Principal Components Instrumental Variable Estimation

Diego Winkelried and Richard J. Smith

31 January 2011

CWPE 1119

Principal Components Instrumental Variable Estimation*

Richard J. Smith

Faculty of Economics, University of Cambridge, Sidgwick Avenue, CB3 9DD, Cambridge, UK

Diego Winkelried†

St. John's College, University of Cambridge, CB2 1TP, Cambridge, UK

January 31, 2011

Abstract

Instrumental variable estimators can be severely biased in finite samples when the degree of overidentification is high or when the instruments are weakly correlated with the endogenous regressors. This paper proposes an estimator based on the use of the principal components of the instruments as a means of dealing with these issues. By promoting parsimony, the proposed estimator can exhibit considerably lower bias, often without giving up asymptotic efficiency. To make the estimator operational, a simple but flexible rule to select the relevant components for estimation is suggested. Simulation evidence shows that this approach yields significant finite sample improvements over other instrumental variable estimators.

JEL Classification : C13, C31, C51.

Keywords : Many instrument asymptotics, principal components.

*We would like to thank Alastair Hall, Melvyn Weeks and seminar participants at the the University of Cambridge for helpful comments. Diego Winkelried gratefully acknowledges the financial support of the ORS award, St John's College Benefactors' Scholarship for Research and the Gates Cambridge Trust. The usual disclaimer applies.

†Corresponding author: +511 6132000, diegowq@cantab.net

1 Introduction

The two stage least squares estimator – or generalised instrumental variable (IV) estimator (IVE hereafter) – is known to be biased in finite samples if the degree of overidentification is high or when the instruments are weakly related to the endogenous variables (Donald and Newey, 2001; Stock, Wright, and Yogo, 2002). A growing body of literature has responded to these issues by proposing alternative estimators based either on large-sample approximations of the asymptotic moments, or on the asymptotic framework advanced in Bekker (1994), where the number of instruments grows proportionately to the sample size. The results are some form of k-class, limited information maximum likelihood (LIML) or Jackknife IV estimators. Many of them successfully reduce the bias of the IVE but at the cost of fatter tails in their sampling distribution, and hence the lack of finite sample moments. This phenomenon implies that extreme estimates are likely to be encountered in actual empirical situations, and hence authors like inter alia Hahn, Hausman, and Kuersteiner (2004) and Davidson and MacKinnon (2006) suggest using estimators known to posses moments instead.

A different way to address the many instrument bias is related to the specification of the reduced form of the endogenous variables. Hahn (2002) determines an efficiency bound for the variance of a general class of estimators in the linear IV framework, and shows that strikingly none of the 'no moments' estimators can achieve it under the many instrument asymptotics. Moreover, Hahn makes a strong case for parsimony by showing that an IVE using only a subset of the available instruments may be consistent and fully efficient. More precisely, for an IVE to be optimal it is required that (**A**) the reduced form of the endogenous variables is parsimonious enough; and (**B**) this simpler specification has asymptotically the same explanatory power as the specification involving the full set of instruments.

Meeting requirements (**A**) and (**B**) can be regarded as a model selection problem: one may think of the many instrument situation as an overparameterised model that contains both relevant IV (that help achieving identification and improve the precision of the estimates) and irrelevant IV (that overfit the regression and add bias to the IVE), and therefore the goal of the researcher would be to retain the relevant IV only. Hall, Rudebusch, and Wilcox (1996) find that determining which instrument among a set is weak with pretesting procedures may paradoxically exacerbate the poor properties of the IVE. Also, simulation evidence in Hall and Peixe (2003) suggests that the inclusion of irrelevant IV can lead to a serious deterioration in the quality of the asymptotic distributions as an approximation to finite sample behaviour (see also Hall, Inoue, Jana, and Shin, 2007, for related

information criteria designed to select relevant moment conditions).

On the other hand, Chao and Swanson (2005) find that it is the combined effect of a large number of possibly weak instruments, and not neccesarily the individual contribution of an instrument, what affects the properties of the IVE, suggesting that model selection may be performed over sets of IV. This task is certainly feasible but computationally cumbersome. Kapetanios (2006) deals with the selection problem by optimising criteria developed in Donald and Newey (2001) among discrete IV sets. Notwithstanding the significant improvements that are obtained for various estimators, the need for non-standard optimisation techniques makes the procedure rather costly.

Another strand of the literature explores alternatives to model selection aimed to parsimoniously summarise large sets of IV, based on the idea of using as much information as possible, while avoiding the possible pitfalls of using too many instruments. Kapetanios and Marcellino (2006) and Bai and Ng (2010) impose a factor structure to the IV set and use a few estimated factors as the IV in the estimation. In a similar fashion, Kapetanios and Marcellino (2008) study the performance of an IVE based on a limited set of instruments obtained as the weighted average of the original IV. Finally, Okui (2010) proposes a shrinkage method to deal with the first stage of the IVE procedure. The method consists of shrinking part of the OLS coefficient estimates from the regression of the endogenous variables on the instruments and then using the predicted values of the endogenous variables, based on the shrunk coefficient estimates, as the instruments (see also Carriero, Kapetanios, and Marcellino, 2008). Consistent with Hahn's findings, these procedures yield estimators with improved finite sample performance.

This paper is concerned with the use of principal components (PC henceforth) as a simple way to address the many (and possibly weak) instrument problem. In particular, we study the properties of a standard IVE that uses few PC of the original IV as instruments (PCIVE). The idea dates back to Kloek and Mennes (1960). Another early account is given in Amemiya (1966) who shows that the use of PC as instruments increases finite sample efficiency by increasing the number of degrees of freedom. More recently, Doran and Schmidt (2006) use PC methods in GMM estimation to attenuate the bias effects of near-singularities in the moment conditions. The novelty of this paper is proving certain optimality results using many (weak) instrument asymptotics as the conceptual framework.

Bai and Ng (2010) show that PC have desirable properties as instruments when the IV set is generated by a factor structure. Nonetheless, we attempt not to impose such explicit structure in our analysis since the resulting IVE can perform badly if the IV do not admit a factor representation

(cf., Kapetanios and Marcellino, 2006). We rely instead on a vague notion of correlation among instruments (assumption A5 or A6 below), as the merits of the PCIVE depend on the instruments displaying *some* correlation.[1] To this end, we also propose using a simple rule to select the relevant components such that conditions (**A**) and (**B**) are satisfied. This rule adjusts to the amount of correlation found in the IV set, and simulation evidence indicates that the PCIVE performs well even when the IV are only slightly correlated.

Even though our theoretical analysis emphasises the dimensionality reduction aspects of PC, we also explore in a simulation study if such technique has also instrument selection advantages. In fact, there may be situations where the PC approach lies between data reduction and model selection. On the one hand, PC effectively reduce the dimensionality of a vector space by means of linear combinations that give rise to maximum variations (a comprehensive treatment on the subject can be found in Jolliffe, 2002). On the other hand, since the PC are just rotations of the original IV that can be unmistakably sorted, it may ease the separation of the irrelevant combinations of IV from the relevant ones. This effect is observed when the irrelevant instruments are close to be orthogonal to the relevant ones.

The rest of the paper is organised as follows. Section 2 describes the setup, briefly summarises the properties of the IVE in highly overidentified models and explores the benefits of parsimony. Then, section 3 presents the PCIVE along with a heuristic rule to retain the relevant components. The PCIVE is found to be consistent and nearly efficient under many instrument asymptotics, even when the instruments are weak. Section 4 presents a Monte Carlo study aimed to cover situations of empirical interest, and evaluates the performance of the PCIVE and the associated retention rule in such situations. Section 5 gives concluding remarks. The derivations of the main results are displayed in two appendices.

2 Econometric Framework

Consider the simultaneous equations model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \qquad \text{and} \qquad \boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{V}\,, \tag{1}$$

[1] An advantage of such general structure is that the 'large-sample' condition used in previous studies $K_n/\sqrt{n} \to 0$, where K_n is the number of instruments and n is the sample size, is not required in our asymptotic approximations.

where $\boldsymbol{y}$ is the $n \times 1$ vector containing n observations of the dependent variable; $\boldsymbol{X}$ is the $n \times G$ matrix with observations of the endogenous regressors; $\boldsymbol{Z}$ is the $n \times K_n$ matrix of IV, where $K_n > G$; $\boldsymbol{\Pi}$ is a $K_n \times G$ matrix of coefficients with rank G; and $\boldsymbol{V}$ is an $n \times G$ matrix of disturbances, whose rows are correlated with the corresponding elements of $\boldsymbol{u}$, the $n \times 1$ vector of disturbances in the structural equation. Note that $\boldsymbol{Z}$ and $\boldsymbol{\Pi}$ are allowed to depend on n but to alleviate the notation this dependence is left implicit and noted only through K_n. The interest lies in estimators of the $G \times 1$ vector of unknown coefficients $\boldsymbol{\beta}$.

For simplicity we take all regressors as endogenous. This is without loss of generality since the matrices of regressors and instruments may be written, respectively, as $\boldsymbol{X} = (\boldsymbol{X}^{\text{END}}, \boldsymbol{X}^{\text{EX}})$ and $\boldsymbol{Z} = (\boldsymbol{X}^{\text{EX}}, \boldsymbol{Z}^{*})$ and the discussion below on variable reduction applies to $\boldsymbol{Z}^{*}$ only. In other words, the relevant quantity is the degree of overidentification $K_n - G$. Alternatively, one may understand (1) as the simultaneous equations system after partialling $\boldsymbol{X}^{\text{EX}}$ out.

We focus our attention on 'many instrument' asymptotic approximations when both K_n and n grow simultaneously. We work with the following assumptions:

A1 Let u_i be the i-th element of $\boldsymbol{u}$ and $\boldsymbol{v}_i{}'$ be the i-th row of $\boldsymbol{V}$. The vectors $(u_i, \boldsymbol{v}_i{}')'$ for $i = 1, 2, \ldots, n$ contain iid normally distributed variables,

$$\begin{bmatrix} u_i \\ \boldsymbol{v}_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_{uu} & \boldsymbol{\rho}' \\ \boldsymbol{\rho} & \boldsymbol{\Omega} \end{bmatrix} \right). \tag{2}$$

A2 $\boldsymbol{Z}$ is a sequence of non-random variables. The smallest eigenvalue of $\boldsymbol{S} = \boldsymbol{Z}'\boldsymbol{Z}/n$ is bounded away from zero as $n \to \infty$. Also, if $\boldsymbol{z}_i{}'$ denotes the i-th row of $\boldsymbol{Z}$, then $\max_{i \leq n} \{\boldsymbol{z}_i{}' \boldsymbol{z}_i / n\} \to 0$.

A3 Define a sequence ψ_n of non-decreasing numbers such that $\psi_n \to \infty$ as $n \to \infty$, and define a $G \times G$ positive definite matrix $\boldsymbol{\Psi}$, such that

$$\boldsymbol{\Psi} = \frac{\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}}{\psi_n} \quad \text{for all } n. \tag{3}$$

A4 The matrix $\boldsymbol{\Pi}$ is such that $\|\boldsymbol{\Pi}\|^2 = (\psi_n / n)\, O(\|\boldsymbol{S}\|^{-1})$ for all n, where $\|\boldsymbol{\Pi}\|^2 = \text{tr}(\boldsymbol{\Pi}'\boldsymbol{\Pi})$ denotes the square of the Frobenius norm for $\boldsymbol{\Pi}$.

These assumptions are made for tractability and can be weakened in several ways. The normality in Assumption A1 may be exchanged, as done in Hansen, Hausman, and Newey (2008), by bounding

the moments of $(u_i, \boldsymbol{v}_i')'$ and by assuming the independence between u_i and $\boldsymbol{v}_i - (\boldsymbol{\rho}/\sigma_{uu})u_i$, the residuals from the population regression of $\boldsymbol{v}_i$ on u_i. As stressed in Bekker (1994), non-normality can affect the moments of the asymptotic distribution of the IVE, when K_n grows in tandem with n, as long as they depend on the ratio $K_n/n \to \alpha_K$ (see also van Hasselt, 2010). The asymptotic distribution of the proposed estimator does not depend on such ratio and thus we conjecture that it is robust to deviations from normality.

The asymptotic non-singularity of $\boldsymbol{S}$ in Assumption A2 is merely a normalisation. Part of the analysis below is made in terms of the eigenvalues of $\boldsymbol{S}$ and no result is altered if some of them are equal to zero. In the case where $\boldsymbol{S}$ is rank deficient, one should use $K_n = \text{rank}(\boldsymbol{Z}'\boldsymbol{Z})$ instead of taking K_n as the number of columns in $\boldsymbol{Z}$. On the other hand, $\boldsymbol{Z}$ may be allowed to be random and one could interpret the results as being conditional on this IV set. With this in mind, we will refer to the degree of collinearity between the columns of $\boldsymbol{Z}$ loosely as 'correlation'. If unconditional inference is pursued, then $\boldsymbol{u}$ and $\boldsymbol{V}$ need to be mean independent from $\boldsymbol{Z}$ to $O_p(1/\sqrt{n})$, a minimum requirement for the asymptotic identification in IV models. In short, we consider the instruments to be *valid* and the reduced form for $\boldsymbol{X}$ to be correctly specified. On the other hand, the condition on the typical row $\boldsymbol{z}_i'$ in Assumption A2 is required for the asymptotic distribution of the IVE to be normal, and is a standard regularity condition (Hahn, 2002; Okui, 2010). It imposes a restriction of a balanced design so that no single observation is dominant.

