

Windmeijer, Frank; Farbmacher, Helmut; Davies, Neil; Smith, George Davey

Conference Paper

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

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2017: Alternative Geld- und Finanzarchitekturen - Session: Treatment Effects, No. G03-V3

Provided in Cooperation with:

Verein für Socialpolitik / German Economic Association

Suggested Citation: Windmeijer, Frank; Farbmacher, Helmut; Davies, Neil; Smith, George Davey (2017) : On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2017: Alternative Geld- und Finanzarchitekturen - Session: Treatment Effects, No. G03-V3, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel, Hamburg

This Version is available at:

<https://hdl.handle.net/10419/168196>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

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

Frank Windmeijer^{a,d,†}, Helmut Farbmacher^b, Neil Davies^{c,d}
George Davey Smith^{c,d}

^aDepartment of Economics
University of Bristol, UK

^bCenter for the Economics of Aging
Max Planck Institute Munich, Germany

^cSchool of Social and Community Medicine
University of Bristol, UK

^dMRC Integrative Epidemiology Unit, Bristol, UK

June 2016

Abstract

We investigate the behaviour 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, Zhang, Cai and Small (2016, Journal of the American Statistical Association). 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 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 a median 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. The methods are applied to a Mendelian randomisation study to estimate the causal effect of

*This research was partly funded by the Medical Research Council (MC_UU_12013/9). Helpful comments were provided by Chirok Han, Whitney Newey, Hyunseung Kang, Chris Skeels, 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.

†f.windmeijer@bristol.ac.uk

BMI on diastolic blood pressure using data on individuals from the UK Biobank, with 96 single nucleotide polymorphisms as potential instruments for BMI.

Key Words: causal inference, instrumental variables estimation, invalid instruments, Lasso, Mendelian randomisation

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’ condition. 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. (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’, 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). 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).

In this paper we consider violations of the exclusion condition of the instruments, following closely the setup of 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) 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. (2016) do 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 the Kang et al. (2016) setup and requires a large number of model evaluations, which has a negative impact on the performance of the selection procedure.

This paper assesses the performance of the Kang et al. (2016) Lasso type selection and estimation procedure. The Lasso can be obtained using a modification of the Least Angle Regression (LARS) algorithm of Efron et al. (2004). By evaluating the LARS/Lasso path 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, 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 relative strength of the 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.¹ Although the less than 50% invalid instruments condition needed for the median estimator is a stronger condition in principle than needed for the Lasso estimator, we derive asymptotic results for a simple model design where the

¹Bowden 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.

parameters are identified when there are more than 50% invalid instruments, but the Lasso method will only select the invalid instruments correctly if there are less than 50% invalid instruments.

Instrument strength is very likely to vary by instruments, so it will be important to consider our adaptive Lasso approach for assessing instrument validity and estimating causal effects. In Mendelian randomisation studies it is clear that genetic markers have differential impacts on exposures from examining the results from genome wide association studies.

Another strand of the literature focuses on instrument selection in potentially 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 find that simply using Hansen’s J -test for a stopping rule performs well. We find that a 10-fold cross-validation method to determine the Lasso penalty parameter selects too many instruments as invalid, which does not improve with increasing sample size, and explain this using distribution theory related to the Hansen J -test. Further, we show and explain that the so-called post-Lasso selection two-stage least squares estimator for the causal effect has smaller finite sample bias than the Lasso estimator.

The next section, Section 2, introduces the model and the Lasso estimator as proposed by Kang et al. (2016). In Section 3, we analyse the behaviour of the Lasso estimator under conditions where it selects the invalid instruments consistently and assess various stopping rules for selecting the number of invalid instruments. In Section 4, we show that the Lasso may select the valid instruments as invalid if the invalid instruments are relatively strong. Section 5 discusses the identification issues that arise from different correlation structures of the instruments. Section 6 introduces the median estimator, establishes its consistency and shows that its asymptotic behaviour is such that the adaptive Lasso estimator enjoys oracle properties. In Section 7 the methods are applied to a Mendelian randomisation study to estimate the causal effect of BMI on diastolic

blood pressure using data on individuals from the UK Biobank, with 96 single nucleotide polymorphisms as potential instruments for BMI. Section 8 concludes.

2 Model and Lasso Estimator

We follow Kang, Zhang, Cai and Small (2016) (KZCS from now on), who considered the following potential outcomes model. For $i = 1, \dots, n$, let $Y_i^{(d, \mathbf{z})}$, 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

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

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

where ϕ measures the direct effect of \mathbf{z} on Y , and ψ 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

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

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

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

and hence $E[\varepsilon_i | \mathbf{Z}_i] = 0$. 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 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

$\mathbf{R} = [\mathbf{d} \ \mathbf{Z}_{sel}]$. The standard Instrumental Variables, or Two-Stage Least Squares (2SLS), estimator is then given by

$$\hat{\theta} = \begin{pmatrix} \hat{\beta} \\ \hat{\alpha}_{sel} \end{pmatrix} = \left(\mathbf{R}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R} \right)^{-1} \mathbf{R}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}. \quad (4)$$

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

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

where ξ_i is defined implicitly, and hence

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

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

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

$$D_i = \mathbf{Z}'_i\gamma + v_i, \quad (6)$$

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

$$\pi_j \equiv \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j}. \quad (7)$$

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

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

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

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

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

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

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 called *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 $A_n = \{j : \widehat{\alpha}_j^\lambda \neq 0\}$. We will first investigate under what conditions the Lasso method consistently selects invalid instruments such that $\lim_{n \rightarrow \infty} P(A_n = A) = 1$, or for a weaker version, such that $\lim_{n \rightarrow \infty} P(A_n \supseteq A) = 1$.

