the case that the LARS/Lasso path is such that it does not include a model where all invalid instruments have been selected as such, which is the case when the final instrument selected as valid is in fact invalid. If that is the case, then there is no value of λ for which $\widehat{\beta}_\lambda$ is consistent.

Below we show under what conditions the large sample, $n \to \infty$, LARS/Lasso path does or does not include models where all invalid instruments have been selected. In simple settings, we show that this does depend on the number of invalid instruments, the relative strengths of the invalid versus the valid instruments and the correlation structure of the instruments. KZCS did show analytically that the performance of the Lasso estimator is influenced by these factors. They derived an estimator performance guarantee condition related to the values of $\mu = \max_{j \neq r} \left| \mathbf{Z}'_{.j} \mathbf{Z}_{.r} \right|$ and $\rho = \max_j \left| \mathbf{Z}'_{.j} \widehat{\mathbf{d}} \right| / \widehat{\mathbf{d}}' \widehat{\mathbf{d}}$. The constant μ measures the maximum correlation between any two columns of the matrix of instruments $\mathbf{Z}$, and ρ is a measures the maximum strength of the individual instruments. Their derived condition on the number of invalid instruments in Corollary 2 is that $s < \min\left(\frac{1}{12\mu}, \frac{1}{10\rho^2}\right)$. KZCS acknowledge the fact that these constraints are very strict. For example, if $\mu = 0.1$, then $s < 10/12$ and no invalid instruments are allowed, although their Monte Carlo results show that a simple correlation structure does not affect the behaviour of the estimator. Similarly for ρ, only a small value is allowed in order to have any invalid instruments allowed in the setup. If we assume that $\text{plim}\,(n^{-1}\mathbf{Z}'\mathbf{Z}) = \mathbf{Q}$, then

$$\text{plim}\left(\frac{\left|\mathbf{Z}'_{.j}\widehat{\mathbf{d}}\right|}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}}\right) = \frac{\left|\mathbf{Q}'_{.j}\gamma\right|}{\gamma'\mathbf{Q}\gamma}.$$

Therefore, if $\mathbf{Q} = \mathbf{I}_L$ this is equal to $\left|\gamma_j\right| / \gamma'\gamma$ and hence ρ is associated with the strongest instrument in terms of γ_j. Our results for consistent selection are based on the relative values of γ for the valid and invalid instruments, where we simply refer to γ_j as the instrument strength for instrument j. We show for uncorrelated instruments, with $\mathbf{Q} = \mathbf{I}_L$, that if invalid instruments are stronger than the valid ones, the selection procedure may select the valid instruments as invalid. Also, for the correlation structure, as in Zou (2006), we show that consistent selection depends on the patterns of correlations, not on the maximum correlation per se. Using our large sample analysis we can find simple configurations where the Lasso selection is inconsistent, which we confirm in some Monte

Carlo studies.

In order to mitigate these problems for the Lasso estimator, one can use 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 above, in the instrumental variables model this option is not available. Let $\widehat{\pi}_j = \widehat{\Gamma}_j/\widehat{\gamma}_j$, where $\widehat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{d}$ and $\widehat{\Gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$. Under standard assumptions as specified below, we show that the median of the $\widehat{\pi}_j$s is a consistent estimator for β when $s < L/2$, without any further restrictions on the relative strengths or correlations of the instruments, and hence this estimator can be used for the Adaptive Lasso.

We finally propose alternative consistent selection methods that are more closely related to the identification results derived from (7). These are the so-called upward and downward selection methods as proposed by Andrews (1999) based on the Hansen (1982) J-test of overidentifying restrictions, using the rank-ordered estimates of $\widehat{\pi}_j$ to guide the search.

As a further contribution, we consider different selection strategies to determine the value of λ for the Lasso estimator (8) when sufficient conditions are met for consistent selection end estimation. We consider a method recently proposed by Belloni, Chen, Chernozhukov and Hansen (2012) and a method based on Hansen's J-test. We show that these methods perform well, but that finite sample estimation results are substantially better for post-Lasso selection 2SLS estimators than for the Lasso estimators.

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

C1. $E[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'] = \mathbf{Q}$ is full rank

C2. $\text{plim}\,(n^{-1}\mathbf{Z}'\mathbf{Z}) = E[\mathbf{Z}_{i.}\mathbf{Z}_{i.}']$; $\text{plim}\,(n^{-1}\mathbf{Z}'\mathbf{d}) = E[\mathbf{Z}_{i.}D_i]$; $\text{plim}\,(n^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}) = E[\mathbf{Z}_{i.}\varepsilon_i] = 0$.

C3. $\gamma = (E[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'])^{-1}E[\mathbf{Z}_{i.}D_i]$, $\gamma_j \neq 0$, $j = 1, ..., L$.

3 Uncorrelated and Equal Strength Instruments

We first consider the conditions under which the Lasso procedure consistently selects the invalid instruments for the case where the instrument strengths are all equal, i.e. $\gamma_j = \widetilde{\gamma}$ for $j = 1, .., L$, and the instruments are uncorrelated, with variances equal to 1,

$E\left[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'\right] = \mathbf{I}_L = \text{plim}\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right)$.

Dividing by the sample size n, incorporating normalisation and noting that $\mathbf{Z}'\mathbf{M}_{\widehat{d}}\mathbf{M}_{\widehat{d}}\mathbf{P}_Z\mathbf{y} = \mathbf{Z}'\mathbf{M}_{\widehat{d}}\mathbf{y}$, the Lasso estimator $\widehat{\alpha}^{\lambda}$ is obtained as

$$\widehat{\alpha}^{\lambda} = \arg\min_{\alpha} \frac{1}{2n}\|\mathbf{y} - \widetilde{\mathbf{Z}}\alpha\|_2^2 + \frac{\lambda}{n}\|\widetilde{\mathbf{\Omega}}_n\alpha\|_1, \tag{11}$$

where $\widetilde{\mathbf{Z}} = \mathbf{M}_{\widehat{d}}\mathbf{Z}$, and $\widetilde{\mathbf{\Omega}}_n$ is an $L \times L$ diagonal matrix with diagonal elements $\widetilde{\omega}_j = \sqrt{\widetilde{\mathbf{Z}}_{j.}'\widetilde{\mathbf{Z}}_{j.}/n} = \sqrt{\mathbf{Z}_{j.}'\mathbf{M}_{\widehat{d}}\mathbf{Z}_{j.}/n}$.

The Lasso path can be obtained using the Lasso modification of the LARS algorithm, see Efron et al. (2004). Starting from the model without any instruments included as explanatory variables, let the L-vector of correlations $\widehat{\mathbf{c}}_n$ be defined as

$$\widehat{\mathbf{c}}_n = n^{-1}\widetilde{\mathbf{\Omega}}_n^{-1}\widetilde{\mathbf{Z}}'\mathbf{y} = n^{-1}\widetilde{\mathbf{\Omega}}_n^{-1}\mathbf{Z}'\mathbf{M}_{\widehat{d}}\mathbf{y},$$

with j-th element

$$\widehat{c}_{n,j} = \frac{n^{-1}\mathbf{Z}_{.j}'\mathbf{M}_{\widehat{d}}\mathbf{y}}{\sqrt{n^{-1}\mathbf{Z}_{.j}'\mathbf{M}_{\widehat{d}}\mathbf{Z}_{.j}}} \tag{12}$$

The first LARS step selects the variable(s) $\widetilde{\mathbf{Z}}_{.j}$ for which $|\widehat{c}_{n,j}|$ is maximum. We have for large samples that

$$\text{plim}\left(\widehat{c}_{n,j}\right) = \frac{\alpha_j - \overline{\alpha}}{\sqrt{(L-1)/L}},$$

as

$$\begin{aligned}
\text{plim}\left(n^{-1}\mathbf{Z}_{.j}'\mathbf{M}_{\widehat{d}}\mathbf{y}\right) &= \text{plim}\left(n^{-1}\mathbf{Z}_{.j}'\mathbf{M}_{\widehat{d}}\mathbf{Z}\alpha\right) \\
&= \text{plim}\left(n^{-1}\mathbf{Z}_{.j}'\mathbf{Z}\alpha\right) - \text{plim}\left(n^{-1}\mathbf{Z}_{.j}'\mathbf{P}_Z\mathbf{d}\left(\mathbf{d}'\mathbf{P}_Z\mathbf{d}\right)^{-1}\mathbf{d}'\mathbf{P}_Z\mathbf{Z}\alpha\right) \\
&= \alpha_j - \gamma_j\frac{\gamma'\alpha}{\gamma'\gamma} = \alpha_j - L^{-1}\sum_{r=1}^{L}\alpha_r \\
&= \alpha_j - \overline{\alpha},
\end{aligned}$$

and

$$\text{plim}\left(n^{-1}\mathbf{Z}_{j.}'\mathbf{M}_{\widehat{d}}\mathbf{Z}_{j.}\right) = 1 - \frac{\gamma_j^2}{\gamma'\gamma} = 1 - \frac{1}{L}$$

using the facts that $\text{plim}\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right) = \mathbf{I}_L$ and that all the γ_js are the same.

There are $s < L$ invalid instruments. If all the invalid instruments have the same effect a, the case considered mostly in the KZCS simulations, then $\overline{\alpha} = sa/L$. We then

get for a valid instrument $\text{plim}\,(\widehat{c}_{n,val}) = -sa/\sqrt{L\,(L-1)}$, and for an invalid instrument $\text{plim}\,(\widehat{c}_{n,inv}) = (L-s)a/\sqrt{L\,(L-1)}$. In large samples, the invalid instruments get therefore selected in the first LARS step if

$$(L-s)\,|a| > s\,|a| \quad \Leftrightarrow \quad s < L/2,$$

so less than 50% of the instruments can be invalid, which is aligned with Theorem 1 and Corollary 1 of KZCS. In practice, of course, the correlations for the invalid (and valid instruments) will not be exactly equal to each other, and the instruments will be selected one at the time, with the LARS update of the predicted mean approaching zero for large sample sizes within the two groups of instruments.

