Journal of Time Series Econometrics

Volume 1, *Issue* 1 2008 *Article* 4

Selecting Instrumental Variables in a Data Rich Environment

Serena Ng* Jushan Bai†

*Columbia University, serena.ng@columbia.edu

†New York University, jushan.bai@nyu.edu

Selecting Instrumental Variables in a Data Rich Environment*

Serena Ng and Jushan Bai

Abstract

Practitioners often have at their disposal a large number of instruments that are weakly exogenous for the parameter of interest. However, not every instrument has the same predictive power for the endogenous variable, and using too many instruments can induce bias. We consider two ways of handling these problems. The first is to form principal components from the observed instruments, and the second is to reduce the number of instruments by subset variable selection. For the latter, we consider boosting, a method that does not require an a priori ordering of the instruments. We also suggest a way to pre-order the instruments and then screen the instruments using the goodness of fit of the first stage regression and information criteria. We find that the principal components are often better instruments than the observed data except when the number of relevant instruments is small. While no single method dominates, a hard-thresholding method based on the *t* test generally yields estimates with small biases and small root-mean-squared errors.

KEYWORDS: relevant instruments, principal components, information criteria, hard-thresholding, boosting, factor model, panel model

---

*This paper was presented at Columbia, Duke, Michigan, Queen's, Yale, UCSD, UCR, Institute of Statistics at Universite Catholique de Louvain, and SETA in Hong Kong. We thank seminar participants for many helpful comments and suggestions. We also acknowledge financial support from the NSF (SES-0551275 and SES-0549978).

1 Introduction

Instrumental variables are widely used in empirical analysis. A good deal of attention has been paid to the situation when a small number of instrumental variables are weakly correlated with the endogenous regressors. An equally practical problem is that many variables in the economic system can be weakly exogenous for the parameters of interest. Even without taking into account that lags and functions of predetermined and exogenous variables are also valid instruments, the number of instruments that practitioners have at their disposal can be quite large. This paper considers the selection of instrumental variables in a 'data rich environment'.

The problem is of empirical relevance for at least two reasons. First, the weak instrument problem may be a consequence of not being able to pick out those observed instruments most relevant for explaining the endogenous regressors. Second, it is well known that the bias of instrumental variable estimators increases with the number of instruments. Thus, irrespective of whether the available instruments are strong or weak, a smaller instrument set may be desirable on bias grounds. The question is how to determine this set.

We evaluate three procedures for forming smaller instrument sets from two larger sets of feasible instruments. These two feasible sets are (i) a set of N observed variables that are weakly exogenous for the parameter of interest, and (ii) the ordered principal components of the N variables. The three selection procedures are (i) boosting, (ii) ranking the predictive ability of the instruments one at a time, and (iii) information criteria applied to the ordered instruments. Throughout, N is assumed to be large, at least 50.

The method of principal components has been used by Kloek and Mennes (1960) to reduce the dimension of the instrument set. These authors were concerned with situations when N is large relative to the given T (in their case, T=30) so that the first stage estimation is inefficient. We are concerned with a data rich environment when N may be smaller or larger than T, but is large enough that a complete evaluation of the 2^N possible instrument sets is computationally burdensome, and that there is no natural ordering of the data that can simplify the model selection process. Amemiya (1966) showed that the method of principal components can be justified from a decision-theoretic basis.

The use of principal components can be motivated from a different perspective. As Connor and Korajzcyk (1986) showed, the first r principal components are consistent estimates of the r common factors underlying the population covariance of the data, if a factor structure indeed exists. Bai and

Ng (2006) showed that when the factors driving the endogenous regressors and the instruments are common, the principal components will estimate the space spanned by the common factors, which in this case are also the ideal instruments. Therefore, when the data admit a factor structure, using the principal components will be more efficient than using the same number of observed instruments directly. However, the factor structure is not necessary to validate the use of principal components as instruments.

Instead of using subsets of the data orthogonalized by the method of principal components, the instrument set can also be made smaller by eliminating the unimportant variables. Amemiya (1966) conjectured that this method may be even better than principal components if it is based on knowledge of the correlation between the endogenous regressors and the instruments. However, few procedures exist to select these instruments when the set of valid instruments is large. We consider two possibilities:- information criteria and hard-thresholding, both conditional on an ordering of the available instruments. We also consider 'boosting' as an instrumental variables selection device. Boosting is a statistical procedure that performs subset variable selection and shrinkage simultaneously to improve prediction. It is a topic that has drawn a good deal of interest in statistics in recent years. It has primarily been used in biostatistics and machine learning analysis as a classification and model fitting device. Consideration of boosting in econometric analysis and especially as a method for selecting instrumental variables appears to be new and is thus of interest in its own right.

The rest of the paper proceeds by first presenting the estimation framework in Section 2. The model selection procedures are discussed in Section 3. Simulation results are presented in Section 4, and an application is given in Section 5. Section 6 concludes.

2 The Econometric Framework

For $t = 1, \ldots T$, let the endogenous variable y_t be a function of a $K \times 1$ vector of regressors x_t:

$$\begin{aligned} y_t &= x_{1t}'\beta_1 + x_{2t}'\beta_2 + \varepsilon_t \\ &= x_t'\beta + \varepsilon_t \end{aligned} \tag{1}$$

The parameter vector of interest is $\beta = (\beta_1', \beta_2')'$ and corresponds to the coefficients on the regressors $x_t = (x_{1t}', x_{2t}')'$, where the exogenous and predetermined regressors are collected into a $K_1 \times 1$ vector x_{1t}, which may include lags of y_t. The $K_2 \times 1$ vector x_{2t} is endogenous in the sense that $E(x_{2t}\varepsilon_t) \neq 0$ and the least squares estimator suffers from endogeneity bias.

DOI: 10.2202/1941-1928.1014

We will be concerned with the properties of the standard GMM estimator. Let Q_t be a vector of instruments, and denote $\varepsilon_t(\beta) = y_t - x_t'\beta$. Given a vector of L moments $g_t(\beta) = Q_t\varepsilon_t(\beta)$, and a $L \times L$ positive definite weighting matrix W_T, the linear GMM estimator is defined as

$$\begin{aligned}\breve{\beta}_{QIV} &= \underset{\beta}{\text{argmin}}\ \bar{g}(\beta)'W_T\bar{g}(\beta) \\ &= (S_{Qx}'W_TS_{Qx})^{-1}S_{Qx}'W_TS_{Qy}\end{aligned}$$

where $\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^T g_t(\beta)$, $S_{Qx} = \frac{1}{T}\sum_{t=1}^T Q_t x_t'$, and S_{Qy} is similarly defined. Let $\breve{\varepsilon}_t = y_t - x_t'\breve{\beta}_{QIV}$ and let $\breve{S} = \frac{1}{T}\sum_{t=1}^T Q_tQ_t'\breve{\varepsilon}_t^2$, where $\breve{\beta}_{QIV}$ is a preliminary consistent estimator for β. Letting $W_T = \breve{S}^{-1}$ gives the efficient GMM estimator

$$\widehat{\beta}_{QIV} = (S_{Qx}'\breve{S}^{-1}S_{Qx})^{-1}S_{Qx}'\breve{S}^{-1}S_{Qy}.$$

This is the estimator whose properties will be evaluated for different choices of Q_t. Other estimators Properties of LIML and JIVE will also be considered.

We assume that there is a panel of valid instruments, $Z_t = (Z_{1t}, \dots Z_{Nt})'$ that are weakly exogenous for β. Here, Z_t can include lags of the endogenous regressors, lags and functions (such as the square) of other predetermined variables. Let $Z = (Z_1, Z_2, ..., Z_T)$ be a $N \times T$ matrix. As many variables in a large simultaneous system can be weakly exogenous for the parameter of interest, N can be large.

The conventional approach is to take some $z_t \subseteq Z_t$ to form a GMM estimator with $Q_t = z_t$,

$$\widehat{\beta}_{ZIV} = (S_{zx}'\breve{S}^{-1}S_{zx})^{-1}S_{zx}'\breve{S}^{-1}S_{zx}.$$

If Z are valid instruments, linear combinations of Z_t are also valid instruments. Thus if $\widetilde{F}$ are the principal components of the $N \times T$ matrix Z, $\widetilde{F}$ is a matrix of valid instruments. Let $\widetilde{f}_t$ be a subset of $\widetilde{F}_t$. For example, $\widetilde{f}_t$ may be the first few principal components of Z. Estimation using the orthogonalized instruments $\widetilde{f}_t$ yields

$$\widehat{\beta}_{FIV} = (S_{\widetilde{f}x}'\breve{S}^{-1}S_{\widetilde{f}x})^{-1}S_{\widetilde{f}x}'\breve{S}^{-1}S_{\widetilde{f}x}.$$

How to choose $\widetilde{f}_t$ from $\widetilde{F}_t$ will be discussed below.

We consider two data generating processes, one assuming that the endogenous variables and the instruments are correlated through common factors, and one without being specific about this.

DGP 1:

$$x_{2t} = \sum_{i=1}^{N} \pi_i Z_{it} + u_{it} \tag{2}$$

where $E(Z_{it}\varepsilon_t) = 0$ and $E(Z_{it}u_{it}) = 0$, $\pi' = (\pi_1, ..., \pi_N)$, $\pi'\pi > 0$ (positive definite matrix). It is possible to have $\pi_i = 0$ for some i. However, the condition $\pi'\pi > 0$ implies that a sufficient number of valid instruments (no smaller than K_2) are contained in Z_t.

DGP 2:

$$x_{2t} = \Psi' F_t + u_t \tag{3}$$
$$Z_{it} = \lambda_i' F_t + e_{it}. \tag{4}$$

where Ψ' is $K_2 \times r$, λ_i is a $r \times 1$ vector of loadings matrix, F_t is a $r \times 1$ vector, and r is a small number. Endogeneity arises when $E(F_t\varepsilon_t) = 0$ but $E(u_t\varepsilon_t) \neq 0$. This induces a non-zero correlation between x_{2t} and ε_t. We assume that $\Psi'\Psi > 0$ and Ψ does not shrink to zero as T increases so that the instruments F_t are 'strong'. One interpretation of the model is that F_t is a vector of r common factors. Then $\lambda_i' F_t$ is referred to as the common component of Z_{it}, e_{it} is an idiosyncratic error that is uncorrelated with x_{2t} and uncorrelated with ε_t. In economic analysis, F_t is a vector of unobservable common shocks that generate comovements amongst economic variables.[1] A theoretical motivation for this interpretation is given by Boivin and Giannoni (2006). Another interpretation is that Z_{it} and x_{2t} are repeated but contaminated measures of the ideal instrument, F_t.