An important difference with the standard Lasso approach for linear models is that the matrix of explanatory variables $\mathbf{M}_{\widehat{\mathbf{d}}} \mathbf{Z}$ 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 in the model 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, $n \rightarrow \infty$, 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 $\mu = \max_{j \neq r} |\mathbf{Z}'_j \mathbf{Z}_r|$ and $\rho = \max_j \left| \frac{\mathbf{Z}'_j \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}' \widehat{\mathbf{d}}} \right|$. 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

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

Therefore, if $\mathbf{Q} = \mathbf{I}_L$ this is equal to $|\gamma_j|/\gamma' \gamma$ 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 $\hat{\pi}_j = \hat{\Gamma}_j / \hat{\gamma}_j$, where $\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{d}$ and $\hat{\Gamma} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y}$. Under standard assumptions as specified below, we show that the median of the $\hat{\pi}_j$ 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 as an initial consistent estimator for the adaptive Lasso.

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

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

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

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

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. $\gamma_j = \tilde{\gamma}$ for $j = 1, \dots, L$, and the instruments are uncorrelated, with variances equal to 1, $E[\mathbf{Z}_i \mathbf{Z}'_i] = \mathbf{I}_L = \text{plim}(n^{-1} \mathbf{Z}'\mathbf{Z})$.

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

$$\hat{\alpha}^\lambda = \arg \min_{\alpha} \frac{1}{2n} \|\mathbf{y} - \tilde{\mathbf{Z}}\alpha\|_2^2 + \frac{\lambda}{n} \|\tilde{\mathbf{\Omega}}_n \alpha\|_1, \quad (11)$$

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

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 $\hat{\mathbf{c}}_n$ be defined as

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

with j -th element

$$\widehat{c}_{n,j} = \frac{n^{-1}\mathbf{Z}'_j\mathbf{M}_{\widehat{d}}\mathbf{y}}{\sqrt{n^{-1}\mathbf{Z}'_j\mathbf{M}_{\widehat{d}}\mathbf{Z}_j}} \quad (12)$$

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

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

as

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

and

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

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

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

$$(L-s)|a| > s|a| \quad \Leftrightarrow \quad s < L/2, \quad (13)$$

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 finite sample 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 . 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. Of interest is the following result.

Proposition 1

Consider observational models (3) and (6) under assumptions C1, with $\mathbf{Q} = \mathbf{I}_L$, C2, C3, and with equal instrument strength $\gamma_j = \tilde{\gamma}$ for $j = 1, \dots, L$. The s invalid instruments have distinct positive effects, ordered in such a way that $\alpha_1 > \alpha_2 > \dots > \alpha_s > \alpha_{s+1} = \dots = \alpha_L = 0$. Then the LARS/Lasso algorithm selects the invalid instruments in the first s steps in large samples if $s < L/2$. For $s > L/2$ the full LARS/Lasso path does not include a model where all invalid instruments have been selected.

Proof: see Appendix.

The result of Proposition 1 is striking, as the parameters are formally identified in this case when $s < L - 1$ and hence the Lasso approach here requires stronger assumptions for selecting the invalid instruments than is needed for identification.

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 Zhao and Yu (2006) and 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

$$\begin{aligned} \begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right); \\ \mathbf{Z}_i &\sim N(0, \mathbf{I}_L); \end{aligned}$$

and we set $\beta = 0$; $\gamma = 0.2\boldsymbol{\iota}_L$ with $\boldsymbol{\iota}_L$ an L -vector of ones; $L = 10$; $\rho = 0.25$; $s = 3$, and the first s elements of α are equal to $a = 0.2$. Note that none of the estimation results presented here and below depend on the value of β . Table 1 presents estimation results for estimators of β 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 $n = 10,000$.

The information content for IV estimation can be summarised by the concentration parameter, see Rothenberg (1984). If all instruments are valid, i.e. $\alpha = 0$, the concentration parameter would be given by $\gamma'\mathbf{Z}'\mathbf{Z}\gamma/\sigma_v^2$. For this data generating process with independent instruments, the expected concentration parameter is therefore given by $n(L-s)(0.2^2)$ and hence equal to 140, 560 and 2800 for the three sample sizes. 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. The F statistic is a test for $H_0 : \gamma_{val} = 0$, where γ_{val} is the $L-s$ sub-vector of γ associated with the valid instruments.

The "2SLS" results are for the 2SLS estimator that treats all instruments as valid. 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 λ has been obtained by 10-fold cross-validation, using the one-standard error rule, as in KZCS. This estimator is denoted "Lasso_{cvse}" 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_{cv}". We further present results 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 α . Further entries in Table 1 are the average number of instruments selected as invalid, 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 instruments selected as invalid include all 3 invalid instruments.