It is clear from the correlations derived above, that many situations can arise in terms of selecting invalid instruments correctly, depending on the values of the α_j's. At the one extreme, it is clear that the first LARS step would correctly select $L-2$ invalid instruments, for L even, when half of them have effect a and the other half $-a$, which is a case where the parameters are in principle not identified. Perhaps more interesting is to consider the situation where the s invalid instruments have distinct positive effects, ordered in such a way that $\alpha_1 > \alpha_2 > ... > \alpha_s > \alpha_{s+1} = ... = \alpha_L = 0$. We show in the Appendix that also for this case the condition that $s < L/2$ need to be satisfied for the LARS algorithm to select the invalid instruments in the first s steps in large samples and that for $s > L/2$ the full LARS path does not include a model where all invalid instruments have been included. This is in contrast to the identification results, as the parameters are identified in this case by Theorem 1 in KZCS as long as $s < L-1$.

For consistent model selection and estimation, the condition that $s < L/2$ is sufficient when instruments are uncorrelated and have equal strengths. We will show below that this condition is no longer sufficient when we allow for differential instrument strengths, especially when invalid instruments are relatively strong. It is also not sufficient under certain correlation structures of the instruments, as observed by Zou (2006) for the standard linear model case. However, before we move to these problems, we will analyse the behaviour of the Lasso estimator of (11) in situations where the condition that $s < L/2$ is sufficient for consistent selection of the invalid instruments.

We start with presenting some estimation results from a simple Monte Carlo exercise,

similar to that in KZCS. The data are generated from

$$\begin{aligned} Y_i &= D_i\beta + \mathbf{Z}_{i.}'\alpha + \varepsilon_i \\ D_i &= \mathbf{Z}_{i.}'\gamma + v_i, \end{aligned}$$

where

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right);$$
$$\mathbf{Z}_{i.} \sim N\left(0, I_L\right);$$

and we set $\beta = 0$; $\gamma_j = 0.2$, $j = 1, ..., L$; $L = 10$; $\rho = 0.25$; $s = 3$, and the first s elements of α are equal to $a = 0.2$. Table 1 presents estimation results for estimators of β in terms of bias, standard deviation, rmse and median absolute deviation (mad) for 1000 replications for sample sizes of $n = 500$, $n = 2000$ and $n = 10,000$. The "2SLS" results are for the 2SLS estimator that treats all instruments as valid. The "2SLS or" is the oracle 2SLS estimator that includes $\mathbf{Z}_{.1}, ., \mathbf{Z}_{.3}$ in the model as explanatory variables. For the Lasso estimates, the value for λ has been obtained by 10-fold cross-validation, using the one-standard error rule, as in KZCS, but we also present results for 10-fold cross validation not using the one-standard error rule. The latter estimator is denoted "Lasso$_{cv}$", whereas the former is denoted "Lasso$_{cvse}$", which is the one produced by sisVIVE. We also present result for the so-called post-Lasso estimator, see e.g. Belloni et al. (2012), which is called the LARS-OLS hybrid by Efron et al. (2004). In this case this is the 2SLS estimator (4), where $\mathbf{Z}_{sel}$ is the set of instruments with non-zero estimated Lasso coefficients. Further entries in the table are the average number of selected instruments, which are the number of instruments with non-zero α coefficients, together with the minimum and maximum number of selected instruments, and the proportion of times the selected instruments include the 3 invalid instruments.

Table 1. Estimation results for β; $L = 10$, $s = 3$

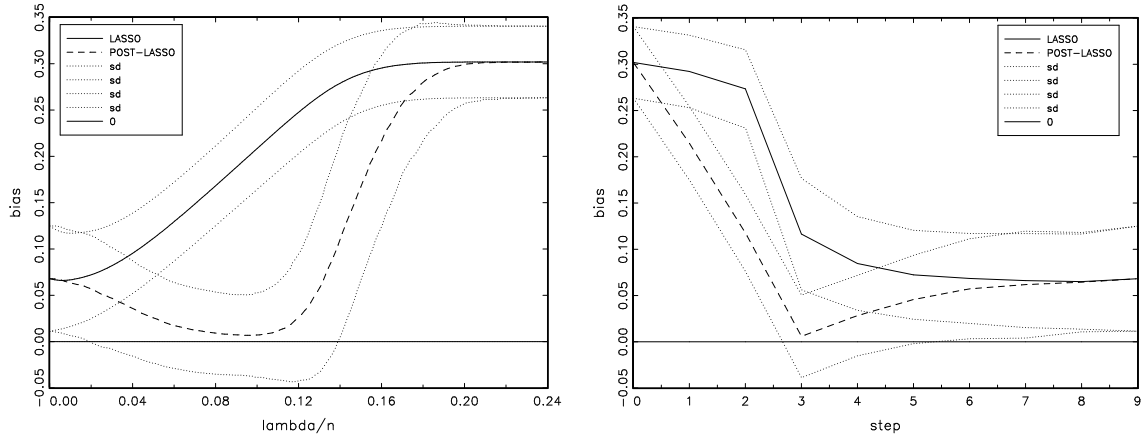
β	bias	std dev	rmse	mad	#sel instr [min, max]	prop $\mathbf{Z}_{.1}, ., \mathbf{Z}_{.3}$ selected
$n = 500$						
2SLS	0.2966	0.0808	0.3074	0.2944	0	0
2SLS or	0.0063	0.0843	0.0845	0.0570	3	1
Lasso$_{cv}$	0.1384	0.0965	0.1687	0.1352	6.41 [2,9]	0.990
Post-Lasso$_{cv}$	0.1169	0.1136	0.1630	0.1143		
Lasso$_{cvse}$	0.2206	0.0847	0.2363	0.2174	3.16 [0,8]	0.664
Post-Lasso$_{cvse}$	0.0905	0.1243	0.1537	0.0994		
$n = 2000$						
2SLS	0.3019	0.0387	0.3044	0.3007	0	0
2SLS or	0.0047	0.0422	0.0424	0.0285	3	1
Lasso$_{cv}$	0.0721	0.0509	0.0882	0.0705	6.56 [3,9]	1
Post-Lasso$_{cv}$	0.0617	0.0577	0.0845	0.0644		
Lasso$_{cvse}$	0.1140	0.0430	0.1218	0.1165	3.76 [3,8]	1
Post-Lasso$_{cvse}$	0.0277	0.0521	0.0590	0.0387		
$n = 10,000$						
2SLS	0.2996	0.0177	0.3002	0.2992	0	0
2SLS or	0.0006	0.0183	0.0183	0.0127	3	1
Lasso$_{cv}$	0.0317	0.0236	0.0395	0.0311	6.44 [3,9]	1
Post-Lasso$_{cv}$	0.0272	0.0267	0.0380	0.0282		
Lasso$_{cvse}$	0.0479	0.0187	0.0514	0.0489	3.81 [3,9]	1
Post-Lasso$_{cvse}$	0.0118	0.0238	0.0265	0.0176		

Notes: Results from 1000 MC replications; $a = 0.2$; $\beta = 0$; $\gamma = 0.2$; $\rho = 0.25$

The results in Table 1 reveal some interesting patterns. First of all, the Lasso$_{cv}$ estimator outperforms the Lasso$_{cvse}$ estimator in terms of bias, rmse and mad for all sample sizes, but this is reversed for the post-Lasso estimators, i.e. the post-Lasso$_{cvse}$ outperforms the post-Lasso$_{cv}$. The Lasso$_{cv}$ estimator selects on average around 6.5 instruments as invalid, which is virtually independent of the sample size. The Lasso$_{cvse}$ estimator selects on average around 3.8 instruments as invalid for $n = 2000$ and $n = 10,000$, but less, 3.17 for $n = 500$. Although the correct instruments are always selected for the larger sample sizes, the Lasso$_{cvse}$ is substantially biased, the biases being larger than twice the standard deviations. The post-Lasso$_{cvse}$ estimator performs best, but is still substantially outperformed by the oracle 2SLS estimator.

Figures 1a and 1b shed some more light on the different behaviour of the Lasso and post-Lasso estimators. Figure 1a shows the bias and standard deviations of the two

estimators for different values of λ/n, for the design above with $n = 2000$, again from 1000 replications. It is clear that the Lasso estimator exhibits a positive bias for all values of λ, declining from that of the naive 2SLS estimator to the minimum bias of 0.0664 at $\lambda/n = 0.0060$. In contrast, the post-Lasso estimator has a (much) smaller bias, obtaining its minimum bias of 0.0068 at the value of λ/n of 0.0965. Figure 1b displays the same information but now as a function of the LARS steps (we have omitted 3 replications where there were Lasso steps). At step 3, the correct 3 invalid instruments have been selected 991 times out of the 997 replications, and the post-Lasso estimator has a bias there of 0.0058, only fractionally larger than that of the oracle 2SLS estimator. In contrast, the Lasso estimator for β still has a substantial upward bias at step 3. Its bias decreases from 0.116 at step 3 to a minimum of 0.0650 at step 8. Interestingly, the bias of the post-Lasso estimator increases again after step 3, reaching the same bias as the Lasso estimator at the last step, as there $\lambda = 0$ and the Lasso and post-Lasso estimators are equal.



Figures 1a and 1b. Bias and standard deviations of Lasso and post-Lasso estimators as functions of λ/n and LARS steps. Same design as in Table 1, $n = 2000$. 3 replications out of 1000 omitted in 1b due to Lasso steps.