Now viewed from either the factor model or the errors-in variables perspective, x_{2t} is just K_2 of the many other variables in the economic system that contain information about F_t. If F_t were observed, $\beta = (\beta_1', \beta_2')'$ could be estimated by using F_t to instrument x_{2t}. But the ideal instrument vector F_t is not observed. In Bai and Ng (2006), we propose to use $\widetilde{F}_t^r$, the first r principal components of $Z'Z$, as instruments, where r is assumed to be known. Using a consistently estimated r will not affect the asymptotic results. Provided N and T tend to infinity, Theorem 1 of Bai and Ng (2006) showed that the FIV (factor instrumental variable estimator) is $\sqrt{T}$ consistent and asymptotically normal. Estimation and inference can proceed as though the ideal instruments were observed. We also showed that if the FIV uses r factors as instruments, it will be more efficient than a ZIV estimator that uses r observed

[1] Some factor loadings can be zero so that the corresponding series are not influenced by the common shocks.

DOI: 10.2202/1941-1928.1014

instruments. While this efficiency argument hinges on a factor structure underlying the data, consistency of the estimator using the principal components as instruments does not.

Whether we use Z_t or $\widetilde{F}_t$ as instruments, and whether or not the data have a factor structure, the factors that best explain the variation in Z may not be the best in terms of explaining the endogenous variables, x_2. For the purpose of selecting instruments, we will simply refer to the instruments as Q_t, which can be $\widetilde{F}_t$ or Z_t.

There are two motivations for considering procedures that select instruments. First, if $Q_t = Z_t$, few would expect every element of Z_t to have the same predictive power for x_{2t}, even though they are all valid. It may also seem that if $Q_t = \widetilde{F}_t$, we can exploit the fact that the first principal component $\widetilde{F}_{1t}$ explains the most variance in Z_{it}, $\widetilde{F}_{rt}$ explains the most variance not explained by $(\widetilde{F}_{1t}, \ldots, \widetilde{F}_{r-1.t})$ and so on, with $\widetilde{F}_{jt}$ orthogonal to $\widetilde{F}_{kt}$ for $j \neq k$. However, of interest is not what explains the panel of instruments Z_{it} per se, but what explains x_{2t}. Factors that have strong explanatory power for Z_t need not be good instruments for x_{2t}. Second, Q_t is a matrix with N columns, and the bias of $\widehat{\beta}_{GMM}$ is known to increase with N. Both considerations motivate forming a set of L instruments with $K_2 \leq L \leq N$ by removing those elements of Q_t that have weak predictive ability for x_{2t}.

We call those $Q_{jt} \subset Q_t$ with small explanatory power for x_{2t} the 'relatively weak' instruments, defined in Hahn and Kuersteiner (2002) as instruments with Ψ_j in the $T^{-\delta}$ neighborhood of zero and $\delta < \frac{1}{2}$. They showed that standard asymptotic results hold when such instruments are used to perform two-stage least squares estimation. Let $q_t \subset Q_t$ be a $L \times 1$ vector of instruments that remain after the relatively weak instruments are removed from Q, where $Q = (Q_1, ..., Q_T)'$ is a $T \times N$ matrix. The q_t will be referred to as the 'relevant instruments'. We do not allow for weak instruments in the sense of Staiger and Stock (1997). This is considered analytically in Kapetanios and Marcellino (2006) within a factor framework. We are interested in comparing the properties of $\widehat{\beta}_{QIV}$ with an estimator that uses q_t as instruments:

$$\widehat{\beta}_{qIV} = (S'_{qx}\breve{S}^{-1}S_{qx})^{-1}S'_{qx}\breve{S}^{-1}S_{qy}.$$

The estimator $\widehat{\beta}_{QIV}$ is a special case of $\widehat{\beta}_{qIV}$ with $L = N$. Both estimators are consistent and asymptotically normal under the assumption that N is fixed. It can also be shown that $\widehat{\beta}_{QIV}$ has a smaller variance, but $\widehat{\beta}_{qIV}$ has smaller bias. There remains the question of how to form q from Q.

Several instrumental variable selection procedures have been considered in the literature. In the case of simultaneous equations, Hall and Peixe (2000)

suggest to first 'structurally order' the instruments as in Fisher (1965), and then determine which instruments satisfy the required orthogonality conditions. Andrews and Lu (2001) propose to use information type criteria to select the number of moment conditions in a GMM setting, which amounts to selecting instruments that satisfy the specified orthogonality conditions. Hall and Peixe (2003) consider a 'canonical correlation information criterion' to select relevant instruments from amongst those that satisfy the orthogonality condition. Our setup is somewhat different as all instruments are valid and thus satisfy the orthogonality condition.

Others have proposed procedures to select q with the goal of improving the properties of $\widehat{\beta}$. Assuming that the instruments can be ordered, Donald and Newey (2001) propose a selection method to minimize the mean-squared error of $\widehat{\beta}$. Carrasco (2006) proposes to replace S_{zz} by a matrix formed from $L \leq N$ of the eigenvectors of S_{zz} whose corresponding eigenvalues exceed a threshold. The method is a form of regularization and is also referred to in the statistics literature as the method of 'supervised principal components', see Bair et al. (2006). The procedures proposed by Donald-Newey and Carrasco amount to minimizing a C_p type criterion. Hall et al. (2007) consider an entropy-based information criterion to select moments for GMM estimation that are both valid and informative. Their criterion is based on the variance of the estimated parameters instead of the fitted regression. Using Monte Carlo simulations, Eryuruk et al. (2008) find that the method of Donald and Newey (2001), Hall and Peixe (2003), and Hall et al. (2007) have similar median bias and finite sample coverage probability.

A maintained assumption in studies that aim to minimize the mean-squared error of $\widehat{\beta}$ is that N is either fixed, and/or the instruments can be ordered. Otherwise, one would need to evaluate 2^N models, which can be computationally prohibitive. In what follows, we will be concerned with forming q_t with the objective of fitting x_{2t} in the presence of x_{1t} when the set of feasible instruments is *large* and that there is *no* natural ordering to these instruments. This amounts to selecting relevant instruments using the first stage regression. This is a more modest task, but as Kapatanios (2006) pointed out, the problem of minimizing the mean-squared error of $\widehat{\beta}_{qIV}$ when N is large is non-standard since the space over which minimization occurs is discrete. He proposes to use combinatorial algorithms to minimize the approximate mean-squared functions over discrete domains. More delicate is the problem that consistency of model selection procedures (such as the C_p) is proved under the assumption that the number of models to be considered is finite. As discussed in Hansen (2008), the chance of selecting a model that overfits increases with the number

DOI: 10.2202/1941-1928.1014

of model considered.

3 Determining the Relevant Instruments

We consider two methods: boosting, and a method that is based on first ordering of the instruments by their relevance.

3.1 Boosting

Boosting was initially introduced by Freund (1995) and Schapire (1990) to the machine learning literature as a classification device. It is recognized for its ability to find predictors that improve prediction without overfitting and for its low mis-classification error rate. Bai and Ng (2008) applied the boosting method to forecasting in the presence of large number of predictors. An introduction to boosting from a statistical perspective can be found in Buhlmann and Hothorn (2006). For our purpose, it is best to think of boosting as a procedure that performs model selection and coefficient shrinkage simultaneously. This means that variables not selected are set to zero, as opposed to being shrunk to zero. In consequence, boosting results in very sparse models.

The specific L_2 boost algorithm we use is based on component-wise least squares. Component-wise boosting was considered by Friedman (2001) and Buhlmann and Yu (2003). Instead of evaluating sets of regressors one set at a time, the regressors are evaluated one at a time. Under component-wise boosting the jth instrument (j is arbitrary) is as likely to be chosen as the first; the ordering does not matter.

We are interested in finding instruments that can explain x_{2t} after controlling for x_{1t}. To this end, let $\widetilde{x}_2 = M_1 x_2$ and $\widetilde{Q} = M_1 Q$, where $M_1 = I - x_1(x_1'x_1)^{-1}x_1$. Thus, $\widetilde{x}_{2t}$ and $\widetilde{Q}_t$ are the residuals from regressing x_{2t} and Q_t, respectively, on x_{1t}. Selecting the best subset q_t out of Q_t after controlling for x_{1t} is then the same as extracting a $\widetilde{q}_t$ from $\widetilde{Q}_t$ that has the most explanatory power for $\widetilde{x}_{2t}$.

Boosting is a multiple-stage procedure. Let $\widehat{\Phi}_m$ be a $T \times 1$ vector with t-th component $\widehat{\Phi}_{t,m}$ $(t = 1, 2, ..., T)$. Define $\widehat{\phi}_m$ similarly. Let $\overline{\widetilde{x}}_2$ denote the sample mean of $\widetilde{x}_{2t}$. The boosting algorithm for fitting the conditional mean of $\widetilde{x}_{2t}$ given a set of N predictors, $\widetilde{Q}_t$, is as follows:

1 Initialize $m = 0$ and let $\widehat{\Phi}_{t,0} = \widehat{\phi}_{t,0} = \overline{\widetilde{x}}_2, \quad (t = 1, 2, ..., T)$;

2 for $m = 1, \ldots M$

  a for $t = 1, \ldots T$, let $u_t = \widetilde{x}_{2t} - \widehat{\phi}_{t,m-1}$ be the 'current residuals';

b for each $i = 1, \dots N$, regress the current residual vector u on $\widetilde{Q}_{.,i}$ (the i-th regressor) to obtain $\widehat{b}_i$. Compute $\widehat{e}_{.,i} = u - \widetilde{Q}_{.,i}\widehat{b}_i$ as well as $SSR_i = \widehat{e}'_{.,i}\widehat{e}_{.,i}$;

c let i_m be such that $SSR_{i_m} = \min_{i\in[1,\dots N]} SSR_i$;

d let $\widehat{\phi}_m = \widetilde{Q}_{.,i_m}\widehat{b}_{i_m}$;

e for $t = 1, \dots T$, update $\widehat{\Phi}_{t,m} = \widehat{\Phi}_{t,m-1} + \nu\widehat{\phi}_{t,m}$, where $0 \leq \nu \leq 1$ is the step length.

Component-wise L_2 boost may seem mysterious as it is relatively new to economists, but it is nothing more than repeatedly fitting least squares to the current residuals and selecting at each step the predictor that minimizes the sum of square residuals. Note that component wise L_2 boost selects one predictor at each iteration, but the same predictor can be selected more than once during the M iterations. This means that boosting makes many small adjustments, rather than accepting the predictor once and for all. This seems to play a role in the ability of boosting not to overfit.