Assumption A3 places a restriction on the informational content of the 'new' instruments as K_n increases. If $\boldsymbol{Z}$ is allowed to be random, this equality can be weakened to a probability limit, without altering the form of asymptotic moments of the IVE. The matrix $\boldsymbol{\Psi}$ in (3) is proportional the *concentration parameter*, a widely accepted measure of the relevance of the IV set, so this assumption implies that it grows at a rate ψ_n as $n \to \infty$. Standard textbook asymptotics implicitly assume that $K_n/\psi_n \to 0$, the many instrument framework of Bekker (1994) sets that both K_n and ψ_n grow at the same rate as n, whereas many weak instrument scenarios (cf. Stock, Wright, and Yogo, 2002; Chao and Swanson, 2005) often impose $\psi_n = O(K_n)$ and $K_n = o(n)$ or even $\psi_n = o(K_n)$.

Finally, Assumption A4, which can be viewed as a refinement to Assumption A3, states that the order of magnitude of the elements of $\boldsymbol{\Pi}$ is inversely proportional to the order of magnitude of the elements of $\boldsymbol{S}$. From the definition of $\boldsymbol{S}$, Assumption A3 implies that $\|\boldsymbol{\Psi}\| \leq (n/\psi_n)\|\boldsymbol{\Pi}\|^2\|\boldsymbol{S}\|$. The left-hand-side of this inequality is $O(1)$ whereas the right-hand-side is $O(1)$ only under Assumption A4. Hence, Assumption A4 simply places a bound for $\|\boldsymbol{\Pi}\|$ such that the implied restriction on the

explanatory power of the instruments in (3) also holds in matrix norms.

2.1 The Case for a Lower-Dimensional IVE

We study the properties of estimators based on an IV set $\bar{\boldsymbol{Z}}$ obtained from linear combinations of the original $\boldsymbol{Z}$, and spanning a lower-dimensional space. More formally, $\bar{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{C}$ where $\boldsymbol{C}$ is a $K_n \times r_n$ matrix of rank r_n ($r_n \geq G$ is required for identifiability). This setup encompasses estimators obtained by selecting a subset of the IV, a case that corresponds to $\boldsymbol{C}$ being formed by r_n columns of a $K_n \times K_n$ identity matrix, and the standard IVE (i.e., the two-stage least squares estimator) when $\boldsymbol{C}$ is of full rank K_n.

Define the projection matrix of rank r_n to be $\boldsymbol{P} = \bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'$, so the IVE of $\boldsymbol{\beta}$ in system (1) associated with the reduced set of instruments is

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{y}\,. \tag{4}$$

The standard IVE, which will be referred to as $\boldsymbol{b}_K$, is based on a set of orthogonality conditions whose sample counterparts are given by $\boldsymbol{Z}'\boldsymbol{u}/n = \boldsymbol{0}$. Similarly, the estimator in (4) is associated with sample moment conditions of the form $\bar{\boldsymbol{Z}}'\boldsymbol{u}/n = \boldsymbol{C}'(\boldsymbol{Z}'\boldsymbol{u})/n = \boldsymbol{0}$, so the lower-dimensional IVE can be thought as the standard IVE after retaining only r_n linear combinations of the original moment conditions or after dropping $K_n - r_n$ of such linear combinations. Which combinations are retained depends on the choice of $\boldsymbol{C}$.

Let $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P}$ and define

$$\boldsymbol{E}_n \equiv \boldsymbol{E}_n(\boldsymbol{C}) = \min_{\gamma}\left\{\frac{(\boldsymbol{Z}\boldsymbol{\Pi} - \bar{\boldsymbol{Z}}\boldsymbol{\gamma})'(\boldsymbol{Z}\boldsymbol{\Pi} - \bar{\boldsymbol{Z}}\boldsymbol{\gamma})}{\psi_n}\right\} = \frac{\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\Pi}}{\psi_n} \tag{5}$$

as the $G \times G$ matrix that measures how well $\boldsymbol{Z}\boldsymbol{C}\boldsymbol{\gamma}$ approximates $\boldsymbol{Z}\boldsymbol{\Pi}$ in a least squares sense. It may be interpreted as the approximation error in explaining the regressors $\boldsymbol{X}$ with the lower-dimensional set $\bar{\boldsymbol{Z}}$ instead of the full K_n-dimensional set $\boldsymbol{Z}$, i.e. a loss (in terms of fit) due to the mispecification of the first stage regression. For all n, $\boldsymbol{E}_n$ is bounded from above by $\boldsymbol{\Psi}$ (trivially, when $\boldsymbol{\gamma} = \boldsymbol{0}$ or $\boldsymbol{M} = \boldsymbol{I}_n$) and from below by $\boldsymbol{0}$ (when $\bar{\boldsymbol{Z}}\boldsymbol{\gamma} = \boldsymbol{Z}\boldsymbol{\Pi}$ or $\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{0}$).

Proposition 1 (see appendix A.1) presents the Bekker-type asymptotic distribution of $\boldsymbol{b}$ for $K_n/n \to \alpha_K \in [0,1]$, $r_n/n \to \alpha \in [0,1]$ and $\boldsymbol{E}_n \to \boldsymbol{E}$ as n grows.

Proposition 1. *Let Assumptions A1, A2 and A3 hold with $\psi_n = n$. In addition, if*

$$\sqrt{n}\left(\frac{r_n}{n} - \alpha\right) = o(1) \qquad \textit{and} \qquad \sqrt{n}\left(\boldsymbol{E}_n - \boldsymbol{E}\right) = o(1), \tag{6}$$

the asymptotic distribution of $\boldsymbol{b}$ is given by

$$\sqrt{n}(\,\boldsymbol{b} - \boldsymbol{\beta} - \alpha \boldsymbol{H}\boldsymbol{\rho}\,) \xrightarrow{d} N(\,\boldsymbol{0}, \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}\,), \tag{7}$$

where

$$\boldsymbol{H} = (\boldsymbol{\Psi} - \boldsymbol{E} + \alpha\boldsymbol{\Omega})^{-1}, \tag{8a}$$

$$\boldsymbol{W} = \bar{\sigma}(\boldsymbol{\Psi} - \boldsymbol{E}) + \alpha(\sigma_{uu} - \alpha\boldsymbol{\rho}'\boldsymbol{H}\boldsymbol{\rho})\boldsymbol{\Omega} - \alpha(\boldsymbol{\Psi} - \boldsymbol{E})\boldsymbol{H}\boldsymbol{\rho}\boldsymbol{\rho}'\boldsymbol{H}(\boldsymbol{\Psi} - \boldsymbol{E}), \tag{8b}$$

$$\bar{\sigma} = \sigma_{uu} - \alpha\boldsymbol{\rho}'\boldsymbol{H}\boldsymbol{\rho} - \alpha\boldsymbol{\rho}'\boldsymbol{H}(\boldsymbol{\Psi} - \boldsymbol{E})\boldsymbol{H}\boldsymbol{\rho}\,. \tag{8c}$$

Corollary 1 (IVE). *The estimator $\boldsymbol{b}_K$ that uses $\boldsymbol{Z}$ as instruments has $\boldsymbol{E}_n = \boldsymbol{0}$ and $\alpha = \alpha_K$.*

For the IVE $\boldsymbol{b}_K$, the conditions in (6) simplify to $K_n/n = \alpha_K + o(1/\sqrt{n}\,)$ and the result is identical to Bekker's. It can be seen that this estimator is inconsistent unless $\alpha_K = 0$, which for $\psi_n = n$ corresponds to the standard large sample asymptotics based on a fixed $K_n = K$.

The reason for the inconsistency is that the least squares estimation used for the projection of the first stage tends to fit too well, rendering an overfitted $\boldsymbol{P}_K\boldsymbol{X}$ which is still correlated with $\boldsymbol{u}$ if the number of instruments is large. The estimating equations akin to the estimator $\boldsymbol{b}_K$ sets (incorrectly) $\boldsymbol{X}'\boldsymbol{P}_K\boldsymbol{u}/n$ equal to zero. The expectation of this random vector equals to $(K_n/n)\boldsymbol{\rho}$, a quantity that does not vanish asymptotically with many instruments (i.e., $\alpha_K \neq 0$) and the inconsistency of the IVE follows.[2] Alternatively, with an increasing number of instruments, the elements of $\boldsymbol{\Pi}$ become incidental parameters. If $\alpha_K = 0$, the sampling variability in the estimation of $\boldsymbol{\Pi}$ in the first stage can be neglected asymptotically, but if $\alpha_K \neq 0$ this uncertainty passes through the asymptotic moments of $\boldsymbol{b}_K$. The following Corollaries discuss conditions for the consistency of $\boldsymbol{b}$.

Corollary 2 (Consistency). *Suppose $r_n/\sqrt{n} \to \bar{\alpha} < \infty$. Then, $\boldsymbol{b} \xrightarrow{p} \boldsymbol{\beta}$ and*

$$\sqrt{n}(\,\boldsymbol{b} - \boldsymbol{\beta}\,) \xrightarrow{d} N(\,\bar{\alpha}(\boldsymbol{\Psi} - \boldsymbol{E})^{-1}\boldsymbol{\rho}, \sigma_{uu}(\boldsymbol{\Psi} - \boldsymbol{E})^{-1}\,). \tag{9}$$

[2] This observation is the basis for the bias-corrected estimator (16) used in section 4.

Contrary to what happens with $\boldsymbol{b}_K$, the estimator (4) is consistent when $r_n = o(n)$. This is clearly seen as $\boldsymbol{b}$ equates to zero estimating equations with expectation $\mathbb{E}[\boldsymbol{X}'\boldsymbol{P}\boldsymbol{u}/n] = (r_n/n)\boldsymbol{\rho} \to \mathbf{0}$. Thus, the consistency under $r_n = o(n)$ can be interpreted as coming from estimating equations that are well 'centered' in the limit. Furthermore, by comparing (9) to (7), it is interesting to note that the influence of overidentification on the asymptotic variance of $\boldsymbol{b}$ vanishes. This is a way to prevent the incidental parameter phenomenon from arising. Even though a fuller comparison between the mean squared errors of $\boldsymbol{b}_K$ and $\boldsymbol{b}$ may be required, one may suspect that the fact that (4) is consistent as $K_n \to \infty$ would be manifested in a superior finite sample performance.

Corollary 3 ($\sqrt{n}$**-Consistency**). *If* $r_n = o(\sqrt{n})$ *or*

$$r_n = o(\sqrt{K_n}), \tag{A}$$

then $\boldsymbol{b}$ *is* $\sqrt{n}$*-consistent and (7) simplifies to*

$$\sqrt{n}(\boldsymbol{b} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma_{uu}(\boldsymbol{\Psi} - \boldsymbol{E})^{-1}). \tag{10}$$

Hahn (2002) shows that the efficiency bound of estimators of $\boldsymbol{\beta}$ in model (1) is equal to $\sigma_{uu}\boldsymbol{\Psi}^{-1}$, which corresponds to the asymptotic variance of $\boldsymbol{b}_K$ when $\alpha_K = 0$. Remarkably, this bound does not depend on α, which suggests it can be achieved if the rate at which the number instruments grow as $n \to \infty$ is controlled without altering the fit of (1), so the conditions under which the standard large-n asymptotic results hold are restored. Under condition (A), $\boldsymbol{b}$ is not only consistent but the expectation of $\sqrt{n}(\boldsymbol{b} - \boldsymbol{\beta})$ becomes $o(1)$. Also, the asymptotic variance of $\boldsymbol{b}$ in Corollary 3 differs from Hahn's efficiency bound by the limit of the error matrix (5). The smaller this matrix, the closer $\boldsymbol{b}$ becomes to an asymptotically efficient estimator. Corollary 4 follows naturally.

Corollary 4 (**Efficiency**). *If condition (A) is satisfied and*

$$\boldsymbol{E}_n \to \boldsymbol{E} = \mathbf{0}, \tag{B}$$

then $\text{Avar}(\boldsymbol{b}) = \sigma_{uu}\boldsymbol{\Psi}^{-1}/n$.

In other words, if the lower-dimensional IV set provides asymptotically the same fit as the original IV set in predicting the endogenous regressors, then (4) achieves the efficiency bound. In line with our previous discussion, when (A) holds ($\bar{\alpha} = 0$), then $\|\boldsymbol{E}_n\| = o(1)$ becomes sufficient for the asymptotic

efficiency of $\boldsymbol{b}$ with many instruments.

2.2 Many Weak Instruments

Within the many instrument framework, where the concentration parameter is set to grow at a linear rate with the sample size ($\psi_n = n$), one may get an insight into the effects of weak identification when the matrix $\boldsymbol{\Psi}$ is of small magnitude. Thus, the bias of $\boldsymbol{b}_K$ (which is a function of $\boldsymbol{\Psi}^{-1}$) can be sizeable with weak instruments even when the degree of overidentification (as measured by α_K) is small.

Corollaries 2 and 3 indicate that either if $r_n = O(\sqrt{K_n})$ or condition (**A**) hold, the lower-dimensional IVE is consistent regardless of the magnitude of $\boldsymbol{\Psi}$. Next, we enquire whether this estimator remains consistent under an asymptotic sequence proposed by Chao and Swanson (2005), which is especially designed to suit many weak instrument situations. Chao and Swanson consider the case where the concentration parameter grows at a slower rate than the number of instruments, $\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi} = o(K_n)$.