We can understand the different behaviour of the Lasso and post-Lasso estimators, which is due to shrinkage of the Lasso estimator for α, as follows. Denote by $\mathbf{Z}_{sel}^{\lambda}$ the matrix of selected instruments for any value of λ, i.e. those instruments with non-zero

values of $\widehat{\alpha}^{\lambda}$. For the Lasso and post-Lasso estimators, $\widehat{\beta}^{\lambda}$ and $\widehat{\beta}$, we have that

$$\widehat{\beta}^{\lambda} = \widehat{\beta} + \frac{\widehat{\mathbf{d}}'\mathbf{Z}^{\lambda}_{sel}\left(\widehat{\alpha}_{sel} - \widehat{\alpha}^{\lambda}_{sel}\right)}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}}.$$

For those values of λ where the correct invalid instruments have been included, the biases of $\widehat{\beta}$ and $\widehat{\alpha}$ are small in large samples. Define $\widehat{\delta}^{\lambda}$ as the shrinkage factor of the Lasso estimator, relative to that of the post-Lasso estimator, i.e. $\widehat{\alpha}^{\lambda}_{sel} \approx \widehat{\delta}^{\lambda}\widehat{\alpha}_{sel}$. We then have approximately

$$\widehat{\beta}^{\lambda} \approx \widehat{\beta} + \left(1 - \widehat{\delta}^{\lambda}\right)\frac{\widehat{\mathbf{d}}'\mathbf{Z}^{\lambda}_{sel}\widehat{\alpha}_{sel}}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}}.$$

Note that in the model without any instruments included as explanatory variables, we have for the 2SLS estimator,

$$\widehat{\beta} = \frac{\widehat{\mathbf{d}}'\mathbf{y}}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}} = \frac{\widehat{\mathbf{d}}'\mathbf{y}}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}} = \beta + \frac{\widehat{\mathbf{d}}'\mathbf{Z}\alpha}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}} + \frac{\widehat{\mathbf{d}}'\boldsymbol{\xi}}{\widehat{\mathbf{d}}'\widehat{\mathbf{d}}}$$

and so we see that the bias of the Lasso estimator due to shrinkage is in the direction of the bias of the 2SLS estimator in the model without any included invalid instruments. As an illustration, for the $n = 2000$ case above, at $\lambda/n = 0.0965$, the means of the first three elements of $\widehat{\alpha}^{\lambda}_{sel}$ are all equal to 0.067, whereas those of the post-Lasso 2SLS estimator are equal to 0.198, hence $1 - \widehat{\delta}^{\lambda} = 0.662$. The bias of the 2SLS estimator treating all instruments as valid is given by 0.302, and $0.662 * 0.302 = 0.200$. This is very close to the difference in the biases of $\widehat{\beta}^{\lambda}$ and $\widehat{\beta}$ at this point, which is given by $0.201 - 0.007 = 0.194$.

3.1 Stopping Rule

It is clear from the results above that the post-Lasso estimator outperforms the Lasso estimator, with the performance of the post-Lasso$_{cvse}$ best, but still some way short of that of the oracle 2SLS estimator, even for $n = 10,000$. It is also clear, and easily understood from Figures 1a and 1b, that the 10-fold cross-validation method selects models that are too large in the sense that valid instruments get selected as invalid over and above the invalid ones, and does not improve with increasing sample size, although the ad hoc one-standard error rule does improve the selection. We next consider two alternative stopping rules, one proposed for the Lasso by Belloni et al. (2012), and one for GMM moment selection by Andrews (1999).

Belloni et al. (2012) explicitly allow for general conditional heteroskedasticity and consider the Lasso estimator defined as

$$\widehat{\alpha}^{\lambda} = \arg\min_{\alpha} \frac{1}{2n}\|\mathbf{y} - \widetilde{\mathbf{Z}}\alpha\|_2^2 + \frac{\lambda}{n}\|\widetilde{\mathbf{\Omega}}_n^*\alpha\|_1,$$

where $\widetilde{\mathbf{\Omega}}_n^*$ is an $L \times L$ diagonal matrix with j-th diagonal element

$$\widetilde{\omega}_{n,j}^* = \sqrt{n^{-1}\sum_{i=1}^{n} \widetilde{z}_{ij}^2 \widetilde{\varepsilon}_i^2},$$

where

$$\widetilde{\varepsilon}_i = y_i - \widetilde{\mathbf{Z}}_{i\cdot}'\alpha.$$

Then let

$$\widehat{\mathbf{c}}_n^* = \frac{1}{n}\sum_{i=1}^{n} \left(\widetilde{\mathbf{\Omega}}_n^*\right)^{-1} \widetilde{\mathbf{Z}}_{i\cdot}\widetilde{\varepsilon}_i.$$

Belloni et al. (2012) apply the moderate deviation theory of Jing, Shao and Wang (2003) to bound deviations of the maximal element of the vector of correlations $\widehat{\mathbf{c}}_n^*$ and hence λ/n for models that have selected the invalid instruments. They establish that

$$P\left(\sqrt{n}\max\left(\widehat{c}_n^*\right) \leq \Phi^{-1}\left(1 - \frac{\tau_n}{2L}\right)\right) \geq 1 - \tau_n + o\left(1\right),$$

where $\Phi^{-1}\left(.\right)$ is the inverse, or quantile function, of the standard normal distribution, and that the penalty level should satisfy

$$P\left(\frac{\lambda}{n} \geq q\max\left(\widehat{c}_n^*\right)\right) \to 1$$

for some constant $q > 1$. Belloni et al. (2012) then recommend selecting

$$\frac{\lambda}{n} = q\Phi^{-1}\left(1 - \frac{\tau_n}{2L}\right)/\sqrt{n}$$

and to set the confidence level $\tau_n = 0.1/\ln\left(n\right)$ and the constant $q = 1.1$. For $n = 2000$, this results in a value for λ/n equal to 0.079, which suggests a good performance of the post-Lasso estimator from Figure 1a, as the design there is conditionally homoskedastic. We obtain the Lasso and post-Lasso estimators using the Belloni et al. (2012) iterative procedure as described in their Appendix A, as the $\widetilde{\varepsilon}_i$ need to be estimated to construct $\widetilde{\mathbf{\Omega}}_n^*$. We use the post-Lasso estimator at every step to estimate the $\widetilde{\varepsilon}_i$.

The second stopping rule we consider is based on the approach of Andrews (1999). We can use this approach because the number of instruments L is fixed and (much) smaller than n. Consider again the model

$$\begin{aligned} \mathbf{y} &= \mathbf{d}\beta + \mathbf{Z}_{sel}\alpha_{sel} + \boldsymbol{\xi} \\ &= \mathbf{R}\theta + \boldsymbol{\xi}. \end{aligned} \tag{13}$$

Let $\mathbf{G}_n(\theta) = n^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{R}\boldsymbol{\theta})$, then the Generalised Methods of Moment (GMM) estimator is defined as

$$\widehat{\theta}_{GMM} = \arg\min_{\theta} \mathbf{G}_n(\theta)\, \mathbf{W}_n^{-1} \mathbf{G}_n(\theta),$$

see Hansen (1982). 2SLS is a one-step GMM estimator, setting $\mathbf{W}_n = n^{-1}\mathbf{Z}'\mathbf{Z}$. Given the moment conditions $E(\mathbf{Z}_{i.}\xi_i) = 0$, 2SLS is efficient under conditional homoskedasticity, $E\left(\xi_i^2 | \mathbf{Z}_{i.}\right) = \sigma_\xi^2$. Under general forms of conditional heteroskedasticity, an efficient two-step GMM estimator is obtained by setting

$$\mathbf{W}_n = \mathbf{W}_n\left(\widehat{\theta}_1\right) = n^{-1}\sum_{i=1}^{n}\left(\left(y_i - \mathbf{R}_{i.}'\widehat{\theta}_1\right)^2 \mathbf{Z}_{i.}\mathbf{Z}_{i.}'\right)$$

where $\widehat{\theta}_1$ is an initial consistent estimator, with a natural choice the 2SLS estimator. Then, under the null that the moment conditions are correct, $E(\mathbf{Z}_{i.}\xi_i) = 0$, the Hansen (1982) J-test statistic and its limiting distribution are given by

$$J_n\left(\widehat{\theta}_1\right) = n\mathbf{G}_n\left(\widehat{\theta}_2\right)\mathbf{W}_n^{-1}\left(\widehat{\theta}_1\right)\mathbf{G}_n\left(\widehat{\theta}_2\right) \xrightarrow{d} \chi^2_{(L-\dim(\mathbf{R}))}.$$

We can now combine the LARS/Lasso algorithm with the Hansen J-test, which is then akin to a directed downward testing procedure in the terminology of Andrews (1999). Let the critical value $\zeta_{n,k} = \chi_k^2(\tau_n)$ be the $1-\tau_n$ quantile of the χ_k^2 distribution, where $k = L - \dim(\mathbf{R})$. Compute at every LARS/Lasso step as described above the Hansen J-test and compare it to the corresponding critical value. We then select the model with the largest degrees of freedom for which the J-test 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. Clearly, this approach is a post-Lasso approach, where the LARS algorithm is purely used for selection of the invalid instruments. For consistent model selection, the critical values $\zeta_{n,k}$ need to satisfy

$$\zeta_{n,k} \to \infty \quad \text{and} \quad \zeta_{n,k} = o(n), \tag{14}$$

see Andrews (1999). We select $\tau_n = 0.1/\ln(n)$ as per the Belloni et al. (2012) method.

Table 2 presents the estimation results using the two alternative stopping rules. The subscripts "bcch" and "ah" denote the Belloni et al. (2012) method and the Andrews/Hansen approach respectively. Both post-Lasso estimators are the simple 2SLS estimators for comparison with the results in Table 1.