After m steps, boosting produces $\widehat{\Phi}_m(\widetilde{Q}) = \overline{\widetilde{x}}_2\,\iota_T + \widetilde{Q}\widehat{\delta}_m$ as an estimate of the conditional mean, where $\widehat{\delta}_m$ is a vector with potentially many zeros, and ι_T is a vector of 1's. Zhang and Yu (2005) and Buhlmann (2006) showed that boosting will consistently estimate the true conditional mean of x_2. The estimator $\widehat{\delta}_m$ can be shown to follow the recursion ($\widehat{\delta}_0 = 0$)

$$\widehat{\delta}_m = \widehat{\delta}_{m-1} + \nu\widehat{b}^{\dagger}_m,$$

where $\widehat{b}^{\dagger}_m$ is an $N \times 1$ vector of zeros, except that its i_mth element equals $\widehat{b}_{i_m}$. Thus, $\widehat{\delta}_m$ and $\widehat{\delta}_{m-1}$ differ only in the i_m-th position. If $\widetilde{x}_2$ is $T \times K_2$, the algorithm needs to be repeated K_2 times. The boosting estimate of the conditional mean can also be rewritten as $\widehat{\Phi}_m(\widetilde{Q}) = B_m\widetilde{x}_2$, where

$$\begin{aligned} B_m &= B_{m-1} + \nu P^{(m)}_{\widetilde{Q}}(I_T - B_{m-1}) \\ &= I_T - (I_T - \nu P^{(0)}_{\widetilde{Q}})(I - \nu P^{(1)}_{\widetilde{Q}}) \cdots (I_T - \nu P^{(m)}_{\widetilde{Q}}) \\ &= I_T - \prod_{j=0}^{m}(I_T - \nu P^{(j)}_{\widetilde{Q}}) \end{aligned} \tag{5}$$

where, for $m \geq 1$, $P^{(m)}_{\widetilde{Q}} = \widetilde{Q}_{.,i_m}(\widetilde{Q}'_{.,i_m}\widetilde{Q}_{.,i_m})^{-1}\widetilde{Q}'_{.,i_m}$ is the projection matrix based upon the regressor that is selected at the m-th step, and $P^{(0)} = \frac{1}{\nu}\iota_T\iota'_T/T$ with ι_T being a $T \times 1$ vector of 1s. This implies that $B_0 = \iota_T\iota'_T/T$ so that $B_0\widetilde{x}_2 = \overline{\widetilde{x}}_2\iota_T$, which is the value used to initialize the boosting procedure. A

DOI: 10.2202/1941-1928.1014

distinctive feature of boosting is that it will produce a sparse solution when the underlying model structure is sparse. In our context, a sparse structure occurs when there are many relatively irrelevant instruments and Ψ has small values in possibly many positions.

If the number of boosting iterations (M) tends to infinity, we will eventually have a saturated model in which case all predictors are used. The sparseness of $\widehat{\delta}_M$ is possible only if boosting is stopped at an 'appropriate' time. In the literature, M is known as the stopping rule. It is often determined by cross-validation in situations when a researcher has access to training samples. But this is often not the case in time series economic applications. Let

$$df_m = trace(B_m).$$

We follow Buhlmann (2006) and let $M = \text{argmin}_{m=1,...\bar{M}} IC(m)$ where

$$IC(m) = \log(\widehat{\sigma}_m^2) + \frac{A_T \cdot df_m}{T}$$

and $\widehat{\sigma}_m^2 = T^{-1}SSR_{i_m}$. The BIC obtains when $A_T = \log(T)$, and AIC obtains when $A_T = 2$. The primary departure from the standard AIC/BIC is that the complexity of the model is measured by the degrees of freedom, rather than by the number of predictors. In our experience, the degrees of freedom in a model with k predictors tends to be higher than k.

Under boosting, $\widehat{\delta}_M$ is expected to be sparse if the number of truly relevant instruments is small. The sparseness of $\widehat{\delta}_M$ is a feature also shared by the LASSO estimator of Tibshirani (1996), defined as

$$\widehat{\delta}_L = \underset{\delta}{\text{argmin}} \left\| \widetilde{x}_2 - \widetilde{Q}\widetilde{\delta} \right\|^2 + \lambda \sum_{j=1}^{N} |\delta_j|.$$

That is, LASSO estimates δ subject to a L_1 penalty. Instrumental variable selection using a ridge penalty has been suggested by Okui (2004) to solve the 'many instrument variable' problem. The ridge estimator differs from LASSO only in that the former replaces the L_1 by an L_2 penalty. Because of the nature of L_2 penalty, the coefficients can only be shrunk towards zero but will not usually be set to zero exactly. As shown in Tibshirani (1996), the L_1 penalty performs both subset variable selection and coefficient shrinkage simultaneously. Efron et al. (2004) showed that certain forward stagewise regressions can produce a solution path very similar to LASSO, and boosting is one such forward-stagewise regression. We consider boosting because the exact LASSO solution is numerically more difficult to solve, and that boosting has been found to have better properties than LASSO in the statistics literature.

3.2 Ranking the Instruments

Conventional variable selection procedures necessitate an evaluation of 2^N possible models if there are N potential instruments. This can be computationally costly if N is large and the instruments do not have a natural ordering. It is, however, useful to compare boosting with alternative variable selection methods. To reduce the number of model evaluations, we use the marginal predictive power of the instruments for the endogenous regressor to rank the instruments. More precisely, for each $i = 1, \dots N$, we consider least squares estimation of

$$x_{2t} = \gamma_0 + x_{1t}'\gamma_1 + \gamma_{2i}Q_{it} + error.$$

This is simply a first stage regression, the same regression that underlies the first-stage F test in determining if an observed instrument is a weak. We use the t statistic for $\widehat{\gamma}_2$ to rank the relative importance of the instruments, though the R^2 can equivalently be used. The ranking yields $\{Q_{[1]t}, \dots Q_{[N]t}\}$, the ordered instrument set. Again, $Q_{[i]t}$ can be $Z_{[i]t}$ or $\widetilde{F}_{[i]_t}$. Given the ordered instruments, the simplest subset-variable selection procedure is to form

$$q_t = \{Q_{[i]t} : |t_{[i],\widehat{\gamma}_2}| > c\}$$

where $t_{[i],\widehat{\gamma}_2}$ is the t-statistic and c is a threshold value. This is essentially a hard thresholding method that keeps as instruments all those variables whose t statistic for $\widehat{\gamma}_2$ exceeds c.

In addition to the t test, we also apply the information criterion to the ordered instrument set. Specifically, We first rank Q_t using the absolute value of the t test. Given $Q_{[1]t}, Q_{[2]t}, \dots Q_{[N]t}$, the set of ranked instruments, we can now obtain $\widehat{\sigma}_l^2 = T^{-1}\sum_{t-1}^{T}\widehat{e}_{tl}^2$, where $\widehat{e}_{tl}$ is the residual from regressing x_{2t} on x_{1t} and the first l of the ordered instruments. Let

$$L = \underset{n}{\operatorname{argmin}} \log(\widehat{\sigma}_n^2) + nA_T/T.$$

The selected instrument set is then

$$q_t = \{Q_{[1]t}, Q_{[2]t}, \dots Q_{[L]t}\}.$$

We use the BIC with $A_T = log(T)$, which is known to select more parsimonious models.

4 Finite Sample Properties

In this section, we study the finite sample properties of the GMM estimator when different methods are used to select the instruments. Estimation proceeds as follows:

DOI: 10.2202/1941-1928.1014

1. Form the potential instrument set, Q_t, which can be $\widetilde{F}_t$ or Z_t;
2. for boosting, partial out the effect of x_{1t} from both x_{2t} and Q_t to yield $\widetilde{x}_{2t}$ and $\widetilde{Q}_t$;
3. use boosting, the t test, or the BIC to determine $l_1, \dots l_L$, the positions in Q_t to keep. Let $q_t = (Q_{tl_1}, \dots Q_{tl_L})$;
4. perform GMM estimation with $q^+ = [x_1 \quad q]$ as instruments and an identity weighting matrix to yield $\breve{\beta}$. Let $\breve{S} = \frac{1}{T}\sum_{t=1}^{T} \breve{\varepsilon}_t^2 q_t^+ q_t^{+\prime}$, where $\breve{\varepsilon}_t = y_t - x_t'\breve{\beta}$;
5. re-do GMM one more time using $W = \breve{S}^{-1}$ to yield $\widehat{\beta}_{qIV}$.

We consider GMM estimation with six sets of instruments:

- FIV_b: $q_t = \widetilde{f}_t \subset (\widetilde{F}_{t,1} \dots \widetilde{F}_{t,N})'$ chosen by boosting
- FIV_t: $q_t = \widetilde{f}_t \subset (\widetilde{F}_{t,1} \dots \widetilde{F}_{t,N})'$ such that $|t_{\widehat{\gamma}_2}| > c$;
- FIV_{ic}: $q_t = \widetilde{f}_t \subset (\widetilde{F}_{t,1} \dots \widetilde{F}_{t,lmax})$ chosen using the BIC;
- IV_b: $q_t = z_t \subset (Z_{t,1}, \dots Z_{t,N})'$ chosen by boosting;
- IV_t: $q_t = z_t \subset (Z_{t,1}, \dots Z_{t,N})'$ chosen such that $|t_{\widehat{\gamma}_2}| > c$;
- IV_{ic} : $q_t = z_t \subset (Z_{t,1}, \dots Z_{t,N})'$ chosen using the BIC.

We set the maximum number of boosting iterations M at $\bar{M} = \min[N^{1/3}, T^{1/3}]$. We follow the literature and set ν to 0.1. The threshold for the t test is set at 2.5, which is slightly above the critical value for the two-tailed one percent level. When the number of t statistics passing the threshold of 2.5 exceeds 20, we take the top twenty variables. An upper bound of 20 variables is also put on the BIC.