Proposition 2. *Let Assumptions* A1*,* A2 *and* A3 *hold and consider that*

$$\frac{K_n}{\psi_n} \to \infty \qquad \textit{and} \qquad \frac{\sqrt{K_n}}{\psi_n} \to c \geq 0 \qquad \textit{as} \qquad n \to \infty\,. \tag{11}$$

(a) *If* $r_n = O(K_n)$*, then* $\boldsymbol{b} \xrightarrow{p} \boldsymbol{\beta} + \boldsymbol{\Omega}^{-1}\boldsymbol{\rho}$.

(b) *If either (i)* $r_n = O(\sqrt{K_n})$ *and* $c = 0$ *or (ii) condition (**A**) holds,* $\boldsymbol{b} \xrightarrow{p} \boldsymbol{\beta}$.

From Proposition 2(a), the many weak instrument situation is rather unfavourable for $\boldsymbol{b}_K$ and any IVE that does not reduce the dimensionality of the IV enough. The probability limit of such estimators equals the probability limit of the OLS estimator. On the contrary, there are conditions under which the lower-dimensional IVE is consistent in this case as well. It is important to mention that the results in Proposition 2 hold regardless of the rate of growth of K_n vis-à-vis the rate of growth of n, we may have either $K_n = O(n)$ or $K_n = o(n)$. Hence, the IVE may be asymptotically biased due to weak identification even if $K_n/n \to 0$, whereas the lower-dimensional IVE has also the ability to overcome such source of inconsistency.

Chao and Swanson (2005) find that under the sequence in (11) for $c = 0$ estimators such as the limited information maximum likelihood (LIML) and the bias-corrected IVE (BCIVE hereafter)

proposed in Donald and Newey (2001), see equation (16) below, are consistent. This is the same condition in Proposition 2(b) for the consistency of the lower-dimensional IVE with $r_n = O(\sqrt{K_n})$ and is a slightly stronger requirement if (**A**) holds (since consistency is achieved even if $c > 0$).

This is a significant result for the practical relevance of the lower-dimensional IVE. Hahn, Hausman, and Kuersteiner (2004) argue that albeit consistent, the LIML estimator and the BCIVE can perform poorly in finite samples. The reason seems to be their lack of finite sample moments. Consequently, Hahn, Hausman, and Kuersteiner and also Davidson and MacKinnon (2006) suggest avoiding these 'no moments' estimators if the sample size is not too large and when the instruments are weak. The lower-dimensional IVE is just an IVE and as such possesses up to $r_n - G - 2$ moments, and thus it may achieve the performance in terms of bias of the 'no moments' estimators without displaying their unduly volatility due to fat tails. The fulfilment of condition (**A**) is key in this respect (and so are related results as in Proposition 4 below).[3]

3 Principal Components IVE

Given the first stage equation in (1), the choice of $\boldsymbol{C}$ that maximises the variance of the sample moment conditions – thereby minimising the variance of the estimator associated to them – subject to a rank r_n constraint, is

$$\boldsymbol{C}_r = \underset{\boldsymbol{C}}{\operatorname{argmax}} \left\{ \frac{\operatorname{var}(\bar{\boldsymbol{Z}}'\boldsymbol{u})}{n} = \sigma_{uu}\boldsymbol{C}'\boldsymbol{S}\boldsymbol{C} \quad \text{subject to} \quad \boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}_{r_n} \right\}. \tag{12}$$

It is not difficult to verify that the columns of $\boldsymbol{C}_r$ are the eigenvectors associated with the largest r_n eigenvalues of $\boldsymbol{S}$. Therefore, the r_n optimal linear combinations contained in $\bar{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{C}_r$ are the first principal components (PC) of $\boldsymbol{Z}$. Furthermore, it can be shown that $\boldsymbol{C} = \boldsymbol{C}_r$ also minimises over $\boldsymbol{C}$ the norm of the sample 'efficiency loss' as defined in (5), i.e. $\boldsymbol{C}_r = \operatorname{argmin}_{\boldsymbol{C}}\{\|\boldsymbol{E}_n(\boldsymbol{C})\|\}$ (see appendix B.1), as well as many other optimality criteria (see Amemiya, 1966).

Consider the lower-dimensional IVE (4) that uses the first r_n PC of $\boldsymbol{Z}$ as instruments, PCIVE henceforth. Its properties will crucially depend on the determination of r_n. It is well-known that each eigenvalue of $\boldsymbol{S}$ equals the variance of the associated PC, and the sum of all eigenvalues – which equals $\operatorname{tr}(\boldsymbol{S})$ – measures the total variation of the IV set $\boldsymbol{Z}$. Thus, by construction small values of r_n

[3] Furthermore, it can be concluded from the derivations in appendix A.1 that if $r_n = o(\sqrt{\psi_n})$, then $\sqrt{\psi_n}(\boldsymbol{b} - \boldsymbol{\beta})$ converges to the normal distribution in (10). We would need to impose more structure to arrive to more definite conclusions, but this result illustrates the desirable effects of parsimony even in the many weak instrument case.

can easily satisfy the parsimony requirement (**A**) but may fail to fulfill condition (**B**) if not enough variation of the IV set is captured by the PC approximation. Similarly, large values of r_n are likely to satisfy with ease the 'goodness-of-fit' condition (**B**) but may violate (**A**).

The ability of PC to successfully reduce the dimensionality of $\boldsymbol{Z}$, i.e. to deliver r_n small relative to K_n while keeping $\boldsymbol{E}_n = o(1)$, depends on the degree of linear dependence among the variables in $\boldsymbol{Z}$. In other words, some restrictions on the structure of $\boldsymbol{Z}$ need to be imposed for PC to be effective. It is important to note that the amount of correlation among the columns in $\boldsymbol{Z}$ (a quantity that will be denoted by the symbol μ) will be reflected in the pattern of the eigenvalues of $\boldsymbol{S}$ (quantities that will be associated with the symbol λ), though the mapping from correlations to eigenvalues may be untractable for general cases (cf. Silverstein and Choi, 1995).

To illustrate this point suppose, with no loss of generality, that the columns of $\boldsymbol{Z}$ have zero mean and that the diagonal elements in $\boldsymbol{S}$ are equal to one, so $\boldsymbol{S}$ can be thought of as a sample correlation matrix. Furthermore, consider the simple case where $\boldsymbol{S}$ can be described by an equicorrelation matrix of the form $\boldsymbol{S} \simeq (1-\theta)\boldsymbol{I}_K + \theta \boldsymbol{J}_K$, where $\boldsymbol{J}_K$ is a $K_n \times K_n$ matrix full of ones and θ is a constant. In this case, the columns of $\boldsymbol{Z}$ can be interpreted as belonging to one 'group' in terms of their correlation structure, with the correlation between any two columns of $\boldsymbol{Z}$ being approximately equal to $\mu \simeq \theta/(1-\theta)$. The maximum eigenvalue of $\boldsymbol{S}$ is $\lambda_1 \simeq K_n\theta + (1-\theta)$ whereas the remaining are $\lambda_k \simeq 1-\theta$ for $1 < k \leq K_n$. Importantly, whereas all eigenvalues of $\boldsymbol{S}$ measure to some extent the degree of association among the columns of $\boldsymbol{Z}$, only the maximum is scaled by the dimension of $\boldsymbol{S}$: $\lambda_1 = O(K_n)$ if $\theta \neq 0$ and $\lambda_k = O(1)$ for $k > 1$. Note that if $\theta = 0$ (or $\mu = 0$), then $\lambda_k = O(1)$ for all k, a situation where no significant association is found among the columns of $\boldsymbol{Z}$ and thus PC will be ineffective as a dimensionality reduction technique.

Similarly, consider now the case where $\boldsymbol{Z}$ can be divided into two nearly orthogonal blocks $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$, i.e. $\boldsymbol{Z}_1{}'\boldsymbol{Z}_2 \simeq \boldsymbol{0}$, of dimensions $n \times s_1$ and $n \times s_2$ ($s_1 + s_2 = K_n$), respectively, such that $\boldsymbol{S}$ is block-diagonal with blocks $\boldsymbol{S}_1 \simeq (1-\theta_1)\boldsymbol{I}_{s_1} + \theta_1 \boldsymbol{J}_{s_1}$ and $\boldsymbol{S}_2 \simeq (1-\theta_2)\boldsymbol{I}_{s_2} + \theta_2 \boldsymbol{J}_{s_2}$ for some constants θ_1 and θ_2. For $f = \{1, 2\}$, the eigenvalues of $\boldsymbol{S}$ are $\lambda_{1f} \simeq s_f\,\theta_f + (1-\theta_f) = O(s_f)$ and $\lambda_{kf} \simeq 1-\theta_f = O(1)$ for $1 < k \leq s_f$. In this case, $\boldsymbol{Z}$ contains variables belonging to two groups, and therefore at most two eigenvalues are scaled by the dimensions of the groups. Define $\mu = \max\{\theta_1/(1-\theta_1), \theta_2/(1-\theta_2)\}$ as the maximum correlation between any two columns of $\boldsymbol{Z}$. At least one eigenvalue of $\boldsymbol{S}$ grows for increasing s_1 or s_2 if either $\theta_1 \neq 0$ or $\theta_2 \neq 0$ or both, i.e. if $\mu \neq 0$, whereas $\lambda_k = O(1)$ for all k when $\mu = 0$. Following this line of reasoning, in more general cases $\boldsymbol{Z}$ may contain several groups of

variables, in which case the existence of a group is signalled by a 'large' $O(s_n)$ eigenvalue (associated with non-zero correlations) and several 'small' $O(1)$ eigenvalues. The ultimate goal of PC is to rotate the data to ease the separation of the linear combinations of $\boldsymbol{Z}$ associated with the large eigenvalues.

We analyse two alternative and somehow complementary generalisations of the group structure described above, and the workings of the PC technique as a means of achieving conditions (**A**) and (**B**) in each case. First, we focus on the eigenvalues of $\boldsymbol{S}$ directly, and then we study the problem by bounding the maximum correlation among the variables in $\boldsymbol{Z}$.

3.1 Analysis Based on Eigenvalues

In this section, the following assumption on $\boldsymbol{S}$ is made:

A5 Consider two non-decreasing sequences $\{s_n\}$ and $\{m_n\}$ such that $s_n = O(K_n^{\varepsilon_s})$ for $0 \leq \varepsilon_s \leq 1$ and $m_n = O(K_n^{\varepsilon_m})$ for $0 \leq \varepsilon_m < 1$. Let $\lambda_k(\boldsymbol{S})$ be the k-th largest eigenvalue of $\boldsymbol{S}$. Then,

$$\lambda_k(\boldsymbol{S}) = O(s_n) \quad \text{for} \quad k = 1, \ldots, m_n\,, \qquad \text{and} \qquad \lambda_k(\boldsymbol{S}) = O(1) \quad \text{for} \quad k = m_n + 1, \ldots, K_n\,.$$

It is important to mention that even though this assumption rules out certain types of instruments (for instance, mutually exclusive dummy variables), it is a flexible formulation which is satisfied by several covariance patterns common in econometrics (e.g., factor structures, random effects, mixture models, among others).

The parameter ε_s is a measure of the strength of the linear dependence between the newly added instruments and the original IV set as $n \to \infty$ and $K_n \to \infty$. The parameter ε_m controls the number of underlying groups contained in $\boldsymbol{Z}$. In a leading case, for instance the two-group example described above, $s_n = O(K_n)$ hence $\lambda_{1f} = O(K_n)$ for at most $m_n = O(1)$ different f. Thus, if $\varepsilon_m = 0$ and $\varepsilon_s = 1$ the underlying group structure observed in the sample is preserved as n and K_n grow. By letting $\varepsilon_s < 1$ we allow the new instruments to have a weaker association with those in the original IV set, whereas the case $0 < \varepsilon_m < 1$ indicates that $\boldsymbol{Z}$ may contain an increasing number of groups that grows at a slower rate than K_n as $K_n \to \infty$. Therefore, under Assumption A5, the increases in the sample size and the number of instruments required to perform the many instrument asymptotic approximations need not to preserve the group structure in the sample.

Consider the PCIVE that uses the first r_n PC of $\boldsymbol{Z}$ as instruments, with $r_n = O(K_n^{\varepsilon_r})$ for $0 < \varepsilon_r < 1$. Proposition 3 (appendix B.2) gives conditions on ε_r, ε_s and ε_m for the PCIVE to satisfy (**B**):

Proposition 3. *Let Assumptions* A4 *and* A5 *hold. Let also* $r_n = O(K_n^{\varepsilon_r})$ *for* $0 < \varepsilon_r < 1$. *If* $1 - 2\varepsilon_s < \varepsilon_m < \varepsilon_r$, *then* $\|\boldsymbol{E}_n\| = o(1)$ *as* $K_n \to \infty$. *Otherwise,* $\|\boldsymbol{E}_n\| = O(1)$.

The main conclusion drawn from Proposition 3 is the intuitive result that PC is effective to satisfy condition (**B**) for small values of ε_m, i.e. either a finite or slowly increasing number of underlying groups in the IV set, and for relatively large values of ε_s. In the leading case where $\varepsilon_m = 0$ and $\varepsilon_s = 1$ (more precisely, $1/2 < \varepsilon_s \leq 1$), *any* choice of $\varepsilon_r > 0$ and hence of r_n delivers the desired result. Moreover, if ε_r is set such that $\varepsilon_r < 1/2$, then condition (**A**) is also satisfied.

