Selecting (In)Valid Instruments

Neil Davies, Helmut Farbmacher, George Davey-Smith,
Ian White, Frank Windmeijer

- Instrumental variables estimation popular method for identifying and estimating the magnitude of the causal effect of a modifiable risk factor on outcomes.

- In epidemiology, the concept of Mendelian randomisation has led to the use of genes as instruments. E.g. some genes have been shown to lead to higher weight, and used to estimate the effect of weight on blood pressure.

- Often, it may be the case that genes don’t satisfy the so-called exclusion restriction, e.g. some may have a direct effect on the outcome.

- Kang et al. (2015) propose use of Lasso type method to identify valid and invalid instruments for 2SLS, sisVIVE.

- Lasso selection of invalid instruments using LARS is similar to forward selection of variables method. We compare using various stopping rules.

- We show that sisVIVE breaks down when invalid instruments are relatively “too” strong.

- Weighting for instrument strength can lead to correct selection of valid instruments when there are more than 50% valid instruments

- This is similar to the Han (2008) method using ℓ_1 -GMM, which results in the median of the IV estimates using all instruments one at the time, which is a consistent estimator when there are more than 50% valid instruments.

- Very Preliminary!

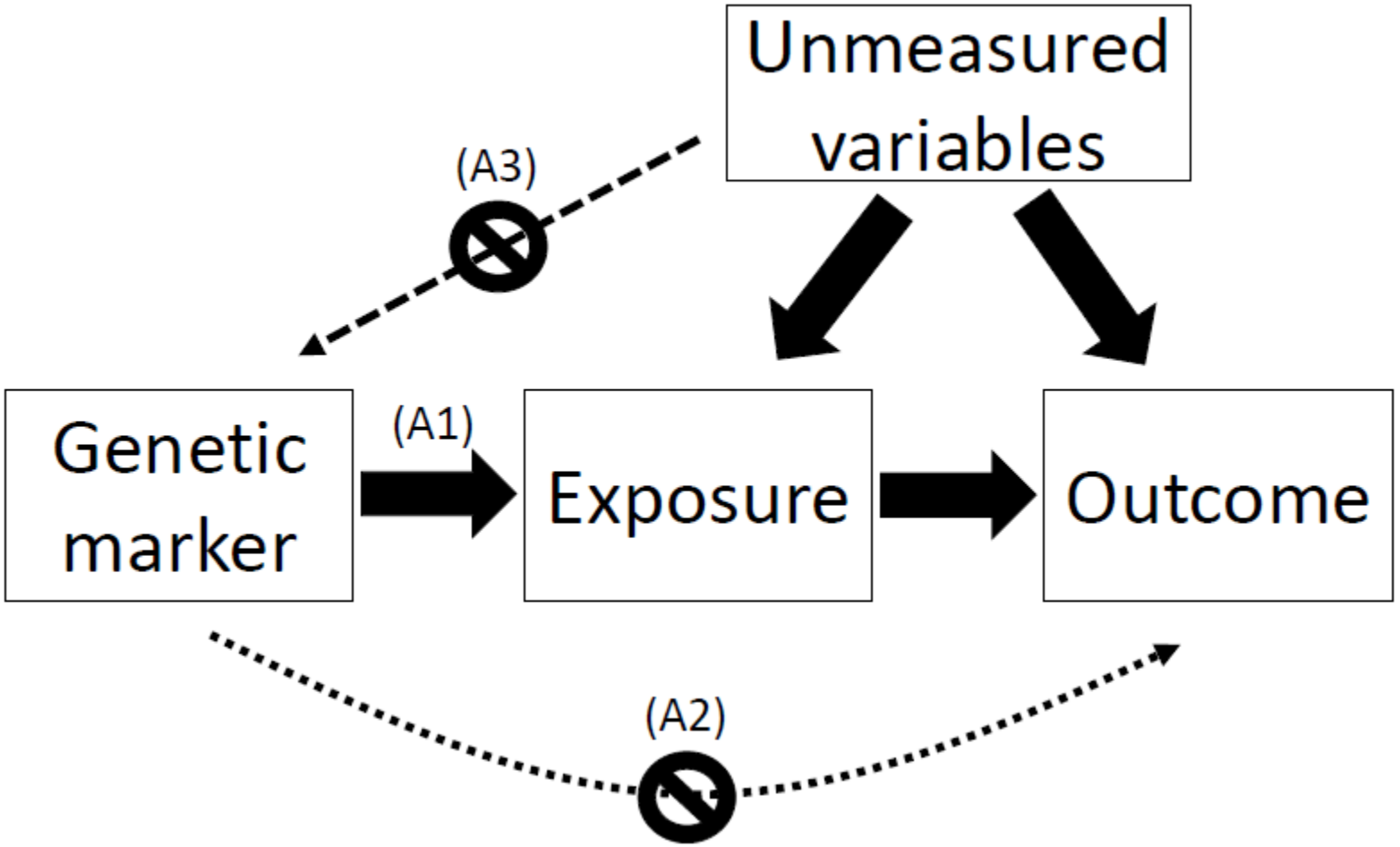

Unmeasured variables
(A3)
Genetic marker
(A1)
Exposure
Outcome
(A2)

Table 6. Estimation results

SMM				
Linear	OLS	2SLS	GMM2	J-test
ψ	0.2009	0.2091	0.2094	0.2965
	(0.0039)	(0.0819)	(0.0819)	
Multiplicative	Gamma	GMM1	GMM2	J-test
θ	0.2974	0.3090	0.3104	0.3071
	(0.0063)	(0.1192)	(0.1192)	
