Division of Economics, EGC
School of Humanities and Social Sciences
Nanyang Technological University
14 Nanyang Drive
Singapore 637332

Identification and Inference in a Simultaneous Equation Under Alternative Information Sets and Sampling Schemes

Jan F. KIVIET

1 November 2012

EGC Report No: 2012/07

HSS-04-88
Tel: +65 67905689
Email: D-EGC@ntu.edu.sg
http://egc.hss.ntu.edu.sg

The author(s) bear sole responsibility for this paper.

Views expressed in this paper are those of the author(s) and not necessarily those of the Economic Growth Centre, NTU.

Identification and inference in a simultaneous equation under alternative information sets and sampling schemes

Jan F. Kiviet*†

1 November 2012
JEL-classification: C12, C13, C15, C26, J31
Keywords: *partial identification, weak instruments, (un)restrained repeated sampling, (un)conditional (limiting) distributions, credible robust inference*

Abstract

In simple static linear simultaneous equation models the empirical distributions of IV and OLS are examined under alternative sampling schemes and compared with their first-order asymptotic approximations. We demonstrate that the limiting distribution of consistent IV is not affected by conditioning on exogenous regressors, whereas that of inconsistent OLS is. The OLS asymptotic and simulated actual variances are shown to diminish by extending the set of exogenous variables kept fixed in sampling, whereas such an extension disrupts the distribution of IV and deteriorates the accuracy of its standard asymptotic approximation, not only when instruments are weak. Against this background the consequences for the identification of parameters of interest are examined for a setting in which (in practice often incredible) assumptions regarding the zero correlation between instruments and disturbances are replaced by (generally more credible) interval assumptions on the correlation between endogenous regressor and disturbance. This yields OLS-based modified confidence intervals, which are usually conservative. Often they compare favorably with IV-based intervals and accentuate their frailty.

1. Introduction

We approach the fundamental problems of simultaneity and identification in econometric linear structural relationships in a non-traditional way. First, we illustrate the effects of the chosen sampling regime on the limiting and empirical finite sample distributions of both OLS (ordinary least-squares) and IV (instrumental variables) estimators. We distinguish sampling schemes

*Division of Economics, School of Humanities and Social Sciences, Nanyang Technological University, 14 Nanyang Drive, Singapore 637332 (jfkiviet@ntu.edu.sg) and Department of Quantitative Economics, Amsterdam School of Economics and Tinbergen Institute, University of Amsterdam, Valckenierstraat 65, 1018 XE Amsterdam, The Netherlands (j.f.kiviet@uva.nl).

†For the many constructive remarks regarding earlier stages of this research I want to thank editor Elie Tamer, two anonymous referees, and seminar participants at the universities of Amsterdam, Montréal, Yale (New Haven), Brown (Providence), Harvard (Cambridge), Carleton (Ottawa), New York, Penn State (University Park), Rutgers (New Brunswick) and Columbia (New York), and those giving feedback at presentations at ESWC (Shanghai, August 2010), EC2 (Toulouse, December 2010), CESifo (Munich, February 2011) and the SER Conference (Singapore, August 2011).

in which exogenous variables are either random or kept fixed. We find that, apart from its shifted location due to its inconsistency, the OLS distribution is rather well-behaved, though mildly dependent on the chosen sampling regime, both in small and in large samples. On the other hand, and not predicted by standard asymptotic theory, the sampling regime is found to have a much more profound effect on the actual IV distribution. Especially for fixed and not very strong instruments, IV is in fact often much less informative on the true parameter value than OLS is. Next, following ideas from the recent partial identification literature[1], we then dare to replace the standard point-identifying but in practice often incredible assumption on the zero correlation between (possibly weak) instruments and the disturbance term by an alternative identifying information set. We demonstrate that identification can also be obtained by abstaining from instruments altogether, by adopting instead a restricted range of possible values for the correlation between endogenous regressor and disturbance. Choosing this range wide leads to a gain in credibility and robustness, but at the expense of precision. In principle this approach and the standard approach are equally (un)feasible, because both are based on statistically untestable presuppositions.