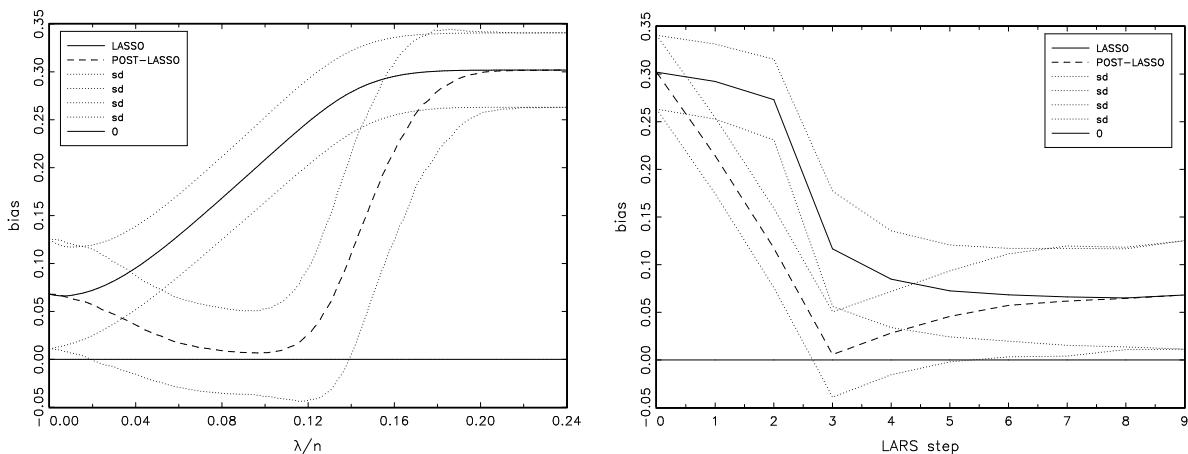
Table 1. Estimation results for β ; $L = 10$, $s = 3$

β	bias	std dev	rmse	mad	av. # instr selected as invalid [min, max]	freq. all s invalid instr 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	6.64 [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 fewer, 3.17 for $n = 500$. Although the 3 invalid instruments are always jointly selected as invalid 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 outperformed by the oracle 2SLS estimator at $n = 10,000$. Although the post-Lasso_{cvse} estimator has a larger standard deviation than the Lasso_{cvse} estimator, it has a smaller bias, rmse and mad for all sample sizes.

Figures 1a and 1b illustrate 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 is (much) less biased, 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 finite sample 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}^λ the matrix of selected instruments for any value of λ , i.e. those instruments with non-zero values of $\hat{\alpha}^\lambda$. For the Lasso and post-Lasso estimators, $\hat{\beta}^\lambda$ and $\hat{\beta}$, we have that

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

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

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

Note that we have for the 2SLS estimator in the model that treats all instruments as valid,

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

Therefore, the bias of the Lasso estimator due to shrinkage is in the direction of the bias of the 2SLS estimator in the model where all instruments are treated as valid. As an illustration, for the $n = 2000$ case above, at $\lambda/n = 0.0965$, the means of the first three elements of $\hat{\alpha}_{sel}^\lambda$ are all equal to 0.067, whereas those of the post-Lasso 2SLS estimator are equal to 0.198, hence $1 - \hat{\delta}^\lambda = 0.662$. 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 $\hat{\beta}^\lambda$ and $\hat{\beta}$ 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, that the 10-fold cross-validation method selects too many valid instruments as invalid over and above the invalid ones. 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

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

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

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

where

$$\tilde{\varepsilon}_i = y_i - \tilde{\mathbf{Z}}_i' \alpha.$$

Then let

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

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 $\tilde{\mathbf{c}}_n^*$ and hence λ/n for models that have selected the invalid instruments. They establish that

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

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

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

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 $\tau_n = 0.1/\ln(n)$ 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 $\tilde{\varepsilon}_i$ need to be estimated to construct $\tilde{\mathbf{\Omega}}_n^*$. We use the post-Lasso estimator at every step to estimate the $\tilde{\varepsilon}_i$.

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

$$\begin{aligned} \mathbf{y} &= \mathbf{d}\beta + \mathbf{Z}_{sel}\alpha_{sel} + \boldsymbol{\xi} \\ &= \mathbf{R}\boldsymbol{\theta} + \boldsymbol{\xi}. \end{aligned} \tag{14}$$

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

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

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, $E(\xi_i^2|\mathbf{Z}_i) = \sigma_\xi^2$. Under general forms of conditional heteroskedasticity, an efficient two-step GMM estimator is obtained by setting

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

where $\widehat{\boldsymbol{\theta}}_1$ 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

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

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 used purely for selection of the invalid instruments. For consistent model selection, the critical values $\zeta_{n,k}$ need to satisfy

$$\zeta_{n,k} \rightarrow \infty \quad \text{and} \quad \zeta_{n,k} = o(n), \tag{15}$$

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

From the distribution theory above we can explain the fact that cross-validation selects too many instruments as invalid. In the 10-fold cross-validation as proposed by KZCS, the model is estimated on a random 90% training sample for a grid of values of λ/n . The estimates are then used in the validation sample to evaluate the minimisation criterion

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

where the subscripts v and t here indicate that the data are from the validation sample and the estimates obtained from the training sample. The value of λ/n is chosen that minimises the average of $S_v \left(\widehat{\beta}_t^\lambda, \widehat{\alpha}_t^\lambda \right)$ over the 10 folds. In this homoskedastic design, note that at the true parameter values