We consider five data generating processes in the simulations:

DGP 1. The model is adapted from Moreira (2003):

$$\begin{aligned} y_t &= x_{1t}\beta_1 + x_{2t}\beta_2 + \sigma_y \varepsilon_t \\ x_{i1t} &= \alpha_x x_{i1,t-1} + e_{ix1t}, \quad i = 1, \dots K_1 \\ x_{i2t} &= \lambda_{i2}' F_t + e_{ix2t}, \quad i = 1, \dots K_2 \\ z_{it} &= \lambda_{iz}' F_t + \sigma_{zi} e_{it}, \quad i = 1, 2, ..., N \\ F_{jt} &= \rho_j F_{jt-1} + \eta_{jt}, \quad j = 1, \dots r \end{aligned}$$

where $\varepsilon_t = \frac{1}{\sqrt{2}}(\widetilde{\varepsilon}_t^2 - 1)$ and $e_{ix2t} = \sqrt{r}\sigma_{x2}\frac{1}{\sqrt{2}}(\widetilde{e}_{ix2t}^2 - 1)$ with $\widetilde{\varepsilon}_t$ and $\widetilde{e}_{ix2t}$ to be defined later; $\alpha_x \sim U(.2, .8)$, $e_{ix1t} \sim N(0,1)$, and both α_x and e_{ix1t} are

uncorrelated with $\widetilde{e}_{jx2t}$ and $\widetilde{\varepsilon}_t$. Furthermore, $e_{it} \sim N(0,1)$, $\eta_{jt} \sim N(0,1)$, $\lambda_{iz} \sim N(0,1)$, and $\rho_j \sim U(.2,.8)$. Finally, $(\widetilde{\varepsilon}_t, \widetilde{e}'_{x2t}) \sim N(0_{K_2+1}, \Sigma)$ where $diag(\Sigma) = 1$,$\Sigma(\ j,1) = \Sigma(1\ ,j) \sim U(.3,.6)$, and zero elsewhere. This means that $\widetilde{\varepsilon}_t$ is correlated with $\widetilde{e}_{ix2t}$ with covariance$\Sigma(1\ ,i)$ but $\widetilde{e}_{ix2t}$ and $\widetilde{e}_{jx2t}$ are uncorrelated $(i \neq j)$. By construction, the errors are heteroskedastic. The parameter σ_y^2 is set to $K_1\bar{\sigma}_{x1}^2 + K_2\bar{\sigma}_{x2}^2$ where $\bar{\sigma}_{xj}^2$ is the average variance of x_{jt}, $j = 1, 2$. This puts the noise-to-signal ratio in the primary equation of roughly one-half. We considered various values of K_2, σ_{x2}, σ_z, and r. The parameter of interest is β_2 whose true value is 2. The results are reported in Table 1 with $K_1 = K_2 = 1$.

DGP 2, 3. The model is based on Carrasco (2006). Let

$$\begin{aligned} y_t &= \beta_2 x_{2t} + \varepsilon_t \\ x_{2t} &= z_t'\pi + u_t \end{aligned}$$

$$\begin{pmatrix} \varepsilon_t \\ u_t \end{pmatrix} \ iid\ N(0,\Sigma), \quad \Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$$

where $z_t \sim N(0, I_N)$, $R_{x2}^2 = \frac{\pi'\pi}{1+\pi'\pi} = \{.9, .75, .5\}$. Two cases are considered

DGP 2: $\pi_j = d(1 - \frac{.5j}{L+1})^4$, d is chosen such that $\pi'\pi = \frac{R_{x2}^2}{1-R_{x2}^2}$.

DGP 3: $\pi_j = (\frac{R_{x2}^2}{N(1-R_{x2}^2)})^{1/2}$.

In both cases, $\beta_2 = 1.0$. Under DGP 2, the instruments are in decreasing order of importance. The smaller is R_{x2}^2, the less predictable is x_2. When $R_{x2}^2 = .9$ and $N = 100$, π ranges from .628 to .047, with about 12% of the π coefficients bigger than .5, and 30% of the π coefficients between .25 and .5. When $R_{x2}^2 = .5$, the coefficients range from .209 to .0139. While none of the coefficients exceed .25, 34% exceed .1. Under DGP 3, the instruments are of equal importance and there is no reason to choose one instrument over another. The π coefficients are .3, .1732, and .1 for $R_{x2}^2 = .9, .75, .5$, respectively. Thus when $R_{x2}^2 = .5$, DGP 2 simulates a case of few relevant instruments while DGP 3 simulates a case of many relatively weak instruments. Arguably, when R_{x2}^2 falls below .5, DGPs 2 and 3 do not constitute a data rich environment as the number of relevant instruments is in fact quite small.

DGPs 4, 5. An undesirable feature of DGPs 2 and 3 is that the degree of endogeneity always increases as the predictability of the endogenous regressor

DOI: 10.2202/1941-1928.1014

decreases. In other words, the instruments are weaker when the endogeneity bias is more severe. We consider two other DGPs that allow the variance of ε_t to be a free parameter. Specifically, we let

$$\Sigma = \begin{pmatrix} \sigma_{11} & .25 \\ .25 & 1 \end{pmatrix}$$

where $\sigma_{11} = (2.5, 1.5, .5)$, holding R^2_{x2} fixed at 0.5. Under DGP 4, 20% of the πs exceed 0.5, while the π coefficients are now .1414 under DGP 5. It should be remarked that under DGPs 3 and 5, some relevant instruments will always be omitted whenever $L < N$, and such omission cannot be justified on a priori grounds.

4.1 Results

We compare the estimators using the mean estimate and the root-mean-squared error in 1000 replications. We also report the average endogeneity bias (ρ) as measured by the correlation between ε_t and x_2, the predictability of the endogenous variable as measured by the $1 - \frac{var(u_2)}{var(x_2)}$, and a measure of commonality in the instruments. In theory, if the data have r factors, the r largest eigenvalues of the population covariance matrix of Z should increase with N while the $r+1$-th eigenvalue is bounded. Therefore we also report $\frac{\mu_{r+1}}{\mu_r}$, the ratio of $(r+1)$th over the rth eigenvalues. The smaller is this quantity, the stronger is the factor structure.

The results for DGP 1 (the factor model) are reported in Table 1. The impact of endogeneity bias on OLS is immediately obvious. The six IV estimators all provide improvements. Of the six methods, boosting and the t test select more variables than the BIC. Under DGP 1, the instruments and the endogenous regressor share common factors. Thus, many observed instruments have strong correlations with x_2. Not surprisingly, the t test detected this. The number of instruments used by the t test is 20 in view of the upper bound. The number of observed instruments being selected is always higher than the number of principal components being selected, indicating that the principal components is effective in compressing information. The BIC tends to select the most parsimonious set of instruments.

The point estimates are closer to the true value when σ_z^2 is small. A higher σ_z^2 increases both the bias and RMSE. Put differently, the estimates are less precise when the common factors have weak explanatory power for the endogenous regressor. In results not reported, the estimates are also more precise when T increases but do not vary systematically with N. The t test on the principal components produces estimates that are closest to the true

value of 2. Although its RMSE is not always the smallest, no one method systematically dominates the other in terms of RMSE. However, using the principal components tends to yield smaller bias than the observed instruments when the instruments are selected by boosting or the t test.

We now turn to DGP 2. Unlike DGP 1, the data no longer have a factor structure. Thus, $\frac{\mu_{r+1}}{\mu_r}$ is much larger in DGPs 2 and 3 than in DGP 1. While $\widetilde{F}_t$ are still valid instruments under DGP 2 (and 3), these are no longer estimates of the ideal instruments. Evidently, all estimators provide significant improvements over OLS when R^2_{x2} (the strength of the instruments) is high. However, as R^2_{x2} falls, the bias in all the estimators increases and the estimates move toward OLS. By the construction of DGP 2, x_2 is predicted only by a few variables. If these variables can be isolated, they should be better instruments than principal components as these are linear combinations of the available instruments, some of which have no explanatory power. The results bear this out. The three methods used to select observed instruments are quite comparable both in terms of bias and RMSE. Of the three methods, the t test (with $c = 2.5$) tends to select the smallest instrument set.

Under DGP 3, all instruments have the same predictive power for x_2 and there is no priori reason to choose one over another. Indeed, given that π is a constant vector, an efficient instrument can be obtained by summing all N series. Such an option is not available, though the principal components should achieve this goal approximately. The simulations reveal that the estimates using the principal components are better than those using a subset of observed instruments. However, as the degree of endogeneity rises while the instruments get weaker, all methods are biased, even though the estimates are closer to the true value than OLS.

DGPs 4 and 5 allow the strength of the instruments to increase with the degree of endogeneity. Evidently, the estimates are much less biased than those reported in Table 2. However, as in DGP 2 when the endogenous regressor is a function of just a few variables, carefully selecting the observed variables as instruments dominates the principal components approach. And as in DGP 3 when all instruments have equal predictive power, the method of principal components achieves dimension reduction and is generally superior to using the observed variables as instruments. However, Table 3 reveals that even when the instruments are strong (with R^2_{x2} well above .66), the bias in all the estimates can still be quite substantial. Both the bias and RMSE are increasing in N, the number of instruments available, and L, the number of instruments chosen. This, however, is an artifact of the DGP which allows the number of explanatory variables for x_2 to increase with N.

DOI: 10.2202/1941-1928.1014

All three instrument selection methods involve pretesting, and it is of some interest to see to what extent this affects the asymptotic distribution of the estimators. Figure 1 presents the standardized distribution of $\widehat{\beta}_2$ for DGP 1, $\sigma_\varepsilon^2 = \sigma_z^2 = 1$. Notably, all distributions appear to be symmetric. The tails of the distribution suggest that there are some large outliers, but the normal distribution seems to be a reasonably good approximation to the finite sample distribution of the estimators. The issue with the IV estimation is not so much its shape, but where it is centered. From this perspective, the $\widehat{\beta}_2$ produced by using the t test to select the principal components (labeled FIV_t) appears to have the best properties.

Overall, using the observed data are better than the principal components only when the endogenous variables can be predicted by a small number of instruments. Or, in other words, when the data rich environment assumption is not appropriate. Otherwise, principal components compress the information in the available instruments to a smaller set of variables and is preferred on bias grounds. The three methods considered have comparable properties when T is large. For small T, the t test has better biased properties but sometimes has large RMSE. Which method one chooses depends on the preference of the practitioner. When (i) the degree of endogeneity is large, (ii) when the instruments are weak or when ideal instruments do not exist, the point estimates can be biased whichever method we use to select the instruments.

4.2 Other Estimators

The two stage least squares estimator is known to have poor finite sample properties, especially when the number of instruments is large. Newey and Smith (2004) showed using higher order asymptotic expansions that the bias of the GMM estimator is linear in $N - K_2$, which is the number of overidentifying restrictions. Phillips (1983) showed that the rate at which $\widehat{\beta}_{2SLS}$ approaches normality depends on the true value of β and N. Hahn and Hausman (2002) showed that the expected bias of 2SLS is linear in N, the number of instruments. The result arises because $E(x'P_Q\varepsilon/T) = E(u'P_Q\varepsilon/T) = \sigma_{u\varepsilon}N/T$. The question then arises as to whether the instrument selection procedures we considered is as effective when other estimators are used.