Achieving identification through replacing orthogonality conditions by simultaneity assumptions yields confidence intervals which are purely based on standard OLS statistics. These have just to be transformed by a very simple formula, which involves: both the endpoints of the assumed interval for the simultaneity correlation coefficient, the sample size, and a standard Gaussian or Student critical value. The simplicity of the procedure is partly based on the following truly remarkable result regarding the asymptotic variance of OLS under simultaneity. This asymptotic variance is smaller (in a matrix sense) than the standard expression that holds under classic assumptions. Under the adopted partial identification assumption, however, the OLS estimator is corrected for its inconsistency by a term based on standard OLS results and on the assumed simultaneity coefficient. Because this correction is random, it yields an increment to the asymptotic variance of inconsistent OLS. Under normality and adopting the conservative unconditional sampling scheme, this modified OLS estimator is found to have an asymptotic variance identical again to the classic well-known simple expression.

The limiting distribution of inconsistent OLS has been examined before (Goldberger, 1964; Rothenberg, 1972; Schneeweiss, 1980; Kiviet and Niemczyk, 2012), but the effects of (not) conditioning on exogenous variables have not yet been studied in a systematic way. Here we provide the framework to do so, but in order to keep the derivations transparent we restrict ourselves to results for single regressor models. Unconditional IV estimation received little attention in the literature, possibly because in the standard case conditioning has no consequences asymptotically. Early studies on the small sample distribution of method of moments estimators in static simultaneous equations (Sawa, 1969; Mariano, 1973; Phillips, 1983), all focussed on the case of fixed instruments. The more recent literature on the effects of weak instruments took off in Nelson and Startz (1990a,b). They adopted even more severe conditioning restrictions, leading to extreme correlation between structural and reduced form disturbances, as noted by Maddala and Jeong (1992). Next a literature developed in which only observed instruments are kept fixed again (Woglom, 2001; Hillier, 2006; Forchini, 2006; Phillips 2006, 2009; Andrews and Stock, 2007a), although in some studies that analyze weak/many instrument asymptotic frameworks attention is also paid to random instruments (Staiger and Stock, 1997; Andrews and Stock, 2007b). Here we will systematically examine and compare the consequences of three stages of conditioning on simple IV and OLS estimation,

[1] See Manski (2003, 2007) and for a recent review see Tamer (2010). Phillips (1989) and others have used 'partial identification' to indicate a fundamentally different situation.

both asymptotically and experimentally.

Concerns about identification and alternative approaches to achieve identification in simultaneous equations models can be found in many econometric studies. Since the validity of instruments can only be tested statistically under the maintained hypothesis of a sufficient number of valid instruments to achieve just identification, other arguments than empirical-statistical ones are required to validate the initial set of identifying conditions. Recently, such argumentation - based either on common sense reasoning or economic theory - has often been used convincingly by designing empirical research in the format of natural or quasi-experiments, see (references in and the discussions on) Angrist and Pischke (2010). Doubts about instrument validity can also be alleviated through replacing exclusion restrictions by point or interval assumptions (Kraay, 2012; Conley *et al.*, 2012), or by allowing instruments to be correlated with structural disturbances, but less so than the endogenous regressor (Nevo and Rosen, 2012), or in the context of treatment response analysis by making so-called monotone instrumental variable assumptions (Manski and Pepper, 2000, 2009). All these approaches enable to obtain inference in the form of confidence intervals which are asymptotically valid under the relaxed alternative assumptions that they require. The inference that we will produce here is unique in the sense that no assumptions on external instruments have to be made at all; the identifying information exploited relates directly to the degree of simultaneity. Whether in a particular situation this approach is preferable does not simply depend on objective characteristics of the case under study, because the actual credibility of alternative assumptions to achieve identification is inherently subjective and cannot be measured objectively on a simple ordinal scale.

Correlation between regressors and disturbances also occurs in models allowing for measurement errors, but then the degree of simultaneity is directly related to the parameters of primary interest and the assumptions made regarding the distribution of the errors in the variables. Studies in that context with similar aims as the present study are Kapteyn and Wansbeek (1984), Erickson (1993) and Hu (2006), but their results cannot directly be converted to the classic simultaneous equations model.

The structure of this paper is as follows. In Section 2 we sketch the various settings that we will examine regarding DGP (data generating process), alternative adopted information sets and different sampling schemes. In Section 3 the limiting and simulated finite sample distributions of IV and OLS are examined under these various settings. Section 4 demonstrates how inference can be based just on OLS results and an assumed interval for the simultaneity coefficient. It also compares its accuracy with best-practice IV-based methods in a simulation study, and illustrates the empirical consequences for the well-known Angrist and Krueger (1991) data. Finally Section 6 concludes.