$$\frac{S_v(\beta, \alpha)}{\sigma_\varepsilon^2} \xrightarrow{d} \chi_L^2.$$

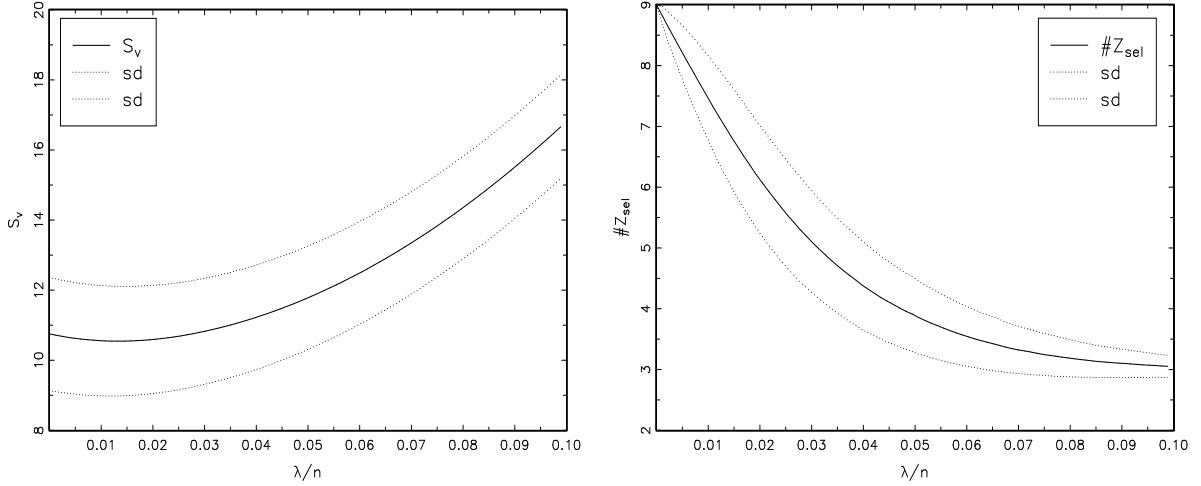
The same result holds for any \sqrt{n} consistent estimator of $(\beta, \alpha)'$, estimated in the training sample. Therefore, for the cross-validation exercise,

$$\frac{S_v \left(\widehat{\beta}_t^\lambda, \widehat{\alpha}_t^\lambda \right)}{\sigma_\varepsilon^2} \stackrel{a}{\sim} \chi_L^2,$$

for values of λ/n such that the invalid instruments have correctly been selected. However, due to the finite sample shrinkage bias, as discussed above and seen in Figure 1a, $S_v \left(\widehat{\beta}_t^\lambda, \widehat{\alpha}_t^\lambda \right)$ will be larger than $\sigma_\varepsilon^2 L$ in expectation. Figure 2a below confirms this. It shows the Monte Carlo mean of the 10-fold average of $S_v \left(\widehat{\beta}_t^\lambda, \widehat{\alpha}_t^\lambda \right)$ as a function of λ/n , for the $n = 2,000$ (and $\sigma_\varepsilon^2 = 1$) case. It clearly mimics the bias results for $\widehat{\beta}^\lambda$ as depicted in Figure 1a and is close to $L = 10$ when the bias of $\widehat{\beta}^\lambda$ is small, and larger everywhere else. This bias is small for small values of λ/n and hence the cross-validation method is conservative in selecting too many instruments as invalid. This is confirmed in Figure 2b, which shows the average number of instruments selected as invalid in the training samples as a function of λ/n .

The Monte Carlo frequency distribution of selecting 3, 4, ..., 9 instruments as invalid for the Lasso_{cv} estimator in Table 1 for the $n = 2,000$ case is given by 0.024, 0.081, 0.194,

0.175, 0.167, 0.177, 0.182. This is an approximate uniform distribution for 5, 6, ..., 9 instruments as expected from Figures 2a and 2b.



Figures 2a and 2b. Means and standard deviations of 10-fold average $S_v \left(\hat{\beta}_t^\lambda, \hat{\alpha}_t^\lambda \right)$, and number of selected instruments as invalid in training samples. Same design as in Table 1, $n = 2000$, 1000 MC replications.

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	av. # instr selected as invalid [min, max]	freq. all s invalid instr selected
$n = 500$						
Lasso _{<i>bcch</i>}	0.2770	0.0846	0.2896	0.2770	1.16 [0,4]	0.068
Post-Lasso _{<i>bcch</i>}	0.1914	0.1324	0.2327	0.2028		
Post-Lasso _{<i>ah</i>}	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 _{<i>bcch</i>}	0.1688	0.0438	0.1744	0.1694	3.11 [3,5]	1
Post-Lasso _{<i>bcch</i>}	0.0091	0.0445	0.0454	0.0294		
Post-Lasso _{<i>ah</i>}	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 _{<i>bcch</i>}	0.0751	0.0180	0.0772	0.0756	3.11 [3,5]	1
Post-Lasso _{<i>bcch</i>}	0.0027	0.0191	0.0193	0.0134		
Post-Lasso _{<i>ah</i>}	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 in this particular case. This is similar to the 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.

Using standard asymptotic robust standard errors for the post-Lasso 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 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 $E(\mathbf{Z}_i \mathbf{Z}_i') = \mathbf{I}_L$,

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

It is clear that allowing for differential instrument strengths, i.e. different values of γ , may result in the LARS/Lasso path not selecting all invalid instruments as invalid. 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,

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

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

$$\frac{st}{\sqrt{st^2 + L - s - 1}} > \frac{L - s}{\sqrt{(s - 1)t^2 + L - s}}. \quad (16)$$

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 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 $\beta + a/\gamma_{inv} = 0 + 0.2/0.6 = 0.33$, which is the value of the causal effect estimated by the invalid instruments. The results in Table 3 confirm that, for large sample sizes, the LARS/Lasso method selects 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	av. # instr selected as invalid [min, max]	freq. all s invalid instr 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; $a = 0.2$; $\beta = 0$; $\gamma_{val} = 0.2$ $\rho = 0.25$

It is clear that various combinations of instrument strengths 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 as invalid.

On the other hand, from (16) 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 $t = 0.5$, $|\text{plim}(\hat{c}_{n,inv})| > |\text{plim}(\hat{c}_{n,val})|$ for $s = 1, \dots, 6$.

