Linear Instrumental Variables Model Averaging Estimation

Luis F. Martins
Department of Quantitative Methods, ISCTE-LUI, Portugal
Centre for International Macroeconomic Studies (CIMS), UK
(luis.martins@iscte.pt)*

Vasco J. Gabriel
CIMS, University of Surrey, UK and NIPE-UM
(v.gabriel@surrey.ac.uk)

This version: April 2013

Abstract

Model averaging (MA) estimators in the linear instrumental variables regression framework are considered. The obtaining of weights for averaging across individual estimates by direct smoothing of selection criteria arising from the estimation stage is proposed. This is particularly relevant in applications in which there is a large number of candidate instruments and, therefore, a considerable number of instrument sets arising from different combinations of the available instruments. The asymptotic properties of the estimator are derived under homoskedastic and heteroskedastic errors. A simple Monte Carlo study contrasts the performance of MA procedures with existing instrument selection procedures, showing that MA estimators compare very favourably in many relevant setups. Finally, this method is illustrated with an empirical application to returns to education.

Keywords: Instrumental Variables; Model Selection; Model Averaging; Model Screening; Returns to Education.

1 Introduction

In this paper, we consider model averaging (MA) estimation methods in the linear instrumental variables (IV) regression context. The model averaging estimator is a weighted average of individual estimates obtained using different lists of valid instruments. We propose obtaining empirical weights based on existing and well-established instrument selection criteria for IV models. This

*Department of Quantitative Methods, ISCTE-IUL, Av. das Forças Armadas, 1649-026 Lisbon. Tel: +351929659233, Fax: +351217903942.

can be achieved by direct smoothing of information criteria arising from the estimation stage, as in Buckland, Burnham, and Augustin (1997), Burnham and Anderson (2002) and Hjort and Claeskens (2003). We show that the MA estimator is consistent and normally distributed with a specific closed-form expression for its asymptotic variance-covariance matrix. The proposed MA estimator and its first-order asymptotic properties are defined under the case of homoskedasticity and for general forms of heteroskedastic errors.

In many applications of IV estimation, there is often a large set of candidate variables that can be used as instruments. However, the properties of IV estimators are very sensitive to the choice (and the characteristics) of the instrument set. Indeed, instruments might be poorly correlated with the endogenous variables, which invalidates conventional inference procedures (Staiger and Stock, 1997 and Stock and Wright, 2000). On the other hand, using many (potentially weak) instruments can improve efficiency and precision, but it can also lead to substantial deviations from the usual Gaussian asymptotic approximation (see Chao and Swanson, 2005, Han and Phillips, 2006, Hansen, Hausman and Newey, 2008 and Newey and Windmeijer, 2009).

Thus, much of the literature has focused on procedures for the selection of the appropriate number and list of instruments. Donald and Newey (2001) propose a selection procedure such that an approximate mean-square error is minimized over all existing instruments deemed to be valid. It includes Two-Stage Least Squares (TSLS) limited information maximum likelihood and a bias adjusted version of the TSLS. On the other hand, Andrews (1999) developed GMM analogues of model selection criteria (MSC) based on the J-statistic in order to consistently select the largest set of valid moment conditions. Hall, Inoue, Jana and Shin (2007) suggest selecting instruments according to the relevant moment selection criterion (RMSC), based on the entropy of the limiting distribution of the estimator, while Hall and Peixe (2003) propose a canonical correlations information criteria (CCIC) for instrument selection (see also Lo and Ronchetti (2012) for an information and entropy-based approach to moment conditions estimation). As an alternative, Pesaran and Smith (1994) define a measure of the goodness of fit for IV regressions. They call it generalized R-squared, GR^2, and show that it ought to be computed based on the prediction errors.

Model selection entails choosing one of the estimated competing models under consideration. Testing competing, non-nested formulations, in which the outcome may not be the selection of one particular model, can be carried out using the tests of Smith (1992) and Smith and Ramalho (2002). Shrinkage methods, on the other hand, are an alternative to model selection. Caner (2009) proposes a LASSO-type GMM estimator, while Canay (2010) and Okui (2011) propose shrinkage-type estimators for linear models with many instruments. In fact, shrinkage estimators can be viewed as a special case of 'instrument averaging', in which some instruments receive weights approaching zero.

Here, we pursue the alternative approach of model averaging, in which parameter estimates are constructed based on a weighted average of estimates from a number of possible specifications. By making use of the information conveyed by otherwise discarded alternative specifications, model

averaging as an estimation strategy may yield some gains in terms of bias and efficiency when compared to procedures that make use of a single set of instruments. Furthermore, our approach can cope with high-dimensional problems arising from large numbers of combinations of instruments, particularly when there is no clear indication as to which instruments should be discarded.

Our work is a natural extension of the literature, in which model averaging usually involve weights obtained from functions of model selection criteria, such as the BIC, AIC, etc. Indeed, there is a large literature on model averaging, both in the Bayesian tradition and in a frequentist context (see Claeskens and Hjort, 2008 for a review). In the latter framework, Hansen (2007) proposed a Mallows criterion for the selection of weights for averaging across least squares estimates obtained from a set of approximating models, in which regressors (or groups of regressors) are added sequentially. Liang, Zou, Wan and Zhang (2011), in turn, discuss optimal weight choice based on an unbiased estimator of the MA estimator's mean squared error (MSE), thus attaining good finite sample properties. On the other hand, Hansen and Racine (2012) consider a jackknife MA estimator, with weights based on a cross-validation criterion, which is asymptotically optimal under bounded heteroskedasticity of unknown form.

Model averaging in the linear IV context has seen some very recent developments. Kuersteiner and Okui (2010) suggest using Hansen's (2007) method as a first step to construct optimal instruments IV estimation with TSLS, LIML and Fuller estimators. The weights are chosen to minimize the approximate mean squared error (AMSE), as in Donald and Newey (2001). Koop, Leon-Gonzalez and Strachan (2012), on the other hand, use a Bayesian model averaging approach to address different sources of uncertainty, such as the set of instruments, exogeneity restrictions, the validity of identifying restrictions and the set of exogenous regressors.

Nevertheless, our approach is distinct in that it averages estimates of the parameters of interest (rather than first-stage results as in Kuersteiner and Okui, 2010) and, in our case, the list of candidate models does not depend on ordered instruments from the full-instrument matrix. Indeed, with m instruments we can consider m models (each model including an extra instrument as in Hansen, 2007), but also any possible combination of these. This makes our approach more general and not restricted by how the instruments are ordered. Also, unlike TSLS kernel-based weighting as proposed by Canay (2010) and Okui (2011), our procedure does not depend on the choice of kernels or arbitrarily user-chosen smoothing parameters. Thus, we are able to combine the estimation of general moment conditions models with one-step and information criteria-based model averaging estimation.

In fact, our paper is related to recent (and parallel) contributions. Lee and Zhou (2012) consider a different 'feasible' weighting scheme based on the strength of subsets of instruments (measured by the ratio of the first-stage R^2 and the Sargan statistic). Similarly, Chen, Jacho-Chàvez and Linton (2012) consider averaging moment condition estimators in a more general conditional estimation setup.

To study the properties of our estimators, we conduct a small-scale Monte Carlo experiment in

which we contrast the performance of an averaging approach and that of an instrument selection strategy, showing that, in several setups, our model averaging estimation procedure outperforms the selection method of Donald and Newey (2001) in terms of median bias, absolute deviation and dispersion. Moreover, we illustrate empirically the use of MA procedures by examining returns to education, in which we show that even when both the sample and the number of instruments is quite large, our methods are flexible enough to cope with high-dimensional problems in an efficient way.

Next, section 2 introduces the linear IV regression model and defines the instrument selection criteria. In section 3, we introduce our model averaging approach, we discuss different procedures to obtain empirical weights and model screening as a strategy to narrow down the list of candidate specifications, thus reducing the computational burden. In section 4, we derive the asymptotic properties of the model averaging estimator. A Monte Carlo simulation study providing evidence in support of our MA procedures is discussed in Section 5. An application to returns to schooling is considered in Section 6 and, finally, Section 7 concludes.

2 Definitions

Following the notation in Staiger and Stock (1997), the linear IV regression model is specified by a structural equation of interest

$$y = Y\beta + X\gamma + u, \tag{1}$$

where y is a $T \times 1$ vector, Y is a $T \times n$ matrix of endogenous regressors, X is a $T \times K_1$ matrix of exogenous regressors, and by a reduced form equation for the endogenous Y

$$Y = Z\Pi + X\Phi + V, \tag{2}$$

where Z is a $T \times K_2$ matrix of instruments, with Y, X and Z full ranked and $K_2 \geq n$. For the sake of simplicity (and in accordance to the application we study in this paper), we let $n = 1$ and assume independent and identically distributed data. The error structure $w_i = (u_i, V_i)'$ satisfies the moment conditions

$$E\left(w_i | X_i, Z_i\right) = 0 \tag{3}$$

and

$$E\left(w_i w_i' | X_i, Z_i\right) = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}. \tag{4}$$

Define the parameter of interest $\theta = (\beta', \gamma')'$ and let $\overline{Z} = [X, Z]$ be a $T \times K$ matrix where $K = K_1 + K_2$ and $\overline{X} = [Y, X]$ is $T \times (1 + K_1)$, so that endogeneity arises if $E\left(Y_i u_i\right) \neq 0$. Throughout the paper we assume that β is identified, i.e. that $E(\overline{Z}_i \overline{X}_i')$ is of full column rank for any choice of instruments such that $K_2 \geq n$. Also, define $\pi = (\Phi', \Pi')'$.