2. DGP, information sets, sampling schemes

The DGP involves just two jointly dependent data series and one external instrument, from which any further exogenous explanatory variables have been partialled out. Generalizations with more than one endogenous regressor, overidentification and explicit exogenous regressors are left for future work. Despite its simplicity, this DGP already embodies various salient features met in practice when analyzing actual relationships, especially when based on a cross-section sample. After its specification, we discuss the various settings under which the relationship might be analyzed in practice. These settings differ regarding two aspects, being characteristics of the adopted information set and of the assumed relevant sampling scheme.

The latter relates to conditioning (or not) on random exogenous variables and the former to the available information in the form of observed data and of adopted parametric restrictions.

2.1. The DGP

Instead of specifying the analyzed model and the underlying DGP top-down, as is usually done in econometrics, we introduce these here bottom-up, following precisely the generating scheme as used in the simulation experiments that will be exploited later. The basic building block of the DGP consists of three series (for $i = 1, ..., n$) of IID zero mean random variables

$$\varepsilon_i \sim IID(0, \sigma_\varepsilon^2),\ v_i^* \sim IID(0, \sigma_{v^*}^2) \text{ and } z_i \sim IID(0, \sigma_z^2), \tag{2.1}$$

which are also mutually independent. To avoid particular pathological cases we assume $\sigma_\varepsilon^2 > 0$ and $\sigma_z^2 > 0$, but allow $\sigma_{v^*}^2 \geq 0$. The ε_i will establish the disturbance term in the single simultaneous equation of interest with just one regressor x_i. The reduced form equation for this regressor x_i has reduced form disturbance v_i, which consists of two independent components, since

$$v_i = v_i^* + \xi\varepsilon_i. \tag{2.2}$$

Variable z_i is to be used as instrument. The reduced form equation for x_i is

$$x_i = \pi z_i + v_i, \tag{2.3}$$

and the simultaneous equation under study is

$$y_i = \beta x_i + \varepsilon_i. \tag{2.4}$$

So, in total we have 6 parameters, namely σ_ε^2, $\sigma_{v^*}^2$, σ_z^2, ξ, π and β, where the latter is the parameter of primary interest. However, some of these parameters can be fixed arbitrarily, without loss of generality. We may take $\beta = 0$, hence acting as if we had first subtracted βx_i form both sides of (2.4). Also, we may take $\sigma_z = 1$ and next interpret π as the original $\pi\sigma_z$. Moreover, scaling equation (2.4) by taking $\sigma_\varepsilon = 1$ has no principle consequences either. So, we find that this DGP has in fact only three fundamental parameters, namely ξ, π and $\sigma_{v^*}^2$. Nevertheless, it represents the essentials of a general linear just identified (provided $\pi \neq 0$) simultaneous equation with one endogenous regressor, where any further exogenous explanatory variables, including the constant, have been netted out.

We shall consider a simple nonlinear reparametrization, because this yields three different basic parameters which are easier to interpret. Not yet imposing the normalizations ($\sigma_z = 1 = \sigma_\varepsilon$) we find

$$Cov(x_i, \varepsilon_i) = \rho_{x\varepsilon}\sigma_x\sigma_\varepsilon = \xi\sigma_\varepsilon^2 \text{ and } Cov(x_i, z_i) = \rho_{xz}\sigma_x\sigma_z = \pi\sigma_z^2. \tag{2.5}$$

The correlations $\rho_{x\varepsilon}$ and ρ_{xz} parametrize simultaneity and instrument strength respectively. Because the instrument obeys the moment restrictions $E(z_i\varepsilon_i) = 0$ (thus $\rho_{z\varepsilon} = 0$) and $E(z_i v_i) = 0$, we also find

$$Var(x_i) = \sigma_x^2 = \pi^2\sigma_z^2 + \sigma_v^2, \tag{2.6}$$

with

$$Var(v_i) = \sigma_v^2 = \sigma_{v^*}^2 + \xi^2\sigma_\varepsilon^2, \tag{2.7}$$

thus

$$Var(v_i^*) = \sigma_{v^*}^2 = \sigma_x^2(1 - \rho_{x\varepsilon}^2 - \rho_{xz}^2), \tag{2.8}$$

and from $\sigma_{v^*}^2 \geq 0$ it follows that

$$\rho_{x\varepsilon}^2 + \rho_{xz}^2 \leq 1. \tag{2.9}$$

So, only combinations of $\rho_{x\varepsilon}$ and ρ_{xz} values in or on a circle with radius unity are admissible. Such a restriction is plausible, because, given that z_i is uncorrelated with ε_i, regressor x_i cannot be strongly correlated with both ε_i and z_i at the same time.