5 Correlated Instruments

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

$$E[\mathbf{Z}_i \mathbf{Z}_i'] = \text{plim}(n^{-1} \mathbf{Z}' \mathbf{Z}) = \mathbf{Q}$$

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

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

$$\text{plim} \left(n^{-1} \mathbf{Z}'_j \mathbf{M}_d \mathbf{y} \right) = \mathbf{Q}'_j \left(\alpha - \gamma \frac{\gamma' \mathbf{Q} \alpha}{\gamma' \mathbf{Q} \gamma} \right) = \mathbf{Q}'_j \left(\alpha - \boldsymbol{\nu}_L \frac{\boldsymbol{\nu}'_L \mathbf{Q} \alpha}{\boldsymbol{\nu}'_L \mathbf{Q} \boldsymbol{\nu}_L} \right),$$

where \mathbf{Q}_j is the j th column of \mathbf{Q} ; $\boldsymbol{\nu}_L$ is an L -vector of ones, and the second result follows because the γ_j are all the same.

For the denominator, we get

$$\text{plim} \left(n^{-1} \mathbf{Z}'_j \mathbf{M}_d \mathbf{Z}_j \right) = 1 - \frac{(\mathbf{Q}'_j \boldsymbol{\nu}_L)^2}{\boldsymbol{\nu}'_L \mathbf{Q} \boldsymbol{\nu}_L} = 1 - \frac{(\mathbf{Q}'_j \boldsymbol{\nu}_L)^2}{\boldsymbol{\nu}'_L \mathbf{Q} \boldsymbol{\nu}_L}.$$

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

$$\text{plim} \left(\widehat{c}_{n,j} \right)_{j \in \{1, \dots, s\}} = \frac{a \mathbf{Q}'_j \left(\mathbf{e}_s - \boldsymbol{\nu}_L \frac{\boldsymbol{\nu}'_L \mathbf{Q} \mathbf{e}_s}{\boldsymbol{\nu}'_L \mathbf{Q} \boldsymbol{\nu}_L} \right)}{\sqrt{1 - \frac{(\mathbf{Q}'_j \boldsymbol{\nu}_L)^2}{\boldsymbol{\nu}'_L \mathbf{Q} \boldsymbol{\nu}_L}}}, \quad (17)$$

and for the valid instruments

$$\text{plim} \left(\widehat{c}_{n,r} \right)_{r \in \{s+1, \dots, L\}} = \frac{a \mathbf{Q}'_r \left(\mathbf{e}_s - \boldsymbol{\nu}_L \frac{\boldsymbol{\nu}'_L \mathbf{Q} \mathbf{e}_s}{\boldsymbol{\nu}'_L \mathbf{Q} \boldsymbol{\nu}_L} \right)}{\sqrt{1 - \frac{(\mathbf{Q}'_r \boldsymbol{\nu}_L)^2}{\boldsymbol{\nu}'_L \mathbf{Q} \boldsymbol{\nu}_L}}}, \quad (18)$$

where $\mathbf{e}_s = \left(\boldsymbol{\nu}'_s \quad \mathbf{0}'_{L-s} \right)'$ 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 $\mathbf{Q}'_j \boldsymbol{\nu}_L = \mathbf{Q}'_r \boldsymbol{\nu}_L$ and the invalid instruments are 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

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

which is the same result as (13) derived for uncorrelated 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 (17) and (18), 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 as invalid, even when $s < L/2$. This has been documented well for the Lasso in the standard linear model, see e.g. Zhao and Yu (2006) and 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 (17) and (18) we get for the invalid and valid instruments

$$\begin{aligned}\text{plim}(\widehat{c}_{n,inv})/a &= 0.5570 \\ \text{plim}(\widehat{c}_{n,val})/a &= -0.5915.\end{aligned}$$

Hence, for this parameter configuration and 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, $E[Y_i^{(0,0)}|\mathbf{Z}_i] = \mathbf{Z}_i'\psi$, 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. Consider changing the conditional mean assumption (2) to one of correlation, i.e.

$$E[Y_i^{(0,0)}\mathbf{Z}_i] = \widetilde{\psi}, \tag{19}$$

with some of the elements of $\widetilde{\psi}$ equal to 0. These are for example the moments considered by Han (2008), 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 the outcome model (3) under condition (19), 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 less than 50% of the potential instruments are invalid. This is the same condition as that for the LARS/Lasso selection to be guaranteed 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 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 2, 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 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 $\widehat{\Gamma} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y}$; $\widehat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{d}$, and let $\widehat{\pi}$ be the L -vector with j -th element

$$\widehat{\pi}_j = \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}, \quad (20)$$

Under the standard assumptions, we show below that the median of the $\widehat{\pi}_j$ 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.

Proposition 2

Under model specifications (3) and (6), assumptions C1-C3, the L -vector $\widehat{\pi}$ with elements as defined in (20) and the condition that $s < L/2$, the estimator $\widehat{\beta}_m$, defined as

$$\widehat{\beta}_m \equiv \text{median}(\widehat{\pi}) \quad (21)$$

is a consistent estimator for β .

Proof: Under the stated assumptions,

$$\begin{aligned}\text{plim}(\widehat{\Gamma}) &= \gamma\beta + \alpha; \\ \text{plim}(\widehat{\gamma}) &= \gamma.\end{aligned}$$

Hence

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

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

$$\text{plim}(\widehat{\beta}_m) = \text{median}\{\text{plim}(\widehat{\pi})\} = \beta.$$

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

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

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

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

and, for given values of v can be estimated straightforwardly using the LARS algorithm, see Zou (2006).