Logistic	Logistic regression	GMM1	GMM2	J-test
ξ	0.9487	1.0409	1.0528	0.2924
	(0.0189)	(0.4220)	(0.4217)	

Notes: Sample size 55,523.

Standard errors in brackets; p-values are reported for the J-test.

In a structural linear model

$$y_i = x_i'\beta + u_i$$

where $E\left[x_i u_i\right] \neq 0$, but $E\left[z_i u_i\right] = 0$, the two-stage least squares estimator (2SLS) for β is given by ($P_Z = Z\left(Z'Z\right)^{-1} Z'$)

$$\hat{\beta} = \left(X'P_Z X\right)^{-1} X'P_Z y$$

The Sargan test for the null $E\left[z_i u_i\right] = 0$ is given by ($\hat{u} = y - X\hat{\beta}$)

$$\hat{u}'P_Z\hat{u} / \hat{\sigma}_u^2 \xrightarrow{d} \chi^2_{k_z - k_x}$$

Kang et al. setup. They consider potential outcomes model for outcome Y, treatment D and instrument vector Z, containing L potential instruments

$$Y_i^{(d'',z'')} - Y_i^{(d,z)} = \left(z'' - z\right)\phi + \left(d'' - d\right)\beta$$
$$E\left[Y_i^{(0,0)} \mid Z_i\right] = Z_i'\psi$$

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

They position their model in the Mendelian randomisation genetic framework, where genes are instruments and independently distributed, i.e. $E\left[Z_i Z_i'\right] = I_L$, after standardisation.

The observed data model is

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

where $\alpha = \phi + \psi$; $u_i = Y_i^{(0,0)} - E\left[Y_i^{(0,0)} \mid Z_i \right]$, and hence $E\left[u_i \mid Z_i\right] = 0$.
A valid instrument is then a Z_j for which $\alpha_j = 0$.

They estimate the parameters α and β by Lasso type method using ℓ_1 penalisation:

$$\left(\hat{\alpha}_\lambda, \hat{\beta}_\lambda\right) = \arg\min_{\alpha,\beta} \frac{1}{2}\left\| P_Z\left(Y - Z\alpha - D\beta\right)\right\|_2^2 + \lambda\left\|\alpha\right\|_1$$