The above derivations yield a more insightful three-dimensional parameterization for this DGP. This is based on simply choosing values for $\sigma_x > 0$, $\rho_{x\varepsilon}$ and ρ_{xz}, respecting (2.9). Next the data can be generated according to the equations (2.1) through (2.4), where

$$\pi = \rho_{xz}\sigma_x/\sigma_z \text{ and } \xi = \rho_{x\varepsilon}\sigma_x/\sigma_\varepsilon, \tag{2.10}$$

with $\sigma_{v^*}^2$ as in (2.8), and taking $\sigma_\varepsilon = \sigma_z = 1$ and $\beta = 0$.

Another basic model characteristic often mentioned in this context is the population concentration parameter (PCP), given by

$$PCP \equiv \frac{nVar(\pi z_i)}{Var(v_i)} = n\frac{\pi^2\sigma_z^2}{\sigma_{v^*}^2 + \xi^2\sigma_\varepsilon^2} = n\frac{\rho_{xz}^2}{1-\rho_{xz}^2}. \tag{2.11}$$

We control this simply by setting values for ρ_{xz} and n. These two, together with $\rho_{x\varepsilon}$ and σ_x, and the further distributional characteristics (third and higher moments) of the IID variables ε_i, v_i^* and z_i, will determine the actual properties of any inference technique for β in this model.

2.2. Alternative information sets

Despite its simplicity the data generation scheme above for the relationship (2.4) is to be used to represent the essentials of particular practically relevant modelling situations, in which the ultimate goal is to produce inferences regarding β. These inferences should preferably be: interpretable (identified), accurate (supplemented by relatively precise probability statements), credible (based on nonsuspect assumptions), robust (valid under a relatively wide set of assumptions) and relatively efficient (small absolute or mean squared errors). Although they have to be obtained from a finite actual sample they will generally be based on asymptotic approximations to the distributions of estimators and test statistics. To derive these we have to make assumptions. Regarding these we will compare a few alternative sets. These information sets are characterized by whether particular variables are supposed to be available or not and by particular adopted parametric restrictions.

We will assume throughout that the variables y_i and x_i have been observed and are available to the practitioner, but not the disturbances ε_i, v_i^* and v_i. We will distinguish settings in which z_i is available too, and is used as an instrument, but also the situation in which it is not used. Instrument z_i may not be used because: (i) it is not available, (ii) because the researcher (possibly erroneously) assumes that $\rho_{x\varepsilon} = 0$, or (iii) the researcher is not willing to make the assumption $\rho_{z\varepsilon} = 0$, because it seems incredible. In these cases OLS may be used, although x_i establishes an invalid instrument when $\rho_{x\varepsilon} \neq 0$.

So, regarding parametric restrictions, we will consider the situation where one is either willing to make the assumption $\rho_{z\varepsilon} = 0$, or not. In addition, we will consider the situation where one is prepared to make an assumption regarding $\rho_{x\varepsilon}$ of the from $\rho_{x\varepsilon} \in \mathcal{R}_{x\varepsilon} \equiv \left[\rho_{x\varepsilon}^L, \rho_{x\varepsilon}^U\right]$, or not. Below we will show that under $\rho_{x\varepsilon} \in \mathcal{R}_{x\varepsilon}$ coefficient β can be identified and analyzed accurately by a modified version of OLS, which does not require any instruments. In a way this brings IV and OLS on a comparable base: IV requires the statistically unverifiable (and

therefore often incredible) assumption $\rho_{z\varepsilon} = 0$, whereas the modified OLS technique requires the generally unverifiable assumption $\rho_{x\varepsilon} \in \left[\rho_{x\varepsilon}^L, \rho_{x\varepsilon}^U\right]$, which can be made more credible (at the expense of precision) by choosing a wider interval. A similar though fundamentally different approach is followed by Nevo and Rosen (2012) who adopt an interval for $\rho_{z\varepsilon}$.