Although Staiger and Stock (1997) define a general k-class of estimators, we simply focus on TSLS. In this case, it can be shown that

$$\widehat{\theta} = \left(\overline{X}'\overline{Z} \left(\overline{Z}'\overline{Z} \right)^{-1} \overline{Z}'\overline{X} \right)^{-1} \left(\overline{X}'\overline{Z} \left(\overline{Z}'\overline{Z} \right)^{-1} \overline{Z}'y \right) \tag{5}$$

and, in particular, the scalar

$$\widehat{\beta} = \left(Y^{\perp\prime} \left(I - M_{Z^{\perp}} \right) Y^{\perp} \right)^{-1} \left(Y^{\perp\prime} \left(I - M_{Z^{\perp}} \right) y^{\perp} \right), \tag{6}$$

where

$$Y^{\perp} = M_X Y;\ y^{\perp} = M_X y;\ Z^{\perp} = M_X Z \tag{7}$$

$$M_X = I - X \left(X'X \right)^{-1} X'; M_{Z^{\perp}} = I - Z^{\perp} \left(Z^{\perp\prime} Z^{\perp} \right)^{-1} Z^{\perp\prime}, \tag{8}$$

(see Staiger and Stock, 1997, for details). This estimator ignores the presence of heteroskedasticity and is a GMM-type of estimator under the population unconditional moment condition $E\left(\overline{Z}_i u_i \left(\theta \right) \right) = 0$. In the general case,

$$\widehat{\theta} = \left(\overline{X}'\overline{Z} \left(\overline{Z}'\widehat{\Sigma}_u \overline{Z} \right)^{-1} \overline{Z}'\overline{X} \right)^{-1} \left(\overline{X}'\overline{Z} \left(\overline{Z}'\widehat{\Sigma}_u \overline{Z} \right)^{-1} \overline{Z}'y \right), \tag{9}$$

where $\widehat{\Sigma}_u = diag\left(\widehat{u}_1^2 \left(\widehat{\theta} \right), ..., \widehat{u}_T^2 \left(\widehat{\theta} \right) \right)$ with the residuals evaluated at $\widehat{\theta}$ in the absence of heteroskedasticity (5). Various tests for functional form and heteroskedasticity for linear IV regressions can be found in Pesaran and Taylor (1999).

Under some mild regularity conditions, $\widehat{\theta}$ in (9) is $\sqrt{T}-$consistent and asymptotically normal, with asymptotic variance

$$V = \left(Q_{\overline{X},\overline{Z}} \widetilde{Q}_{\overline{Z},\overline{Z}}^{-1} Q_{\overline{Z},\overline{X}} \right)^{-1}, \tag{10}$$

where $Q_{\overline{X},\overline{Z}} = E\left(\overline{X}_i \overline{Z}_i' \right), Q_{\overline{Z},\overline{X}} = Q_{\overline{X},\overline{Z}}'$ and $\widetilde{Q}_{\overline{Z},\overline{Z}} = E\left(u_i^2 \overline{Z}_i \overline{Z}_i' \right)$, whereas, under homoskedasticity, $\widehat{\theta}$ in (5) has asymptotic variance

$$V = \sigma_u^2 \left(Q_{\overline{X},\overline{Z}} Q_{\overline{Z},\overline{Z}}^{-1} Q_{\overline{Z},\overline{X}} \right)^{-1}, \tag{11}$$

with $Q_{\overline{Z},\overline{Z}} = E\left(\overline{Z}_i \overline{Z}_i' \right)$, $Q_{\overline{X},\overline{Z}}$ is finite and full ranked, for the purpose of identification (see Hall, 2005, inter alia). These are efficient GMM-type estimators.

Given that the rejection of the Sargan J-statistic is an indicator that some instruments are invalid, and acknowledging the usual trade-off between bias and efficiency when picking a particular list of instruments, we take all possible combinations of instruments when estimating the structural equation. Let $\mathcal{M}$ be the collection of candidate instruments. Here, $\mathcal{M}$ is a countable/finite set, such that model M_i belongs to the family of models $\mathcal{M} : M_i \in \mathcal{M}$. For now, take any particular model, M_i, which is characterized by a particular set of instruments. Then, following Andrews (1999), one can define a selection vector $c \in \Re^K$ that represents a list of "selected" instruments. Defining the unit-simplex set

$$C = \left\{ c \in \Re^K \backslash \{0\} : c_j = 0 \text{ or } 1, \forall 1 \leq j \leq K, \text{ where } c = (c_1, ..., c_K)' \right\}, \tag{12}$$

c is a vector of zeros (excluded instruments) and ones (included instruments) and $|c| = \sum_j^K c_j \leq K$ for $c \in C$ denotes the number of the selected instruments c. Also, define $\overline{c} = \iota_K$, a vector of ones, which implies using the whole set of instruments. Thus, quantities such as Z_c, $\overline{Z}_c$, $\widehat{\theta}_c$, $\widehat{\beta}_c$, $Z_c^{\perp}$, $M_{Z_c^{\perp}}$, $\widehat{\Sigma}_{u_c}$, V_c, $Q_{\overline{X},\overline{Z}_c}$, $\widetilde{Q}_{\overline{Z},\overline{Z}_c}$ and $Q_{\overline{Z},\overline{Z}_c}$ are obtained after deleting the instruments j corresponding to $c_j = 0$. Take, for the sake of simplicity, the homoskedastic case (5). Then,

$$\widehat{\theta}_c = \left(\overline{X}'\overline{Z}_c\left(\overline{Z}_c'\overline{Z}_c\right)^{-1}\overline{Z}_c'\overline{X}\right)^{-1}\left(\overline{X}'\overline{Z}_c\left(\overline{Z}_c'\overline{Z}_c\right)^{-1}\overline{Z}_c'y\right), \tag{13}$$

with $\overline{Z}_c = [X, Z_c]$ where Z_c is a $T \times |c|$ matrix, $|c| \leq K_2$, that only includes instruments associated with $1's$ at vector c. To make it clear, we are selecting only over the available K_2 instruments and therefore keeping all K_1 exogenous regressors (X) in the estimation procedure.

There are several procedures for the selection of the appropriate instruments c_0 over the full list of candidate models $c \in \mathcal{C}$, where $\mathcal{C} \subset C$, with $\{0\} \in \mathcal{C}$, is some parameter space for the instrument selection vector. Donald and Newey (2001) propose a selection procedure such that an approximate mean-square error, $AMSE$, is minimized over all existing instruments deemed to be valid. The $AMSE$ criterion is defined as

$$\widehat{c}_{AMSE} = \arg\min_{c\in C} AMSE_T(c) = \arg\min_{c\in C}\left(\widehat{\sigma}_{au}^2\frac{|c|^2}{T} + \widetilde{\sigma}_u^2\left(\widehat{R}(c) - \widehat{\sigma}^2\frac{|c|}{T}\right)\right), \tag{14}$$

where $\widehat{\sigma}_{au}^2 = T^{-1}\tilde{a}'\tilde{u}$, $\tilde{a} = \tilde{e}[\tilde{\pi}\overline{Z}_{\overline{c}}'\overline{Z}_{\overline{c}}\tilde{\pi}/T]^{-1}\tilde{\lambda}$, $\tilde{e} = [I_T - \overline{Z}_{\overline{c}}(\overline{Z}_{\overline{c}}'\overline{Z}_{\overline{c}})^{-1}\overline{Z}_{\overline{c}}']\overline{X}$, $\tilde{u} = u(\tilde{\theta})$, $\widetilde{\sigma}_u^2 = \tilde{u}'\tilde{u}/T$, $\widehat{\sigma}^2 = \tilde{a}'\tilde{a}/T$, $\tilde{\pi}$ and $\tilde{\theta}$ are preliminary estimators of π and θ, respectively, (say, those for which $c = \overline{c}$), $\tilde{\lambda}$ is some vector of linear combination coefficients $\tilde{\theta}$ (i.e., $\tilde{\lambda}'\tilde{\theta}$) and $\widehat{R}(c)$ is a measure of the goodness of fit of the reduced form model (for instance, based on cross-validation or Mallows criteria).

Andrews (1999) developed GMM analogues of model selection criteria based on the overidentifying restrictions J statistic in order to consistently select the largest set of valid moment conditions. The model selection criteria is defined as $MSC_T(c) = J_T(c) - \kappa_T(|c| - p)$, where $J_T(c)$ is computed with the relevant selection vector c, $|c| - p$ is the number of over-identifying restrictions and $\kappa_T = o(T)$ is a sequence that defines the selection criterion ($\kappa_T = 2$ for the AIC; $\kappa_T = \log T$ for the BIC; and $\kappa_T = Q\log\log T$ for some $Q > 2$ for the HQ-type criterion). In our setup, $p = 1 + K_1$ and, whenever all K_1 exogenous variables are used as instruments, $|c| - p$ will be equal to the number of instruments Z in model c minus one. In the context of linear models, we consider Sargan's statistic:

$$\widehat{c}_{MSC} = \arg\min_{c\in C}\left(T\frac{\left(\widehat{Q}_{\overline{Z}_c,y} - \widehat{Q}_{\overline{Z}_c,\overline{X}}\widehat{\theta}_c\right)'\widehat{Q}_{\overline{Z}_c,\overline{Z}_c}^{-1}\left(\widehat{Q}_{\overline{Z}_c,y} - \widehat{Q}_{\overline{Z}_c,\overline{X}}\widehat{\theta}_c\right)}{\widehat{\sigma}_{u,c}^2} - \kappa_T(|c| - p)\right), \tag{15}$$

where

$$\widehat{Q}_{\overline{Z}_c,y} = \frac{\overline{Z}_c'y}{T}; \widehat{Q}_{\overline{Z}_c,\overline{X}} = \frac{\overline{Z}_c'\overline{X}}{T}; \widehat{Q}_{\overline{Z}_c,\overline{Z}_c} = \frac{\overline{Z}_c'\overline{Z}_c}{T}; \widehat{\sigma}_{u,c}^2 = \frac{u(\widehat{\theta}_c)'u(\widehat{\theta}_c)}{T}. \tag{16}$$

Alternative procedures have been developed in the literature. Hall et al. (2007) proposed a criterion based on the entropy of the limiting distribution of the GMM estimator, in which the

focus is the *relevance* of instruments. The relevant moment selection criterion $RMSC$ is defined as

$$\widehat{c}_{RMSC} = \arg\min_{c \in C} \left(\ln\left(\left| \widehat{V}_c \right| \right) + \kappa_T \left(|c| - p \right) \right), \tag{17}$$

where

$$\widehat{V}_c = \widehat{\sigma}^2_{u,c} \left(\widehat{Q}_{\overline{X},\overline{Z}_c} \widehat{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c} \widehat{Q}_{\overline{Z}_c,\overline{X}} \right)^{-1}. \tag{18}$$

On the other hand, Hall and Peixe (2003), consider the problem of instrument selection based on a combination of the efficiency and non-redundancy conditions

$$\widehat{c}_{CCIC} = \arg\min_{c \in C} \left(T \sum_{j=1}^{p} \ln\left[1 - r^2_{j,T}(c) \right] + \kappa_T \left(|c| - p \right) \right), \tag{19}$$

where $r_{j,T}(c)$ is the j^{th} sample canonical correlation between $d_i(\tilde{\theta})$ and $\overline{Z}_{i_c}$, with $d_i(\theta) = \frac{\partial u_i(\theta)}{\partial \theta} = -\overline{X}_i$ and $\tilde{\theta}$ is a $\sqrt{T}-$ consistent preliminary estimator. That is, $r_T(c)$ is a correlation between $\overline{X}$ and $\overline{Z}_c$. See Eryuruk et al. (2009) for a Monte Carlo comparative study of the $AMSE$, $RMSC$ and $CCIC$ data-based methods of instrument selection.