Table 2. Estimation results for β; $L = 10$, $s = 3$

β	bias	std dev	rmse	mad	#sel instr [min, max]	prop $\mathbf{Z}_{.1}, ., \mathbf{Z}_{.3}$ selected
$n = 500$						
Lasso$_{bcch}$	0.2770	0.0846	0.2896	0.2770	1.16 [0,4]	0.068
Post-Lasso$_{bcch}$	0.1914	0.1324	0.2327	0.2028		
Post-Lasso$_{ah}$	0.0896	0.1252	0.1539	0.1007	2.56 [0,5]	0.391
2SLS or	0.0063	0.0843	0.0845	0.0570	3	1
$n = 2000$						
Lasso$_{bcch}$	0.1688	0.0438	0.1744	0.1694	3.11 [3,5]	1
Post-Lasso$_{bcch}$	0.0091	0.0445	0.0454	0.0294		
Post-Lasso$_{ah}$	0.0055	0.0430	0.0434	0.0286	3.02 [3,5]	1
2SLS or	0.0047	0.0422	0.0424	0.0285	3	1
$n = 10,000$						
Lasso$_{bcch}$	0.0751	0.0180	0.0772	0.0756	3.11 [3,5]	1
Post-Lasso$_{bcch}$	0.0027	0.0191	0.0193	0.0134		
Post-Lasso$_{ah}$	0.0009	0.0186	0.0186	0.0129	3.02 [3,5]	1
2SLS or	0.0006	0.0183	0.0183	0.0127	3	1

Notes: Results from 1000 MC replications; $a = 0.2$; $\beta = 0$; $\gamma = 0.2$ $\rho = 0.25$

For $n = 500$, we find that the value of λ/n as determined by the BCCH method is too large and the method selects too few instruments as invalid, resulting in severely biased Lasso and post-Lasso estimates. The AH approach behaves better for $n = 500$, but it also selects too few invalid instruments, resulting in an upward bias for this particular case, which is similar to results for the post-Lasso$_{cvse}$ estimator in Table 1. For $n = 2000$ and $n = 10,000$, both post-Lasso procedures perform very well with properties very similar to that of the oracle 2SLS estimator, with the AH approach marginally outperforming the BCCH approach for this design. Also, using standard asymptotic robust standard errors for the 2SLS estimators, Wald tests for the null $H_0 : \beta = 0$, at the 10% level, reject 11.9% (10.9%) and 10.8% (9.4%) for the BCCH and AH methods respectively for

$n = 2000$ ($n = 10,000$), indicating that their distributions are very well approximated by the standard limiting distribution of the 2SLS estimator.

4 Varying Instrument Strength

As derived above, for the first step of the LARS algorithm we have, still assuming that $E(\mathbf{Z}_{i.}\mathbf{Z}_{i.}') = \mathbf{I}_L$,

$$\text{plim}\,(\widehat{c}_{n,j}) = \frac{\alpha_j - \gamma_j \frac{\gamma'\alpha}{\gamma'\gamma}}{\sqrt{1 - \frac{\gamma_j^2}{\gamma'\gamma}}}.$$

It is clear that allowing for differential instrument strengths, i.e. different values of γ, may result in the LARS/Lasso path not including all invalid instruments. For example, consider again the situation where all s invalid instruments have the same direct effect a. The valid instruments all have strength γ_{val}, whereas the invalid instruments all have strength $\gamma_{inv} = t\gamma_{val}$, with $t > 0$ Then for an invalid and a valid instrument we get respectively,

$$\begin{aligned}
\text{plim}\,(\widehat{c}_{n,inv}) &= \frac{1}{\sqrt{st^2 + L - s}} \frac{a(L-s)}{\sqrt{(s-1)t^2 + L - s}}; \\
\text{plim}\,(\widehat{c}_{n,val}) &= -\frac{1}{\sqrt{st^2 + L - s}} \frac{ast}{\sqrt{st^2 + L - s - 1}},
\end{aligned}$$

and hence we see that the valid instruments get selected as being invalid in large samples if

$$\frac{st}{\sqrt{st^2 + L - s - 1}} > \frac{L - s}{\sqrt{(s-1)t^2 + L - s}}. \tag{15}$$

For example, when $L = 10$ and $s = 3$, this happens when $t > 2.7$. As all the invalid instruments in this case are "valid" for a causal estimate of $\beta + a/\gamma_{inv}$, the Lasso and post-Lasso estimators will select the $L - s$ valid instruments as invalid. Table 3 presents estimation results for the same Monte Carlo design as in Table 1, with $\gamma_{val} = 0.2$, but $\gamma_{inv} = 3\gamma_{val}$. For brevity, we only present results for the post-Lasso$_{cvse}$ and post-Lasso$_{ah}$ estimators. Note that the behaviour of the oracle 2SLS estimator is the same as in Table 1. In this case $\beta + \alpha/\gamma_{inv} = 0 + 0.2/0.6 = 0.33$, and the results confirm that the LARS/Lasso method now selects in the larger samples the valid instruments as invalid because of the relative strength of the invalid instruments. For $n = 500$ the algorithm cannot separate the instruments and selects only very few as invalid.

Table 3. Estimation results for β; $L = 10$, $s = 3, \gamma_{inv} = 3\gamma_{val}$

β	bias	std dev	rmse	mad	#sel instr [min, max]	prop $\mathbf{Z}_{.1}, ., \mathbf{Z}_{.3}$ selected
$n = 500$						
Post-Lasso$_{cvse}$	0.2658	0.0428	0.2692	0.2651	0.44 [0,8]	0
Post-Lasso$_{ah}$	0.2651	0.0485	0.2695	0.2666	0.76 [0,6]	0
$n = 2000$						
Post-Lasso$_{cvse}$	0.2911	0.0352	0.2932	0.2933	6.58 [0,9]	0.00
Post-Lasso$_{ah}$	0.2803	0.0399	0.2831	0.2845	5.05 [1,9]	0
$n = 10,000$						
Post-Lasso$_{cvse}$	0.3202	0.0122	0.3204	0.3205	8.70 [7,9]	0
Post-Lasso$_{ah}$	0.3233	0.0131	0.3236	0.3242	8.09 [6,9]	0

Notes: Results from 1000 MC replications; $a = 0.2$; $\beta = 0$; γ_{val}= 0.2 $\rho = 0.25$

It is clear that various combinations of instrument strength can lead to inconsistent selection and estimation. One simple, but quite interesting example is the following. Let $L = 5$ and $s = 2$. Let $\alpha = (0.2, 0.15, 0, 0, 0)'$ and $\gamma = (0.8, 0.7, 1, 0.25, 0.15)'$, so the strongest instrument and the two weakest instruments are valid, and the two invalid instruments are relatively strong. The large sample LARS/Lasso path can be calculated in this case to be $\{3, 1, 4, 5\}$, i.e. the strong valid instrument gets selected as invalid first, and the full LARS path does not include a model where the two invalid instruments are selected.

On the other hand, from (15) it is also easily seen that when the valid instruments are stronger than the invalid ones, the LARS algorithm may select the correct invalid instruments also when $s \geq L/2$. For example, again for $L = 10$, when $t = 0.5$, $|\text{plim}\,(\widehat{c}_{n,inv})| > |\text{plim}\,(\widehat{c}_{n,val})|$ for $s = 1, ..., 6$.

5 Correlated Instruments

We revert back to the case where all γs are the same, but we now allow the instruments to be correlated such that

$$E\left[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'\right] = \text{plim}\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right) = \mathbf{Q}$$

with all diagonal elements of $\mathbf{Q}$ equal to 1.

For the numerator of $\widehat{c}_{n,j}$ as defined in (12) we get

$$\text{plim}\left(n^{-1}\mathbf{Z}'_{.j}\mathbf{M}_{\widehat{d}}\mathbf{y}\right) = \mathbf{Q}'_{.j}\left(\alpha - \gamma\frac{\gamma'\mathbf{Q}\alpha}{\gamma'\mathbf{Q}\gamma}\right) = \mathbf{Q}'_{.j}\left(\alpha - \boldsymbol{\iota}_L\frac{\boldsymbol{\iota}'_L\mathbf{Q}\alpha}{\boldsymbol{\iota}'_L\mathbf{Q}\boldsymbol{\iota}_L}\right),$$

where $\mathbf{Q}_{.j}$ is the jth column of $\mathbf{Q}$; $\boldsymbol{\iota}_L$ is an L-vector of ones, and the second result follows because the γs are all the same.

For the denominator, we get

$$\text{plim}\left(n^{-1}\mathbf{Z}'_{.j}\mathbf{M}_{\widehat{d}}\mathbf{Z}_{.j}\right) = 1 - \frac{\left(\mathbf{Q}'_{.j}\gamma\right)^2}{\gamma'\mathbf{Q}\gamma} = 1 - \frac{\left(\mathbf{Q}'_{.j}\boldsymbol{\iota}_L\right)^2}{\boldsymbol{\iota}'_L\mathbf{Q}\boldsymbol{\iota}_L}.$$

If we denote again the first s instruments to be the invalid ones, and when all the α_js for the invalid instruments are the same and equal to a, then for the invalid instruments we have that

$$\text{plim}\left(\widehat{c}_{n,j}\right)_{j\in\{1,..,s\}} = \frac{\mathbf{Q}'_{j.}\left(\mathbf{e}_s - \boldsymbol{\iota}_L\frac{\boldsymbol{\iota}'_L\mathbf{Q}\mathbf{e}_s}{\boldsymbol{\iota}'_L\mathbf{Q}\boldsymbol{\iota}_L}\right)}{\sqrt{1 - \frac{\left(\mathbf{Q}'_{j.}\boldsymbol{\iota}_L\right)^2}{\boldsymbol{\iota}'_L\mathbf{Q}\boldsymbol{\iota}_L}}}, \tag{16}$$

and for the valid instruments

$$\text{plim}\left(\widehat{c}_{n,r}\right)_{r\in\{s+1,..,L\}} \frac{\mathbf{Q}'_{.r}\left(\mathbf{e}_s - \boldsymbol{\iota}_L\frac{\boldsymbol{\iota}'_L\mathbf{Q}\mathbf{e}_s}{\boldsymbol{\iota}'_L\mathbf{Q}\boldsymbol{\iota}_L}\right)}{\sqrt{1 - \frac{\left(\mathbf{Q}'_{.r}\boldsymbol{\iota}_L\right)^2}{\boldsymbol{\iota}'_L\mathbf{Q}\boldsymbol{\iota}_L}}} \tag{17}$$

where $\mathbf{e}_s = \begin{pmatrix} \boldsymbol{\iota}'_s & \mathbf{0}'_{L-s} \end{pmatrix}'$ and $\mathbf{0}_{L-s}$ is an $L-s$ vector of zeros.