By the same token, $\|\boldsymbol{E}_n\| = O(1)$, a violation of (**B**), if the number of groups grows faster than the number of selected PC $\varepsilon_m > \varepsilon_r$, since in the limit the variation of $\boldsymbol{Z}$ captured by the r_n first PC will not fully represent the original IV set; or when $\varepsilon_s < (1 - \varepsilon_m)/2$, since under these circumstances the association among the columns of $\boldsymbol{Z}$ is too weak to separate them into a small number of groups. Cases where $\varepsilon_s \simeq 0$ or $\varepsilon_m \simeq 1$ imply $\lambda_k(\boldsymbol{S}) \simeq O(1)$ for *all* k, which occur for instance when $\boldsymbol{Z}$ contain a covariance structure resembling that of mutually exclusive dummy variables. Needless to say, PC would fail to meet condition (**B**), regardless on whether (**A**) is satisfied or not.

As a practical implication, Proposition 3 suggests performing an analysis on the eigenvalues of $\boldsymbol{S}$ prior to the IV estimation, as to determine likely values for ε_m and ε_s. If the IV set can be viewed as being represented closely enough by $\varepsilon_m \simeq 0$ and $\varepsilon_s > 1/2$, then the conclusions in section 2 suggest that the PCIVE will deliver finite sample improvements over alternative IVE.

3.2 Analysis Based on Correlations

Next, we take a different viewpoint to evaluate the usefulness of PC for IV estimation. In keeping with standard approaches in multivariate analysis, we focus the attention on the correlations among the columns of $\boldsymbol{Z}$. Since our analysis is based on large samples, we make no distinction between the population variance of the typical row of $\boldsymbol{Z}$ and its sample counterpart, $\boldsymbol{S}$. We work with the following assumption:

A6 The typical row of $\boldsymbol{Z}$, $\boldsymbol{z}_i$, is drawn from a distribution with zero mean and covariance matrix $\boldsymbol{S} = \boldsymbol{I}_{K_n} + \mu\boldsymbol{\Upsilon}$, where $0 < \mu < 1$ is a constant and where $\boldsymbol{\Upsilon}$ is a symmetric matrix with zero diagonal entries and non-diagonal entries within the interval $[-1, 1]$.

The parameter μ is the absolute value of the largest correlation between any two elements of $\boldsymbol{z}_i$.

The determination of r_n can be made through a heuristic rule designed to satisfy (**A**) and (**B**). This is inspired by usual practices in the selection of those PC regarded as important in explaining the main features of the original data, and are conveniently based on the eigenvalues of $\boldsymbol{S}$. The proposed rule dictates that only those PC accounting individually for at least a given share of the total variation of the IV are to be retained. Formally,

Retention Rule. *Retain the principal components of $\boldsymbol{Z}$ associated with eigenvalues greater than $K_n^{-\delta}\text{tr}(\boldsymbol{S})$.*

Hence, the problem of determining r_n translates into the selection of a suitable value for δ.[4] If $\delta = 1$, this criterion boils down to the so-called *Kaiser rule*, which is a useful benchmark due to its popularity (Jolliffe, 2002, ch. 6): only those PC with eigenvalues greater than the average are worth selecting because such PC summarise individually more information than any single original variable. Indeed, as mentioned earlier, if a subset of variables in the original IV can be represented as coming from a single group, then this subset will produce a single eigenvalue of $\boldsymbol{S}$ that is greater than the average $\text{tr}(\boldsymbol{S})/K_n$ and the remaining eigenvalues below this threshold.[5]

Proposition 4. *Let Assumptions A4 and A6 hold, and the Retention Rule is used:*

(a) *If $\delta > 1/2$, then condition (**B**) is met.*

(b) *For μ close to zero, there exists a scalar $\bar{\delta} \leq 1$ such that using the Retention Rule with $\delta < \bar{\delta}$ satisfies (**A**).*

The condition $\delta > 1/2$ in Proposition 4(a) implies an increasing r_n. Establishing the precise rate at which r_n grows under the proposed rule is a complicated problem. Even so, although the asymptotic distribution of the eigenvalues of $\boldsymbol{S}$ for $K_n/n \to \alpha$ is known to exist, no closed-form is available unless for the special case when $\boldsymbol{S} = \boldsymbol{I}_{K_n}$ (see Silverstein and Choi, 1995, and the references therein). In face of this, Proposition 4(b) present a *necessary* condition to regulate r_n as K_n grows.

If the columns of $\boldsymbol{Z}$ are uncorrelated ($\mu = 0$), then (**A**) cannot be met using PC. The number of PC in such situation is $r_n = K_n$ and the PC set is exactly $\boldsymbol{Z}$. At the other end, if the columns of $\boldsymbol{Z}$ are highly correlated, condition (**A**) is satisfied trivially as only a few PC can represent the entire IV set. The parameter μ in Proposition 4(b) places an upper bound to the correlations among the columns

[4] Strictly speaking, if $\tilde{r}_n$ is the number of components retained by this rule, then $r_n = \max\{\tilde{r}_n, G\}$ components are to be used to identify $\boldsymbol{\beta}$. However, since G is fixed the lower bound for r_n can be safely ignored in the asymptotic analysis.

[5] Another simple criterion for estimating the number of nontrivial PC is to include components that account jointly for a given proportion of the total variance. The proposed rule can be easily adapted to such a criterion.

of $\boldsymbol{Z}$. This bound is *close to zero* (the proof of Proposition 4(b) in appendix B.4 relies on this fact), stating that (**A**) can be satisfied even when the large IV set displays a small degree of correlation.

Remarkably, Proposition 4(b) rules out selecting the PC with the Kaiser rule ($\delta = 1$) when μ is small. This finding resembles the critique that this criterion tends to retain too many components (Jolliffe, 2002, ch. 6). Indeed, if $\boldsymbol{S}$ is nearly a diagonal matrix ($\mu \simeq 0$), it is expected the Kaiser rule to retain about half of the PC, making $r_n = O(K_n)$ and thus violating (**A**). However, for $\delta < 1$ the threshold in the Kaiser rule, i.e. $\text{tr}(\boldsymbol{S})/K_n$, is increased by a factor of $K_n^{1-\delta}$, therefore increasingly penalising the rate at which the PC are retained and promoting parsimony. Proposition 4(b) states that a value $\delta < 1$ exists such that (**A**) is reestablished as long as $\mu \neq 0$.

If $\bar{\delta}$ is viewed as an upper bound for δ, one may conjecture that it depends positively on μ, so an increase in μ would bring $\bar{\delta}$ closer to one. Recall that the scope of Proposition 4(b) is local, for small values of μ, and so it may be the case that $\delta \geq 1$ satisfies condition (**A**) for relatively large values of μ. Besides, since, as found in Silverstein and Choi (1995), the asymptotic distribution of the eigenvalues of $\boldsymbol{S}$ depends on α, so does $\bar{\delta}$. One may intuitively expect this relationship to be inverse, as for a given μ a larger IV set (associated with larger α) is likely to require a larger 'penalty' term $K_n^{1-\bar{\delta}}$ (and so a smaller $\bar{\delta}$) to achieve (**A**).

We explore these conjectures numerically by performing a small simulation experiment to evaluate the relationship $\bar{\delta} = \bar{\delta}(\mu, \alpha)$. We generate random matrices $\boldsymbol{Z}$ of dimension $n \times K_n$, with each row drawn independently from a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{S}$ as described in Assumption A6. The matrix $\boldsymbol{\Upsilon}$ is symmetric with zero diagonal entries, and its non-zero entries are drawn from a uniform distribution with support $[-1, 1]$. Hence, this setup implies that the elements of $\boldsymbol{z}_i$ are uncorrelated 'on average', though some correlation will arise within a particular draw. As mentioned, the parameter μ controls the strength of such correlations. Having generated $\boldsymbol{Z}$ we next seek for the maximum value of δ that makes the number of retained components equal to $r_n = \lfloor \sqrt{K_n} \rfloor$, where $\lfloor A \rfloor$ rounds A down to the nearest integer. This specification is probably the most natural way to interpret condition (**A**) in practice, though it is certainly arbitrary. We perform the evaluation for an equidistant 20-point grid $\mu \in [0.01, 0.20]$ and selected values of α, and report the average of $\bar{\delta}$ over 10,000 replications.[6]

[Insert Figure 1 here]

[6] We also use $n = 300$. However, it is important to note that we found this function to depend on the ratio α regardless of the absolute values of K_n or n. Furthermore, the results in Figure 1 where remarkably robust to the number of replications used in the numerical evaluation and to the distributional assumption made about the rows of $\boldsymbol{Z}$.

Figure 1 shows the results. For a given α the bound $\bar{\delta}$ is increasing in μ, whereas for a given μ it is decreasing in α. For a given α, values of δ below the corresponding line (but above 1/2) render a PCIVE satisfying conditions (**A**) and (**B**). These results confirm the statement of Proposition 4(b) that $\bar{\delta} \leq 1$ for small μ. Also, the positive slopes of $\bar{\delta}(\mu, \alpha = \bar{\alpha})$ suggest that as μ increases beyond what may be considered small, the bound for δ exceeds one for some values of α. This is particularly true for very small values of α. Finally, a value of $\delta \simeq 0.8$ seems to be a safe choice for (**A**) to hold in parameterisations of interest.

4 Monte Carlo Simulations

Next we carry out a simulation study to assess the finite sample behaviour of the PCIVE relative to other IVE, and under different PC retention criteria. To this end, we have slightly adapted the designs in Hahn, Hausman, and Kuersteiner (2004) and in Davidson and MacKinnon (2006) to suit our purposes.

4.1 Data Generation

We consider the single regressor case

$$y_i = x_i \beta + u_i \qquad \text{and} \qquad x_i = \boldsymbol{z}_i{}' \boldsymbol{\pi} + v_i \qquad \text{with} \qquad \beta = 1\,, \tag{13}$$

where u_i and v_i are drawn independently from a multivariate normal distribution with zero means, $\text{var}(u_i) = \text{var}(v_i) = 1$ and $\text{corr}(u_i, v_i) = \rho$.

The vectors of instruments $\boldsymbol{z}_i$ and first-stage coefficients $\boldsymbol{\pi}$ are partitioned as $\boldsymbol{z}_i = (\boldsymbol{z}_i^*, \boldsymbol{z}_i^a)$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_*{}', \mathbf{0}_a{}')'$. The first block $\boldsymbol{z}_i^*$ is a set of $\dim(\boldsymbol{z}_i^*) = K_*$ relevant instruments, whereas the second contains $\dim(\boldsymbol{z}_i^a) = a$ irrelevant instruments. Thus, the data generation process of x_i involves redundant instruments in the sense of Breusch, Qian, Schmidt, and Wyhowski (1999) and Hall, Inoue, Jana, and Shin (2007). We take iid draws $\boldsymbol{z}_i^* \sim N(\mathbf{0}, \boldsymbol{S}^*)$ where, as in Assumption A6, $\boldsymbol{S}_* = \boldsymbol{I}_{K_*} + \mu \boldsymbol{\Upsilon}$ with $\boldsymbol{\Upsilon}$ containing draws from a uniform distribution (we set the conservative value of $\mu = 0.10$), and $\boldsymbol{z}_i^a \sim N(\mathbf{0}, \boldsymbol{I}_a)$ drawn independently from $\boldsymbol{z}_i^*$. The uncorrelatedness between $\boldsymbol{z}_i^*$ and $\boldsymbol{z}_i^a$ implies that PC will be effective to reduce the dimensionality of $\boldsymbol{z}_i$ only under a stringent Retention Rule.

It is convenient to describe the properties PC as instruments in this setup. Let the matrix of instruments be partitioned as $\boldsymbol{Z} = [\, \boldsymbol{Z}^* : \boldsymbol{Z}^a \,]$, so its sample correlation matrix will be approximately equal to $\boldsymbol{S} = \text{diag}(\boldsymbol{S}^*, \boldsymbol{I}_a)$. As discussed in section 3, matrix $\boldsymbol{S}^*$ contains, say, s eigenvalues above 1 and the remaining $K^* - s$ eigenvalues below unity. Thus, the spectral decomposition of $\boldsymbol{S}$ will be $\boldsymbol{S} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'$ with $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{I}_a, \boldsymbol{\Lambda}_2)$, where $\boldsymbol{\Lambda}_1$ is the $s \times s$ diagonal matrix that collects the s eigenvalues of $\boldsymbol{S}$ above unity and $\boldsymbol{\Lambda}_2$ is the diagonal matrix whose entries are the $K^* - s$ eigenvalues of $\boldsymbol{S}$ below unity. It follows from the equality $\boldsymbol{\Lambda} = \boldsymbol{C}'\boldsymbol{S}\boldsymbol{C}$ that the $K \times K$ matrix of eigenvectors (the so-called 'loadings') has the form

$$\boldsymbol{C} = \begin{bmatrix} \bar{\boldsymbol{C}}_{K^* \times s} & \boldsymbol{0}_{K^* \times a} & \hat{\boldsymbol{C}}_{K^* \times (K^* - s)} \\ \boldsymbol{0}_{a \times s} & \boldsymbol{I}_a & \boldsymbol{0}_{a \times (K^* - s)} \end{bmatrix} \quad \text{therefore} \quad \boldsymbol{Z}\boldsymbol{C} = [\, \boldsymbol{Z}^*\bar{\boldsymbol{C}} \,:\, \boldsymbol{Z}^a \,:\, \boldsymbol{Z}^*\hat{\boldsymbol{C}} \,]. \tag{14}$$

The PC of $\boldsymbol{Z}$ used as instruments are $\bar{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{C}_r$, the first r columns of $\boldsymbol{Z}\boldsymbol{C}$. From (14) it is clear that if $r \leq s$, $\bar{\boldsymbol{Z}}$ will contain linear combinations of the relevant instruments $\boldsymbol{Z}^*$ only, whereas if $r > s$ the IV set $\bar{\boldsymbol{Z}}$ also includes $r - s$ irrelevant instruments. Thus, retaining too many PC will have adverse effects on the bias of the PCIVE because of overidentification and instrument weakness.