In addition, a measure of the goodness of fit for IV regressions was proposed by Pesaran and Smith (1994). According to these authors, model selection follows from observing the largest generalized R-squared, GR^2, a measure that is based on the prediction errors. More specifically,

$$\widehat{c}_{GR^2} = \arg\max_{c \in C} \left(1 - \frac{\sum_{i=1}^{T} \left(y_i - \widehat{\overline{X}}_i \widehat{\theta}_c \right)^2}{\sum_{i=1}^{T} \left(y_i - \overline{y} \right)^2} \right), \tag{20}$$

where $y - \widehat{\overline{X}} \widehat{\theta}_c = \widetilde{u}$ is the residual from the second step regression, $\widehat{\overline{X}} = \overline{Z} \left(\overline{Z}' \overline{Z} \right)^{-1} \overline{Z}' \overline{X}$, and $\overline{y}$ is the sample mean.

3 Linear IV Model Averaging Estimators

3.1 The Procedure

In this section we present MA estimation methods where the empirical weights are based on the above mentioned instrument selection criteria for IV models. Consider K and c as defined in (12) and the relevant objects indexed by c. Now, let $\omega = \left(\omega_1, ..., \omega_{|C|} \right)'$ be a weight vector in the unit-simplex in $\Re^{|C|}$:

$$H_K = \{ \omega \in [0,1]^{|C|} : \sum_{c \in C} \omega_c = 1 \}. \tag{21}$$

Although the weights need not be restricted, as in Kuersteiner and Okui (2010), we only consider weights in the unit-simplex. In our approach, by assuming $n = 1$ and that all K_1 exogenous variables are kept in any model c, we combine over the $K_2 \geq 1$ instruments Z, which implies $|C| = 2^{K_2} - 1$ different elements in C. By the same token, if we partition Z in two blocks, Z_F of dimension K_{2F} that is always kept in model c and Z_V of dimension K_{2V} that is left free to combine,

such that $K_2 = K_{2F} + K_{2V}$, then $|C| = 2^{K_{2V}} - 1$. Moreover, $c \in \mathcal{C}$ is restricted to $c_j = 1$ for all j corresponding to X and Z_F and $c_s, s \neq j$ equals either one or zero, depending on whether the correspondent instrument from Z_V stays in the model or not. In practice, Z_V can itself be a block of instruments and the averaging scheme is over estimators that follow from models with distinct blocks of instruments. In this case, the definitions are straightforward and should not lead the reader to confusion.

Thus, a model averaging estimator of the unknown $(1 + K_1) \times 1$ vector θ is

$$\widehat{\theta}(\omega) = \sum_{c \in C} \omega_c \widehat{\theta}_c, \tag{22}$$

and, in particular, for the scalar β it equals $\widehat{\beta}(\omega) = \sum_{c \in C} \omega_c \widehat{\beta}_c$. Clearly, the post-model selection estimator is a special case for which no averaging occurs: $\omega_{c^*} = 1$ for some selected model c^* and $\omega_{c'} = 0$ for $c' \neq c^*$ and $\widehat{\theta}(\omega) = \widehat{\theta}_{c^*}$. In general, the vector ω will be unknown. As in much of the literature on model averaging, a data-dependent procedure will have to be used to determine the weights in order to implement estimation according to (22). Thus, we suggest linking the problem of selecting empirical weights $\widehat{\omega}$ with model selection criteria obtained in the estimation stage by doing direct 'smoothing' in line with Buckland, Burnham and Augustin (1997), which developed an idea originally suggested by Akaike (1979) (see also Burnham and Anderson, 2002 and Hjort and Claeskens, 2003).

Let ISC_c denote the 'instrument selection criterion' for candidate model $M \in \mathcal{M}$ that is defined by $c \in C$, according to our notation. Here, ISC may represent $AMSE$, MSC, $RMSC$ or $CCIC$ as described earlier in the paper and GR^2 is the appropriate goodness of fit measure. The averaging scheme is obtained by using weights proportional to the exponential form of a given ISC or GR^2 :

$$\widehat{\omega}_c(ISC) = \frac{\exp(-\frac{1}{2} ISC_c)}{\sum_{c' \in C} \exp(-\frac{1}{2} ISC_{c' \in C})}, \tag{23}$$

$$\widehat{\omega}_c(GR^2) = \frac{\exp(\frac{1}{2} GR_c^2)}{\sum_{c' \in C} \exp(\frac{1}{2} GR_{c' \in C}^2)}, \tag{24}$$

where the sum term encompasses all, not necessarily nested, $M' \in \mathcal{M}$ models of interest.

3.2 Computational Issues and Model Screening

An important issue that arises in this framework is that in some cases, the number of potential combinations is inevitably quite large and increases very fast with the number of available instruments - for example, five instruments generate 31 different combinations, while 10 instruments allow for 1023 combinations. Averaging over many combinations is not in itself a problem, but a large number of models implies that many weights will be effectively zero. Therefore, it makes sense to consider a smaller number of specifications for averaging, by removing the poorest performing models, as done in the regression literature, but so far unexplored in an IV setting.

We suggest that model screening can take place at different stages of the estimation procedure. An initial form of screening can be achieved by incorporating the information that certain instru-

ments are assumed to be valid (which could be based on formal testing) or based on instrument strength, for example by looking at the first-stage R^2. Moreover, one can exploit the fact that certain blocks of instruments are either valid or invalid block by block, rather than instrument by instrument, as suggested by Andrews (1999) in the context of model selection and as discussed in sections 5 and 6.

Another possibility would be to consider a 'backward elimination' procedure as in Cleaskens, Croux and Venkerckhoven (2006) (see also Zhang, Wan and Zhou, 2012). One can start from the specification containing all instruments K, then remove one instrument at a time such that this deletion leads to the smallest value for the information criterion. The procedure continues until $p \leq K^* < K$ instruments are obtained. The MA estimator then averages the estimates arising from combinations of these K^* instruments, which can reduce substantially the number of combinations, as pointed out above. This is slightly different from Claeskens et al (2006) and Zhang et al (2012) because in their case sub-models are nested after the selection of the most relevant regressors.

Furthermore, one can consider a screening method adapted from Yuan and Yang (2005). This involves splitting the sample into two parts $W^{(1)} = (y_i, Y_i, X_i, Z_i), 1 \leq i \leq T/2$ and $W^{(2)} = (y_i, Y_i, X_i, Z_i), T/2+1 \leq i \leq T$, and compute the selection criteria for each model $c \in C$ based on the first half of the sample, retaining the top m models (set C_s with M elements). Then, using the remaining half of the data, and for each model $c \in C_s$, define $\widehat{\omega}_c^* = \frac{1}{T/2}\sum_{i=T/2+1}^{T}\widehat{\omega}_{c,i}$ for $c \in C_s$ where $\widehat{\omega}_{c,T/2+1} = 1/M$ and $\widehat{\omega}_{c,i}$ is calculated from $\widehat{\omega}_c$ for each criterion for the sample $T/2+1, ..., i$. There are several possibilities for the choice of m (Yuan and Yang, 2005 set $m = 40$, for example), but we suggest using a concave function of the number of combinations such as $m = \lfloor P^{1/2} \rfloor$, where P denotes the number of instrument combinations and $\lfloor . \rfloor$ denotes the integer part.

Finally, we consider a simplified screening procedure, which we designate by 'trimming', in which selection criteria are computed for all possible combinations, but only the top m models are retained for averaging. Again, we experimented with different choices for m's, but we opted for the sample dependent choice $m = \lfloor P^{1/2} \rfloor$.

Nonetheless, we should stress that there are important differences regarding model screening in an IV setup. Contrary to LS, in IV regression the screening is applied to the reduced form equation and not to the structural equation of interest. This may give less relevance to screening in IV since instruments are not part of the structural equation and, thus, omitted variable bias due to averaging will not be a factor.

4 Properties of the Linear IV Estimator

Given that the MA estimator $\widehat{\theta}(\omega) = \sum_{c \in C} \omega_c \widehat{\theta}_c$ is averaging over a list of candidate TSLS estimators for a given ω, its limit statistical properties depend on a linear combination of the random processes $\widehat{\theta}_c, c \in C$, possibly containing common instruments, which are $\sqrt{T}$-gaussian with asymptotic variance V_c under standard regularity conditions, as stated in the following Assumption (implicit

at all Theorems and Corollaries of this paper):

Assumption 1 *(Asymptotic normality of $\widehat{\theta}_c$)*

Assume that the data is random; $E\left(\overline{Z}_i u_i\right) = 0$; $E\left(\overline{Z}_i Y_i\right)$ is full column rank n, and $\widetilde{Q}_{\overline{Z},\overline{Z}} = E\left(u_i^2 \overline{Z}_i \overline{Z}_i'\right)$ is nonsingular.

Hence, we show in the next theorem that $\widehat{\theta}(\omega)$ is also consistent and $\sqrt{T}$-gaussian. Note, however, that the asymptotic variance will include covariance terms associated to $\widehat{\theta}_c$'s with common instruments, which could complicate the derivation of its limiting behavior. We circumvent this problem by defining a selection matrix that contains certain rows with zeros, operating on the full list of instruments, $\overline{Z}_{\overline{c}}$ (here, $c = \overline{c} = \iota_K$ and $|c| = K$), as in Domowitz and White (1982), see also Newey (1985). Let Λ_c be a matrix of dimension K by $|c|$, such that each column $j = 1, ..., |c|$ contains zeros, except a single "1" at position i that corresponds to the instrument as defined in model $c = \iota_K$. Then,

$$\overline{Z}_c = \overline{Z}_{\overline{c}} \Lambda_c, \tag{25}$$

and in this way we obtain the limiting distribution of our MA estimator in any general form of heteroskedasticity, as shown in the following theorem.

Theorem 1 *(Distribution of the MA estimator): Assume that the model is correctly specified. As $T \to \infty$, for any $\omega \in H_K$,*

$$\widehat{\theta}(\omega) = \sum_{c \in C} \omega_c \widehat{\theta}_c \xrightarrow{p} \theta, \tag{26}$$

where $\widehat{\theta}_c$ is the TSLS estimator for model $c \in C$. Moreover,

$$\sqrt{T}\left(\widehat{\theta}(\omega) - \theta\right) \xrightarrow{d} N(0, V_\omega), \tag{27}$$

where

$$V_\omega = \left(\sum_{c \in C} \omega_c V_c Q_{\overline{X},\overline{Z}_c} \widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c} \Lambda_c'\right) \widetilde{Q}_{\overline{Z}_{\overline{c}},\overline{Z}_{\overline{c}}} \left(\sum_{c \in C} \omega_c \Lambda_c \widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c} Q_{\overline{Z}_c,\overline{X}} V_c\right). \tag{28}$$

Note that the variance matrix in the middle corresponds to employing all instruments (i.e., $c = \overline{c} = \iota_m$).