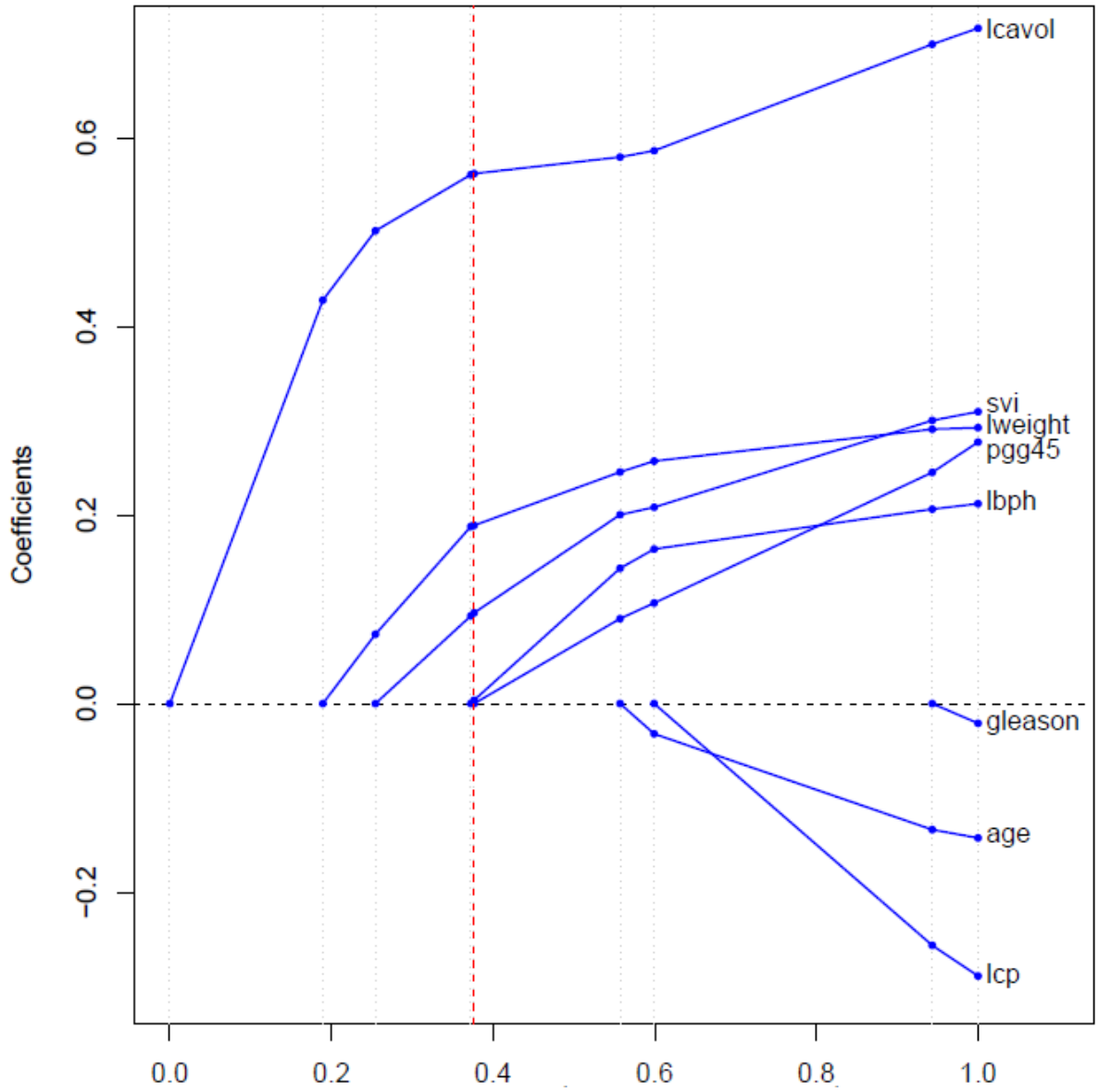
A simple two-step algorithm, using the LARS routine, is programmed up in R and called sisVIVE (*some invalid and some valid IV estimator)*. It uses a cross-validation technique for finding the value of λ, but no inference. It works when there are less than 50% invalid instruments.

A two-step method is employed: for a given λ find $\hat{\alpha}_\lambda$ from

$$\arg\min_{\alpha} \frac{1}{2}\left\| M_{\hat{D}} P_Z Y - M_{\hat{D}} Z\alpha \right\|_2^2 + \lambda \left\| \alpha \right\|_1$$

$\hat{D} = P_Z D$, $M_{\hat{D}} = I - \hat{D}\left(\hat{D}'\hat{D}\right)^{-1}\hat{D}$, then

$$\hat{\beta}_\lambda = \frac{\hat{D}'\left(Y - Z\hat{\alpha}_\lambda\right)}{\hat{D}'\hat{D}}$$


Coefficients
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp

LAR is similar to forward stagewise linear regression (matching pursuits), adding the instruments sequentially to the model according to the magnitude of their correlation with the residual. Considering one instrument at the time, the 2SLS estimator for α_j in the model

$$Y = D\beta + Z_j\alpha_j + u$$

is given by

$$\hat{\alpha}_j = \left(Z_j' M_{\hat{D}} Z_j\right)^{-1} Z_j' M_{\hat{D}} P_Z Y$$

and forward selection is based on the standardised version

$$\hat{\alpha}_j^s = \left(Z_j' M_{\hat{D}} Z_j\right)^{-1/2} Z_j' M_{\hat{D}} P_Z Y .$$

We will compare sisVIVE with this forward selection method, using cross-validation (10-fold) to determine the value *ac*, such that Z_j gets selected when $\hat{\alpha}_j^s > ac$.