On the other hand, the population R^2 of the equation for x_i in (13), which measures the strength of the IV set, is

$$R^2 = \frac{\boldsymbol{\pi}_*{}'\boldsymbol{S}_*\boldsymbol{\pi}_*}{1 + \boldsymbol{\pi}_*{}'\boldsymbol{S}_*\boldsymbol{\pi}_*}. \tag{15}$$

By setting $\boldsymbol{\pi}_* = (\pi, \pi, \ldots, \pi)'$, then π and hence $\boldsymbol{\pi}_*$ can be determined from a particular choice of R^2. In this setup, Hahn's efficiency bound is given by $\sigma_{uu}/\Psi = (1 - R^2)/R^2$.

4.2 Estimators

We consider the standard IVE $\boldsymbol{b}_K$, where the IV set include both relevant and irrelevant instruments ($\boldsymbol{z}_i$). As a useful benchmark we also consider an IVE, dubbed IVE*, that includes only relevant instruments in the estimation ($\boldsymbol{z}_i^*$). In a sense this estimator is unfeasible because the data generating process and hence the exact set of relevant instruments are not known in practice. However, as discussed in the introduction, model selection procedures may render IVE* as a final choice.

In addition, we also include a bias-corrected version of the IVE (BCIVE),

$$\boldsymbol{b}_{\text{BC}} = (\boldsymbol{X}'\boldsymbol{P}_K\boldsymbol{X} - \hat{\alpha}\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{P}_K\boldsymbol{y} - \hat{\alpha}\boldsymbol{X}'\boldsymbol{y}), \tag{16}$$

where $\hat{\alpha} = (K_n - G - 1)/n$. This estimator is consistent even when the instruments are weak in the sense of section 2.2, and shares many desirable finite sample properties with more complex estimators such as the LIML estimator, while being easier to compute. Besides, it has proven to work well in other simulation studies (Donald and Newey, 2001; Kapetanios, 2006) and in general constitutes an attractive choice in practice.

Finally, following the results in section 3.2 we consider two PCIVE. When $\delta = 1$ the PC selection rule becomes the Kaiser rule, which as discussed has the tendency of retaining (many) more components than necessary and is likely to violate condition (**A**). Such an estimator is denoted PCIVE_1. By virtue of the results in Figure 1, we also consider the case when $\delta = 0.8$, which raises the threshold in the selection rule by a factor of $K_n^{0.2}$ with respect to the Kaiser rule. We call this estimator $\text{PCIVE}_{0.8}$. It is important to mention that for a given sampling scenario simulated, PCIVE_1 tended to retain about half of the K available PC, whereas $\text{PCIVE}_{0.8}$ often selected only from 3 to 7.

4.3 Experiments and Results

The aim of the Monte Carlo study is to be as comprehensive and illustrative as possible within some reasonable bounds. We run a total of 500 experiments, each using $50,000$ replications, that are divided according to which parameter, besides n and a, varies:

1. Set $K_* = 10$ and $R^2 = 0.10$, and evaluate ρ at each of 50 equidistant values from 0.10 to 0.90.
2. Set $K_* = 10$ and $\rho = 0.90$, and evaluate R^2 at each of 50 equidistant values from 0.01 to 0.20.
3. Set $\rho = 0.90$ and $R^2 = 0.10$, and perform simulations for all 25 values of K_* from 6 to 30.

Each experiment is performed four times for $(n, a) = \{(100, 10), (100, 30), (300, 10), (300, 30)\}$ (the result of further experiments with different values for n and a are available upon request). Note that when one parameter varies, the others are set to values that make the estimation of β challenging. In particular, $K_* = 10$ implies significant finite sample biases for IVE and IVE*.

As opposed to the IVE (and by extension to the PCIVE), the BCIVE is known not have moments so its empirical distribution over replications displays fat tails, making summary statistics such as the mean squared error meaningless. For this reason, we report results on the median absolute error

(MAE, henceforth).[7] To save space we do not discuss the results on bias as they were as expected: IVE performed extremely poor due to the presence of $\boldsymbol{z}_i^a$; IVE* does better but the benchmark value of $K_* = 10$ and the relatively small samples used induced some noticeable finite sample bias; BCIVE reported almost no (median) bias across experiments; and $\text{PCIVE}_{0.8}$ did almost as well as the BCIVE. In fact, the main source of MAE for the IVE, IVE* and $\text{PCIVE}_{0.8}$ is its bias, the variances of these estimators being relatively small, whereas for the BCIVE the main source is its variance. The PCIVE_1 is a compromise between the IVE and the $\text{PCIVE}_{0.8}$. The results on the MAE are displayed in Figures 2 to 4. In all cases, the vertical scales are comparable only between panels with the same value of n, as increasing the sample size improves the performance of all estimators.

[Insert Figure 2 here]

We observe in Figure 2 that the MAE of the IVE, IVE* and PCIVE_1 increases quickly with ρ. Recall that since $K_* = 10$, the number of instruments used in the estimation (but in the case of IVE*) is $K = 20$ when $a = 10$ and $K = 40$ when $a = 30$. The performance of PCIVE_1 is close to that of the IVE* only when $a = 10$, so the complete IV set is not too polluted with irrelevant instruments. This illustrates that even though the dimensionality of the IV set is reduced and hence some improvements are observed with respect to the IVE, this PCIVE still exhibits the finite sample problems of the IVE, as expected from the findings in Figure 1. Additionally, it is interesting to note that the PC technique renders an estimator that performs closely to what a model selection procedure would produce, insofar as the IV set is not too weak (the PCIVE_1 deteriorates dramatically when $a = 30$, an effect that is also clearly observed in Figure 4 below).

The behaviour of the BCIVE is different. Its MAE increases with ρ at a much lower rate than the IVE*. The performance of the BCIVE is poor when $n = 100$ and improves relatively more than the others when $n = 300$, even though in this last case the BCIVE seems to be far more volatile. Conversely, the MAE of the $\text{PCIVE}_{0.8}$ is quite insensitive to ρ and increases little with a. In a comparison between IVE* and this PCIVE, we see that either dominates the other for all values of ρ. The $\text{PCIVE}_{0.8}$ dominates the IVE* when identification becomes difficult, i.e. for values of ρ higher than, say, 0.40 or 0.50. For smaller values, IVE* is dominant. However, since the IVE* deteriorates rapidly as ρ increases, the gap between the PCIVE and the IVE* is larger when the former performs better than when the IVE* dominates.

[7] Of course, for the IVE and the PCIVE the results using the root mean squared error are almost identical.

[Insert Figure 3 here]

In Figure 3 we change the value of R^2 setting $\rho = 0.90$, and hence the results lie in the area in which the $\text{PCIVE}_{0.8}$ outperforms all other estimators. Note that this is true for all the values of R^2 considered, though the differences are less clear as the IV become stronger (as R^2 approaches 0.2). The improvements brought by the $\text{PCIVE}_{0.8}$ can be sizeable when the IV set is very weak, and again it is observed that the performance of this estimator worsens only slightly with a larger a. This is a consequence of the consistency result in Proposition 2.

[Insert Figure 4 here]

Figure 4 (p. 34) shows the dependence of the MAE on the number of overidentifying restrictions. What is remarkable is the behaviour of the $\text{PCIVE}_{0.8}$. Whereas the MAE in other estimators increase with the degree of overidentification, in a lesser extent for the case of the BCIVE, it decreases for the $\text{PCIVE}_{0.8}$ and tends to stabilise for $K_* \geq 15$. In our setup, an increase in K_* causes two effects. Firstly, for a given a, it reduces the proportion of irrelevant instruments so the IV set becomes stronger as a whole. Secondly, it further overfits the first stage regression which translates into larger finite sample biases. The first effect dominates for the $\text{PCIVE}_{0.8}$, whereas the second is most important for the other estimators. The PC obtained from stronger IV would deliver an estimator with low variance, and almost unbiased when the number of PC is small compared to the number of IV. This is achieved with $\delta = 0.8$.

4.4 Back to the Retention Rule

In Figure 4 the MAE of PCIVE_1 runs in parallel with the MAE of the IVE. This is a manifestation that for the low value of μ in the simulations, $r_n = O(K_n)$ when $\delta = 1$. To illustrate the connection between the Retention Rule as computed in Figure 1 and the performance of the PCIVE, we run an additional set of experiments whose results are depicted in Figure 5 (p. 34). We set $n = 200$ and $a = 0$ and vary the value of K_* while computing various PCIVE for different values of δ and μ. To ease visualisation the results on the IVE and IVE* are omitted, but the results of the BCIVE – that do not depend on μ – are included for reference.

[Insert Figure 5 here]

We observe that for small values of μ (0.05 in the Figure), values for δ of about 0.90 or higher give estimators that at some stage become sensitive to the degree of overidentification. An increase in μ (to 0.15) widens the range of values δ can take while delivering a PCIVE with good sampling properties. Indeed, the increase in μ betters the performance of all PCIVE in Figure 1, and as a result the behaviour of the estimator with $\delta = 0.9$ gets closer to that of $\delta = 0.8$.

5 Concluding Remarks

Previous research (cf. Hahn, 2002) has established that within the linear IV framework in the presence of many instruments, an IVE coming from a parsimonious specification that does not sacrifice the explanatory power of the system using the full set of instruments may be asymptotically optimal.

This paper has explored the use of principal components as a simple mean of achieving such an estimator. In particular, we proposed a rule to select the appropriate components to be used as instruments, in order to deliver an IVE that is consistent and asymptotically efficient when both the number of instruments and the sample size go to infinity. The PCIVE is also consistent even if the many instruments are weak in the sense of Chao and Swanson (2005). The curves depicted in Figure 1 provide a guide to implement the rule in practice. Simulation evidence shows that the PCIVE performs well in many circumstances of empirical interest and that it can mark a significant improvement over alternative estimators. Moreover, the proposed estimator is free from the 'no moments' problem that some alternatives display.

It is important to recall that PC exploits the relationship between the instruments alone, without reference to their correlation with the endogenous variables. Given the asymptotic sequences used in the theoretical analysis, and also the data generating process in our simulations, it can be argued that an implicit assumption throughout the paper has been that these correlations are evenly distributed across instruments: valid instruments are "equally relevant". Hence, the performance of the PCIVE may be affected in situations where there is a clear ranking among instruments based on relevance. A promising route for further research would be to consider such setup and to analyse the use of canonical correlations between $\boldsymbol{Z}$ and $\boldsymbol{X}$ (instead of PC of $\boldsymbol{Z}$), in the spirit of Hall and Peixe (2003) and Hall, Inoue, Jana, and Shin (2007), within a many instrument framework.

Finally, our analysis has ruled out cases where a set of mutually exclusive dummy variables

are used as instruments, since in this case the instruments are orthogonal and the use of PC as a variable reduction technique is futile. Nonetheless, the use of such instruments is often encountered in applications and thus it may be fruitful to extend the methods discussed here to such situations.