The following Corollary is under the assumption of homoskedasticity noting that, in this case, $\widetilde{Q}_{\overline{Z}_c,\overline{Z}_c} = \sigma_u^2 Q_{\overline{Z}_c,\overline{Z}_c}$.

Corollary 1 *(Distribution of the MA estimator under homoskedasticity): Assume that the model is correctly specified. As $T \to \infty$, for any $\omega \in H_K$,*

$$\widehat{\theta}(\omega) = \sum_{c \in C} \omega_c \widehat{\theta}_c \xrightarrow{p} \theta, \tag{29}$$

where $\widehat{\theta}_c$ is the TSLS estimator for model $c \in C$. Moreover,

$$\sqrt{T}\left(\widehat{\theta}(\omega) - \theta\right) \xrightarrow{d} N(0, V_\omega), \tag{30}$$

where

$$V_\omega = \left(\sum_{c \in C} \omega_c V_c Q_{\overline{X},\overline{Z}_c} Q^{-1}_{\overline{Z}_c,\overline{Z}_c} \Lambda_c' \right) \frac{Q_{\overline{Z}_{\widetilde{c}},\overline{Z}_{\widetilde{c}}}}{\sigma_u^2} \left(\sum_{c \in C} \omega_c \Lambda_c Q^{-1}_{\overline{Z}_c,\overline{Z}_c} Q_{\overline{Z}_c,\overline{X}} V_c \right). \tag{31}$$

Remark. A post-model-selection estimator, PMSE, is indeed a special case of MA. Whenever $\omega_{\widetilde{c}} = 1$ and $\omega_{c'} = 0$, for all $c' \neq \widetilde{c}$, for some model $c = \widetilde{c}$, we have $\widehat{\theta}(\omega) = \widehat{\theta}_{\widetilde{c}}$ and $V_\omega = V_{\widetilde{c}}$.

Thus, for a given ω, and noting that Λ_c is known for all $c \in C$, a consistent estimator of V_ω can be obtained using consistent estimators for $Q_{\cdot_c}$, for all $c \in C$, and for σ_u^2 as well, and inference can be carried out in the usual way. In the previous section, we recommend the use of a few criteria for selecting ω. Clearly, the asymptotic covariance matrix V_ω will differ across methods for obtaining $\widehat{\omega}$. Still, notice that the expression for V_ω is derived independently on the criteria in section 3 that we pick for replacing ω by $\widehat{\omega}$.

Now, we provide asymptotic results for the MA estimator evaluated at $\widehat{\omega}$. This is possible due to the fact that we have closed form expressions for $\widehat{\omega}(ISC)$ and $\widehat{\omega}(GR^2)$ and, for illustration purposes, we now only consider the case of the $RMSC$ using the AIC penalty term. For this particular case,

$$\widehat{\omega}_c(RMSC) = \frac{\left|\widehat{V}_c\right|^{-\frac{1}{2}} \exp(p - |c|)}{\sum_{c' \in C} \left|\widehat{V}_{c'}\right|^{-\frac{1}{2}} \exp(p - |c'|)}, \tag{32}$$

which converges in probability to

$$\omega_c(RMSC) = \frac{|V_c|^{-\frac{1}{2}} \exp(p - |c|)}{\sum_{c' \in C} |V_{c'}|^{-\frac{1}{2}} \exp(p - |c'|)}, \tag{33}$$

as $T \to \infty$. Recall that, $p = 1 + K_1$ and $|c| = K_{1c} + K_{2c}$, where K_{1c}, K_{2c} are the number of exogenous variables and instruments used in model c, respectively. Hence, $|c| - p = K_{1c} + K_{2c} - 1 - K_1$. For simplicity, once we restrict to models that have exactly the same number of instruments (not necessarily the same instruments), as $T \to \infty$,

$$\widehat{\omega}_c(RMSC) = \frac{\left|\widehat{V}_c\right|^{-\frac{1}{2}}}{\sum_{c' \in C^*} \left|\widehat{V}_{c'}\right|^{-\frac{1}{2}})} \xrightarrow{p} \frac{|V_c|^{-\frac{1}{2}}}{\sum_{c' \in C^*} |V_{c'}|^{-\frac{1}{2}})} = \omega_c(RMSC), \tag{34}$$

where

$$C^* = \left\{ c \in \Re^K \backslash \{0\} : c_j = 0 \text{ or } 1, \forall 1 \leq j \leq K, c = (c_1, ..., c_K)', \text{ s.t. } |c| = K_{1c} + K_{2c} = K^* < K, \forall c \right\}, \tag{35}$$

with, $C^* \subset C$ and $\sum_{c \in C} \omega_c(RMSC) = 1$.

Note that the empirical weights $\widehat{\omega}$ follow from existing instrument selection criterion or a measure of the goodness of fit: $AMSE$, MSC, $RMSC$, $CCIC$ and GR^2. Hence, in order to study the properties of the MA estimator for each criterion, we need the following additional assumption for Theorem 2:

Assumption 2 *(Regularity conditions for instrument selection criteria and goodness of fit)*

Depending on the chosen MA approach, assume either the conditions ***(A2-MSC)*** *for the MSC as in Andrews (1999); or* ***(A2-RMSC)*** *for the RMSC as in Hall et al. (2007); or* ***(A2-CCIC)*** *for the CCIC as in Hall and Peixe (2003); or* ***(A2-GR^2)*** *for the GR^2 as in Pesaran and Smith (1994).*

By a similar token, we can consider any information criteria such that $\widehat{\omega}(ISC) \xrightarrow{p} \omega(ISC)$, as $T \to \infty$, for some well defined $\omega(ISC)$ (similarly for $\widehat{\omega}(GR^2)$) to establish the next Theorem.

Theorem 2 *(Distribution of the MA estimator evaluated at $\widehat{\omega}$): Assume that the model is correctly specified. As $T \to \infty$, for any $\omega \in H_K$,*

$$\widehat{\theta}\left(\widehat{\omega}(ISC)\right) = \sum_{c \in C} \widehat{\omega}_c(ISC)\widehat{\theta}_c \xrightarrow{p} \theta, \tag{36}$$

where $\widehat{\theta}_c$ and $\widehat{\omega}_c(ISC)$ are the TSLS estimator and the empirical weight, respectively, for model $c \in C$. Moreover,

$$\sqrt{T}\left(\widehat{\theta}\left(\widehat{\omega}(ISC)\right) - \theta\right) \xrightarrow{d} N\left(0, V_\omega(ISC)\right), \tag{37}$$

where

$$V_\omega(ISC) = \left(\sum_{c \in C} \omega_c(ISC) V_c Q_{\overline{X},\overline{Z}_c} \widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c} \Lambda_c'\right) \widetilde{Q}_{\overline{Z}_{\overline{c}},\overline{Z}_{\overline{c}}} \left(\sum_{c \in C} \omega_c(ISC) \Lambda_c \widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c} Q_{\overline{Z}_c,\overline{X}} V_c\right), \tag{38}$$

for well a defined $\omega(ISC)$.

Note that for the AMSE, we cannot directly use the conditions in Donald and Newey (2001), as their framework allows for K_2 to increase with T at a suitable rate. However, for the purposes of model averaging, their regularity conditions remain applicable for a fixed dimension of the candidate set Z_c.

5 Monte Carlo Study

In this section, we report results of a simple Monte Carlo study assessing the finite sample properties of the proposed IV model averaging estimators. To facilitate comparisons, we base our experiments on a modified version of the design used in Donald and Newey (2001), Eryuruk et al. (2009), Kuersteiner and Okui (2010) and Okui (2011), for example, by also allowing for an exogenous variable and the presence of heteroskedastic errors, which are relevant features in most applications.

The data generating process is

$$y_i = \beta_0 Y_i + \gamma_0 X_i + \varepsilon_i(1 + \phi|z_i^*|), \qquad Y_i = \pi' Z_i + u_i + \eta_i, \qquad i = 1, ..., T, \tag{39}$$

where the true parameter of interest is the scalar β_0, which is fixed at 0.1. Y_i and X_i are scalars, with $X_i = \nu_i + \eta_i$, ν_i and η_i being independently distributed $N(0,1)$ random variables. Note that

Y_i and X_i are correlated via η_i (which itself is independent from u_i). Furthermore,

$$\left(\varepsilon_i, u_i, Z_i'\right)' \sim i.i.d.N\left(0, \Sigma\right), \text{ where } \Sigma = \begin{pmatrix} 1 & 0.5 & 0_{1\times M} \\ 0.5 & 1 & 0_{1\times M} \\ 0_{M\times 1} & 0_{M\times 1} & I_M \end{pmatrix}. \tag{40}$$

The degree of endogeneity is 0.5 and we define $z_i^* = z_{1i} + \eta_i$, where z_{1i} is the first column of Z. Thus, the error term is heteroskedastic when $\phi \neq 0$, so we set $\phi \in \{0, 0.1\}$. We also consider cases with and without the exogenous variable X_i, that is, $\gamma_0 \in \{0, 0.1\}$. The number of observations is $T \in \{100, 250\}$ and the number of replications is 5000.

We set the maximum number of instruments M to 10 and 20 and allow for different combinations of instruments. We fix a block of moment conditions (M_{fixed}), assumed to be valid, namely $M_{fixed} = 2$ when $M = 10$ (255 combinations) and $M_{fixed} = 10$ when $M = 20$ (1023 combinations). For each replication, we select the instruments for the fixed block that maximize the correlation with the endogenous regressor Y_i, while using all possible combinations of the remaining instruments, up to a maximum of 255/1023 combinations (i.e., all possible combinations of the 'free' instruments).

In terms of specifications for π, we have, for $j = 1, ..., M$,

$$\text{Model A (equal coefficients)} \quad : \quad \pi_j = \sqrt{\frac{R_f^2}{M\left(1 - R_f^2\right)}}, \tag{41}$$

$$\text{Model B (declining coefficients)} \quad : \quad \pi_j = c\left(M\right)\left(1 - \frac{j}{M+1}\right)^4, \tag{42}$$

where $c\left(M\right)$ is set so that π satisfies $\pi'\pi = R_f^2/\left(1 - R_f^2\right)$, where $R_f^2 \in \{0.1, 0.01\}$.