To gain some insight into this question, we also evaluate the finite sample properties of LIML and the estimator by Fuller, both considered in Donald and Newey (2001) and Hausman et al (2007), among others. These results are reported in Tables 4, 5, and 6 for Models 1, 2, and 3 respectively. Compared with results for the IV reported in Tables 1 to 3, LIML and FULLER have smaller biases, with FULLER being the most accurate. However, as far as

selecting instruments is concerned, the general picture holds up. Specifically, for Model 1, using the t test to select the instruments gives more precise estimates. For Models 2 and 3, all three methods (boosting, IC, and t test) have rather similar performances, with no particular method dominating the others. Comparing the LIML and FULLER results with those of the IV, the advantage of using principal components over the observed instruments is not as strong for Model 1, though for Models 2 and 3, forming principal components still seems desirable.

5 Application

Consider estimating the elasticity of intertemporal substitution, denoted, ψ. Let r_{t+1} be a measure of real asset returns, which for the purpose of illustration, we will use the real interest rate. As discussed in Yogo (2004), ψ can be estimated from

$$r_{t+1} = b + \frac{1}{\psi}\Delta c_{t+1} + e_{r,t+1}$$

by instrumenting Δc_{t+1}, where Δc_t is consumption growth. It can also be estimated from the reverse regression

$$\Delta c_{t+1} = a + \psi r_{t+1} + e_{c,t+1}$$

by instrumenting r_{t+1}. Whereas r_{t+1} is persistent and thus predictable, Δc_{t+1} tends to be difficult to predict. The first formulation is thus more susceptible to the weak instrument problem, especially when the number of instruments considered is small.

We take the data used in Yogo (2004), who used the nominal interest rate, inflation, consumption growth, and the log dividend price ratio lagged twice as instruments. His results will be labeled IV. Additionally, we use data collected in Ludvigson and Ng (2007), which consist of quarterly observations on a panel of 209 macroeconomic series for the sample 1960:1-2002:4, where some are monthly series aggregated to quarterly levels. Following Yogo (2004), we only use data from 1970:3-1998:4 for the analysis for a total of 115 observations. Principal components are estimated from the 209 series, all lagged twice.[2] Thus Z_t is of dimension 209+4, while $\widetilde{F}_t$ is of dimension $\min[N,T]$+4, since the four instruments used in Yogo (2004) are always included in the instrument set, and there are $min[N,T]$ non-zero eigenvalues. We can give some interpretation to the principal components by considering the marginal R^2 of each factor in each series. This is obtained by regressing each series on the principal

[2]The PCP criterion developed in Bai and Ng (2002) chooses 8 factors.

DOI: 10.2202/1941-1928.1014

components one at a time. This exercise reveals that the highest marginal R^2 associated with $\widetilde{F}_{1t}$ is UTIL1 (capacity utilization), $\widetilde{F}_{2t}$ is PUXHS (CPI excluding shelter), and $\widetilde{F}_{3t}$ is DLAGG (a composite index of seven leading indicators).

The results are reported in Tables 7a and 7b. The upper bound on the instruments considered by BIC is set at 20, and a threshold of 2.5 is used for the t test. Notably, it is much easier to predict real interest rate than consumption. Whichever is the endogenous regressor, there is predictability beyond the instruments used in Yogo (2004). The R^2_{x2} increases from .08 to as much as .375 in the regression for estimating $1/\psi$, and from .279 to .528 in the regression for estimating ψ.

For the first regression when the parameter of interest is $1/\psi$, $\widetilde{F}_{1t}, \widetilde{F}_{6t}$, and $\widetilde{F}_{7t}$ along with 13 other factors *and* lagged consumption were selected by boosting. The t test selects a principal component that corresponds to those of the largest ten eigenvalues, along with lagged consumption. The BIC selects 4 principal components along with lagged consumption growth. The first three variables selected by boosting are GNSQF, GYFIR, and LBCPU, though nominal interest rate was also significant. The top three observed variables selected by the t test are PUXM, GNET, and PWIMSA. Lagged consumption growth was also chosen. The BIC selects GDCD, along with lagged consumption growth used in Yogo (2004). However, none of the first stage R^2_{x2} exceeds 0.4.

For the second regression when the parameter of interest is ψ, $\widetilde{F}_{2t}, \widetilde{F}_{5t}, \widetilde{F}_{6t}$ and 8 other principal components are selected by boosting, along with inflation and lagged consumption growth. The t test chooses $\widetilde{F}_{2t}$, consumption growth and one other factor, while the BIC additionally selects seven other principal components. However, these two methods do not select the variables used in Yogo (2004). Along with the nominal interest and lagged inflation, boosting selects FMD, GFINO, and GYDPCQ as additional observed instruments. The t test selects all four variables used in Yogo (2004), along with BPB. The IC selects the nominal interest rate along with GDFSFC, PUCX, HS6FR, and LIPM. The picture is thus quite clear that information is available in predicting the endogenous regressor beyond what was used in Yogo (2004).

As far as the estimates go, using the principal components as instruments all come out much better determined than using the observed instruments. The point estimate of $1/\psi$ ranges from .470 to 1.453. But the endogenous regressor here is consumption growth, which has little predictability. Using the simulation results of DGP 2 and 3 as guide, one might expect these estimates to be strongly biased. The endogenous regressor associated with ψ in the second

equation is the real interest rate, which is persistent and can be predicted by many variables. Arguably, the real interest rate can be predicted by common factors in the economy. Thus if we use DGP 1 as a guide, the estimates using the principal components should be quite precise. The estimates for ψ range from .066 to .109. They are slightly higher than the IV with four observed instruments used in Yogo (2004). As the endogenous regressor is quite well predicted by the principal components, the weak instrument problem is of smaller concern. Our estimates suggest that the intertemporal elasticity of substitution is around .1 and is only marginally significant.

6 Conclusion

It is not uncommon for practitioners to have at their disposal a large number of variables that are weakly exogenous for the parameter of interest, some of which are especially informative about the endogenous regressor. Yet, there are few guides in the literature for forming subsets of instruments for estimation. This paper considers three methods, and also evaluates the effectiveness of using the principal components instead of the observed instruments directly. The results suggest that when the instruments have a factor structure, the principal components approach is unambiguously preferred to using the raw data as instruments. Ranking the t statistics of the first stage regression is effective in picking out those principal components most useful for estimation. When there are truly only a few instruments that have predictive power for the endogenous regressor, or in other words, the practitioner does not actually have many variables informative about the endogenous regressor, the observed data are better instruments than the principal components. When there is a large number of instruments each contributing to a small fraction of the variation in the endogenous regressor, the method of principal components is still preferred, even when an ideal instrument may not exist. In such a case, all three methods for selecting instruments have comparable properties.

DOI: 10.2202/1941-1928.1014

Table 1: Finite Sample Properties of $\widehat{\beta}_2$, GMM

T	$N\sigma_{x2}$	σ_z	ρ	$\frac{\mu_{k+1}}{\mu_k}$	$\widehat{R}^2_{x2}$	FIV_b	FIV_t	FIV_{ic}	IV_b	IV_t	IV_{ic}	OLS
						DGP 1: $\beta_2^0 = 2.0$						
200	50	1.0	1.0	0.253	0.053	0.527	2.037	2.024	2.038	2.055	2.075	2.041	2.389
							0.154	0.155	0.155	0.157	0.166	0.163	0.420
							5	4	5	9	20	4	
200	50	1.0	5.0	0.633	0.325	0.241	2.333	2.148	2.205	2.421	2.380	2.283	3.178
							0.436	0.346	0.375	0.493	0.463	0.427	1.202
							6	3	4	10	16	2	
200	100	1.0	1.0	0.429	0.042	0.527	2.156	2.012	2.102	2.129	2.122	2.053	2.722
							0.231	0.172	0.203	0.209	0.205	0.181	0.752
							8	3	5	13	20	4	
200	100	1.0	5.0	0.316	0.143	0.353	2.200	2.049	2.117	2.185	2.151	2.078	2.511
							0.274	0.213	0.230	0.259	0.240	0.227	0.542
							10	3	5	15	20	4	
200	50	3.0	1.0	0.539	0.057	0.297	2.184	2.070	2.113	2.165	2.238	2.102	2.842
							0.276	0.240	0.247	0.259	0.309	0.246	0.865
							6	3	4	8	20	2	
200	50	3.0	5.0	0.508	0.245	0.300	2.176	2.085	2.117	2.219	2.220	2.093	2.783
							0.268	0.236	0.246	0.295	0.296	0.237	0.807
							6	3	4	10	19	2	
200	100	3.0	1.0	0.295	0.030	0.322	2.164	2.043	2.097	2.079	2.107	2.034	2.418
							0.234	0.192	0.206	0.191	0.196	0.198	0.447
							10	3	5	9	20	2	
200	100	3.0	5.0	0.296	0.169	0.375	2.154	2.039	2.093	2.149	2.113	2.072	2.417
							0.218	0.170	0.183	0.209	0.189	0.181	0.443
							10	3	5	17	20	4	

Note: T is the sample size and N is the number of observed valid instruments; σ_{x2} and σ_z are scale parameters for innovations to x_2 and z; ρ is the correlation coefficient between the regression error and the endogenous regressor x_2, while $\bar{R}^2_{x2}$ measures the predictability of x_2. Furthermore, μ_k is the k-th largest eigenvalue of the $z'z$ matrix. FIV uses estimated factors as instruments while IV uses observed variables as instruments. The instruments are selected by boosting (b), t test (t), and BIC (ic).