A Asymptotics

This appendix presents proofs of Propositions 1 and 2. The following lemma is an adaptation of Lemma 2 in Bekker (1994, p. 678), which is based on the moment generating function of Wishart matrices:

Lemma 1. *Let $\boldsymbol{U} = \bar{\boldsymbol{U}} + \boldsymbol{U}^*$, where $\boldsymbol{U}^*$ is an $n \times p$ matrix whose rows are iid normally distributed with zero mean and nonsingular covariance matrix $\boldsymbol{\Xi}$ and $\bar{\boldsymbol{U}} = \mathbb{E}[\,\boldsymbol{U}\,]$ is a fixed $n \times p$ matrix. Also, let $\boldsymbol{d}$ be a $p \times 1$ fixed vector and $\boldsymbol{P}$ be an $n \times n$ fixed idempotent matrix of rank r_n. Then,*

$$\mathbb{E}[\,\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}\,] = \boldsymbol{R}\boldsymbol{d} + r_n\boldsymbol{\Xi}\boldsymbol{d}\,, \tag{A1a}$$

$$\mathrm{var}(\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}) = \boldsymbol{d}'\boldsymbol{\Xi}\boldsymbol{d}(\boldsymbol{R} + r_n\boldsymbol{\Xi}) + \boldsymbol{d}'\boldsymbol{R}\boldsymbol{d}\,\boldsymbol{\Xi} + r_n\boldsymbol{\Xi}\boldsymbol{d}\boldsymbol{d}'\boldsymbol{\Xi} + \boldsymbol{\Xi}\boldsymbol{d}\boldsymbol{d}'\boldsymbol{R} + \boldsymbol{R}\boldsymbol{d}\boldsymbol{d}'\boldsymbol{\Xi}\,, \tag{A1b}$$

where $\boldsymbol{R} = \bar{\boldsymbol{U}}'\boldsymbol{P}\bar{\boldsymbol{U}}$. Furthermore, consider a non-decreasing sequence θ_n such that $\theta_n \to \infty$ as $n \to \infty$. If $\boldsymbol{R}/\theta_n$ and r_n/θ_n converge as $\theta_n \to \infty$ so that $\mathrm{var}(\,\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}\,)/\theta_n \to \boldsymbol{W}$*,*

$$\frac{\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d} - \mathbb{E}[\,\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}\,]}{\sqrt{\theta_n}} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{W})\,. \tag{A2}$$

To apply the Lemma to the simultaneous equations model (1), let $\boldsymbol{U} = (\boldsymbol{y}\,, \boldsymbol{X})$, so that $\bar{\boldsymbol{U}} = \boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}\,, \boldsymbol{I}_G)$ and $\boldsymbol{U}^* = (\boldsymbol{u} + \boldsymbol{V}\boldsymbol{\beta}\,, \boldsymbol{V})$. From (2), the covariance matrix of the typical row of $\boldsymbol{U}^*$ is

$$\boldsymbol{\Xi} = \begin{bmatrix} \sigma_{uu} + 2\boldsymbol{\rho}'\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta} & \boldsymbol{\rho}' + \boldsymbol{\beta}'\boldsymbol{\Omega} \\ \boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\beta} & \boldsymbol{\Omega} \end{bmatrix}. \tag{A3}$$

Let $\boldsymbol{L} = (\boldsymbol{0}\,, \boldsymbol{I}_G)$ and $\boldsymbol{d} = (1\,, -\bar{\boldsymbol{b}}')'$ where $\bar{\boldsymbol{b}}$ is the probability limit of $\boldsymbol{b}$. It is easy to verify that

$$\boldsymbol{b} - \bar{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}\,, \tag{A4}$$

and therefore the asymptotic distribution of $\boldsymbol{b}$ can be derived from that of $\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}$.

For brevity, call $\boldsymbol{\Psi}_n = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{\Pi}/\psi_n = \boldsymbol{\Psi} - \boldsymbol{E}_n$ and its limit $\boldsymbol{\Psi}_n \to \bar{\boldsymbol{\Psi}} \equiv \boldsymbol{\Psi} - \boldsymbol{E}$. Also, let $\boldsymbol{\delta} = \boldsymbol{\beta} - \bar{\boldsymbol{b}}$. Using these definitions, note that $\boldsymbol{L}\bar{\boldsymbol{U}}' = (\boldsymbol{Z}\boldsymbol{\Pi})'$, $\bar{\boldsymbol{U}}\boldsymbol{d} = \boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{\delta}$, $\boldsymbol{L}\boldsymbol{\Xi}\boldsymbol{d} = \boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta}$, $\boldsymbol{L}'\bar{\boldsymbol{U}}'\boldsymbol{P}\bar{\boldsymbol{U}}\boldsymbol{d} = \psi_n\boldsymbol{\Psi}_n\boldsymbol{\delta}$, $\boldsymbol{L}\bar{\boldsymbol{U}}'\boldsymbol{P}\bar{\boldsymbol{U}}\boldsymbol{L}' = \psi_n\boldsymbol{\Psi}_n$ and $\boldsymbol{L}\boldsymbol{\Xi}\boldsymbol{L}' = \boldsymbol{\Omega}$. It follows from (A1) that

$$\mathbb{E}[\,\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}\,] = (\psi_n\boldsymbol{\Psi}_n + r_n\boldsymbol{\Omega})\boldsymbol{\delta} + r_n\boldsymbol{\rho}\,. \tag{A5a}$$

$$\begin{aligned}\mathrm{var}(\,\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}\,) = {} & (\,\sigma_{uu} + 2\boldsymbol{\rho}'\boldsymbol{\delta} + \boldsymbol{\delta}'\boldsymbol{\Omega}\boldsymbol{\delta}\,)(\psi_n\boldsymbol{\Psi}_n + r_n\boldsymbol{\Omega}) + \psi_n(\boldsymbol{\delta}'\boldsymbol{\Psi}_n\boldsymbol{\delta})\boldsymbol{\Omega} + \ldots \\ & \ldots + r_n(\boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta})(\boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta})' + \psi_n\left\{(\boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta})\boldsymbol{\delta}'\boldsymbol{\Psi}_n + \boldsymbol{\Psi}_n\boldsymbol{\delta}(\boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta})'\right\}.\end{aligned} \tag{A5b}$$

A.1 Many Instruments - Proposition 1

Under many instrument asymptotics, $\psi_n = n$ and $K_n = O(n)$. Since at most $r_n = O(K_n)$ the result in (A5b) indicates that var$(\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}) = O(n)$ and thus we set $\theta_n = n$ in Lemma 1. Furthermore,

$$\frac{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}}{n} = \boldsymbol{\Psi}_n + \frac{r_n}{n}\boldsymbol{\Omega} + O_p\left(\frac{1}{\sqrt{n}}\right) \qquad \text{and} \qquad \frac{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{u}}{n} = \frac{r_n}{n}\boldsymbol{\rho} + O_p\left(\frac{1}{\sqrt{n}}\right), \tag{A6}$$

so Slutsky's theorem gives

$$\boldsymbol{b} = \boldsymbol{\beta} + \frac{r_n}{n}\left(\boldsymbol{\Psi}_n + \frac{r_n}{n}\boldsymbol{\Omega}\right)^{-1}\boldsymbol{\rho} + o_p(1) \xrightarrow{p} \bar{\boldsymbol{b}} \equiv \boldsymbol{\beta} + \alpha(\bar{\boldsymbol{\Psi}} + \alpha\boldsymbol{\Omega})^{-1}\boldsymbol{\rho}\,, \tag{A7}$$

where $\bar{\boldsymbol{b}}$ is the probability limit of $\boldsymbol{b}$. From (A5a) and (A7),

$$\mathbb{E}\left[\frac{\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}}{n}\right] = \left(\boldsymbol{\Psi}_n + \frac{r_n}{n}\boldsymbol{\Omega}\right)\boldsymbol{\delta} + \frac{r_n}{n}\boldsymbol{\rho} = \left[\left(\frac{r_n}{n} - \alpha\right)\bar{\boldsymbol{\Psi}} + \alpha(\bar{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_n)\right](\bar{\boldsymbol{\Psi}} + \alpha\boldsymbol{\Omega})^{-1}\boldsymbol{\rho}\,. \tag{A8}$$

Since $\bar{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_n = \boldsymbol{E}_n - \boldsymbol{E}$, then $\mathbb{E}[\,\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}/\sqrt{n}\,] = o(1)$ if the conditions (6) in Proposition 1 hold.

On the other hand, using (A7) it is straightforward to verify that $\boldsymbol{\Psi}_n\boldsymbol{\delta} = -(r_n/n)(\,\boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta}\,) + o_p(1)$.

Plugging this equality into (A5b) yields

$$\begin{aligned}\text{var}\left(\frac{\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}}{\sqrt{n}}\right) = (\,\sigma_{uu} + 2\boldsymbol{\rho}'\boldsymbol{\delta} + \boldsymbol{\delta}'\boldsymbol{\Omega}\boldsymbol{\delta}\,)\left(\boldsymbol{\Psi}_n + \frac{r_n}{n}\boldsymbol{\Omega}\right) + \ldots \\ \ldots + (\,\boldsymbol{\delta}'\boldsymbol{\Psi}_n\boldsymbol{\delta}\,)\boldsymbol{\Omega} - \frac{r_n}{n}(\,\boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta}\,)(\,\boldsymbol{\rho} + \boldsymbol{\Omega}\boldsymbol{\delta}\,)' + o_p(1).\end{aligned} \tag{A9}$$

Therefore, using (A4), (A2), the Cramér-Wold device and under conditions (6),

$$\begin{aligned}\sqrt{n}(\,\boldsymbol{b} - \bar{\boldsymbol{b}}\,) &= \left(\frac{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}}{n}\right)^{-1}\frac{\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}}{\sqrt{n}} \\ &= (\,\boldsymbol{H} + o_p(1)\,)\frac{\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d} - \mathbb{E}[\,\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}\,]}{\sqrt{n}} + o_p(1) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}),\end{aligned} \tag{A10}$$

where $\boldsymbol{H} = \text{plim}(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}/n)^{-1}$, cf. (A6), and $\boldsymbol{W}$ is the limit of (A9) with $\boldsymbol{\delta} = -\alpha\boldsymbol{H}\boldsymbol{\rho}$. Straightforward but tedious manipulations give (8).

Corollaries

Corollaries 1 and 3 are direct implications of the previous analysis, for $\alpha = \alpha_K$ and $\boldsymbol{E}_n = \boldsymbol{E} = \boldsymbol{0}$, and $\alpha = 0$ respectively.

For a proof to Corollary 2, note that since $r_n = O(\sqrt{n}\,)$, then $r_n/n \to \alpha = 0$ yielding a consistent estimator, cf.

(A7). Thus, we have now that $\bar{\boldsymbol{b}} = \boldsymbol{\beta}$, $\boldsymbol{d} = (1\,, -\boldsymbol{\beta}')'$ and $\boldsymbol{\delta} = \mathbf{0}$. Furthermore, from (A8)

$$\mathbb{E}\left[\frac{\boldsymbol{L}\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\boldsymbol{d}}{\sqrt{n}}\right] = \frac{r_n}{\sqrt{n}}\boldsymbol{\rho} = \bar{\alpha}\boldsymbol{\rho} + o(1)\,, \tag{A11}$$

and, therefore,

$$\sqrt{n}(\,\boldsymbol{b} - \boldsymbol{\beta}\,) = (\,\boldsymbol{H} + o_p(1)\,)\cdot N(\mathbf{0}, \boldsymbol{W}) + \bar{\alpha}\boldsymbol{H}\boldsymbol{\rho} + o_p(1) \xrightarrow{d} N(\,\bar{\alpha}\boldsymbol{H}\boldsymbol{\rho}\,,\, \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}\,)\,, \tag{A12}$$

where $\boldsymbol{W} = \sigma_{uu}\bar{\boldsymbol{\Psi}}$ is the limit of (A9) under the conditions of Corollary 2. Finally, under Corollary 4, $\bar{\boldsymbol{\Psi}} = \boldsymbol{\Psi}$.

A.2 Many Weak Instruments - Proposition 2

Proposition 2(a): In this case $r_n = O(K_n)$ and hence $\psi_n = o(r_n)$ as an implication of sequence (11). Thus, from (A5b) we have that $\text{var}(\,\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\,) = O(r_n)$ and so

$$\frac{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}}{r_n} = \frac{\psi_n}{r_n}\boldsymbol{\Psi}_n + \boldsymbol{\Omega} + O_p\left(\frac{1}{\sqrt{r_n}}\right) \qquad \text{and} \qquad \frac{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{u}}{r_n} = \boldsymbol{\rho} + O_p\left(\frac{1}{\sqrt{r_n}}\right). \tag{A13}$$

The result follows from Slutsky's theorem:

$$\boldsymbol{b} = \boldsymbol{\beta} + (\,\boldsymbol{\Omega} + o_p(1)\,)^{-1}(\,\boldsymbol{\rho} + o_p(1)\,) = \boldsymbol{\beta} + (\,\boldsymbol{\Omega} + o_p(1)\,)^{-1}\boldsymbol{\rho} + o_p(1) \xrightarrow{p} \boldsymbol{\beta} + \boldsymbol{\Omega}^{-1}\boldsymbol{\rho}\,. \tag{A14}$$

Propositions 2(b): Given either (i) $r_n = O(\sqrt{K_n}\,)$ and $\sqrt{K_n} = o(\psi_n)$, or (ii) $r_n = o(\sqrt{K_n}\,)$,

$$\text{(i)}\ \frac{r_n}{\psi_n} = \frac{r_n}{\sqrt{K_n}}\cdot\frac{\sqrt{K_n}}{\psi_n} = O(1)o(1) = o(1) \quad \text{or} \quad \text{(ii)}\ \frac{r_n}{\psi_n} = \frac{r_n}{\sqrt{K_n}}\cdot\frac{\sqrt{K_n}}{\psi_n} = o(1)O(1) = o(1)\,. \tag{A15}$$

Therefore, (A5b) indicates that $\text{var}(\,\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\,) = O(\psi_n)$. Hence,

$$\frac{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}}{\psi_n} = \boldsymbol{\Psi}_n + \frac{r_n}{\psi_n}\boldsymbol{\Omega} + O_p\left(\frac{1}{\sqrt{\psi_n}}\right) \qquad \text{and} \qquad \frac{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{u}}{\psi_n} = \frac{r_n}{\psi_n}\boldsymbol{\rho} + O_p\left(\frac{1}{\sqrt{\psi_n}}\right). \tag{A16}$$

An application of Slutsky's theorem gives

$$\boldsymbol{b} = \boldsymbol{\beta} + \frac{r_n}{\psi_n}\left(\boldsymbol{\Psi}_n + \frac{r_n}{\psi_n}\boldsymbol{\Omega}\right)^{-1}\boldsymbol{\rho} + o_p(1) = \boldsymbol{\beta} + o_p(1) \xrightarrow{p} \boldsymbol{\beta}\,. \tag{A17}$$

Finally, the assertion in footnote 3 follows from the consistency of $\boldsymbol{b}$ (thus, $\boldsymbol{\delta} = \mathbf{0}$) and similar results to equations (A11) and (A12) with ψ_n in lieu of n, because $\text{var}(\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}) = O(\psi_n)$.