Note that in model A, all instruments are equally important (and relatively weak), which means that instrument selection methods may not be very effective. In model B, the strength of the instruments declines gradually, but the ordering matters. The value of $R_f^2 = 0.01$ can be interpreted as the "weak instruments" case, quite common in empirical applications.

As in Kuersteiner and Okui (2010) and Okui (2011), we use the selection method of Donald and Newey (2001) as the benchmark against which our procedures are compared. Following Donald and Newey (2001), we compute their estimator (TSLS-DN) using a cross-validation criterion for the first-stage reduced-form model. Notice that for Model B, the instruments should be included in the order of their explanatory power, which corresponds to the case in which the practitioner knows the relative importance of the instruments. Thus, these conditions may favour selection methods such as that of Donald and Newey (2001) - the experiments in Okui (2011) and Eryuruk et al. (2009) suggest that sequentially adding the instruments in the 'wrong' order causes the performance of selection methods to deteriorate.

On the other hand, we consider smooth MA estimators using different $\widehat{\omega}(.)$ based on the different criteria discussed in section 2 (denoted as MA-DN, MSC-BIC, RMSC, CCIC and GR^2) and for each estimator we compute the median bias (MB) and the median absolute deviation (MAD) relative to

Table 1: Model Averaging Estimators, Model A

$M = 10$, $M_{fixed} = 2$		DNTSLS	MA-DN	MSC-BIC	$RMSC$	$CCIC$	GR^2
$T = 100$, $R_f^2 = 0.1$	MB	0.1918	0.1812	0.1670	0.1917	0.1829	0.1791
	$[RMB]$		[0.9446]	[0.8708]	[0.9992]	[0.9534]	[0.9337]
	MAD	0.2725	0.1966	0.1943	0.2147	0.2068	0.1965
	$[RMAD]$		[0.7217]	[0.7133]	[0.7881]	[0.7591]	[0.7214]
	IDR	1.0117	0.4928	0.5580	0.5687	0.5526	0.5152
$R_f^2 = 0.01$	MB	0.2421	0.2424	0.2384	0.2503	0.2437	0.2420
	$[RMB]$		[1.0014]	[0.9848]	[1.0341]	[1.0067]	[0.9998]
	MAD	0.2614	0.2523	0.2570	0.2652	0.2651	0.2535
	$[RMAD]$		[0.9650]	[0.9830]	[1.0145]	[1.0141]	[0.9697]
	IDR	0.5869	0.5698	0.6482	0.6444	0.6770	0.5894
$T = 250$, $R_f^2 = 0.1$	MB	0.1365	0.1235	0.1060	0.1391	0.1436	0.1170
	$[RMB]$		[0.9051]	[0.7772]	[1.0197]	[1.0521]	[0.8571]
	MAD	0.1733	0.1424	0.1388	0.1663	0.1660	0.1416
	$[RMAD]$		[0.8215]	[0.8011]	[0.9597]	[0.9581]	[0.8171]
	IDR	0.4545	0.3947	0.4299	0.4894	0.4581	0.4109
$R_f^2 = 0.01$	MB	0.2300	0.2113	0.2162	0.2289	0.2227	0.2192
	$[RMB]$		[0.9189]	[0.9400]	[0.9952]	[0.9682]	[0.9530]
	MAD	0.2473	0.2219	0.2337	0.2497	0.2510	0.2316
	$[RMAD]$		[0.8973]	[0.9448]	[1.0096]	[1.0150]	[0.9365]
	IDR	0.5421	0.6094	0.6040	0.6426	0.6796	0.5646
$M = 20$, $M_{fixed} = 10$							
$T = 100$, $R_f^2 = 0.1$	MB	0.2214	0.2088	0.2008	0.2079	0.2088	0.2079
	$[RMB]$		[0.9431]	[0.9073]	[0.9391]	[0.9431]	[0.9390]
	MAD	0.2970	0.2103	0.2057	0.2170	0.2126	0.2102
	$[RMAD]$		[0.7079]	[0.6925]	[0.7305]	[0.7156]	[0.7075]
	IDR	0.4795	0.3690	0.4171	0.4949	0.3998	0.3743
$R_f^2 = 0.01$	MB	0.2505	0.2515	0.2506	0.2504	0.2460	0.2504
	$[RMB]$		[1.0040]	[1.0007]	[0.9996]	[0.9823]	[0.9995]
	MAD	0.2584	0.2524	0.2541	0.2541	0.2491	0.2516
	$[RMAD]$		[0.9768]	[0.9831]	[0.9832]	[0.9639]	[0.9736]
	IDR	0.4315	0.3973	0.4651	0.5156	0.4372	0.4075
$T = 250$, $R_f^2 = 0.1$	MB	0.1781	0.1611	0.1539	0.1607	0.1703	0.1591
	$[RMB]$		[0.9046]	[0.8641]	[0.9020]	[0.9561]	[0.8929]
	MAD	0.2272	0.1644	0.1600	0.1788	0.1768	0.1616
	$[RMAD]$		[0.7236]	[0.7045]	[0.7869]	[0.7785]	[0.7112]
	IDR	0.4364	0.3339	0.3572	0.4904	0.3948	0.3442
$R_f^2 = 0.01$	MB	0.2480	0.2461	0.2475	0.2504	0.2449	0.2420
	$[RMB]$		[0.9925]	[0.9983]	[1.0097]	[0.9877]	[0.9758]
	MAD	0.2559	0.2473	0.2497	0.2639	0.2489	0.2440
	$[RMAD]$		[0.9665]	[0.9760]	[1.0313]	[0.9729]	[0.9535]
	IDR	0.4067	0.3872	0.4219	0.5446	0.4697	0.4076

Notes: numbers in square brackets are measures relative to the Donald-Newey estimator; standard deviations in round brackets; (R)MD and (R)MAD denote (Relative) Median Bias and Median Absolute Deviation.

Table 2: Model Averaging Estimators, Model B

$M = 10$, $M_{fixed} = 2$		DNTSLS	MA-DN	MSC-BIC	$RMSC$	$CCIC$	GR^2
$T = 100$, $R_f^2 = 0.1$	MB	0.1714	0.1754	0.1627	0.1892	0.1698	0.1768
	[RMB]		[1.0228]	[0.9490]	[1.1038]	[0.9904]	[1.0315]
	MAD	0.1996	0.1884	0.1903	0.2067	0.1898	0.1895
	[$RMAD$]		[0.9439]	[0.9532]	[1.0351]	[0.9508]	[0.9490]
	IDR	0.4895	0.4771	0.5404	0.5499	0.5330	0.4953
$R_f^2 = 0.01$	MB	0.2368	0.2441	0.2411	0.2441	0.2427	0.2444
	[RMB]		[1.0309]	[1.0179]	[1.0308]	[1.0246]	[1.0321]
	MAD	0.2598	0.2552	0.2570	0.2615	0.2627	0.2560
	[$RMAD$]		[0.9823]	[0.9894]	[1.0066]	[1.0113]	[0.9856]
	IDR	0.5853	0.5770	0.6509	0.6539	0.6747	0.5990
$T = 250$, $R_f^2 = 0.1$	MB	0.1038	0.1070	0.1014	0.1419	0.0927	0.1212
	[RMB]		[1.0317]	[0.9773]	[1.3677]	[0.8930]	[1.1679]
	MAD	0.1388	0.1278	0.1290	0.1653	0.1257	0.1399
	[$RMAD$]		[0.9207]	[0.9295]	[1.1905]	[0.9057]	[1.0076]
	IDR	0.4172	0.3757	0.4111	0.4659	0.4107	0.3988
$R_f^2 = 0.01$	MB	0.2194	0.2270	0.2344	0.2251	0.2181	0.2241
	[RMB]		[1.0347]	[1.0685]	[1.0258]	[0.9942]	[1.0213]
	MAD	0.2348	0.2468	0.2449	0.2553	0.2541	0.2401
	[$RMAD$]		[1.0512]	[1.0428]	[1.0874]	[1.0824]	[1.0226]
	IDR	0.5513	0.5828	0.6196	0.6375	0.6550	0.5679
$M = 20$, $M_{fixed} = 10$							
$T = 100$, $R_f^2 = 0.1$	MB	0.1995	0.2048	0.1971	0.2100	0.2010	0.2051
	[RMB]		[1.0266]	[0.9878]	[1.0529]	[1.0076]	[1.0283]
	MAD	0.2095	0.2073	0.2012	0.2160	0.2051	0.2068
	[$RMAD$]		[0.9896]	[0.9602]	[1.0310]	[0.9791]	[0.9873]
	IDR	0.3912	0.3650	0.4127	0.4647	0.3872	0.3721
$R_f^2 = 0.01$	MB	0.2478	0.2521	0.2504	0.2483	0.2487	0.2510
	[RMB]		[1.0172]	[1.0103]	[1.0021]	[1.0035]	[1.0128]
	MAD	0.2571	0.2528	0.2534	0.2527	0.2508	0.2521
	[$RMAD$]		[0.9833]	[0.9856]	[0.9828]	[0.9757]	[0.9805]
	IDR	0.4325	0.3971	0.4653	0.5090	0.4304	0.4053
$T = 250$, $R_f^2 = 0.1$	MB	0.1505	0.1534	0.1419	0.1675	0.1493	0.1553
	[RMB]		[1.0194]	[0.9429]	[1.1129]	[0.9918]	[1.0315]
	MAD	0.1652	0.1577	0.1491	0.1781	0.1552	0.1584
	[$RMAD$]		[0.9544]	[0.9024]	[1.0776]	[0.9392]	[0.9589]
	IDR	0.3673	0.3193	0.3443	0.4386	0.3542	0.3291
$R_f^2 = 0.01$	MB	0.2195	0.2067	0.2141	0.2278	0.2208	0.2175
	[RMB]		[0.9417]	[0.9752]	[1.0378]	[1.0059]	[0.9910]
	MAD	0.2402	0.2175	0.2335	0.2479	0.2484	0.2311
	[$RMAD$]		[0.9055]	[0.9718]	[1.0319]	[1.0339]	[0.9621]
	IDR	0.5431	0.6029	0.6048	0.6445	0.6719	0.5617

See notes to Table 1.

that of TSLS-DN (RMB and RMAD), as well as the inter-decile range (IDR, the difference between the 10% and 90% deciles), as in Kuersteiner and Okui (2010).