Table 2: Finite Sample Properties of $\widehat{\beta}_2$, GMM

T	N	R^x_{x2}	ρ	$\frac{\mu_{k+1}}{\mu_k}$	$\widehat{R}^2_{x2}$	FIV_b	FIV_t	FIV_{ic}	IV_b	IV_t	IV_{ic}	OLS
						DGP 2: $\beta_2^0 = 1.0$						
200	50	0.900	0.157	0.941	0.899	1.010	1.010	1.010	1.012	1.012	1.012	1.050
						0.029	0.032	0.028	0.030	0.030	0.028	0.054
						11	8	17	14	13	20	
200	50	0.750	0.249	0.941	0.750	1.028	1.029	1.028	1.035	1.035	1.034	1.125
						0.055	0.060	0.054	0.057	0.061	0.054	0.129
						11	7	14	14	12	19	
200	50	0.500	0.351	0.941	0.495	1.076	1.074	1.075	1.090	1.090	1.089	1.249
						0.111	0.126	0.115	0.117	0.124	0.119	0.253
						11	5	9	14	8	12	
200	100	0.900	0.157	0.952	0.899	1.017	1.017	1.019	1.015	1.020	1.020	1.050
						0.033	0.038	0.034	0.032	0.037	0.035	0.055
						20	8	20	20	15	19	
200	100	0.750	0.248	0.954	0.748	1.049	1.050	1.055	1.044	1.055	1.054	1.124
						0.066	0.078	0.071	0.063	0.075	0.072	0.129
						20	7	20	20	12	16	
200	100	0.500	0.352	0.953	0.497	1.128	1.117	1.131	1.119	1.139	1.137	1.249
						0.145	0.156	0.151	0.138	0.166	0.160	0.253
						20	5	14	20	8	10	
						DGP 3: $\beta_2^0 = 1.0$						
200	50	0.900	0.158	0.942	0.899	1.010	1.010	1.010	1.023	1.023	1.022	1.050
						0.031	0.034	0.030	0.038	0.040	0.037	0.055
						11	8	17	18	14	19	
200	50	0.750	0.250	0.943	0.748	1.029	1.027	1.030	1.058	1.061	1.058	1.126
						0.059	0.062	0.058	0.077	0.085	0.079	0.131
						11	7	14	18	11	17	
200	50	0.500	0.351	0.941	0.497	1.073	1.072	1.071	1.126	1.137	1.135	1.249
						0.107	0.121	0.109	0.146	0.170	0.162	0.253
						11	5	9	16	7	10	
200	100	0.900	0.157	0.952	0.899	1.018	1.017	1.020	1.029	1.030	1.030	1.050
						0.033	0.037	0.034	0.044	0.047	0.046	0.054
						20	8	20	20	13	14	
200	100	0.750	0.248	0.952	0.749	1.048	1.046	1.053	1.072	1.073	1.075	1.124
						0.066	0.077	0.070	0.087	0.094	0.093	0.129
						20	7	20	20	11	12	
200	100	0.500	0.354	0.954	0.497	1.133	1.127	1.135	1.163	1.170	1.172	1.251
						0.150	0.167	0.156	0.179	0.198	0.196	0.255
						20	5	14	20	7	8	

DOI: 10.2202/1941-1928.1014

Table 3: Finite Sample Properties of $\widehat{\beta}_2$, GMM

T	N	R^x_{x2}	ρ	$\frac{\mu_{k+1}}{\mu_k}$	$\widehat{R}^2_{x2}$	FIV_b	FIV_t	FIV_{ic}	IV_b	IV_t	IV_{ic}	OLS
						DGP 4: $\beta_2^0 = 1.0$						
200	50	2.500	0.130	0.941	0.284	1.031	1.036	1.033	1.035	1.034	1.035	1.070
						0.081	0.116	0.100	0.074	0.104	0.094	0.080
						9	3	5	13	5	6	
200	50	1.500	0.157	0.941	0.401	1.035	1.038	1.037	1.044	1.045	1.045	1.099
						0.085	0.112	0.094	0.085	0.103	0.094	0.109
						11	4	7	14	6	8	
200	50	0.500	0.201	0.941	0.663	1.040	1.038	1.039	1.046	1.049	1.046	1.165
						0.095	0.105	0.094	0.093	0.099	0.092	0.175
						11	6	12	14	10	17	
200	100	2.500	0.134	0.952	0.282	1.049	1.042	1.046	1.046	1.048	1.049	1.072
						0.082	0.126	0.103	0.080	0.116	0.107	0.082
						16	3	7	20	5	6	
200	100	1.500	0.156	0.954	0.397	1.060	1.057	1.059	1.055	1.056	1.058	1.099
						0.090	0.126	0.101	0.088	0.116	0.109	0.108
						20	4	10	20	6	8	
200	100	0.500	0.202	0.953	0.664	1.070	1.068	1.077	1.065	1.081	1.080	1.165
						0.104	0.123	0.108	0.100	0.124	0.116	0.175
						20	6	20	20	11	14	
						DGP 5: $\beta_2^0 = 1.0$						
200	50	2.500	0.134	0.942	0.281	1.033	1.033	1.034	1.044	1.046	1.047	1.072
						0.085	0.119	0.104	0.082	0.123	0.109	0.081
						9	3	5	14	4	5	
200	50	1.500	0.160	0.943	0.397	1.037	1.032	1.035	1.053	1.060	1.060	1.101
						0.091	0.114	0.097	0.094	0.124	0.115	0.111
						10	4	7	15	6	7	
200	50	0.500	0.202	0.941	0.665	1.038	1.037	1.038	1.077	1.082	1.078	1.165
						0.091	0.102	0.091	0.112	0.129	0.117	0.174
						11	6	12	17	10	15	
200	100	2.500	0.132	0.952	0.284	1.049	1.040	1.046	1.052	1.053	1.052	1.070
						0.079	0.122	0.095	0.082	0.121	0.107	0.080
						16	3	7	20	4	5	
200	100	1.500	0.157	0.952	0.397	1.059	1.055	1.058	1.069	1.068	1.069	1.100
						0.091	0.125	0.101	0.097	0.123	0.117	0.109
						20	4	10	20	6	7	
200	100	0.500	0.206	0.954	0.665	1.074	1.073	1.081	1.104	1.107	1.111	1.168
						0.109	0.130	0.113	0.135	0.151	0.148	0.178
						20	6	20	20	9	11	

Table 4a: Finite Sample Properties of $\widehat{\beta}_2$, LIML

T	$N\sigma$	$_{x2}$	σ_z	ρ	$\frac{\mu_{k+1}}{\mu_k}$	R^2_{x2}	FLM_b	FLM_t	FLM_{ic}	LM_b	LM_t	LM_{ic}	OLS
				DGP 1: $\beta_2^0 = 2.0$									
200	50	1.0	1.0	0.253	0.053	0.527	2.025	2.017	2.027	2.029	2.011	2.030	2.389
							0.153	0.154	0.153	0.153	0.158	0.161	0.420
							5	4	5	9	20	4	
200	50	1.0	5.0	0.633	0.325	0.241	2.174	2.123	2.136	2.272	2.113	2.276	3.178
							0.342	0.320	0.331	0.382	0.322	0.418	1.202
							6	3	4	10	16	2	
200	100	1.0	1.0	0.429	0.042	0.527	2.077	2.002	2.050	2.054	1.997	2.039	2.722
							0.185	0.174	0.180	0.178	0.177	0.178	0.752
							8	3	5	13	20	4	
200	100	1.0	5.0	0.316	0.143	0.353	2.154	2.039	2.090	2.122	2.041	2.066	2.511
							0.248	0.211	0.219	0.230	0.218	0.224	0.542
							10	3	5	15	20	4	
200	50	3.0	1.0	0.539	0.057	0.297	2.103	2.053	2.072	2.085	2.014	2.099	2.842
							0.236	0.234	0.231	0.231	0.236	0.241	0.865
							6	3	4	8	20	2	
200	50	3.0	5.0	0.508	0.245	0.300	2.104	2.068	2.080	2.129	2.035	2.087	2.783
							0.233	0.227	0.229	0.244	0.230	0.232	0.807
							6	3	4	10	19	2	
200	100	3.0	1.0	0.295	0.030	0.322	2.125	2.034	2.075	2.041	1.999	2.035	2.418
							0.215	0.192	0.199	0.187	0.201	0.196	0.447
							10	3	5	9	20	2	
200	100	3.0	5.0	0.296	0.169	0.375	2.119	2.032	2.072	2.094	2.025	2.063	2.417
							0.199	0.170	0.176	0.184	0.178	0.179	0.443
							10	3	5	17	20	4	

Table 4b: Finite Sample Properties of $\widehat{\beta}_2$, Fuller

T	$N\sigma_{x2}$	σ_z	ρ	$\frac{\mu_{k+1}}{\mu_k}$	R^2_{x2}	FFU_b	FFU_t	FFU_{ic}	FU_b	FU_t	FU_{ic}	OLS
			DGP 1: $\beta_2^0 = 2.0$									
200	50	1.0	1.0	0.253	0.053	0.527	2.022	2.013	2.023	2.026	2.007	2.026	2.389
							0.153	0.154	0.154	0.153	0.159	0.162	0.420
							5	4	5	9	20	4	
200	50	1.0	5.0	0.633	0.325	0.241	2.149	2.090	2.106	2.252	2.085	2.245	3.178
							0.339	0.322	0.331	0.373	0.323	0.407	1.202
							6	3	4	10	16	2	
200	100	1.0	1.0	0.429	0.042	0.527	2.071	1.995	2.044	2.048	1.991	2.032	2.722
							0.183	0.175	0.180	0.178	0.179	0.178	0.752
							8	3	5	13	20	4	
200	100	1.0	5.0	0.316	0.143	0.353	2.149	2.031	2.084	2.118	2.034	2.058	2.511
							0.246	0.212	0.219	0.228	0.219	0.225	0.542
							10	3	5	15	20	4	
200	50	3.0	1.0	0.539	0.057	0.297	2.090	2.037	2.057	2.072	1.999	2.084	2.842
							0.234	0.237	0.233	0.231	0.242	0.241	0.865
							6	3	4	8	20	2	
200	50	3.0	5.0	0.508	0.245	0.300	2.091	2.053	2.066	2.118	2.021	2.071	2.783
							0.231	0.228	0.229	0.241	0.234	0.232	0.807
							6	3	4	10	19	2	
200	100	3.0	1.0	0.295	0.030	0.322	2.122	2.027	2.069	2.036	1.992	2.028	2.418
							0.214	0.194	0.199	0.188	0.205	0.198	0.447
							10	3	5	9	20	2	
200	100	3.0	5.0	0.296	0.169	0.375	2.116	2.026	2.068	2.091	2.020	2.058	2.417
							0.197	0.171	0.175	0.183	0.180	0.179	0.443
							10	3	5	17	20	4	