Table 4 presents the estimation results for the adaptive Lasso for the design as in Table 3, setting $v = 1$. 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.

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

β	bias	std dev	rmse	mad	av. # instr selected as invalid [min, max]	freq. all s invalid instr selected
$n = 500$						
$\widehat{\beta}_m$	0.1126	0.0935	0.1463	0.1129		
Post-ad Lasso _{cvse}	0.2426	0.0787	0.2550	0.2568	0.46 [0,6]	0.04
Post-ad Lasso _{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-ad Lasso _{cvse}	0.0283	0.0774	0.0824	0.0348	3.07 [0,6]	0.89
Post-ad Lasso _{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-ad Lasso _{cvse}	0.0011	0.0185	0.0185	0.0128	3.02 [3,6]	1
Post-ad Lasso _{ah}	0.0009	0.0185	0.0185	0.0128	3.01 [3,5]	1

Notes: Results from 1000 MC replications; $a = 0.2$; $\beta = 0$; $\gamma_{val} = 0.2$; $\rho = 0.25$; $\nu = 1$

As can be seen in Table 4, the adaptive Lasso 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. This is as expected from the theoretical results of Zou (2006), who shows that the adaptive Lasso has oracle properties in terms of consistency of variable selection and asymptotic normality. Following Remark 1 in Zou (2006, p. 1420), $\widehat{\alpha}_m$ is not required to be \sqrt{n} consistent. As the results show, there is still some bias in $\widehat{\beta}_m$, even at the largest sample size. This is due to the fact that the median estimator, whilst converging at the \sqrt{n} rate, has an asymptotic bias in this design because its limiting distribution is that of an order statistic which does not have zero mean. It follows that $\widehat{\alpha}_m$ converges at the \sqrt{n} rate, but with an asymptotic bias. This implies that $\sqrt{n}(\widehat{\alpha}_m - \alpha) = O_p(1)$ and hence the conditions in Remark 1 in Zou (2006) hold for the adaptive Lasso proposed here to enjoy oracle properties.

To derive the asymptotic bias of the median estimator $\widehat{\beta}_m$, let δ be the L -vector with elements

$$\delta_j = \frac{\alpha_j}{\gamma_j}.$$

Partition δ as $\delta = (\delta'_s \ \delta'_0)'$, where δ_s contains the elements of δ that are not equal to 0, and $\delta_0 = \mathbf{0}_{L-s}$. Partition $\widehat{\pi}$ accordingly as $\widehat{\pi} = (\widehat{\pi}'_s \ \widehat{\pi}'_0)'$. Under the standard

conditions and using the delta method, the limiting distribution of $\widehat{\pi}$ is given by

$$\sqrt{n}(\widehat{\pi} - (\beta\boldsymbol{\iota}_L + \delta)) \xrightarrow{d} N(0, \Sigma_\pi).$$

As $\widehat{\beta}_m = \text{median}(\widehat{\pi})$,

$$\begin{aligned} \sqrt{n}(\widehat{\beta}_m - \beta) &= \sqrt{n}(\text{median}(\widehat{\pi}) - \beta) \\ &= \text{median}(\sqrt{n}(\widehat{\pi} - \beta\boldsymbol{\iota}_L)). \end{aligned}$$

As

$$\sqrt{n}(\widehat{\pi} - \beta\boldsymbol{\iota}_L) = \begin{pmatrix} \sqrt{n}(\widehat{\pi}_s - (\beta\boldsymbol{\iota}_s + \delta_s)) + \sqrt{n}(\delta_s) \\ \sqrt{n}(\widehat{\pi}_0 - \beta\boldsymbol{\iota}_{L-s}) \end{pmatrix},$$

it then follows that

$$\sqrt{n}(\widehat{\beta}_m - \beta) = \text{median}(\sqrt{n}(\widehat{\pi} - \beta\boldsymbol{\iota}_L)) \xrightarrow{d} q_{[l], L-s}$$

where for L odd, $q_{[l], L-s}$ is the l -th order-statistic of the limiting distribution of $\sqrt{n}(\widehat{\pi}_0 - \beta\boldsymbol{\iota}_{L-s})$, where l is determined by L , s and the signs of the elements of δ_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.

The design in Table 4 has $L = 10$, $s = 3$ and $\alpha_1 = \alpha_2 = \alpha_3 = a > 0$. The median is defined here as $\widehat{\beta}_m = (\widehat{\pi}_{[5]} + \widehat{\pi}_{[6]})/2$. We therefore get that

$$\sqrt{n}(\widehat{\beta}_m - \beta) = \text{median}(\sqrt{n}(\widehat{\pi} - \beta\boldsymbol{\iota}_{10})) \xrightarrow{d} q_{[5,6],7}$$

where $q_{[5,6],7}$ is the limiting distribution of the average of the fifth and sixth order statistic of

$$\sqrt{n} \left(\begin{pmatrix} \widehat{\pi}_4 \\ \vdots \\ \widehat{\pi}_{10} \end{pmatrix} - \beta\boldsymbol{\iota}_7 \right) \xrightarrow{d} N(0, \Sigma_\pi^*),$$

where for the design in Table 4, $\Sigma_\pi^* = 25\mathbf{I}_7$, as $\sigma_\varepsilon^2 = 1$ and $1/\gamma_j^2 = 25$ for $j = 4, \dots, 10$. From a simple simulation, drawing repeatedly from the $N(0, 25\mathbf{I}_7)$ distribution, we find that $E[q_{[5,6],7}] = 2.78$. Therefore $E[q_{[5,6],7}]/\sqrt{n} = 0.0278$ for $n = 10,000$, exactly the result found for the bias of $\widehat{\beta}_m$ in Table 4.