Tables 1 and 2 contain results for Models A and B, respectively. A few interesting conclusions emerge from this simple study. First, the MA estimators perform almost uniformly better than the DN selection procedure across all specifications in terms of median absolute deviations, with substantial reductions in this measure. The gains in terms of bias can be sizeable, although they depend on the setting at hand. Second, the MSC-based MA estimator appears to be the most robust in all settings, but the differences between estimators are generally small. Third, even in the case where one would expect the DN procedure to dominate (Model B), MA estimators have comparable MB performances, usually bettering the DN method in terms MAD. In fact, for this specification the comparison favors some MA methods, namely the MSC-BIC procedure, as the sample size and the number of instruments increase. Fourth, varying the strength of the instruments through R_f^2 has the effect of equalizing the performance of selection and MA procedures for $R_f^2 = 0.01$ for both models A and B. Though less striking, the performance of the MA estimators is slightly superior to the DN estimator in many instances, displaying a better balance between bias and dispersion.

Concerning estimation with screening procedures, we focus on the case of Model B with $M = 20$ and $M_{fixed} = 10$, such that the strength of the instruments varies and the number of combinations is the largest, and therefore screening methods might be most useful. Results in Table 3, however, are mixed, as there is no clear pattern across specifications or MA procedures. For example, the 'trimming' approach improves the performance of all MA estimators in terms of bias when $T = 250$ and $R^2 = 0.1$, but for $R^2 = 0.01$ no screening is a better choice, while the converse is true for $T = 100$. In fact, of the three screening procedures, the simple 'trimming' seems to behave the best, while the performance of the MA estimators deteriorates most with the split-sample screening of Yuan and Yang (2005), particularly in terms of MAD. Interestingly, the 'backward elimination' procedure can lead to significant gains in bias for the MA-DN and RMSC approaches, but not in a consistent way. Overall, it appears that non-screened estimators display a well balanced finite sample bias-variance trade-off, suggesting that the averaging schemes studied here are able to correctly filter poorer models by appropriately weighing them down, while averaging over a larger number of models helps to decrease dispersion.

Finally, in Table 4 we observe that the performance of the estimators in terms of median bias relative to DN is largely unaffected by the presence of an endogenous variables and heteroskedasticity when $M = 10$ and $M_{fixed} = 2$, but these additional features improve the relative behavior of MA estimators when $M = 20$ and $M_{fixed} = 10$. Note, however, that the advantages of MA estimators are substantial in terms of the RMAD measure when an exogenous variable is included. Moreover, the MA estimators show, in general, a great deal less dispersion, as can be seen by the IDR rows, with the DN selection procedures performing a lot worse when the exogenous variable is in place. This seems to imply that MA procedures do indeed attenuate the trade-off between bias and dispersion relative to instrument selection procedures.

Table 3: MA Estimation with Screening, Model B, $M = 20$, $M_{fixed} = 10$

$T = 100$, $R_f^2 = 0.1$		DNTSLS	MA-DN	MSC-BIC	$RMSC$	$CCIC$	GR^2
Trimming	MB	0.1995	0.2138	0.2109	0.2191	0.2144	0.2156
	$[RMB]$		[1.0692]	[1.0546]	[1.0956]	[1.0721]	[1.0779]
	MAD	0.2095	0.2180	0.2202	0.2223	0.2162	0.2186
	$[RMAD]$		[1.0383]	[1.0484]	[1.0584]	[1.0297]	[1.0411]
	IDR	0.3912	0.3653	0.3876	0.3843	0.3751	0.3661
Screening	MB		0.2130	0.2100	0.2168	0.2223	0.1253
	$[RMB]$		[1.0652]	[1.0498]	[1.0840]	[1.1114]	[0.6264]
	MAD		0.2227	0.2222	0.2340	0.2279	0.1391
	$[RMAD]$		[1.0607]	[1.0579]	[1.1141]	[1.0853]	[0.6625]
	IDR		0.4556	0.4354	0.4756	0.4934	0.3248
Backward elimination	MB		0.1065	0.1897	0.0983	0.1965	0.4626
	$[RMB]$		[0.5858]	[1.0435]	[0.5411]	[1.0815]	[2.5455]
	MAD		0.2235	0.1938	0.2848	0.2394	0.4707
	$[RMAD]$		[1.1564]	[1.0025]	[1.4730]	[1.2382]	[2.4350]
	IDR		0.7864	0.3807	1.0106	0.7171	0.4603
$R_f^2 = 0.01$							
Trimming	MB	0.2478	0.2423	0.2432	0.2498	0.2455	0.2418
	$[RMB]$		[0.9691]	[0.9729]	[0.9992]	[0.9819]	[0.9672]
	MAD	0.2571	0.2426	0.2474	0.2529	0.2482	0.2422
	$[RMAD]$		[0.9331]	[0.9517]	[0.9726]	[0.9546]	[0.9316]
	IDR	0.4325	0.3879	0.3889	0.4270	0.4162	0.3875
Screening	MB		0.2638	0.2651	0.2609	0.2539	0.1602
	$[RMB]$		[1.0554]	[1.0605]	[1.0437]	[1.0155]	[0.6409]
	MAD		0.2692	0.2720	0.2628	0.2609	0.1655
	$[RMAD]$		[1.0355]	[1.0462]	[1.0109]	[1.0036]	[0.6365]
	IDR		0.5047	0.4803	0.5227	0.5262	0.3542
Backward elimination	MB		0.1839	0.2399	0.2095	0.2546	0.5318
	$[RMB]$		[0.7608]	[0.9926]	[0.8670]	[1.0537]	[2.2005]
	MAD		0.2768	0.2439	0.3677	0.2934	0.5329
	$[RMAD]$		[1.1010]	[0.9700]	[1.4627]	[1.1670]	[2.1197]
	IDR		0.8503	0.4217	1.0830	0.7154	0.4798
$T = 250$, $R_f^2 = 0.1$							
Trimming	MB	0.1505	0.1537	0.1451	0.1536	0.1503	0.1546
	$[RMB]$		[0.9604]	[0.9071]	[0.9601]	[0.9396]	[0.9663]
	MAD	0.1652	0.1554	0.1499	0.1568	0.1533	0.1555
	$[RMAD]$		[0.9144]	[0.8818]	[0.9221]	[0.9015]	[0.9145]
	IDR	0.3673	0.3060	0.3136	0.3323	0.3228	0.3042
Screening	MB		0.2149	0.2040	0.2198	0.2110	0.1822
	$[RMB]$		[1.4231]	[1.3508]	[1.4554]	[1.3976]	[1.2063]
	MAD		0.2169	0.2057	0.2256	0.2174	0.1848
	$[RMAD]$		[1.2759]	[1.2103]	[1.3270]	[1.2786]	[1.0872]
	IDR		0.4565	0.4260	0.4960	0.4858	0.4109
Backward elimination	MB		0.1527	0.1587	0.1005	0.1773	0.4127
	$[RMB]$		[1.0408]	[1.0821]	[0.6853]	[1.2084]	[2.8135]
	MAD		0.2325	0.1611	0.2407	0.2241	0.4151
	$[RMAD]$		[1.4530]	[1.0070]	[1.5043]	[1.4002]	[2.5939]
	IDR		0.7758	0.3278	0.9231	0.6258	0.3832
$R_f^2 = 0.01$							
Trimming	MB	0.2195	0.2304	0.2336	0.2426	0.2345	0.2290
	$[RMB]$		[1.0474]	[1.0617]	[1.1028]	[1.0658]	[1.0408]
	MAD	0.2402	0.2306	0.2338	0.2433	0.2372	0.2301
	$[RMAD]$		[0.9601]	[0.9733]	[1.0130]	[0.9875]	[0.9580]
	IDR	0.5431	0.3844	0.4051	0.4492	0.4447	0.3887
Screening	MB		0.2336	0.2110	0.2398	0.2251	0.2209
	$[RMB]$		[1.0620]	[0.9590]	[1.0898]	[1.0233]	[1.0042]
	MAD		0.2833	0.2469	0.3084	0.2963	0.2685
	$[RMAD]$		[1.1803]	[1.0288]	[1.2851]	[1.2344]	[1.1189]
	IDR		0.8668	0.7513	0.9472	0.9873	0.8176
Backward elimination	MB		0.2564	0.2452	0.2376	0.2651	0.5340
	$[RMB]$		[1.0560]	[1.0102]	[0.9789]	[1.0919]	[2.1996]
	MAD		0.3296	0.2464	0.3856	0.2984	0.5360
	$[RMAD]$		[1.3082]	[0.9779]	[1.5304]	[1.1844]	[2.1270]
	IDR		0.9468	0.3892	1.0593	0.7887	0.4557

See notes to Table (1); Trimming corresponds to the case where the $m = \lfloor 1023^{1/2} \rfloor$ best models are used for averaging; Screening denotes the split-sample procedure as in Yuan and Yang (2005); 'Backward elimination' designates the screening procedure following Claeskens et al. (2006).

Table 4: Model A, $R^2 = 0.1$, with an Exogenous Regressor and/or Heteroskedasticity