2.3. Alternative sampling schemes

When running the experiments required to obtain the finite sample distribution of estimators of β we have the choice to draw new independent series for the three random variables of (2.1) every replication, or to keep the series for either z_i or v_i^*, or for both, fixed. By keeping none of the series which are exogenous regarding ε_i fixed, we will examine unconditional distributions. This seems especially relevant in the context of a cross-section analysis where the random sample of n observations is obtained from a much larger population. Each Monte Carlo experiment should then mimic that a new independently drawn sample does not only yield different values for the disturbances, but also for the instruments, and for the component of the reduced form disturbances which is not associated with ε_i.

However, in the other extreme, when the sample covers the whole population and hence establishes a census, the situation is different. Then the ε_i, which represent the effects of variables not explicitly taken into account into the model, should still be redrawn every experiment, but there are good reasons to condition on one specific realization of the z_i, in order to keep close to the population under study. And since the v_i^* represent all the further typical determinants of x_i, in addition to πz_i, which are not associated with ε_i, one can argue that one would even keep closer to the characteristics of the actual population by keeping v_i^* fixed as well, despite the fact that in practice v_i^* is not observed. That would lead then to conditioning on latent variable x_i^*, where

$$x_i^* = \pi z_i + v_i^*. \tag{2.12}$$

When the data establish time-series there is no clear-cut population and there are good reasons to condition inference on the actual realized values of all the exogenous determinants of the endogenous variables. Note that $x_i = x_i^* + \xi\varepsilon_i$ with $E(x_i^*\varepsilon_i) = 0$, so x_i^* is exogenous, and – if available – would establish the strongest possible instrument for x_i, since $E(x_i^* x_i) = \sigma_{x^*}^2 = \sigma_x^2(1 - \rho_{x\varepsilon}^2)$ and hence, due to (2.9),

$$\rho_{x^*x}^2 = \frac{\sigma_x^4(1-\rho_{x\varepsilon}^2)^2}{\sigma_x^4(1-\rho_{x\varepsilon}^2)} = 1 - \rho_{x\varepsilon}^2 \geq \rho_{xz}^2. \tag{2.13}$$

In the present context, as will be shown in the next section, the limiting distribution of consistent estimators is invariant with respect to whether or not one conditions on exogenous variables. Therefore, inference based on asymptotic arguments is not directly affected, which explains why the issue of sampling schemes is not often discussed in the econometrics literature. However, we shall also demonstrate that often – also for consistent estimators – finite sample distributions may be affected substantially by such conditioning. Therefore, the actual accuracy of asymptotic inference may be highly dependent on whether it is used to provide information on a much larger population, or whether it is just used to answer questions about, say, a given set of countries or states, or a particular time period over which the exogenous variables had particular values and conditioning on them seems useful, simply to make the inferences more relevant. In addition to all that, we will also demonstrate in the next section that the limiting distribution of inconsistent estimators is not invariant with respect to conditioning.

For the simple DGP examined here, we will distinguish three different conditioning sets, indicated by $\mathcal{C}_{\varnothing}$, $\mathcal{C}_z$ and $\mathcal{C}_{x^*}$ respectively. They are defined as follows[2]:

$$\left.\begin{array}{l} \mathcal{C}_{\varnothing}: \text{ conditioning set is empty (unconditional)} \\ \mathcal{C}_z: \text{ conditioning set is } \{z_1, ..., z_n\} \\ \mathcal{C}_{x^*}: \text{ conditioning set is } \{x_1^*, ..., x_n^*\} \end{array}\right\} \tag{2.14}$$

Obviously, when conditioning simulation results on a single arbitrary realization of a random exogenous series, this will yield specific results, since a different realization will yield different results. Below we will cope with that in two ways, namely by producing results for a few different arbitrary realizations, and also by stylizing the "arbitrary" realizations in a particular way, and then producing results for a few stylized arbitrary realizations.

The actual effects on estimators of conditioning on exogenous variables have not yet been fully documented in the literature. In studies primarily aiming to understand the effects of weakness of instruments Nelson and Startz (1990a,b) examine IV in the present simple model exclusively under $\mathcal{C}_{x^*}$, whereas Maddala and Jeong (1992), Woglom (2001), Hillier (2006) and Forchini (2006) focus on $\mathcal{C}_z$. So does Phillips (2006, 2009), who also examines OLS, but especially under further parametric restrictions which bring in an identity and lead to $\mathcal{C}_z = \mathcal{C}_{x^*}$. Then the DGP refers to the classic consumption function model $c_i = \gamma x_i + \varepsilon_i$ (where x_i is income, $0 < \gamma < 1$), supplemented and closed by the budget restriction equation $x_i = c_i + z_i$, where z_i is exogenous. Defining $y_i = c_i - \gamma x_i$ this yields $y_i = \beta x_i + \varepsilon_i$ (with true value $\beta = 0$) and reduced form equation