For this design, the asymptotic bias of the median estimator is affected by the number of invalid instruments in the following way. For $n = 10,000$ we get for $s = 4, 2, 1, 0$

respectively $E [q_{[5,6],6}] / \sqrt{n} = 0.0477$; $E [q_{[5,6],8}] / \sqrt{n} = 0.0156$; $E [q_{[5,6],9}] / \sqrt{n} = 0.0069$; and $E [q_{[5,6],10}] / \sqrt{n} = 0$.

Having all elements of δ_s with the same sign is clearly the worst case scenario for the asymptotic bias of the median estimator. The best case scenario is for even s , if half the elements in δ_s are positive and half negative, as we then have that $\sqrt{n} (\hat{\beta}_m - \beta)$ converges to the median of the limiting distribution of $\sqrt{n} (\hat{\pi}_0 - \beta \mathbf{1}_{L-s})$, and therefore has no asymptotic bias.

For the results in Table 4, for $n = 2000$, the means of the estimates for the positive $\alpha_j = 0.2$ are approximately 0.161 whereas the means of the estimates for the $\alpha_j = 0$ are approximately 0.0186. For $n = 10,000$, these are approximately 0.183 and 0.0085. The ratios of the biases for $n = 10,000$, relative to those of $n = 2000$ are approximately 0.44 which is equal to $\sqrt{2000}/\sqrt{10,000}$, confirming that the bias in $\hat{\alpha}_m$ decreases at the \sqrt{n} rate.

Using standard asymptotic robust standard errors for the post-adaptive Lasso_{ah} 2SLS estimator, the Wald test for the null $H_0 : \beta = 0$, at the 10% level, rejects 15.4% and 9.2% for $n = 2000$ and $n = 10,000$ respectively, confirming the oracle property for the large sample size.

7 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. (2016) 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 randomisation studies the SNPs used as potential instruments can be invalid for various reasons, such as linkage disequilibrium, population stratification and pleiotropy, see e.g. von Hinke et al. (2016). For example, a SNP has pleiotropic effects if it not only affects the exposure but also has a direct effect on the outcome. Whilst we guard against population stratification by considering only Caucasian UK 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, age² and sex, together with 15 principal components of the genetic relatedness matrix as additional explanatory variables. Table 5 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 critical value for the Hansen test based procedures we take again $0.1/\ln(n) = 0.0086$.

Table 5. Estimation results, the effect of $\ln(BMI)$ on $\ln(DBP)$

	estimate	st err	# instr selected as invalid	p-value J -test
OLS	0.206	0.003		
2SLS	0.087	0.016	0	0.0000
Lasso _{cvse}	0.110		18	
Post-Lasso _{cvse}	0.147	0.020	18	0.4889
Post-Lasso _{ah}	0.122	0.018	12	0.0122
median, $\hat{\beta}_m$	0.148			
ad Lasso _{cvse}	0.131		17	
Post-ad Lasso _{cvse}	0.151	0.019	17	0.4091
Post-ad Lasso _{ah}	0.163	0.018	11	0.0102

Notes: sample size $n = 105,276$; $L = 96$

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). The J -test, however, rejects the null that all the instruments are valid. The Lasso_{cvse} procedure identifies 18 instruments as invalid and the Lasso_{cvse} estimate is equal to 0.110. The Post-Lasso_{cvse} estimate is equal to 0.147, which is in line with our findings above that the Lasso estimator is biased towards the 2SLS estimator that treats all instruments as valid due to shrinkage. The Post-Lasso_{ah} procedure selects a subset of 12 instruments as invalid, and the Post-Lasso_{ah} 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 *cvse* method selecting 17 instruments as invalid. The adaptive Lasso_{ah}

method selects a subset of 11 instruments as invalid. The adaptive Lasso_{cvse}, post-Lasso_{cvse} and post-Lasso_{ah} estimates are equal to 0.131, 0.151 and 0.163 respectively. The median and the two post-adaptive Lasso estimators 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 5 it is selected as an invalid instrument. The value for $\hat{\pi}_{FTO} = -0.009$, i.e. negative, which is contrary to the direction of the found causal effect.

The F statistic for $H_0 : \gamma_{val} = 0$ for the model resulting from the adaptive Lasso_{ah} 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 equal to 0.159 (s.e. 0.019), indicating that there is not a many weak instruments problem here, see Davies et al. (2015).

8 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 non-random selection of exposure. The main identifying assumption is that the instruments satisfy the exclusion restriction, i.e. they only affect the outcomes through their relationship with the exposure. In an important contribution, Kang et al. (2016) show 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, i.e. 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 (2016) and we show that it performs well for larger sample sizes in our simulations, solving the inconsistency problem of the Lasso method when the relative strength of the invalid instruments is such that Lasso method selects the valid instruments as invalid.

Whilst less than 50% invalid instruments is a sufficient condition for identification, in principle the parameters are identified if the valid instruments form the largest group. Instruments form a group if they have the same estimate for the causal effect. Future research will therefore focus on how to obtain consistent results when more than 50% of the instruments are invalid, but the parameters are such that they are asymptotically identified.

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, (2015), A Review of Instrumental Variable Estimators for Mendelian Randomization, *Statistical Methods in Medical Research*, in Press.
- [6] Collins, R., (2012), What makes UK Biobank special? *The Lancet* 379, 1173–1174.