B Principal Components IVE

This appendix is concerned with the results related to the PCIVE and the use of the Retention Rule. Proofs of Propositions 3 and 4 are displayed below.

B.1 PC Approximations

Let $\boldsymbol{C}$ be the $K_n \times K_n$ matrix whose columns are given by the orthonormal eigenvectors of $\boldsymbol{S}$ and let $\boldsymbol{\Lambda}$ be the diagonal matrix whose diagonal entries are the corresponding eigenvalues sorted in decreasing order. Let $\boldsymbol{C}_r$ be the $K_n \times r_n$ matrix that contains the first r_n columns of $\boldsymbol{C}$ and let $\boldsymbol{C}_e$ be the $K_n \times (K_n - r_n)$ matrix formed by the remaining columns of $\boldsymbol{C}$: $\boldsymbol{C}_e{}'\boldsymbol{C}_r = \boldsymbol{0}$, $\boldsymbol{C}_e\boldsymbol{C}_e{}' = \boldsymbol{C}_r\boldsymbol{C}_r{}' = \boldsymbol{I}_{K_n}$, $\boldsymbol{C}_e{}'\boldsymbol{C}_e = \boldsymbol{I}_{K_n - r_n}$ and $\boldsymbol{C}_r{}'\boldsymbol{C}_r = \boldsymbol{I}_{r_n}$. Partition $\boldsymbol{\Lambda}$ conformably into two diagonal matrices $\boldsymbol{\Lambda}_r$ and $\boldsymbol{\Lambda}_e$ with diagonal entries equal, respectively, to the largest r_n and the smallest $K_n - r_n$ eigenvalues of $\boldsymbol{S}$. By the spectral decomposition of $\boldsymbol{S}$ we have that

$$\boldsymbol{S} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}' = \boldsymbol{C}_r\boldsymbol{\Lambda}_r\boldsymbol{C}_r{}' + \boldsymbol{C}_e\boldsymbol{\Lambda}_e\boldsymbol{C}_e{}' = \boldsymbol{S}_r + \boldsymbol{S}_e\,, \tag{B1}$$

so that $\boldsymbol{S}_r$ is the r_n-rank approximation to $\boldsymbol{S}$ and $\boldsymbol{S}_e$ is the residual from such approximation. Notice also that since $\boldsymbol{S}\boldsymbol{C}_r = \boldsymbol{C}_r\boldsymbol{\Lambda}_r$,

$$\frac{\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}}{n} = \boldsymbol{S} - \boldsymbol{S}\boldsymbol{C}_r(\boldsymbol{C}_r{}'\boldsymbol{S}\boldsymbol{C}_r)^{-1}\boldsymbol{C}_r{}'\boldsymbol{S} = \boldsymbol{S} - \boldsymbol{C}_r\boldsymbol{\Lambda}_r\boldsymbol{C}_r{}' = \boldsymbol{S} - \boldsymbol{S}_r = \boldsymbol{S}_e\,. \tag{B2}$$

Therefore, using this fact together with the definition $\boldsymbol{E}_n = (n/\psi_n)\boldsymbol{\Pi}'\boldsymbol{S}_e\boldsymbol{\Pi}$, cf. (5), and Assumption A4,

$$\|\boldsymbol{E}_n\| \leq (n/\psi_n)\cdot\|\boldsymbol{S}_e\|\cdot\|\boldsymbol{\Pi}\|^2 = (\,\|\boldsymbol{S}_e\|/\|\boldsymbol{S}\|\,)O(1)\,. \tag{B3}$$

B.2 Analysis Based on Eigenvalues - Proposition 3

From Assumption A5, $m_n = O(K_n^{\varepsilon_m}) = o(K_n)$ since $\varepsilon_m < 1$. Thus, by the spectral decomposition of $\boldsymbol{S}$,

$$\|\boldsymbol{S}\|^2 = \sum_{k=1}^{K_n}\lambda_k(\boldsymbol{S})^2 = m_n O(s_n^2) + (K_n - m_n)O(1) = O(K_n^{\varepsilon_m+2\varepsilon_s}) + O(K_n) = O(K_n^q)\,, \tag{B4}$$

where $q = \max\{\varepsilon_m + 2\varepsilon_s, 1\}$.

Suppose first that $\varepsilon_m > \varepsilon_r$ such that $r_n/m_n = O(K_n^{\varepsilon_r-\varepsilon_m}) = o(1)$ or, in other words, that there exists a finite number n_0 such that $m_n > r_n$ for all $n > n_0$. Thus,

$$\|\boldsymbol{S}_e\|^2 = \sum_{k=r_n+1}^{K_n}\lambda_k(\boldsymbol{S})^2 = (m_n - r_n)O(s_n^2) + (K_n - m_n)O(1) = O(K_n^{\varepsilon_m+2\varepsilon_s}) + O(K_n) = O(K_n^q)\,. \tag{B5}$$

From (B4) and (B5) we obtain $\|\boldsymbol{S}_e\|/\|\boldsymbol{S}\| = O(1)$. Plugging this result into (B3) gives $\boldsymbol{E}_n = O(1)$.

On the other hand, consider the opposite case $\varepsilon_r > \varepsilon_m$ such that $m_n/r_n = o(1)$ or that there exists a finite number n_0 such that $r_n > m_n$ for all $n > n_0$. Recall that $r_n = O(K_n^{\varepsilon_r}) = o(K_n)$ since $\varepsilon_r < 1$. Then,

$$\|\boldsymbol{S}_e\|^2 = \sum_{k=r_n+1}^{K_n} \lambda_k(\boldsymbol{S})^2 = (K_n - r_n)O(1) = O(K_n)\,. \tag{B6}$$

Plugging (B4) and (B6) into (B3) gives $\boldsymbol{E}_n \leq O(K_n^{(1-q)/2})$. Therefore, if $\varepsilon_m < 1 - 2\varepsilon_s$ then $q = 1$ and $\boldsymbol{E}_n = O(1)$; on the contrary, if $1 - 2\varepsilon_s < \varepsilon_m$ then $q > 1$ and we conclude that $\boldsymbol{E}_n = o(1)$. Gathering all these findings gives Proposition 3.

B.3 Analysis Based on Correlations - Proposition 4(a)

Consider Assumption A6. Since all entries of $\boldsymbol{\Upsilon}$ are less than one in absolute value and their diagonal elements are all equal to zero, it is easy to verify that

$$\|\boldsymbol{S}\|^2 = \|\boldsymbol{I}_{K_n}\|^2 + \mu^2\|\boldsymbol{\Upsilon}\|^2 = O(K_n^2) \qquad \text{and} \qquad \text{tr}(\boldsymbol{S}) = O(K_n)\,. \tag{B7}$$

By the Retention Rule, if r_n PC are retained then all diagonal entries of $\boldsymbol{\Lambda}_r$ in (B1) must be greater than $K_n^{-\delta}\text{tr}(\boldsymbol{S})$, whereas all entries of $\boldsymbol{\Lambda}_e$ are less than (or equal to) $K_n^{-\delta}\text{tr}(\boldsymbol{S})$. Thus, from (B7)

$$\|\boldsymbol{S}_e\|^2 \leq (K_n - r_n)K_n^{-2\delta}\text{tr}(\boldsymbol{S})^2 \leq K_n^{1-2\delta}O(K_n^2)\,, \tag{B8}$$

for any $r_n = o(K_n)$. Plugging (B7) and (B8) into (B3) gives $\boldsymbol{E}_n \leq K_n^{1/2-\delta}O(1)$. Proposition 4(a) follows.

B.4 Rate of Growth of Retained PC - Proposition 4(b)

We omit the n script for convenience. With no loss of generality normalise all the diagonal elements of the $\boldsymbol{S}$ matrix to one, so that threshold value in the Retention Rule becomes $K^{-\delta}\text{tr}(\boldsymbol{S}) = K^{1-\delta}$. We proceed by induction to establish a bound for δ and thus a necessary condition for (**A**).

Consider that we add one instrument to the IV set at a time, which we call a trial. Besides, suppose that K is large enough such that $(K+1)^{1-\delta} \simeq K^{1-\delta}$. We begin with an initial situation where the eigenvalues of $\boldsymbol{S}$ are $\lambda_1 \geq \ldots \geq \lambda_r > K^{1-\delta} \geq \lambda_{r+1} \geq \ldots \geq \lambda_K$, so $r_0 = O(1)$ PC are retained.

In the first trial we obtain the matrix

$$\boldsymbol{S}^* = \begin{bmatrix} \boldsymbol{S} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{s} \\ \boldsymbol{s}' & 0 \end{bmatrix}, \tag{B9}$$

where $\boldsymbol{s}$ is a $K \times 1$ vector containing the 'correlations' between the newly added instrument and the previous IV set. Let λ_i^* denote the i-th eigenvalue of $\boldsymbol{S}^*$. By the interlacing property of the eigenvalues (Horn and Johnson,

1985, Theorem 4.3.8, p. 185), $\lambda_r^* \geq \lambda_r > K^{1-\delta}$ which implies that the number of retained components after the first trial is no less than r_0. Also, $\lambda_{r+1}^* \geq \lambda_{r+1}$ so the interest lies in determining whether $\lambda_{r+1}^* > K^{1-\delta}$ or, in words, whether the inclusion of a new instrument increases the number of retained PC. Note that $\boldsymbol{S}^*$ can be thought of as the perturbation of the first block diagonal matrix in (B9). By the perturbation theorem for symmetric matrices (Horn and Johnson, 1985, Corollary 6.3.4, p. 367) we have that

$$\lambda_{r+1}^* = \lambda_{r+1} + \epsilon\,, \quad \text{where} \quad \epsilon \leq O(\sqrt{\boldsymbol{s}'\boldsymbol{s}}\,) = \mu\, O(\sqrt{K}\,), \tag{B10}$$

where μ is the absolute value of the largest element of $\boldsymbol{s}$. Importantly, (B10) is a valid expression if $\mu = O(1)$ is small, as it comes from a local approximation of $\boldsymbol{S}^*$ around $\text{diag}(\boldsymbol{S}, 1)$. Then, by the Markov inequality we have that the probability of an extra PC is at most

$$\Pr[\,\mu O(\sqrt{K}\,) \geq K^{1-\delta}\,] = \Pr[\,\mu \geq O(K^{1/2-\delta})\,] = O(K^{\delta-1/2})\,. \tag{B11}$$

It follows that after the first trial, the expected number of retained components is $r_0 + O(K^{\delta-1/2})$. After t trials,

$$r_t = r_0 + \sum_{\tau=1}^{t} O(K^{\delta-1/2}) = r_0 + O(K^{\delta-1/2}) = r_0 + o(K^{\bar{\delta}-1/2})\,, \tag{B12}$$

where $\bar{\delta}$ is strictly greater than δ. For $r_t = o(\sqrt{K}\,)$, then $\bar{\delta} \leq 1$ is required as stated in Proposition 4(b).

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.

BAI, J., AND S. NG (2010): "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26(forthcoming).

BEKKER, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 63(3), 657–681.

BREUSCH, T., H. QIAN, P. SCHMIDT, AND D. WYHOWSKI (1999): "Redundancy of Moment Conditions," *Journal of Econometrics*, 91(1), 89 – 111.

CARRIERO, A., G. KAPETANIOS, AND M. MARCELLINO (2008): "A Shrinkage Instrumental Variable Estimator for Large Datasets," Queen Mary (University of London) Working Paper No. 626.

CHAO, J. C., AND N. R. SWANSON (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73(5), 1673–1692.

DAVIDSON, R., AND J. G. MACKINNON (2006): "The Case Against JIVE," *Journal of Applied Econometrics*, 21(6), 827–833.

DONALD, S. G., AND W. K. NEWEY (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–91.

DORAN, H. E., AND P. SCHMIDT (2006): "GMM Estimators with Improved Finite Sample Properties using Principal Components of the Weighting Matrix, with an Application to the Dynamic Panel Data Model," *Journal of Econometrics*, 133(1), 387–409.

HAHN, J. (2002): "Optimal Inference with Many Instruments," *Econometric Theory*, 18(1), 140–168.

HAHN, J., J. HAUSMAN, AND G. KUERSTEINER (2004): "Estimation with Weak Instruments: Accuracy of Higher-order Bias and MSE Approximations," *Econometrics Journal*, 7(1), 272–306.

HALL, A. R., A. INOUE, K. JANA, AND C. SHIN (2007): "Information in Generalized Method of Moments Estimation and Entropy-based Moment Selection," *Journal of Econometrics*, 138(2), 488 – 512, 'Information and Entropy Econometrics' - A Volume in Honor of Arnold Zellner.

HALL, A. R., AND F. P. M. PEIXE (2003): "A Consistent Method for the Selection of Relevant Instruments.," *Econometric Reviews*, 22(3), 269–287.

HALL, A. R., G. D. RUDEBUSCH, AND D. W. WILCOX (1996): "Judging Instrument Relevance in Instrumental Variables Estimation," *International Economic Review*, 37(2), 283–298.