Table 5a: Finite Sample Properties of $\widehat{\beta}_2$, LIML

T	N	R^x_{x2}	ρ	$\frac{\mu_{k+1}}{\mu_k}$	$\widehat{R}^2_{x2}$	FLM_b	FLM_t	FLM_{ic}	LM_b	LM_t	LM_{ic}	OLS
						DGP 2: $\beta_2^0 = 1.0$						
200	50	0.900	0.157	0.941	0.899	1.006	1.007	1.005	1.008	1.008	1.007	1.050
						0.029	0.032	0.028	0.029	0.030	0.028	0.054
						11	8	17	14	13	20	
200	50	0.750	0.249	0.941	0.750	1.019	1.023	1.017	1.025	1.026	1.021	1.125
						0.053	0.059	0.051	0.053	0.058	0.050	0.129
						11	7	14	14	12	19	
200	50	0.500	0.351	0.941	0.495	1.052	1.064	1.057	1.065	1.075	1.068	1.249
						0.102	0.124	0.108	0.104	0.117	0.108	0.253
						11	5	9	14	8	12	
200	100	0.900	0.157	0.952	0.899	1.012	1.014	1.014	1.010	1.016	1.015	1.050
						0.032	0.039	0.032	0.031	0.037	0.035	0.055
						20	8	20	20	15	19	
200	100	0.750	0.248	0.954	0.748	1.034	1.044	1.042	1.029	1.046	1.043	1.124
						0.060	0.077	0.064	0.057	0.072	0.067	0.129
						20	7	20	20	12	16	
200	100	0.500	0.352	0.953	0.497	1.099	1.107	1.110	1.089	1.126	1.122	1.249
						0.125	0.153	0.138	0.118	0.160	0.152	0.253
						20	5	14	20	8	10	
						DGP 3: $\beta_2^0 = 1.0$						
200	50	0.900	0.158	0.942	0.899	1.007	1.008	1.005	1.019	1.019	1.017	1.050
						0.031	0.034	0.030	0.037	0.040	0.037	0.055
						11	8	17	18	14	19	
200	50	0.750	0.250	0.943	0.748	1.019	1.021	1.019	1.047	1.053	1.047	1.126
						0.057	0.061	0.056	0.072	0.082	0.074	0.131
						11	7	14	18	11	17	
200	50	0.500	0.351	0.941	0.497	1.049	1.062	1.052	1.102	1.126	1.121	1.249
						0.098	0.119	0.102	0.131	0.166	0.154	0.253
						11	5	9	16	7	10	
200	100	0.900	0.157	0.952	0.899	1.013	1.015	1.016	1.024	1.026	1.027	1.050
						0.033	0.037	0.033	0.044	0.048	0.046	0.054
						20	8	20	20	13	14	
200	100	0.750	0.248	0.952	0.749	1.033	1.040	1.040	1.059	1.066	1.067	1.124
						0.059	0.076	0.064	0.081	0.092	0.090	0.129
						20	7	20	20	11	12	
200	100	0.500	0.354	0.954	0.497	1.104	1.118	1.115	1.139	1.161	1.162	1.251
						0.131	0.164	0.144	0.164	0.195	0.191	0.255
						20	5	14	20	7	8	

DOI: 10.2202/1941-1928.1014

Table 5b: Finite Sample Properties of $\widehat{\beta}_2$, Fuller

T	N	R^x_{x2}	ρ	$\frac{\mu_{k+1}}{\mu_k}$	$\widehat{R}^2_{x2}$	FFU_b	FFU_t	FFU_{ic}	FU_b	FU_t	FU_{ic}	OLS
						DGP 2: $\beta^0_2 = 1.0$						
200	50	0.900	0.157	0.941	0.899	1.006	1.007	1.004	1.008	1.008	1.007	1.050
						0.029	0.032	0.028	0.029	0.030	0.028	0.054
						11	8	17	14	13	20	
200	50	0.750	0.249	0.941	0.750	1.018	1.021	1.016	1.024	1.025	1.020	1.125
						0.053	0.059	0.051	0.053	0.058	0.049	0.129
						11	7	14	14	12	19	
200	50	0.500	0.351	0.941	0.495	1.049	1.059	1.054	1.063	1.072	1.066	1.249
						0.101	0.124	0.108	0.103	0.116	0.107	0.253
						11	5	9	14	8	12	
200	100	0.900	0.157	0.952	0.899	1.011	1.014	1.014	1.009	1.016	1.015	1.050
						0.032	0.039	0.032	0.031	0.037	0.035	0.055
						20	8	20	20	15	19	
200	100	0.750	0.248	0.954	0.748	1.034	1.043	1.042	1.028	1.045	1.043	1.124
						0.060	0.077	0.064	0.057	0.072	0.066	0.129
						20	7	20	20	12	16	
200	100	0.500	0.352	0.953	0.497	1.097	1.104	1.109	1.087	1.124	1.120	1.249
						0.124	0.152	0.137	0.118	0.159	0.151	0.253
						20	5	14	20	8	10	
						DGP 3: $\beta^0_2 = 1.0$						
200	50	0.900	0.158	0.942	0.899	1.006	1.007	1.005	1.019	1.018	1.017	1.050
						0.031	0.034	0.030	0.037	0.040	0.037	0.055
						11	8	17	18	14	19	
200	50	0.750	0.250	0.943	0.748	1.018	1.020	1.018	1.046	1.052	1.046	1.126
						0.057	0.061	0.056	0.072	0.082	0.074	0.131
						11	7	14	18	11	17	
200	50	0.500	0.351	0.941	0.497	1.046	1.058	1.049	1.101	1.123	1.119	1.249
						0.097	0.118	0.102	0.130	0.165	0.154	0.253
						11	5	9	16	7	10	
200	100	0.900	0.157	0.952	0.899	1.013	1.014	1.015	1.024	1.026	1.027	1.050
						0.033	0.037	0.033	0.044	0.048	0.046	0.054
						20	8	20	20	13	14	
200	100	0.750	0.248	0.952	0.749	1.032	1.039	1.040	1.058	1.065	1.066	1.124
						0.059	0.076	0.063	0.081	0.092	0.090	0.129
						20	7	20	20	11	12	
200	100	0.500	0.354	0.954	0.497	1.102	1.115	1.113	1.137	1.159	1.160	1.251
						0.130	0.164	0.143	0.163	0.194	0.191	0.255
						20	5	14	20	7	8	

Table 6a: Finite Sample Properties of $\widehat{\beta}_2$, LIML

T	N	R^x_{x2}	ρ	$\frac{\mu_{k+1}}{\mu_k}$	$\widehat{R}^2_{x2}$	FLM_b	FLM_t	FLM_{ic}	LM_b	LM_t	LM_{ic}	OLS
						DGP 4						
200	50	2.500	0.130	0.941	0.284	1.025	1.035	1.030	1.028	1.032	1.032	1.070
						0.088	0.114	0.102	0.080	0.105	0.097	0.080
						9	3	5	13	5	6	
200	50	1.500	0.157	0.941	0.401	1.025	1.035	1.031	1.034	1.040	1.038	1.099
						0.090	0.114	0.097	0.088	0.105	0.097	0.109
						11	4	7	14	6	8	
200	50	0.500	0.201	0.941	0.663	1.027	1.031	1.025	1.032	1.037	1.029	1.165
						0.095	0.107	0.094	0.090	0.097	0.089	0.175
						11	6	12	14	10	17	
200	100	2.500	0.134	0.952	0.282	1.043	1.042	1.043	1.038	1.046	1.047	1.072
						0.089	0.124	0.108	0.088	0.117	0.110	0.082
						16	3	7	20	5	6	
200	100	1.500	0.156	0.954	0.397	1.051	1.055	1.054	1.043	1.053	1.053	1.099
						0.092	0.128	0.104	0.091	0.121	0.113	0.108
						20	4	10	20	6	8	
200	100	0.500	0.202	0.953	0.664	1.051	1.061	1.061	1.044	1.071	1.067	1.165
						0.098	0.124	0.103	0.095	0.124	0.113	0.175
						20	6	20	20	11	14	
						DGP 5						
200	50	2.500	0.134	0.942	0.281	1.027	1.033	1.032	1.038	1.044	1.045	1.072
						0.093	0.118	0.106	0.089	0.122	0.112	0.081
						9	3	5	14	4	5	
200	50	1.500	0.160	0.943	0.397	1.027	1.028	1.029	1.043	1.057	1.055	1.101
						0.096	0.116	0.101	0.098	0.128	0.118	0.111
						10	4	7	15	6	7	
200	50	0.500	0.202	0.941	0.665	1.024	1.030	1.024	1.061	1.072	1.064	1.165
						0.092	0.103	0.091	0.108	0.130	0.115	0.174
						11	6	12	17	10	15	
200	100	2.500	0.132	0.952	0.284	1.043	1.039	1.042	1.046	1.052	1.051	1.070
						0.084	0.121	0.100	0.090	0.124	0.111	0.080
						16	3	7	20	4	5	
200	100	1.500	0.157	0.952	0.397	1.049	1.053	1.052	1.060	1.065	1.066	1.100
						0.095	0.127	0.106	0.102	0.127	0.120	0.109
						20	4	10	20	6	7	
200	100	0.500	0.206	0.954	0.665	1.054	1.066	1.065	1.087	1.100	1.103	1.168
						0.103	0.132	0.107	0.132	0.152	0.148	0.178
						20	6	20	20	9	11	

DOI: 10.2202/1941-1928.1014

Table 6b: Finite Sample Properties of $\widehat{\beta}_2$, Fuller

T	N	R^x_{x2}	ρ	$\frac{\mu_{k+1}}{\mu_k}$	$\widehat{R}^2_{x2}$	FFU_b	FFU_t	FFU_{ic}	FU_b	FU_t	FU_{ic}	OLS
						DGP 4: $\beta^0_2 = 1.0$						
200	50	2.500	0.130	0.941	0.284	1.024	1.034	1.029	1.027	1.030	1.031	1.070
						0.089	0.120	0.105	0.080	0.109	0.100	0.080
						9	3	5	13	5	6	
200	50	1.500	0.157	0.941	0.401	1.024	1.033	1.030	1.033	1.039	1.037	1.099
						0.091	0.117	0.098	0.088	0.106	0.097	0.109
						11	4	7	14	6	8	
200	50	0.500	0.201	0.941	0.663	1.025	1.029	1.023	1.031	1.036	1.028	1.165
						0.095	0.107	0.094	0.090	0.097	0.089	0.175
						11	6	12	14	10	17	
200	100	2.500	0.134	0.952	0.282	1.043	1.040	1.042	1.038	1.045	1.046	1.072
						0.090	0.130	0.110	0.088	0.122	0.113	0.082
						16	3	7	20	5	6	
200	100	1.500	0.156	0.954	0.397	1.050	1.054	1.053	1.042	1.051	1.052	1.099
						0.092	0.131	0.105	0.091	0.123	0.114	0.108
						20	4	10	20	6	8	
200	100	0.500	0.202	0.953	0.664	1.050	1.059	1.060	1.043	1.069	1.066	1.165
						0.098	0.124	0.102	0.094	0.124	0.113	0.175
						20	6	20	20	11	14	
						DGP 5: $\beta^0_2 = 1.0$						
200	50	2.500	0.134	0.942	0.281	1.026	1.031	1.031	1.038	1.043	1.044	1.072
						0.095	0.124	0.111	0.090	0.128	0.115	0.081
						9	3	5	14	4	5	
200	50	1.500	0.160	0.943	0.397	1.026	1.026	1.027	1.042	1.055	1.054	1.101
						0.097	0.119	0.103	0.098	0.131	0.120	0.111
						10	4	7	15	6	7	
200	50	0.500	0.202	0.941	0.665	1.022	1.028	1.022	1.060	1.071	1.063	1.165
						0.092	0.104	0.091	0.108	0.130	0.115	0.174
						11	6	12	17	10	15	
200	100	2.500	0.132	0.952	0.284	1.043	1.037	1.042	1.046	1.051	1.050	1.070
						0.085	0.127	0.102	0.090	0.129	0.114	0.080
						16	3	7	20	4	5	
200	100	1.500	0.157	0.952	0.397	1.049	1.051	1.051	1.060	1.064	1.065	1.100
						0.095	0.130	0.107	0.102	0.130	0.122	0.109
						20	4	10	20	6	7	
200	100	0.500	0.206	0.954	0.665	1.053	1.064	1.064	1.086	1.098	1.101	1.168
						0.103	0.132	0.107	0.132	0.153	0.148	0.178
						20	6	20	20	9	11	