$M = 10$, $M_{fixed} = 2$		DNTSLS	MA-DN	MSC-BIC	$RMSC$	$CCIC$	GR^2
$T = 100$							
$\gamma_0 = 0, \phi = 0.1$	MB	0.2163	0.2019	0.1881	0.2162	0.2053	0.2016
	$[RMB]$		[0.9335]	[0.8694]	[0.9995]	[0.9490]	[0.9319]
	MAD	0.3068	0.2198	0.2183	0.2388	0.2327	0.2200
	$[RMAD]$		[0.7165]	[0.7114]	[0.7783]	[0.7585]	[0.7170]
	IDR	0.6103	0.5507	0.6256	0.6385	0.6185	0.5761
$\gamma_0 = 0.1, \phi = 0.1$	MB	0.2544	0.2413	0.2248	0.2715	0.2530	0.2409
	$[RMB]$		[0.9485]	[0.8837]	[1.0673]	[0.9948]	[0.9471]
	MAD	0.6289	0.2713	0.2577	0.2895	0.3082	0.2651
	$[RMAD]$		[0.4313]	[0.4098]	[0.4604]	[0.4901]	[0.4216]
	IDR	3.7933	0.7081	0.7160	0.7064	0.9051	0.6794
$T = 250$							
$\gamma_0 = 0, \phi = 0.1$	MB	0.1512	0.1395	0.1184	0.1551	0.1595	0.1308
	$[RMB]$		[0.9226]	[0.7827]	[1.0253]	[1.0544]	[0.8652]
	MAD	0.1942	0.1604	0.1553	0.1866	0.1864	0.1574
	$[RMAD]$		[0.8261]	[0.7998]	[0.9609]	[0.9601]	[0.8105]
	IDR	0.4877	0.4457	0.4837	0.5467	0.5179	0.4619
$\gamma_0 = 0.1, \phi = 0.1$	MB	0.1625	0.1363	0.1340	0.1828	0.1723	0.1437
	$[RMB]$		[0.8387]	[0.8244]	[1.1248]	[1.0602]	[0.8846]
	MAD	0.5139	0.1814	0.1754	0.2172	0.2333	0.1805
	$[RMAD]$		[0.3530]	[0.3414]	[0.4227]	[0.4540]	[0.3512]
	IDR	3.1121	0.6275	0.5474	0.6279	0.7278	0.5388
$M = 20$, $M_{fixed} = 10$							
$T = 100$							
$\gamma_0 = 0, \phi = 0.1$	MB	0.2468	0.2339	0.2248	0.2333	0.2346	0.2331
	$[RMB]$		[0.9475]	[0.9108]	[0.9450]	[0.9503]	[0.9444]
	MAD	0.3382	0.2360	0.2303	0.2429	0.2380	0.2360
	$[RMAD]$		[0.6979]	[0.6810]	[0.7184]	[0.7037]	[0.6980]
	IDR	0.5318	0.4124	0.4702	0.5486	0.4464	0.4186
$\gamma_0 = 0.1, \phi = 0.1$	MB	0.3460	0.3004	0.2887	0.3042	0.2977	0.3001
	$[RMB]$		[0.8681]	[0.8344]	[0.8791]	[0.8602]	[0.8672]
	MAD	0.6201	0.3052	0.2909	0.3091	0.3024	0.3020
	$[RMAD]$		[0.4922]	[0.4690]	[0.4984]	[0.4877]	[0.4870]
	IDR	3.3227	0.4917	0.5516	0.5868	0.5553	0.4831
$T = 250$							
$\gamma_0 = 0, \phi = 0.1$	MB	0.1512	0.1404	0.1199	0.1547	0.1606	0.1320
	$[RMB]$		[0.9283]	[0.7926]	[1.0230]	[1.0623]	[0.8727]
	MAD	0.1942	0.1621	0.1595	0.1866	0.1876	0.1604
	$[RMAD]$		[0.8350]	[0.8213]	[0.9609]	[0.9663]	[0.8262]
	IDR	0.4877	0.4517	0.4962	0.5555	0.5200	0.4761
$\gamma_0 = 0.1, \phi = 0.1$	MB	0.3012	0.2280	0.2194	0.2349	0.2275	0.2277
	$[RMB]$		[0.7568]	[0.7284]	[0.7799]	[0.7554]	[0.7560]
	MAD	0.6092	0.2297	0.2221	0.2426	0.2424	0.2293
	$[RMAD]$		[0.3770]	[0.3646]	[0.3982]	[0.3979]	[0.3763]
	IDR	3.0986	0.4491	0.4645	0.5923	0.5542	0.4483

See notes to Table 1.

These results suggest that for higher dimensional problems, when there are more instruments to choose from (and to combine), selection procedures may lead to worse estimates, whereas a model averaging approach may use all available information more efficiently. The same appears to be true for different sample sizes, i.e., the relative performance of MA estimators improves when $T = 250$, particularly for $M = 20$ and $M_{fixed} = 10$. Interestingly, of the additional features we considered, heteroskedasticity seems to have a lighter impact, whereas the presence of an exogenous variable in the model leads to better results for MA estimators.

6 Empirical Illustration: Returns to Education

This section provides an empirical example of our linear IV model averaging approach, in which we reexamine the analysis in Donald and Newey (2001) concerning instrument selection in the well known study of returns to schooling of Angrist and Krueger (1991). Using the quarter of birth and its interactions with other covariates as instruments, Angrist and Krueger (1991) report that their TSLS estimates are close to standard least squares estimates when the log of weekly wages of 329,509 men born between 1930-1939 is regressed on the number of years of schooling, thus suggesting that bias in conventional estimates is negligible (the original dataset and estimation replication files are available from Joshua Angrist's webpage).

Donald and Newey (2001) consider a particular version of the model in which an intercept plus nine year-of-birth dummies and 50 state-of-birth dummies are also used as explanatory variables. These authors then consider the problem of optimal instrument selection by defining 8 different instrument sets, treating each set of dummies as complete blocks. Following Donald and Newey's (2001) notation, the blocks of instruments are: 3 quarter-of-birth dummies (denoted as Q); 27 dummies resulting from the interaction of quarter-of-birth with 9 year-of-birth dummies (denoted as Q^*Y); 150 dummy variables obtained by interacting quarter-of-birth dummies with 50 state-of-birth dummies (Q^*S). In addition, Donald and Newey (2001) consider a set of regional dummies based on the 1980 Census classification of states into four ($R4$) and nine ($R9$) regions, interacted with quarter-of-birth dummies, thus resulting in a set of 9 instruments (Q^*R4) and 24 instruments (Q^*R9), respectively. All sets include block of quarter-of-birth dummies Q, as well as the complete set of 60 exogenous variables (intercept, Y and S), so that the smallest instruments set comprises 63 instruments. Each possible combination of these blocks was then considered, using instrument selection criteria to determine the 'optimal' instrument set.

Interestingly, the different estimators considered by Donald and Newey (2001) - TSLS, LIML and bias-corrected TSLS - yielded distinct optimal sets, although the estimates of returns to schooling differ very little across different instrument sets. Our approach, on the other hand, rather than picking one particular version of the competing models, allow us to obtain an averaged estimate of the different specifications.

For simplicity and to save space, we consider TSLS MA estimation using the AMSE, the BIC-based RMSC-BIC and the GR^2 criteria. The Donald and Newey (2001) AMSE criterion is com-

puted by using a first-stage Mallows criterion based on the largest instrument set. Though slightly different from our results, it would be straightforward to obtain MA estimates by combining the information on selection criteria in Table VII and the estimates for each set in Table VIII of Donald and Newey (2001). Each criteria is calculated for each set of instruments and the weights for averaging are then constructed by using the methods described in section 3. Table 5 contains estimates for each set (standard errors in brackets), alongside selection criteria and the corresponding weights for each specification. The final row shows the averaged estimate for each procedure.

Table 5: Model Averaging Estimates of Returns to Schooling

Instrument Set	K	TSLS	AMSE	$\widehat{\omega}_{AMSE}$	RMSC	$\widehat{\omega}_{RMSC}$	GR^2	$\widehat{\omega}_{GR^2}$
Q	63	0.1077	4.8153	0.1240	203.42	0.019	0.0289	0.125
		(0.0195)						
$Q + Q^*Y$	90	0.0869	4.8148	0.1241	201.45	0.050	0.0289	0.125
		(0.0157)						
$Q + Q^*S$	213	0.0991	4.8119	0.1242	202.82	0.025	0.0291	0.125
		(0.0099)						
$Q + Q^*Y + Q^*S$	240	0.0928	4.8112	0.1243	202.45	0.031	0.0291	0.125
		(0.0093)						
$Q + Q^*R4$	72	0.0520	4.7890	0.1257	198.55	0.215	0.0292	0.125
		(0.0046)						
$Q + Q^*Y + Q^*R4$	99	0.0518	4.7885	0.1257	198.83	0.187	0.0292	0.125
		(0.0045)						
$Q + Q^*R9$	87	0.0636	4.7846	0.1260	198.22	0.253	0.0295	0.125
		(0.0042)						
$Q + Q^*Y + Q^*R9$	114	0.0632	4.7841	0.1260	198.50	0.220	0.0295	0.125
		(0.0042)						
Model Averaging estimates				0.0770		0.0626		0.0771
				(0.0017)		(0.0011)		(0.0017)

Note: standard errors in brackets.

First, it is interesting to note that all selection criteria tend to favor the specifications that include regional dummies, in particular the 9-region classification interactions (unlike the results in Donald and Newey, 2001). Consequently, the 'smoothed' weights are larger for estimates from instrument sets containing these variables, in the case of RMSC considerably so. The weights from the AMSE and GR^2 measures tend to weigh equally all models, given that the statistics are similar (and low, in the case of GR^2), although they agree with the results of RMSC concerning models with regional dummies.

The resulting MA estimates can be read in the last row of Table 5. Given that the estimates of returns to education involving $R4$ and $R9$ are somewhat lower than the other specifications, the MA estimates inevitably reflect that. However, the results for the three different criteria are quite similar and close to the values obtained by OLS, which is consistent with the thrust of Angrist and Krueger (1991). Noticeably, the standard errors are considerably smaller, which suggests that MA estimation may yield substantial efficiency gains when compared to standard procedures.

7 Conclusion

This paper develops novel model averaging estimators in the linear instrumental variables regression framework. It is not confined to homoskedastic errors but allows for general forms of heteroskedasticity. Moreover, the approach is suitable in the context of many of instruments as it is the case when distinct blocks of instruments are at competition. We use different selection criteria to select weights for averaging across estimates. This is achieved by direct smoothing of information criteria arising from the estimation stage. We study the asymptotic properties of the resulting estimators.

A simple Monte Carlo experiment shows that our MA estimators compare very favourably in many relevant setups. It suggests that for higher dimensional problems a model averaging approach may use all available information more efficiently than standard selection procedures. In particular, the MA estimator based on a BIC selection criterion is an easily implementable and robust choice, as it appears to provide the best balance in terms of reducing both bias and dispersion in different settings. Also, we illustrate our method with an empirical application to the well known study of returns to schooling of Angrist and Krueger (1991). The results are quite close to the values obtained by OLS, which is consistent with the thrust of Angrist and Krueger (1991).

There are a few aspects that should deserve further attention. We have considered a fixed number K of instruments to combine. An alternative would be to allow K to grow with T at a certain rate, in which case the number of candidate models also increases with T and at a faster rate. Thus, it would be interesting to determine the optimal rate and to evaluate the efficiency gains of this case when compared to the fixed-K framework. Another important issue is to study the behavior of the MA estimator under different settings, such as the cases of weak or invalid instruments. In either case, consistency of the TSLS estimator is no longer achieved and nonstandard asymptotic theory is required. In addition, one should further investigate the statistical properties of the MA estimators for models with irrelevant instruments since it would simply make estimation less efficient. In this case (and unlike this paper, in which we assume instruments are valid), pretesting as in Andrews (1999) could be useful in putting our methodology into practice. Furthermore, using invalid instruments in combinations will result in poorer specifications, so model screening can play a more significant role under these conditions. We leave these topics for future research.

Acknowledgments: Financial support from the Fundação para a Ciência e Tecnologia under grant PTDC/EGE-ECO/122093/2010 is gratefully acknowledged.