Andrews (1999) and Andrews and Lu (2001) propose a downward testing procedure, starting from the model with the largest degrees of freedom, estimating all possible models and selecting the one with the largest degrees of freedom that passes the Sargan test (Hansen for heteroskedasticity).

A directed search with stopping rule would then be to select the Z_j sequentially with the largest $\hat{\alpha}_j^s$ until the Sargan test does not reject.

Consider the Kang et al. MC design, $n = 2000,\ L = 10,\ s = 3,\ \beta = 1$ and the α's for the 3 invalid instruments are all equal to 1. All reduced form parameters π_j are equal to $1/\sqrt{20}$. $\rho = 0.8$.

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

$$D_i = Z_i'\pi + v_i$$

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

Some Monte Carlo results: $\beta = 1,\ L = 10,\ s = 3(\alpha = 1),\ n = 2000$, 1000 reps

	mean	sd	# Inv Selected
2SLS	2.337	0.084	
sisVIVE	1.096	0.034	3.88, 3-9
FS, CV	1.015	0.048	3.68, 3-9
FS, Sar (0.05)	1.006	0.040	3.06, 3-5
2SLS oracle	1.007	0.038	

Some Monte Carlo results: $\beta = 1,\ L = 20,\ s = 6(\alpha = 0.2),\ n = 500,\ 1000$ reps

	mean	sd	# Inv Selected
2SLS	1.287	0.042	
sisVIVE	1.217	0.047	6.63, 0-18
FS, CV	1.063	0.074	8.33, 0-19
FS, Sar (0.05)	1.046	0.053	5.55, 3-10
2SLS oracle	1.025	0.048	

Maintaining throughout that $E\left(Z_i Z_i'\right) = I_L$ and $\text{plim}\left(n^{-1} Z'Z\right) = I_L$, it is easily shown that

$$\text{plim}\left(\hat{\alpha}_j^s\right) = \left(\alpha_j - \pi_j \frac{\pi'\alpha}{\pi'\pi}\right) \Big/ \left(1 - \frac{\pi_j^2}{\pi'\pi}\right)^{1/2}$$

with the nonzero α_js all equal (to α) and the π_js all equal, this becomes

$$\alpha\left(1 - \frac{s}{L}\right) \Big/ \left(1 - \frac{1}{L}\right)^{1/2} \text{ and } \left(-\alpha \frac{s}{L}\right) \Big/ \left(1 - \frac{1}{L}\right)^{1/2}$$

for the invalid and valid instruments respectively. Hence it is bigger in absolute value for the valid instruments if $\left(L - s\right) > s$, more than 50% of the instruments are valid.

Now, if the invalid instruments are stronger, $\pi_{inv} = c\pi_{val}$ then we get for $\text{plim}\left(\hat{\alpha}_j^s\right)$, α times

$$\left(1-\frac{c^2 s}{c^2 s+L-s}\right)/\left(1-\frac{c^2}{c^2 s+L-s}\right)^{1/2};$$

$$\left(-\frac{cs}{c^2 s+L-s}\right)/\left(1-\frac{1}{c^2 s+L-s}\right)^{1/2}$$

for valid and invalid instruments respectively. When $L=10$ and $s=3$, then this is larger in absolute value for the invalid instruments when $c>2.65$

Some Monte Carlo results: $\beta = 1,\ L = 10,\ s = 3(\alpha = 1),\ \pi_{inv} = 3.5\pi_{val}$
$n = 2000$, 1000 reps

	mean	sd	# Inv Selected
2SLS	2.073	0.016	
sisVIVE	2.252	0.014	7.74, 7-9
FS, CV	2.276	0.016	7.29, 7-9
FS, Sar (0.05)	2.278	0.014	7.05, 7-8
2SLS oracle	1.005	0.037	

We can weigh the moment conditions in order to avoid this problem.

A Generalised Method of Moment (GMM) estimator minimises

$$\left(n^{-1}u'Z\right)W_n^{-1}\left(n^{-1}Z'u\right)$$

Let $u = Y - D\beta$, then $\hat{\beta} = \left(D'ZW_n^{-1}Z'D\right)^{-1} D'ZW_n^{-1}Z'y$, and

$$\text{plim}\left(n^{-1}Z'\hat{u}\right) = \alpha - \pi\frac{\pi'W^{-1}\alpha}{\pi'W^{-1}\pi}$$

where $W = \text{plim}\left(W_n\right)$, again maintaining that $\text{plim}\left(n^{-1}Z'Z\right) = I_L$.