HANSEN, C., J. HAUSMAN, AND W. NEWEY (2008): "Estimation with Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26(4), 398–422.

HORN, R. A., AND C. R. JOHNSON (1985): *Matrix Analysis*. Cambridge University Press.

JOLLIFFE, I. T. (2002): *Principal Component Analysis*. Springer-Verlag, New York, 2nd edn.

KAPETANIOS, G. (2006): "Choosing the Optimal Set of Instruments from Large Instrument Sets," *Computational Statistics & Data Analysis*, 51(2), 612–620.

KAPETANIOS, G., AND M. MARCELLINO (2006): "Factor-GMM Estimation with Large Sets of Possibly Weak Instruments," Queen Mary (University of London) Working Paper No. 577.

——— (2008): "Cross-Sectional Averaging and Instrumental Variable Estimation with Many Weak Instruments," Queen Mary (University of London) Working Paper No. 627.

KLOEK, T., AND L. B. M. MENNES (1960): "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables," *Econometrica*, 28(1), 45–61.

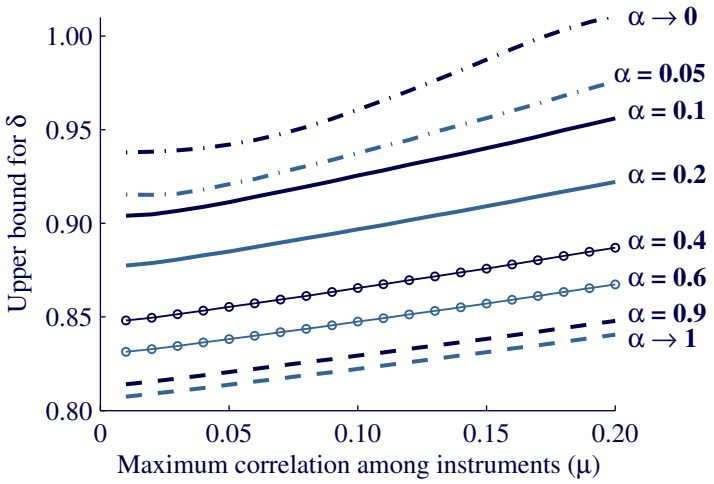
OKUI, R. (2010): "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *Journal of Econometrics*, 154(forthcoming).

SILVERSTEIN, J. W., AND S.-I. CHOI (1995): "Analysis of the Limiting Spectral Distribution of Large Dimensional Random Matrices," *Journal of Multivariate Analysis*, 54(2), 295 – 309.

STOCK, J. H., J. H. WRIGHT, AND M. YOGO (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20(4), 518–29.

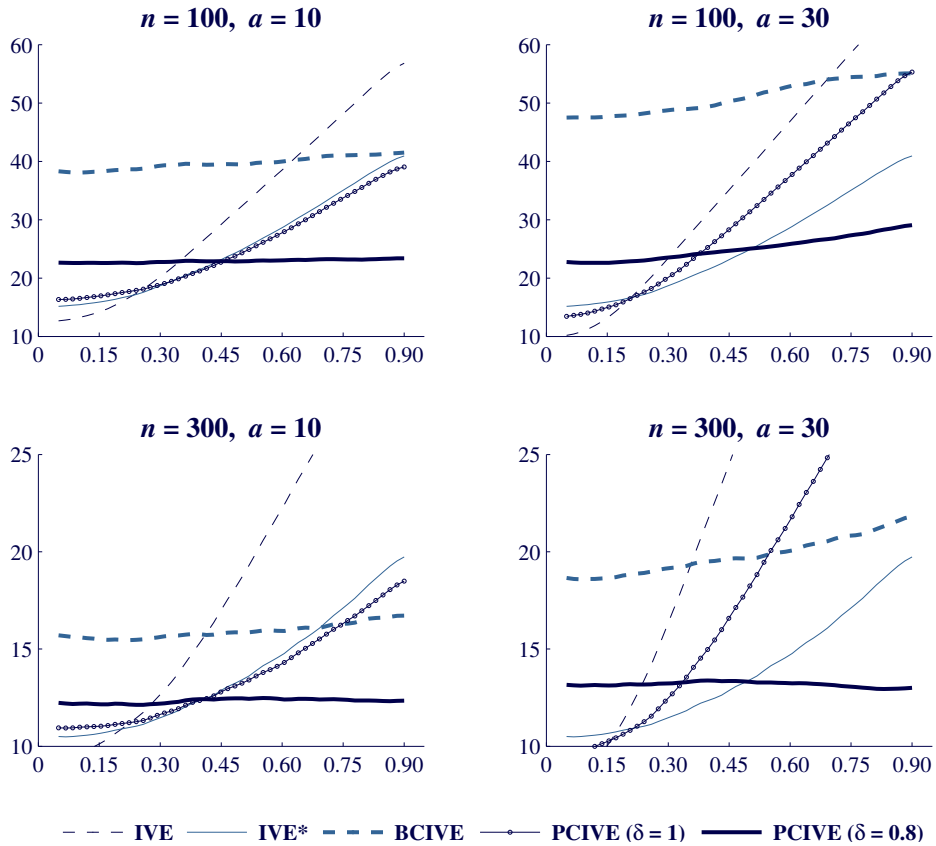
VAN HASSELT, M. (2010): "Many Instruments Asymptotic Approximations Under Nonnormal Error Distributions," *Econometric Theory*, 26(2), 633–645.

Figure 1. The Function $\bar{\delta} = \bar{\delta}(\mu, \alpha)$ for Small μ and Selected Values of α



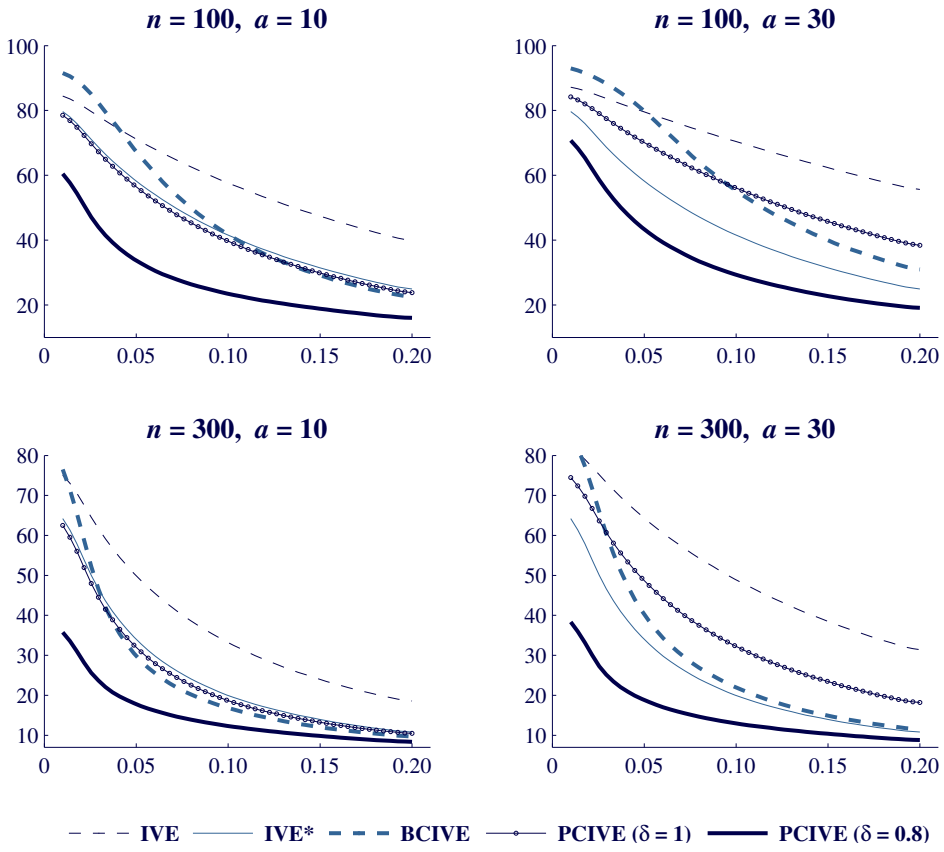
NOTES: Each experiment uses 10,000 replications. $\bar{\delta}$ is the maximum value of δ that satisfies (**A**) for given μ and α.

Figure 2. Median Absolute Error of IV Estimators as a Function of ρ, n and a



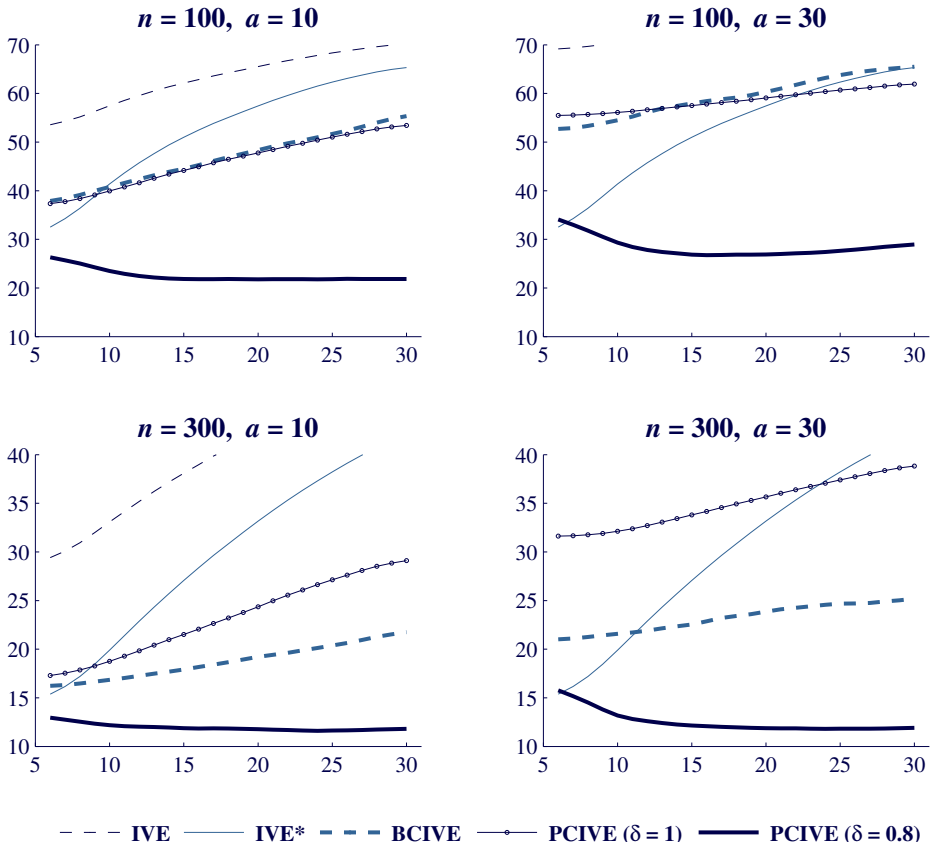
NOTES: The horizontal axis shows values of ρ, the correlation coefficient between u and v, and ranges from 0.10 to 0.90. Also, $R^2 = 0.10$, $K_* = 10$ and $\mu = 0.10$. Each experiment uses 50,000 replications.

Figure 3. Median Absolute Error of IV Estimators as a Function of R^2, n and a



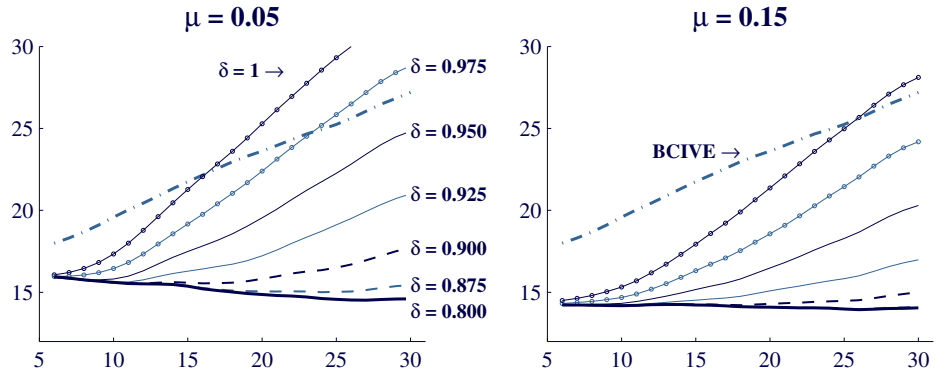
NOTES: The horizontal axis shows values of R^2, the coefficient of determination of the first stage regression, and ranges from 0.01 to 0.20. Also, $\rho = 0.90$, $K_* = 10$ and $\mu = 0.10$. Each experiment uses 50,000 replications.

Figure 4. Median Absolute Error of IV Estimators as a Function of K_*, n and a



Notes: The horizontal axis shows values of K_*, the number of relevant instruments, and ranges from 6 to 30. Also, $\rho = 0.90$, $R^2 = 0.10$ and $\mu = 0.10$. Each experiment uses 50,000 replications.

Figure 5. Median Absolute Error of IV Estimators as a Function of K_*, μ and δ



Notes: The horizontal axis shows values of K_*, the number of relevant instruments, and ranges from 6 to 30. Also, $\rho = 0.90$, $R^2 = 0.10$, $n = 200$ and $a = 0$. Each experiment uses 50,000 replications.