References

[1] Akaike, H. (1979), "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting," *Biometrika,* 66, 237-242.

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

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

[4] Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997), "Model Selection: an Integral Part of Inference", *Biometrics,* 53, 603-618.

[5] Burnham, K. P and Anderson, D. R. (2002), *Model Selection and Multimodal Inference: a Practical Information-Theoretic Approach*, Berlin, Springer-Verlag.

[6] Canay, I. A. (2010), "Simultaneous Selection and Weighting of Moments in GMM Using a Trapezoidal Kernel," *Journal of Econometrics,* 156, 284-303.

[7] Caner, M. (2009), "LASSO Type GMM Estimator," *Econometric Theory*, 25, 1-23.

[8] Chao, J. C. and Swanson, N. R. (2005), "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673-1692.

[9] Chen, X., Jacho-Chàvez, D. T. and Linton, O. (2012), "Averaging of Moment Condition Estimators," cemmap Working Papers CWP26/12, September.

[10] Claeskens, G., Croux, C. and Venkerckhoven, J. (2006), "Variable Selection for Logit Regression Using a Prediction-Focused Information Criterion", *Biometrics*, 62, 972-979.

[11] Claeskens, G. and Hjort, N. L. (2008), *Model Selection and Model Averaging*, Cambridge University Press.

[12] Domowitz, I. and White, H. (1982), "Misspecified Models with Dependent Observations," *Journal of Econometrics*, 20, 35-58.

[13] Donald, S. G., Newey, W. K. (2001), "Choosing the Number of Instruments," *Econometrica,* 69, 1161-1192.

[14] Eryuruk, G., Hall, A. R., and Janas, K. (2009), "A Comparative Study of Three Data-Based Methods of Instrument Selection," *Economics Letters,* 105, 280-283.

[15] Hall, A. R. (2005), "Generalized Method of Moments," Oxford, UK: Oxford University Press.

[16] Hall, A. R., Inoue, A., Jana, K., Shin, C. (2007), "Information in Generalized Method of Moments Estimation and Entropy Based Moment Selection," *Journal of Econometrics,* 138, 488-512.

[17] Hall, A. R. and Peixe, F. P. M. (2003), "A Consistent Method for the Selection of Relevant Instruments," *Econometric Reviews,* 22, 269-287.

[18] Han, C. and Phillips, P. C. B. (2006), "GMM with Many Moment Conditions," *Econometrica* 74, 147-182.

[19] Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175-1189.

[20] Hansen, B. E. and Racine, J. S. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38-46.

[21] Hansen, C., Hausman, J. and Newey, W. K. (2008), "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26, 398-422.

[22] Hjort, N. L. and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879-899.

[23] Koop, G., Leon-Gonzalez, R., Strachan, R. (2012), "Bayesian Model Averaging in the Instrumental Variable Regression Model", *Journal of Econometrics*, 171, 237-250.

[24] Kuersteiner, G. and Okui, R. (2010), "Constructing Optimal Instruments by First-Stage Prediction Averaging," *Econometrica*, 78, 697-718.

[25] Lee, Y. and Zhou, Y. (2012), "Averaged Instrumental Variables Estimator", Working Paper, University of Michigan.

[26] Lo, S. N. and Ronchetti, E. (2012), "Robust small sample accurate inference in moment condition models," *Computational Statistics and Data Analysis*, 56, 3182-3197.

[27] Newey, W. K. (1985), "Generalized Method of Moments Specification Testing," *Journal of Econometrics,* 29, 229-256.

[28] Newey, W. K. and Windmeijer, F. (2009), "Generalized Method of Moments with Many Weak Moment Conditions," *Econometrica*, 77, 687-719.

[29] Okui, R. (2011), "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *Journal of Econometrics*, 165, 70-86.

[30] Pesaran, M. H. and Smith, R. J. (1994), "A Generalized R^2 Criterion for Regression Models Estimated by the Instrumental Variables Method," *Econometrica*, 62, 705-710.

[31] Pesaran, M. H. and Taylor, L. W. (1999), " Diagnostics for IV Regressions," *Oxford Bulletin of Economics and Statistics*, 61, 255-281.

[32] Smith, R. J. (1992), "Non-nested Tests for Competing Models Estimated by Generalized Method of Moments," *Econometrica*, 60, 973-980.

[33] Smith, R. J. and Ramalho, J. S. (2002), "Generalized Empirical Likelihood Non-Nested Tests," *Journal of Econometrics*, 107, 99-125.

[34] Staiger, D. and Stock, J. H. (1997), "Instrumental Variables Regression with Weak Instruments, *Econometrica*, 65, 557-86.

[35] Stock, J. H. and Wright, J. H. (2000), "GMM With Weak Identification," *Econometrica*, 68, 1055-1096.

[36] Yuan, Z. and Yang, Y. (2005), "Combining Linear Regression Models: When and How?", *Journal of the American Statistical Association*, 100, 1202-1214.

[37] Zhang, X., Wan, A. T. K. and Zhou, S. Z. (2012), "Focused Information Criteria, Model Selection, and Model Averaging in a Tobit Model with Nonzero Threshold", *Journal of Business and Economic Statistics*, 30, 132-142.

A Proofs

A.1 Proof of Theorem 1

Consistency follows from

$$\widehat{\theta}(\omega) = \sum_{c\in C} \omega_c \widehat{\theta}_c \xrightarrow{p} \sum_{c\in C} \omega_c \theta = \theta, \tag{43}$$

because $\sum_{c\in C} \omega_c = 1$. The asymptotic distribution follows from the limiting law for $\sqrt{T}\left(\widehat{\theta}_c - \theta\right)$, noting that
$\sqrt{T}\left(\widehat{\theta}(\omega) - \theta\right) = \sum_{c\in C} \omega_c \sqrt{T}\left(\widehat{\theta}_c - \theta\right)$ where

$$\widehat{\theta}_c - \theta = \left(\overline{X}'\overline{Z}_c \left(\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c\right)^{-1} \overline{Z}_c'\overline{X}\right)^{-1} \left(\overline{X}'\overline{Z}_c \left(\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c\right)^{-1} \overline{Z}_c'u\right).$$

For a given $c \in C$,

$$\begin{aligned}
\sqrt{T}\left(\widehat{\theta}_c - \theta\right) &= \left(\overline{X}'\overline{Z}_c \left(\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c\right)^{-1} \overline{Z}_c'\overline{X}\right)^{-1} \overline{X}'\overline{Z}_c \left(\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c\right)^{-1} \sqrt{T}\overline{Z}_c'u && (44)\\
&= \left(\frac{\overline{X}'\overline{Z}_c}{T} \left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1} \frac{\overline{Z}_c'\overline{X}}{T}\right)^{-1} \frac{\overline{X}'\overline{Z}_c}{T} \left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1} \frac{\overline{Z}_c'u}{\sqrt{T}} && (45)\\
&= \left(\frac{\overline{X}'\overline{Z}_c}{T} \left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1} \frac{\overline{Z}_c'\overline{X}}{T}\right)^{-1} \frac{\overline{X}'\overline{Z}_c}{T} \left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1} \Lambda_c' \frac{\overline{Z}_{\overline{c}}'u}{\sqrt{T}} && (46)
\end{aligned}$$

with $\frac{\overline{Z}_{\overline{c}}'u}{\sqrt{T}}$ not indexed by c and asymptotically normal with zero mean and variance $\widetilde{Q}_{\overline{Z}_{\overline{c}},\overline{Z}_{\overline{c}}}$. Hence,

$$\sqrt{T}\left(\widehat{\theta}(\omega) - \theta\right) = \left[\sum_{c\in C} \omega_c \left(\frac{\overline{X}'\overline{Z}_c}{T} \left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1} \frac{\overline{Z}_c'\overline{X}}{T}\right)^{-1} \frac{\overline{X}'\overline{Z}_c}{T} \left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1} \Lambda_c'\right] \frac{\overline{Z}_{\overline{c}}'u}{\sqrt{T}} \tag{47}$$

is also asymptotically normal with zero mean and the asymptotic variance-covariance matrix is given by

$$V_\omega = \Gamma_1 \widetilde{Q}_{\overline{Z}_{\overline{c}},\overline{Z}_{\overline{c}}} \Gamma_2, \tag{48}$$

where

$$\Gamma_1 = \left(\sum_{c\in C}\omega_c\left(Q_{\overline{X},\overline{Z}_c}\widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c}Q_{\overline{Z}_c,\overline{X}}\right)^{-1}Q_{\overline{X},\overline{Z}_c}\widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c}\Lambda_c'\right) \tag{49}$$

$$\Gamma_2 = \left(\sum_{c\in C}\omega_c\Lambda_c\widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c}Q_{\overline{Z}_c,\overline{X}}\left(Q_{\overline{X},\overline{Z}_c}\widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c}Q_{\overline{Z}_c,\overline{X}}\right)^{-1}\right) \tag{50}$$

such that

$$V_\omega = \left(\sum_{c\in C}\omega_c V_c Q_{\overline{X},\overline{Z}_c}\widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c}\Lambda_c'\right)\widetilde{Q}_{\overline{Z}_{\overline{c}},\overline{Z}_{\overline{c}}}\left(\sum_{c\in C}\omega_c\Lambda_c\widetilde{Q}^{-1}_{\overline{Z}_c,\overline{Z}_c}Q_{\overline{Z}_c,\overline{X}}V_c\right) \tag{51}$$

QED.

A.2 Proof of Theorem 2

Proof: Consistency follows from Slutsky as

$$\widehat{\theta}\left(\widehat{\omega}(ISC)\right) = \sum_{c\in C}\widehat{\omega}_c(ISC)\widehat{\theta}_c \xrightarrow{p} \sum_{c\in C}\omega_c(ISC)\theta = \theta, \tag{52}$$

because $\sum_{c\in C}\omega_c(ISC) = 1$. The asymptotic distribution follows from

$$\sqrt{T}\left(\widehat{\theta}\left(\widehat{\omega}(ISC)\right) - \theta\right) \tag{53}$$

$$= \left[\sum_{c\in C}\widehat{\omega}_c(ISC)\left(\frac{\overline{X}'\overline{Z}_c}{T}\left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1}\frac{\overline{Z}_c'\overline{X}}{T}\right)^{-1}\frac{\overline{X}'\overline{Z}_c}{T}\left(\frac{\overline{Z}_c'\widehat{\Sigma}_{u_c}\overline{Z}_c}{T}\right)^{-1}\Lambda_c'\right]\frac{\overline{Z}_{\overline{c}}'u}{\sqrt{T}}. \tag{54}$$

QED.