Under the same circumstances as before, if $W = diag\left(\pi_j\right)$ then the correlations of the invalid instruments with the residuals is stronger than those of the valid ones.

We set $W_n = diag\left(\left|n^{-1}Z_j'D\right|\right)$.

This is the same as the weightmatrix proposed by Han (2008) for ℓ_1 GMM

$$\hat{\beta}_m = \arg\min_{\beta}\left\|W_n^{-1}n^{-1}Z'\left(Y - D\beta\right)\right\|.$$

$\hat{\beta}_m$ is the median of the L IV estimates $\hat{\beta}_j\left(Z_j\right)$ and is a consistent estimator for β as long as more than 50% of the instruments are valid.

Use the GMM estimator for selection of invalid instruments only, and estimate final selected model by 2SLS.

Some Monte Carlo results: $\beta = 1,\ L = 10,\ s = 3 (\alpha = 1),\ n = 2000$

	mean	sd	# Inv Selected
2SLS	2.337	0.084	
sisVIVE	1.096	0.034	3.88, 3-9
FS, CV	1.015	0.048	3.68, 3-9
FS, Sar (0.05)	1.006	0.040	3.06, 3-5
2SLS oracle	1.007	0.038	
FSW, CV	1.015	0.046	3.66, 3-9
FSW, Sar (0.05)	1.006	0.040	3.06, 3-5
ℓ_1 - Han	1.103	0.083	

Some Monte Carlo results: $\beta = 1,\ L = 10,\ s = 3(\alpha = 1),\ \pi_{inv} = 3.5\pi_{val}$
$n = 2000$, 1000 reps

	mean	sd	# Inv Selected
2SLS	2.073	0.016	
sisVIVE	2.252	0.014	7.74, 7-9
FS, CV	2.276	0.016	7.29, 7-9
FS, Sar (0.05)	2.278	0.014	7.05, 7-8
2SLS oracle	1.005	0.037	
FSW,CV	1.015	0.045	3.55, 3-9
FSW, Sar (0.05)	1.006	0.040	3.06, 3-5
ℓ_1 - Han	1.095	0.077	

Discussion

We have investigated alternative ways of selecting invalid instruments, taking the Lasso approach of Kang et al. as a starting point.

Combining selection with a stopping rule based on Sargan statistic seems a sensible approach with good properties. Also, in first design, Wald rejection probabilities are 6.7% at the 5% level for this estimator.

We have proposed a new selection method that allows for different instrument strengths and results in selecting the invalid instruments as long as there are more than 50% of the instruments valid. This is similar to, but behaves better than, Han's ℓ_1 estimator, which is a consistent estimator.

The situation covered assumed independent instruments, which seems ok for Mendelian randomisation. Things are more complicated (and more interesting) when instruments are correlated.

In that case, there are differences whether an instrument has a direct effect or whether they are correlated with the error in a different way from $u = \alpha_j Z_j + e_j$. That is, sometimes an instrument is invalid, but should *not* be included in the main model, i.e. it should just be discarded.

For example, the Han estimator only works with correlated instruments if they are not to be included in the model. If the model is

$$Y = D\beta + u$$

If Z_1, Z_2 are the invalid, valid instruments, then if $E\left[Z_1 u\right] \neq 0$, but $E\left[Z_2 u\right] = 0$, the Han estimator works, but not the Kang et al. approach.

However, if $u = Z_1\alpha_1 + \varepsilon$, then the valid instruments must have that $E\left[Z_2 \varepsilon\right] = 0$, but then $E\left[Z_2 u\right] \neq 0$ due to the correlation of the valid instruments with the invalid ones. So, here Han doesn't work, but Kang et al.'s approach does.

When the instruments are independent this problem doesn't arise, only efficiency is affected.

To deal with correlated instruments then could be a search strategy a la Andrews (1999). Sequentially drop the instruments from the instrument set using the correlations $Z'\hat{u}$ or $\left(Z'Z\right)^{-1}Z'\hat{u}$ and record the Sargan statistic, stop when Sargan passes.

Then add the instrument with the strongest correlation (or $\hat{\alpha}_s^j$) to the model and repeat sequentially dropping instruments from the instrument set until Sargan passes.

Repeat till the end, and then select the model where the Sargan passed with the largest degrees of freedom, $k_z - k_x$.

There are various related papers in the econometrics domain. For example,

Caner, M, Han, X, and Lee, Y, (2013), Adaptive Elastic Net GMM Estimator with many Invalid Moment Conditions: A Simultaneous Model and Moment Selection, mimeo, University of Michigan.