KZCS first of all set all pairwise correlations of the instruments equal to a single value η. In that case $\mathbf{Q}'_{.j}\boldsymbol{\iota}_L = \mathbf{Q}'_{.r}\boldsymbol{\iota}_L$ and hence we get that the invalid instruments get selected if

$$\left|1 + (s-1)\,\eta - \left((1 + (L-1)\,\eta)\,\frac{s}{L}\right)\right| > \left|s\eta - (1 + (L-1)\,\eta)\,\frac{s}{L}\right|,$$

or

$$(L-s)\,(1-\eta) > |-s\,(1-\eta)| \quad \Longleftrightarrow \quad L > 2s,$$

i.e. the same result as before. This type of correlation does therefore not affect the Lasso selection results as derived above for the case of uncorrelated equal strength instruments.

KZCS considered 2 alternative designs, one with the same pairwise correlation η within the valid and invalid instruments but no correlation between the valid and invalid instruments, and one with only pairwise correlation η between valid and invalid instruments. As above, from (16) and (17), it can be shown that both these designs do not

alter the results derived above for equal strength instruments when the instruments are uncorrelated .

There are however correlation structures that affect the selection process in such a way that the large sample LARS/Lasso path does not include a model where all invalid instruments are selected, even when $s < L/2$. This has been documented well for the Lasso in the standard linear model, see e.g. Zou (2006). As a simple example, if η_1 is the pairwise correlation between the invalid instruments, η_{12} that between the valid and invalid ones, and η_2 that between the valid instruments, then e.g. for $L = 10$, $s = 3$, and values of $\eta_1 = -0.22$, $\eta_{12} = -0.11$ and $\eta_2 = 0.85$, from (16) and (17) we get for the invalid and valid instruments

$$\begin{aligned} \text{plim}\,(\widehat{c}_{n,inv}) &= 0.1475 \\ \text{plim}\,(\widehat{c}_{n,val}) &= -0.1506. \end{aligned}$$

Hence, for this correlation structure, the valid instruments will be selected as invalid in large samples.

There is an important conceptual issue when instruments are correlated in the sense that for general correlation structures valid instruments are only valid after inclusion of the invalid instruments in the model. This is unlike the case of uncorrelated instruments, where inclusion of invalid instruments in the model or dropping them from the instrument set both lead to a consistent 2SLS estimator. Therefore, assumption (2) about the relationship between the instrument and the confounders, $E\left[Y_i^{(0,0)}|\mathbf{Z}_{i.}\right] = \mathbf{Z}_{i.}'\psi$, is essential for the identification and estimation of the parameters when instruments are correlated. As can be seen from the observational model (3), the direct effect assumption (1) and the conditional mean assumption (2) are observationally equivalent. If one were to change the conditional mean assumption (2) to one of correlation, i.e.

$$E\left[Y_i^{(0,0)}\mathbf{Z}_{i.}\right] = \widetilde{\psi}, \tag{18}$$

with some of the elements of $\widetilde{\psi}$ are equal to 0, which are for example the moments considered by Liao (2013) and Cheng and Liao (2015), then model (3) no longer follows unless instruments are uncorrelated. For general correlation structures all instruments would enter (3) under condition (18), or in other words, all α_j coefficients would be

unequal to 0 and the causal effect parameter would therefore not be identified using the selection methods based on model specification (3).

6 A Consistent Estimator when $s < L/2$ and Adaptive Lasso

As the results above highlight, the LARS/Lasso path may not include the correct model, leading to an inconsistent estimator of β. This is the case even if more than 50% of the instruments are valid because of differential instrument strength and/or correlation patterns of the instruments. In this section we present an estimation method that consistently selects the invalid instruments when more than 50% of the potential instruments are valid. This is the same condition as that for the LARS/Lasso selection to be consistent 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 build on the result of Han (2008), who shows that the median of the L IV estimates of β using one instrument at the time is a consistent estimator of β in a model with invalid instruments, but where the instruments cannot have direct effects on the outcome, unless the instruments are uncorrelated.

Define

$$\begin{aligned} \widehat{\Gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}; \\ \widehat{\gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{d}, \end{aligned}$$

and let $\widehat{\pi}$ be the L vector with j-th element

$$\widehat{\pi}_j = \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}, \tag{19}$$

then, under model specification (3), assumptions C1-C3 and the condition that $s < L/2$,

$$\widehat{\beta}_m \equiv \text{median}\,(\widehat{\pi}) \tag{20}$$

is a consistent estimator for β.

The proof is simply established, as under the stated assumptions,

$$\begin{aligned}
\operatorname{plim}\left(\widehat{\Gamma}\right) &= \operatorname{plim}\left(\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right)^{-1} n^{-1}\mathbf{Z}'\mathbf{y}\right) \\
&= \operatorname{plim}\left(\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right)^{-1} n^{-1}\mathbf{Z}'\left(\mathbf{d}\beta + \mathbf{Z}\alpha + \boldsymbol{\varepsilon}\right)\right) \\
&= \left(E\left[\mathbf{Z}_i.\mathbf{Z}_i'.\right]\right)^{-1} E\left[\mathbf{Z}_i.D_i\right]\beta + \alpha \\
&= \gamma\beta + \alpha; \\
\operatorname{plim}\left(\widehat{\gamma}\right) &= \operatorname{plim}\left(\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right)^{-1} n^{-1}\mathbf{Z}'\mathbf{d}\right) \\
&= \left(E\left[\mathbf{Z}_i.\mathbf{Z}_i'.\right]\right)^{-1} E\left[\mathbf{Z}_i.D_i\right] = \gamma.
\end{aligned}$$

Hence

$$\operatorname{plim}\left(\widehat{\pi}_j\right) = \frac{\gamma_j\beta + \alpha_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j}.$$

As $s < L/2$, more than 50% of the αs are equal to zero and hence it follows that more than 50% of the elements of $\operatorname{plim}\left(\widehat{\pi}\right)$ are equal to β. Using a continuity theorem, it then follows that

$$\operatorname{plim}\left(\widehat{\beta}_m\right) = \operatorname{median}\left\{\operatorname{plim}\left(\widehat{\pi}\right)\right\} = \beta.$$

Given the consistent estimator derived above for β, we can obtain a consistent estimator for α

$$\widehat{\alpha}_m = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\left(\mathbf{y} - \mathbf{d}\widehat{\beta}_m\right),$$

which can then be used for the adaptive Lasso specification of (11) as proposed by Zou (2006). The adaptive Lasso estimator for α is defined as

$$\widehat{\alpha}_{ad}^{\lambda} = \arg\min_{\alpha} \frac{1}{2n}\|\mathbf{y} - \widetilde{\mathbf{Z}}\alpha\|_2^2 + \frac{\lambda}{n}\sum_{l=1}^{L}\frac{|\widetilde{\omega}_l\alpha_l|}{|\widehat{\alpha}_{m,l}|^{\nu}}.$$

Table 4 present the estimation results for the adaptive Lasso for the design as in Table 3, setting $\nu = 1$. For $n = 500$ the adaptive Lasso estimators are severely biased upwards, but for the larger sample sizes they perform well.

Table 4. Estimation results for β, Adaptive Lasso; $L = 10$, $s = 3, \gamma_{inv} = 3\gamma_{val}$

β	bias	std dev	rmse	mad	#sel instr [min, max]	prop $\mathbf{Z}_{.1}, ., \mathbf{Z}_{.3}$ selected
$n = 500$						
$\widehat{\beta}_m$	0.1126	0.0935	0.1463	0.1129		
Post-ALasso$_{cvse}$	0.2426	0.0787	0.2550	0.2568	0.46 [0,6]	0.04
Post-ALasso$_{ah}$	0.2173	0.1091	0.2432	0.2471	0.85 [0,5]	0.07
$n = 2000$						
$\widehat{\beta}_m$	0.0636	0.0503	0.0811	0.0649		
Post-ALasso$_{cvse}$	0.0283	0.0774	0.0824	0.0348	3.07 [0,6]	0.89
Post-ALasso$_{ah}$	0.0172	0.0673	0.0694	0.0302	3.05 [1,5]	0.94
$n = 10,000$						
$\widehat{\beta}_m$	0.0278	0.0226	0.0358	0.0285		
Post-ALasso$_{cvse}$	0.0011	0.0185	0.0185	0.0128	3.02 [3,6]	1
Post-ALasso$_{ah}$	0.0009	0.0185	0.0185	0.0128	3.01 [3,5]	1

Notes: Results from 1000 MC replications; $a = 0.2$; $\beta = 0$; γ_{val}= 0.2 $\rho = 0.25$

7 An Alternative Selection Method

As discussed in Section 2, Theorem 1 in KZCS states that the parameters in model (3) are identified if the valid instruments form the largest group, where instruments form a group if they have the same value of π as defined in (7). In the previous section we derived a consistent estimator of the parameters for the case where $s < L/2$, which is a sufficient but not necessary condition for identification.

The Andrews (1999) method in principle could be applied directly, but that would entail estimating all possible configurations of model (3). This would lead to an impractical number of models, even with a moderate number of potential instruments. As highlighted by Andrews and Lu (2001), this would lead to a quite poor performance of the selection method based on Hansen's J-statistic. However, given the simple model structure, we can use the consistent estimates of the π_j as given in (19) to consistently select the correct model. Following Andrews (1999), selection can be done in two fashions, an upward and a downward selection method.