Table 7a: Estimates of $1/\psi$

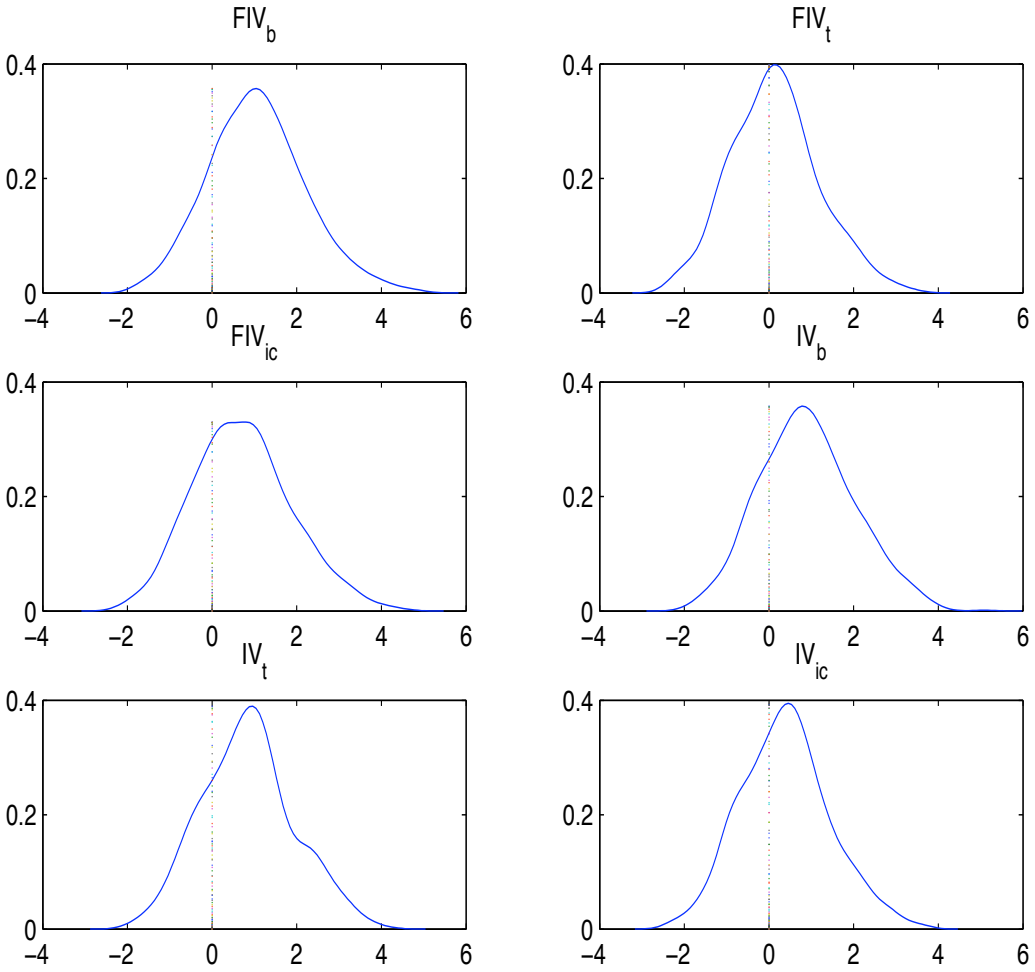
	$\widehat{1/\psi}$	$s.e.$	t	N	*	R^2_{x2}
FIV_b	0.470	0.245	1.918	11		0.361
FIV_t	1.453	0.501	2.900	2		0.130
FIV_{ic}	0.769	0.346	2.225	5		0.223
IV_b	0.150	0.222	0.675	14		0.375
IV_t	-0.271	0.274	-0.987	12		0.204
IV_{ic}	-0.281	0.375	-0.750	2		0.200
IV	0.526	0.498	1.055	3		0.082
OLS	0.409	0.168	2.427	1		1.000

Table 7b: Estimates of ψ

	$\widehat{\psi}$	$s.e.$	t	N	*	R^2_{x2}
FIV_b	0.089	0.056	1.589	13		0.515
FIV_t	0.066	0.101	0.649	3		0.231
FIV_{ic}	0.109	0.058	1.895	9		0.458
IV_b	-0.004	0.059	-0.063	17		0.528
IV_t	0.010	0.054	0.176	20		0.354
IV_{ic}	-0.121	0.099	-1.221	5		0.288
IV	0.058	0.090	0.645	3		0.279
OLS	0.120	0.050	2.427	1		1.000

DOI: 10.2202/1941-1928.1014

Figure 1: Finite Sample Distribution of $t_{\widehat{\beta}}$



References

Amemiya, T. 1966, On the Use of Principal Components of Independent Variables in Two-Stage Least Squares Estimation, *International Economic Review* **7:3**, 283–303.

Andrews, D. W. K. and Lu, B. 2001, Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models, *Journal of Econometrics* **12**, 123–164.

Bai, J. and Ng, S. 2002, Determining the Number of Factors in Approximate Factor Models, *Econometrica* **70:1**, 191–221.

Bai, J. and Ng, S. 2006, Instrumental Variable Estimation in a Data Rich Environment, Department of Economics, University of Columbia.

Bai, J. and Ng, S. 2008, Boosting Diffusion Indices, *forthcoming in Journal of Applied Econometrics*.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. 2006, Prediction by Supervised Principal Components, *Journal of the American Statistical Association* **101:473**, 119–137.

Boivin, J. and Giannoni, M. 2006, DSGE Models in a Data Rich Environment, NBER WP 12772.

Buhlmann, P. 2006, Boosting for High-Dimensional Linear Models, *Annals of Statistics* **54:2**, 559–583.

Buhlmann, P. and Hothorn, T. 2006, Boosting: A Statistical Perspective, mimeo.

Buhlmann, P. and Yu, B. 2003, Boosting with the L_2 Loss: Regression and Classification, *Journal of the American Statistical Assocation* **98**, 324–339.

Carrasco, M. 2006, A Regularization Approach to the Many Instruments Problem, mimeo, Université de Montreal.

Connor, G. and Korajzcyk, R. 1986, Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis, *Journal of Financial Economics* **15**, 373–394.

Donald, S. and Newey, W. 2001, Choosing the Number of Instruments, *Econometrica* **69:5**, 1161–1192.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004, Least Angle Regression, *Annals of Statistics* **32**, 407–499.

Eryuruk, G., Hall, A. and Jana, K. 2008, A Comparative Study of Three Data-Based Methods of Instrument Selection, mimro, NC State.

DOI: 10.2202/1941-1928.1014

Fisher, F. 1965, The Choice of instrumental Variables in the Estimation of Economy-Wide Econometric Models, *International Economic Reivew* **6**, 245–74.

Freund, Y. 1995, Boosting a Weak Learning Algorithm by Majority, *Information and Computation* **121**, 256–285.

Friedman, J. 2001, Greedy Function Approxmiation: a Gradient Boosting Machine, *The Annals of Statistics* **29**, 1189–1232.

Hahn, J. and Hausman, J. 2002, Notes on Bias in Estimators for Simultaneous Equations Models, *Economics Letters* **75**, 237–241.

Hahn, J. and Kuersteiner, G. 2002, Discontinuities of Weak Instrument Limiting Distributions, *Economics Letters* **75**, 325–331.

Hall, A. and Peixe, F. 2000, Data Mining and the Selection of Instruments, *Journal of Economic Methodology* pp. 265–277.

Hall, A. and Peixe, F. 2003, A Consistent Method for the Selection of Relevant Instruments, *Econometric Reviews* pp. 269–288.

Hall, A., Inoue, A., Jana, K. and Shin, C. 2007, Information in Generalized Method of Moments Estimation Entropy Based Moment Selection, *Journal of Econometrics* **138**, 488–512.

Hansen, P. 2008, In-Sample and Out-of-Sample Fit: Their Joint Distribution and its Implications for Model Selection, manuscript, Stanford University.

Hausman, J. W, Newey., T. Woutersen, J. Chao and N. R. Swanson, 2007, *Instrumental Variable Estimation with Heteroskedasticity and Many Instruments*, Department of Economics, MIT. Unpublishded manuscript.

Kapatanios, G. 2006, Choosing the Optimal Set of Instruments From Large Instrument Sets, *Computational Statistics and Data Analysis,* **15:2**, 612–620.

Kapetanios, G. and Marcellino, M. 2006, Factor-GMM Estimation with Large Sets of Possibly Weak Instruments, mimeo.

Kloek, T. and Mennes, L. 1960, Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables, *Econometrica* **28**, 46–61.

Ludvigson, S. and Ng, S. 2007, The Empirical Risk Return Relation: A Factor Analysis Approach, *Journal of Financial Economics* **83**, 171–222.

Moreira, M. 2003, A Conditional Likelihood Ratio Test for Structural Models, *Econometrica* **71:4**, 1027–1048.

Newey, W. and Smith, R. 2004, Higher Order Properties of GMM and Generalized Empirical Likeliood Estimators, *Econometrica* **71:1**, 219–255.

Okui, R. 2004, Shrinkage Methods for Instrumental Variable Estimation, mimeo.

Phillips, P. C. B. 1983, Exact Small Sample Theory in the Simultaneous Equations Models, *in* M. D. Intriligator and Z. Grilches (eds), *Handbook of Econometrics, Volume 1*, North Holland.

Schapire, R. E. 1990, The Strenght of Weak Learnability, *Machine Learning* **5**, 197–227.

Staiger, D. and Stock, J. H. 1997, Instrumental Variables Regression with Weak Instruments, *Econometrica* **65:3**, 557–586.

Tibshirani, R. 1996, Regression Shrinkage and Selection via the Lasso, *Journal of Royal Statistical Society Series B* **58:1**, 267–288.

Yogo, M. 2004, Estimating the Elasticity of Intertemporal Substitution When Instruments are Weak, *Review of Economics and Statistics* **86:3**, 797:810.

Zhang, T. and Yu, B. 2005, Boosting with Early Stopping: Convergence and Consistency, *Annals of Statistics* **33:4**, 1538–1579.

DOI: 10.2202/1941-1928.1014