- [7] 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.
- [8] Clarke, P.S. and F. Windmeijer, (2012), Instrumental Variable Estimators for Binary Outcomes, *Journal of the American Statistical Association* 107, 1638-1652.
- [9] Davies, N.M., S. von Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer and G. Davey Smith, (2015), The Many Weak Instruments Problem and Mendelian Randomization, *Statistics in Medicine* 34, 454-468.
- [10] Efron, B., T. Hastie, I. Johnstone and R. Tibshirani, (2004), Least Angle Regression, *The Annals of Statistics* 32, 407-451.
- [11] Greenland, S., (2000), An introduction to instrumental variables for epidemiologists, *International Journal of Epidemiology* 29, 722-729.
- [12] Han, C., (2008), Detecting Invalid Instruments using L_1 -GMM, *Economics Letters* 101, 285-287.
- [13] Hansen, L.P., (1982), Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* 50, 1029-1054.
- [14] von Hinke, S., G. Davey Smith, D.A. Lawlor, C. Propper and F. Windmeijer, (2016), Genetic Markers as Instrumental Variables, *Journal of Health Economics* 45, 131-148.
- [15] Imbens, G.W., (2014), Instrumental Variables: An Econometrician's Perspective, *NBER Working Paper 19983*.
- [16] 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.
- [17] Kang, H., A. Zhang, T.T. Cai and D.S. Small, (2016), Instrumental Variables Estimation with some Invalid Instruments and its Application to Mendelian Randomization, *Journal of the American Statistical Association* 111, 132-144.

- [18] 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* 33, 474-484.
- [19] 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.
- [20] Liao, Z., (2013), Adaptive GMM Shrinkage Estimation with Consistent Moment Selection, *Econometric Theory* 29, 857-904.
- [21] 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.
- [22] Locke, A. E., et al., (2015), Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology. *Nature* 518, 197–206.
- [23] Rothenberg, T.J., (1984), Approximating the Distributions of Econometric Estimators and Test Statistics. In Z. Griliches and M.D. Intriligator (Eds.), *Handbook of Econometrics, Volume 2*, 881-935. Amsterdam: North Holland.
- [24] Sargan, J. D., (1958), The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica* 26, 393–415.
- [25] Staiger, D. and J.H. Stock, (1997), Instrumental Variables Regression with Weak Instruments, *Econometrica* 65, 557-586.
- [26] 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.
- [27] Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh et al., (2015), UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* 12: e1001779.

[28] Zhao, P. and B. Yu (2006), On Model Selection Consistency of Lasso, *Journal of Machine Learning Research* 7, 2541-2563.

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

9 Appendix

9.1 LARS Steps for Instrument Selection

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

$$\hat{\mathbf{c}}_n = n^{-1} \tilde{\mathbf{Z}}' (\mathbf{y} - \hat{\boldsymbol{\mu}}_A) \quad (22)$$

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

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

Define

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

and

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

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

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

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

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

where

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

Further, define

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

with j -th element $b_{n,j}^A$.

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

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

where

$$\hat{\kappa}_A = \min_{j \in A^c}^+ \left\{ \frac{\hat{C}_n - \hat{c}_{n,j}}{B_{n,A} - b_{n,j}^A}, \frac{\hat{C}_n + \hat{c}_{n,j}}{B_{n,A} + b_{n,j}^A} \right\}, \quad (23)$$

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

Assuming that all γ_j are the same and that $E[\mathbf{Z}_i \mathbf{Z}_i'] = \text{plim}(n^{-1} \mathbf{Z}' \mathbf{Z}) = \mathbf{I}_L$, we have that

$$\text{plim} \left(n^{-1} \tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} \right) = \mathbf{I} - L^{-1} \boldsymbol{\nu}_L \boldsymbol{\nu}_L',$$

and hence

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

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

As

$$\text{plim} \left(n^{-1} \tilde{\mathbf{Z}}_A' \tilde{\mathbf{Z}}_A \right) = \mathbf{I}_A - L^{-1} \boldsymbol{\nu}_A \boldsymbol{\nu}_A',$$

it follows that

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

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

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

and

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

where

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

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

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

and

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

Then

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

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

$$\text{plim}(\hat{\mathbf{c}}_n) = \text{plim} \left(n^{-1} \tilde{\mathbf{Z}}' \mathbf{y} \right) = \alpha - \bar{\alpha}. \quad (24)$$

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

$$\min(\hat{\kappa}_{A_1, inv}) = \frac{(\alpha_1 - \bar{\alpha}) - (\alpha_2 - \bar{\alpha})}{B_{A_1} - b_2^{A_1}} = \frac{\alpha_1 - \alpha_2}{B_{A_1} - b_2^{A_1}}$$

and for the valid instruments,

$$\min(\hat{\kappa}_{A_1, val}) = \frac{(\alpha_1 - \bar{\alpha}) + (-\bar{\alpha})}{B_{A_1} + b_{val}^{A_1}} = \frac{\alpha_1 - 2\bar{\alpha}}{B_{A_1} + b_{val}^{A_1}}$$

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

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

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

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

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

$$\mathbf{w}_A = B_A \left(1 + \frac{|A|}{L - |A|} \right) \mathbf{1}_A = B_A \left(\frac{L}{L - |A|} \right) \mathbf{1}_A$$

$$b_{j+1}^{A_j} = b_{val}^{A_j} = -B_{A_j} \left(\frac{|A_j|}{L - |A_j|} \right).$$

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

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

For the valid instruments it is given by

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

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

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

where

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

Then the next invalid instrument gets selected if

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

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

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

For L even, if $s = L/2$ then $\min(\hat{\kappa}_{A_s, inv}) = \min(\hat{\kappa}_{A_s, val})$ 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.