The upward selection method is as follows. Rank-order the $\widehat{\pi}_j$s, and denote the rank-ordered sequence by $\widehat{\pi}_{[j]}$, with $\widehat{\pi}_{[1]}$ the smallest and $\widehat{\pi}_{[L]}$ the largest values of $\widehat{\pi}$. Denote the instruments associated with $\widehat{\pi}_{[j]}$ by $\mathbf{Z}_{.[j]}$. Then, starting with $\mathbf{Z}_{.[1]}$ and $\mathbf{Z}_{.[2]}$, estimate

model (3) by two-step GMM treating $\mathbf{Z}_{.[1]}$ and $\mathbf{Z}_{.[2]}$ as the valid instruments and compute Hansen's J-statistic. Using the same setup as in Section 3, $\mathbf{Z}_{.[1]}$ and $\mathbf{Z}_{.[2]}$ form a group of instruments if the J-test is smaller than the corresponding critical value $\zeta_{n,1}$. If that is the case, we then add in subsequent steps $\mathbf{Z}_{.[3]}, \mathbf{Z}_{.[4]}...$ until the Hansen J-test rejects. If this happens at the l-th step, then $\left\{\mathbf{Z}_{.[1]}, ..., \mathbf{Z}_{.[l-1]}\right\}$ forms a group of instruments. We then repeat this sequence starting with $\mathbf{Z}_{.[2]}$ and $\mathbf{Z}_{.[3]}$, until the last pair $\mathbf{Z}_{.[L-1]}$ and $\mathbf{Z}_{.[L]}$ has been evaluated. We then select as the valid instruments the group with the largest number of instruments, or, as in Andrews (1999), if there are multiple groups with the largest number of instruments, we select within this set the one with the smallest value of the J-test. If there are no groups of size two and above, then none of the instruments can be classified as valid.

The downward selection method is similar, but starts from the model with the set of valid instruments being all instruments, $\left\{\mathbf{Z}_{.[1]}, ..., \mathbf{Z}_{.[L]}\right\}$. If this model is rejected, consider the set $\left\{\mathbf{Z}_{.[1]}, ..., \mathbf{Z}_{.[L-1]}\right\}$ etc. until the Hansen $J-$test does not reject, or the set $\left\{\mathbf{Z}_{.[1]}, \mathbf{Z}_{.[2]}\right\}$ has been evaluated. Then repeat the procedure starting from $\left\{\mathbf{Z}_{.[2]}, ..., \mathbf{Z}_{.[L]}\right\}$, until the final set $\left\{\mathbf{Z}_{.[L-1]}, \mathbf{Z}_{.[L]}\right\}$ has been reached. The final model selection is as described above for the upward testing procedure. Clearly, for both procedures models with a smaller number of valid instruments than the largest group already identified need not be evaluated.

Under the assumption that the group of valid instruments is the largest group, and for the critical values of the J-test obeying assumptions (14) as in Andrews (1999), these procedure will consistently select the correct model. Table 5 presents results for the 2SLS estimator after applying the upward selection method for the same design as in Tables 1/2 and 3/4. The results of the downward procedure are virtually identical. It is clear that for these designs, this sequential approach performs very similar to the best performing (adaptive) post-Lasso estimators.

Table 5. Estimation results for β, Upward selection procedure; $L = 10$, $s = 3$

$2SLS_{ah,up}$ β	bias	std dev	rmse	mad	#sel instr [min, max]	prop $\mathbf{Z}_{.1}, ., \mathbf{Z}_{.3}$ selected
$\gamma_{inv} = \gamma_{val}$						
$n = 500$	0.0568	0.1230	0.1354	0.0816	2.50 [0,4]	0.54
$n = 2000$	0.0043	0.0426	0.0428	0.0285	3.01 [3,5]	1
$n = 10,000$	0.0006	0.0183	0.0183	0.0127	3.01 [3,4]	1
$\gamma_{inv} = 3\gamma_{val}$						
$n = 500$	0.2396	0.1034	0.2610	0.2616	0.81 [0,4]	0.05
$n = 2000$	0.0161	0.0751	0.0768	0.0285	2.97 [1,4]	0.94
$n = 10,000$	0.0006	0.0184	0.0184	0.0127	3.01 [3,4]	1

Notes: Results from 1000 MC replications; $a = 0.2$; $\beta = 0$; γ_{val}= 0.2 $\rho = 0.25$

Again, for $n = 500$, the method produces a substantially upward biased estimator when $\gamma_{inv} = 3\gamma_{val}$. This behaviour can be understood as follows. As

$$\begin{aligned} Y_i &= D_i\beta + \mathbf{Z}_{i.}'\alpha + \varepsilon_i = \mathbf{Z}_{i.}'(\gamma\beta + \alpha) + \varepsilon_i + v_i\beta \\ &= \mathbf{Z}_{i.}'\Gamma + \varepsilon_i + v_i\beta, \end{aligned}$$

we have for this particular design that, under standard regularity conditions, the limiting distribution of $\widehat{\Gamma}$ is given by

$$\sqrt{n}\left(\widehat{\Gamma} - \Gamma\right) \xrightarrow{d} N(0, I).$$

As

$$\widehat{\pi} = T_{\widehat{\gamma}}^{-1}\widehat{\Gamma}$$

where $T_{\widehat{\gamma}} = \text{diag}\left(\widehat{\gamma}_j\right)$, i.e. a diagonal matrix with diagonal elements $\left\{\widehat{\gamma}_j\right\}$, it follows that

$$\sqrt{n}\left(\widehat{\pi} - \pi\right) \xrightarrow{d} N\left(0, T_{\gamma}^{-2}\right),$$

or

$$\widehat{\pi} \overset{a}{\sim} N\left(\pi, n^{-1}T_{\gamma}^{-2}\right). \tag{21}$$

Using this asymptotic distribution to approximate the finite sample distribution, when $a = 0.2$ and $\gamma_{inv} = \gamma_{val} = 0.2$, we have that $a/\gamma_{inv} = 1$, and for a valid and an invalid instrument, we have

$$P\left(\widehat{\pi}_{val} < \widehat{\pi}_{inv}\right) = P\left(\frac{\widehat{\pi}_{val} - \widehat{\pi}_{inv} + 1}{2\left(0.2\sqrt{n}\right)^{-1}} < \frac{1}{2\left(0.2\sqrt{n}\right)^{-1}}\right) \approx \Phi\left(0.1\sqrt{n}\right),$$

which is equal to 0.987 for $n = 500$. When $\gamma_{inv} = 3\gamma_{val} = 0.6$, we have that $a/\gamma_{inv} = 1/3$, and

$$P\left(\widehat{\pi}_{val} < \widehat{\pi}_{inv}\right) = P\left(\frac{\widehat{\pi}_{val} - \widehat{\pi}_{inv} + 1/3}{\left(0.2\sqrt{n}\right)^{-1} + \left(0.6\sqrt{n}\right)^{-1}} < \frac{1/3}{4\left(0.6\sqrt{n}\right)^{-1}}\right) \approx \Phi\left(0.05\sqrt{n}\right),$$

which is equal to 0.868 for $n = 500$. When taking $100,000$ draws from the distribution (21) we find that when $\gamma_{inv} = \gamma_{val} = 0.2$, and $n = 500$, for 98.6% of the draws the $\widehat{\pi}_{[L-2]}, ...\widehat{\pi}_L$ are those associated with the 3 invalid instruments, i.e. with those that have nonzero α. When $\gamma_{inv} = 3\gamma_{val} = 0.6$, this percentage drops to 43.4%. So we see that the relative strength of the invalid instruments affects the ability to segregate the valid and invalid instruments by the values of $\widehat{\pi}$ in finite samples.

8 An application, the effect of BMI on school outcomes using genetic markers as instruments

We have data on $111,091$ individuals from the UK Biobank and investigate the effect of BMI on school outcomes. We use 95 genetic markers as instruments for BMI as identified in independent GWAS studies. The linear model specification includes age, age2 and sex, together with 15 principal components of the genetic relatedness matrix as additional explanatory variables. Table 6 presents the estimation results for the causal effect parameter of BMI. As critical value for the test based procedures we take again $0.1/\ln(n) = 0.0086$.

Table 6. Estimation results

	estimate	st err	# sel instr	p-value J-test	$j : \mathbf{Z}_{.j}$ selected
OLS	-0.300	0.012			
2SLS	-0.203	0.083	0	0.0008	
Lasso$_{cv}$	-0.206		19		
Post-Lasso$_{cv}$	-0.323	0.105	19	0.9945	
Post-Lasso$_{cvse}$	-0.203	0.083	0	0.0008	
Post-Lasso$_{bcch}$	-0.203	0.083	0	0.0008	
Post-Lasso$_{ah}$	-0.136	0.084	2	0.0316	78, 86
ALasso$_{cvse}$	-0.193		3		
Post-ALasso$_{cvse}$	-0.172	0.085	3	0.0894	5, 78, 86
Post-ALasso$_{ah}$	-0.204	0.084	2	0.0272	5, 86
2SLS$_{ah,up}$	-0.168	0.083	4	0.0107	32, 38, 51, 86
2SLS$_{ah,down}$	-0.161	0.083	2	0.0097	51, 86

Note: Coefficients and standard errors multiplied by 10.

The J-test rejects the model treating all 95 instruments as valid, but from the various selection methods, we find that selecting 2 instruments as invalid is sufficient for the J-test to not reject the null of valid instruments. The Post-Lasso$_{ah}$, Post-ALasso$_{ah}$ and the 2SLS$_{ah,down}$ estimators all find 2 instruments as invalid, with the Post-Lasso$_{ah}$ estimator having the smallest value of the J-test. The Lasso$_{cv}$ procedure selects too many, 19, instruments as invalid, whereas both the Lasso$_{cvse}$ and Lasso$_{bcch}$ procedures don't select any instruments as invalid.

All selection procedures that do select instruments as invalid select instrument number 86. This is rs1808579_CT. The Post-Lasso$_{ah}$ estimator further selects rs3888190_CA, whereas the Post-ALasso$_{ah}$ procedure selects rs3101336_TC instead. Interestingly, the ALasso$_{cvse}$ selects the union of these two sets of instruments, which is the third step in the LARS algorithm for *both* the Lasso and adaptive Lasso procedures. Multiplying the estimates by 10 for ease of exposition, we see that the potentially confounded OLS estimate is equal to -0.30 (se 0.012), whereas the 2SLS estimate treating all instruments as valid is smaller and equal to -0.20 (se 0.083). The Post-Lasso$_{ah}$ estimate is equal to -0.14 (se 0.084), whereas the the Post-ALasso$_{ah}$ estimate is equal to -0.20 (se 0.084) and that of the Post-ALasso$_{cvse}$ estimator is equal to -0.172 (se 0.085). So we see that there is considerable heterogeneity in the results, although these are all pointing to a smaller effect than the observational one. This is also the case for the 2SLS$_{ah,down}$ and 2SLS$_{ah,up}$ results of -0.161 (se 0.083) and -0.168 (se 0.083) respectively.

References

[1] Andrews, D.W.K., (1999), Consistent Moment Selection Procedures for Generalized Method of Moments Estimation, *Econometrica* 67, 543-564.

[2] Angrist, J.D., and A.B. Krueger (1991), Does Compulsory School Attendance Affect Schooling and Earnings?, *Quarterly Journal of Economics* 106, 979-1014.

[3] Belloni, A., D. Chen, V. Chernozhukov and C. Hansen, (2012), Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain, *Econometrica* 80, 2369-2429.

[4] Bowden, J., G.D. Smith, S. Burgess, (2015), Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression, *International Journal of Epidemiology* 44, 512-525.

[5] Burgess, S., D.S. Small and S.G. Thompson, A Review of Instrumental Variable Estimators for Mendelian Randomization, *Statistical Methods in Medical Research*, in Press.

[6] Cheng, X., Z. Liao, (2015), Select the Valid and Relevant Moments: An Information-based LASSO for GMM with Many Moments, *Journal of Econometrics* 186, 443-464.

[7] Clarke, P.S. and F. Windmeijer, (2012), Instrumental Variable Estimators for Binary Outcomes, *Journal of the American Statistical Association* 107, 1638-1652.

[8] Efron, B., T. Hastie, I. Johnstone and R. Tibshirani, (2004), Least Angle Regression, *The Annals of Statistics* 32, 407-451.

[9] Greenland, S., (2000), An introduction to instrumental variables for epidemiologists, *International Journal of Epidemiology* 29, 722-729.

[10] Han, C., (2008), Detecting Invalid Instruments using L_1-GMM, *Economics Letters* 101, 285-287.

[11] Hansen, L.P., (1982), Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* 50, 1029-1054.

[12] Imbens, G.W., (2014), Instrumental Variables: An Econometrician's Perspective, *NBER Working Paper 19983*.

[13] Jing, B.-Y., Q.-M. Shao and Q. Wang, (2003), Self-Normalized Cramér-Type Large Deviations for Independent Random Variables, *The Annals of Probability* 31, 2167-2215.

[14] Kang, H., A. Zhang, T.T. Cai and D.S. Small, (2015), Instrumental Variables Estimation with some Invalid Instruments and its Application to Mendelian Randomization, *Journal of the American Statistical Association*, in Press.

[15] Kolesar, M., R. Chetty, J. Friedman, E. Glaeser, G.W. Imbens, (2015), Identification and Inference with Many Invalid Instruments, *Journal of Business and Economic Statistics*, in press.

[16] Lawlor, D.A., R.M. Harbord, J.A.C. Sterne,N. Timpson and G. Davey Smith, (2008), Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology, *Statistics in Medicine* 27, 1133-1163.

[17] Liao, Z., (2013), Adaptive GMM Shrinkage Estimation with Consistent Moment Selection, *Econometric Theory* 29, 857-904.

[18] Lin, W., R. Feng, H. Li, (2015), Regularization Methods for High-Dimensional Instrumental Variables Regression With an Application to Genetical Genomics, *Journal of the American Statistical Association* 110, 270-288.

[19] Sargan, J. D., (1958), The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica* 26, 393–415.

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

[21] Stock, J.H. and M. Yogo, (2005), Testing for weak instruments in linear IV regression. In D.W.K. Andrews and J.H. Stock (Eds.), *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, 80-108. New York: Cambridge University Press.

[22] Zou, H., (2006), The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association* 101, 1418-1429.

9 Appendix

9.1 LARS

Following Efron et al. (2004), the LARS algorithm begins at $\widehat{\boldsymbol{\mu}}_0 = \mathbf{0}$ and builds up $\widehat{\boldsymbol{\mu}}$ by steps. Suppose that $\widehat{\boldsymbol{\mu}}_A$ is the current LARS estimate and that

$$\widehat{\mathbf{c}}_n = n^{-1}\widetilde{\mathbf{Z}}'\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_A\right) \tag{22}$$

is the vector of current correlations. The active set A is the set of indices corresponding to covariates with the greatest absolute current correlations

$$\widehat{C}_n = \max_j \left\{|\widehat{c}_{n,j}|\right\} \quad \text{and} \quad A = \left\{j : |\widehat{c}_{n,j}| = \widehat{C}_n\right\}.$$

Define

$$s_j = \text{sign}\left\{\widehat{c}_{n,j}\right\} \quad \text{for} \quad j \in A$$

and

$$\widetilde{\mathbf{Z}}_A^s = \left(\cdots s_j \widetilde{\mathbf{Z}}_{\cdot j} \cdots\right)_{j\in A} = \widetilde{\mathbf{Z}}_A \mathbf{S}_A$$

with $\mathbf{S}_A = \text{diag}\left(s_j\right)$. Further, define

$$\mathbf{G}_{n,A} = n^{-1}\widetilde{\mathbf{Z}}_A^{s\prime}\widetilde{\mathbf{Z}}_A^s = \mathbf{S}_A\widetilde{\mathbf{Z}}_A'\widetilde{\mathbf{Z}}_A\mathbf{S}_A.$$

and

$$B_{n,A} = \left(\boldsymbol{\iota}_A'\mathbf{G}_{n,A}^{-1}\boldsymbol{\iota}_A\right)^{-1/2}$$

where $\boldsymbol{\iota}_A$ is a vector of ones of length $|A|$, the size of A. Define the equiangular vector

$$\mathbf{u}_{n,A} = \widetilde{\mathbf{Z}}_A\mathbf{S}_A\mathbf{w}_{n,A},$$

where

$$\mathbf{w}_{n,A} = B_{n,A}\mathbf{G}_{n,A}^{-1}\boldsymbol{\iota}_A.$$

Further, define

$$\mathbf{b}_n^A = n^{-1}\widetilde{\mathbf{Z}}'\mathbf{u}_{n,A},$$

with j-th element $b^A_{n,,j}$.

Then the next step of the LARS algorithm updates $\widehat{\boldsymbol{\mu}}_A$ to

$$\widehat{\boldsymbol{\mu}}_{A+} = \widehat{\boldsymbol{\mu}}_A + \widehat{\kappa}_A \mathbf{u}_{n,A}$$

where

$$\widehat{\kappa}_A = \min_{j \in A^c}{}^{+} \left\{ \frac{\widehat{C}_n - \widehat{c}_{n,j}}{B_{n,A} - b^A_{n,j}}, \frac{\widehat{C}_n + \widehat{c}_{n,j}}{B_{n,A} + b^A_{n,j}} \right\} \tag{23}$$

where $\min^{+}$ indicates that the minimum is taken over only positive components within each choice of j. $\widehat{\kappa}_A$ is the smallest positive value of κ_A such some new index $\widehat{j}$ joins the active set; $\widehat{j}$ is the minimizing index in (23) and the new active set A_+ is $A \cup \left\{\widehat{j}\right\}$. The updated correlations are equal to $\widehat{c}_{n,j} - \widehat{\kappa}_A b^A_{n,j}$, the new maximum absolute correlation is $\widehat{C}_{n,+} = \widehat{C}_n - \widehat{\kappa}_A B_{n,A}$, which is the value of the correlations for the active set A_+.

Assuming that all γ_js are the same and that $E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right] = \operatorname{plim}\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right) = \mathbf{I}_L$, we have that

$$\operatorname{plim}\left(n^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right) = \mathbf{I} - L^{-1}\boldsymbol{\iota}_L\boldsymbol{\iota}'_L,$$

and hence

$$\operatorname{plim}\left(\widetilde{\boldsymbol{\Omega}}_n\right) = \operatorname{diag}\left(\sqrt{1 - L^{-1}}\right)$$

and so we can ignore $\widetilde{\boldsymbol{\Omega}}$ asymptotically for this case and focus on $\widehat{\mathbf{c}}_n$ as defined above in (22).

As

$$\operatorname{plim}\left(n^{-1}\widetilde{\mathbf{Z}}'_A\widetilde{\mathbf{Z}}_A\right) = \mathbf{I}_A - L^{-1}\boldsymbol{\iota}_A\boldsymbol{\iota}'_A,$$

it follows that

$$\begin{aligned}\mathbf{G}_A &= \operatorname{plim}\left(n^{-1}G_{n,A}\right) = \mathbf{S}'_A\left(\mathbf{I}_A - L^{-1}\boldsymbol{\iota}_A\boldsymbol{\iota}'_A\right)\mathbf{S}_A \\ &= \mathbf{I}_A - L^{-1}\mathbf{s}_A\mathbf{s}'_A,\end{aligned}$$

where $\mathbf{s}_A$ is the $|A|$ vector of signs $\{s_j\}$. Hence,

$$\mathbf{G}_A^{-1} = \mathbf{I}_A + \left(L - |A|\right)^{-1}\mathbf{s}_A\mathbf{s}'_A$$

and

$$B_A = \operatorname{plim}\left(B_{n,A}\right) = \left(\boldsymbol{\iota}'_A\mathbf{G}_A^{-1}\boldsymbol{\iota}_A\right)^{-1/2} = \left(|A| + \left(L - |A|\right)^{-1}q_A^2\right)^{-1/2}$$

where

$$q_A = \boldsymbol{\iota}_A' \mathbf{s}_A$$

is the difference in the numbers of $+1$ and -1 in $\mathbf{s}_A$. Further,

$$\mathbf{w}_A = \text{plim}\left(\mathbf{w}_{n,A}\right) = B_A \mathbf{G}_A^{-1} \boldsymbol{\iota}_A = B_A \left(\boldsymbol{\iota}_A + \frac{q_A}{(L - |A|)} \mathbf{s}_A \right)$$

and

$$\text{plim}\left(n^{-1} \mathbf{S}_A \widetilde{\mathbf{Z}}_A' \mathbf{u}_A \right) = B_A \boldsymbol{\iota}_A.$$

Then

$$\mathbf{b}^A = \text{plim}\left(\mathbf{b}_n^A\right) = \begin{bmatrix} \text{plim}\left(n^{-1} \widetilde{Z}_A' \widetilde{Z}_A \mathbf{S}_A \mathbf{w}_A \right) \\ \text{plim}\left(n^{-1} \widetilde{Z}_{A^c}' \widetilde{Z}_A \mathbf{S}_A \mathbf{w}_A \right) \end{bmatrix} = \begin{bmatrix} \left(\mathbf{I}_A - L^{-1} \boldsymbol{\iota}_A \boldsymbol{\iota}_A' \right) \mathbf{S}_A \mathbf{w}_A \\ -L^{-1} \boldsymbol{\iota}_{A^c} \boldsymbol{\iota}_A' \mathbf{S}_A \mathbf{w}_A \end{bmatrix}$$

Consider the case as described in Section 2 with all non-zero αs being positive, and ordered such that $\alpha_1 > \alpha_2 > .. > \alpha_s > \alpha_{s+1} = ... = \alpha_L = 0$. We have at $\widehat{\boldsymbol{\mu}}_0 = \mathbf{0}$,

$$\text{plim}\left(\widehat{\mathbf{c}}_n\right) = \text{plim}\left(n^{-1} \widetilde{\mathbf{Z}}' \mathbf{y} \right) = \alpha - \overline{\alpha}. \tag{24}$$

It follows that if $(\alpha_1 - \overline{\alpha}) > \overline{\alpha}$, then $A = A_1 = \{1\}$ and $\widehat{C} = |\alpha_1 - \overline{\alpha}|$. The minimum $\widehat{\kappa}_{A_1}$ for the invalid instruments is given by

$$\min\left(\widehat{\kappa}_{A_1,inv}\right) = \frac{(\alpha_1 - \overline{\alpha}) - (\alpha_2 - \overline{\alpha})}{B_{A_1} - b_2^{A_1}} = \frac{\alpha_1 - \alpha_2}{B_{A_1} - b_2^{A_1}}$$

and for the valid instruments,

$$\min\left(\widehat{\kappa}_{A_1,val}\right) = \frac{(\alpha_1 - \overline{\alpha}) + (-\overline{\alpha})}{B_{A_1} + b_{val}^{A_1}} = \frac{\alpha_1 - 2\overline{\alpha}}{B_{A_1} + b_{val}^{A_1}}$$

and so the invalid $\widetilde{\mathbf{Z}}_{.2}$ enters the active set if

$$\frac{\alpha_1 - \alpha_2}{B_{A_1} - b_2^{A_1}} < \frac{\alpha_1 - 2\overline{\alpha}}{B_{A_1} + b_{val}^{A_1}},$$

and then $\widehat{\kappa}_{A_1} = \min\left(\widehat{\kappa}_{A_1,inv}\right)$.

At step m, assume that $m < s$ invalid instruments have been selected, hence $A = A_m = \{1, 2, ..., m\}$. For all $A_1 ... A_m$ we have that all correlations are positive, so $\mathbf{S}_A = \mathbf{I}_A$, $\mathbf{s}_A = \boldsymbol{\iota}_A$, $q_A = |A|$, and

$$B_A = \left(\frac{L - |A|}{L\,|A|} \right)^{1/2}$$

$$\mathbf{w}_A = B_A \left(1 + \frac{|A|}{L - |A|}\right) \boldsymbol{\iota}_A = B_A \left(\frac{L}{L - |A|}\right) \boldsymbol{\iota}_A$$
$$b_{j+1}^{A_j} = b_{val}^{A_j} = -B_{A_j} \left(\frac{|A_j|}{L - |A_j|}\right).$$

As there are no Lasso steps, $|A_j| = j$.

By repeated substitution, the minimum $\widehat{\kappa}_{A_m}$ for the invalid instruments is then given by

$$\begin{aligned} \min \widehat{\kappa}_{A_m,inv} &= \frac{\left(\alpha_1 - \overline{\alpha} - \sum_{j=1}^{m-1} \widehat{\kappa}_{A_j} B_{A_j}\right) - \left(\alpha_{m+1} - \overline{\alpha} - \sum_{j=1}^{m-1} \widehat{\kappa}_{A_j} b_{m+1}^{A_j}\right)}{B_{A_m} - b_{m+1}^{A_m}} \\ &= \frac{\alpha_1 - \alpha_{m+1} - \sum_{j=1}^{m-1} (\alpha_j - \alpha_{j+1})}{B_{A_m} - b_{m+1}^{A_m}} = \frac{\alpha_m - \alpha_{m+1}}{B_{A_m} - b_{m+1}^{A_m}}. \end{aligned}$$

For the valid instruments it is given by

$$\begin{aligned} \min \widehat{\kappa}_{A_m,val} &= \frac{\left(\alpha_1 - \overline{\alpha} - \sum_{j=1}^{m-1} \widehat{\kappa}_{A_j} B_{A_j}\right) + \left(-\overline{\alpha} - \sum_{j=1}^{m-1} \widehat{\kappa}_{A_j} b_{val}^{A_j}\right)}{B_{A_m} + b_{val}^{A_m}} \\ &= \frac{\alpha_1 - 2\overline{\alpha} - \sum_{j=1}^{m-1} (\alpha_j - \alpha_{j+1}) \frac{B_{A_j} + b_{val}^{A_j}}{B_{A_j} - b_{j+1}^{A_j}}}{B_{A_m} + b_{val}^{A_m}} \end{aligned}$$

as

$$\begin{aligned} B_{A_j} + b_{val}^{A_j} &= B_{A_j} \left(1 - \frac{|A_j|}{L - |A_j|}\right) \\ &= B_{A_j} \left(\frac{L - 2|A_j|}{L - |A_j|}\right) \end{aligned}$$

and

$$\begin{aligned} B_{A_j} - b_{j+1}^{A_j} &= B_{A_j} \left(1 + \frac{|A_j|}{L - |A_j|}\right) \\ &= B_{A_j} \left(\frac{L}{L - |A_j|}\right) \end{aligned}$$

it follows that

$$\frac{B_{A_j} + b_{val}^{A_j}}{B_{A_j} - b_{j+1}^{A_j}} = \frac{L - 2|A_j|}{L} = \frac{L - 2j}{L}.$$

Therefore

$$\begin{aligned} \alpha_1 - 2\overline{\alpha} - \sum_{j=1}^{m-1} (\alpha_j - \alpha_{j+1}) \frac{B_{A_j} + b_{val}^{A_j}}{B_{A_j} - b_{j+1}^{A_j}} &= \alpha_1 - 2\overline{\alpha} - \sum_{j=1}^{m-1} (\alpha_j - \alpha_{j+1}) \frac{L - 2j}{L} \\ &= \frac{L - 2(m-1)}{L} \left(\alpha_m - 2\overline{\alpha}_{(m-1)} \frac{L - (m-1)}{L - 2(m-1)}\right) \end{aligned}$$

where

$$\overline{\alpha}_{(m-1)} = \frac{1}{L-(m-1)} \sum_{j=m}^{L} \alpha_j.$$

Then the next invalid instrument gets selected if

$$\begin{aligned}
\frac{\alpha_m - \alpha_{m+1}}{B_{A_m} - b_{m+1}^{A_m}} &< \frac{\frac{L-2(m-1)}{L}\left(\alpha_m - 2\overline{\alpha}_{(m-1)}\frac{L-(m-1)}{L-2(m-1)}\right)}{B_{A_m} + b_{val}^{A_m}} \\
\alpha_m - \alpha_{m+1} &< \frac{L-2\,(m-1)}{L}\left(\alpha_m - 2\overline{\alpha}_{(m-1)}\frac{L-(m-1)}{L-2\,(m-1)}\right)\left(\frac{L}{L-2m}\right) \\
\alpha_{m+1} &> 2\overline{\alpha}_{(m)}\frac{L-m}{L-2m}.
\end{aligned}$$

Hence the LARS algorithm selects the last invalid instrument at step s if

$$\begin{aligned}
\alpha_s &> 2\overline{\alpha}_{(s-1)}\frac{L-(s-1)}{L-2\,(s-1)} = 2\frac{\alpha_s}{L-(s-1)}\frac{L-(s-1)}{L-2\,(s-1)} \\
L-2s+2 &> 2 \quad \Leftrightarrow \quad s < L/2.
\end{aligned}$$

For L even, if $s = L/2$ then $\min\left(\widehat{\kappa}_{A_s,inv}\right) = \min\left(\widehat{\kappa}_{A_s,val}\right)$ and all remaining instruments get in principle selected as invalid. In practice therefore, the invalid instrument may or may not be selected as invalid. If $s > L/2$, the valid instruments get selected as invalid before all invalid ones have been selected, and hence there is no path that includes all invalid instruments.