$$x_i = \frac{1}{1-\gamma} z_i + \frac{1}{1-\gamma} \varepsilon_i. \tag{2.15}$$

So, this represents the very special parameterization $\pi = \xi = 1/(1-\gamma)$ with $\sigma_{v^*}^2 = 0$. Instead of three parameters, this DGP has just one free parameter σ_x, since from $\rho_{x\varepsilon} = \rho_{xz}$ and $\rho_{x\varepsilon}^2 + \rho_{xz}^2 = 1$ it follows that $\rho_{x\varepsilon} = \rho_{xz} = \frac{1}{2}\sqrt{2} \approx 0.7$, whilst $PCP = n$. Hence, the simultaneity is always serious (in fact the conditional simultaneity coefficient is unity), and, according to the rule of thumb devised by Staiger and Stock (1997), the instrument is not weak when $n > 10$.

The above should illustrate that our present-day understanding of the finite sample distribution of IV estimators largely refers just to very particular cases, mostly in a setting in which some kind of conditioning on exogenous variables has been accommodated. We will demonstrate that a major finding in the established literature on the shape of IV probability densities (often bimodal and even zero between the modes) is not typical for the three parameter model examined here, but only occurs in very specific cases under a conditional sampling scheme. However, before we examine simulated distributions also for less restricted settings, we will first design a generic approach by which we can derive for both consistent and inconsistent estimators what the effects of conditioning are on their limiting distributions.

3. Limiting and simulated distributions of IV and OLS

Throughout the paper we make the following basic assumption:

Assumption 1 (basic DGP). *The single simple structural equation (2.4) for y_i is completed by the reduced form equation (2.3) for x_i, whereas the external instrumental variable z_i and*

[2] It does not seem to make sense to condition just on v_i^* and not on z_i. Further motivation to examine the consequences of conditioning on x_i^* are given in Appendix A.

the two disturbance terms are characterized by (2.1) and (2.2). Moreover, $0 < \sigma_\varepsilon^2 < \infty$, $0 < \sigma_{v^*}^2 < \infty$, $0 < \sigma_z^2 < \infty$, $0 < \sigma_x^2 < \infty$ *and* $\pi \neq 0$.

Hence, we restrict ourselves to the standard case where the variable z_i is an appropriate instrument in the sense that it is both valid and relevant, because the orthogonality condition $E[z_i(y_i - \beta x_i)] = 0$ and the identification rank condition $E(z_i x_i) \neq 0$ hold. Hence, β is point-identified, although the identification may be weak when ρ_{xz} is close to zero. In the asymptotic analysis to follow we do not consider consequences of instrument weakness in which one assumes $\pi = \rho_{xz}\sigma_x/\sigma_z = O(n^{-1/2})$, as in Staiger and Stock (1997). To maintain generality of the analytic results, we will not impose yet the normalization restrictions to be used later in the simulations, which are: $\beta = 0$, $\sigma_\varepsilon^2 = 1$ and $\sigma_z^2 = 1$.

For particular results we specify the following further regularity assumptions:

Assumption 2 (regularity ε). *For the structural form disturbances* ε_i *we have* $E(\varepsilon_i^4) = 3\sigma_\varepsilon^4$.

Assumption 3 (regularity v). *For the reduced form disturbances* v_i *we have* $E(v_i^4) = 3\sigma_v^4$.

Assumption 4 (regularity x). *For the regressor* x_i *we have* $E(x_i^4) = 3\sigma_x^4$.

In the next two subsections we examine the limiting and the finite sample distributions of the IV and OLS estimators of β, which are given by

$$\hat{\beta}_{IV} = \textstyle\sum_{i=1}^n z_i y_i / \sum_{i=1}^n z_i x_i = \beta + \sum_{i=1}^n z_i \varepsilon_i / \sum_{i=1}^n z_i x_i \tag{3.1}$$

and

$$\hat{\beta}_{OLS} = \textstyle\sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2 = \beta + \sum_{i=1}^n x_i \varepsilon_i / \sum_{i=1}^n x_i^2. \tag{3.2}$$

3.1. Limiting distributions

The following lemma allows to investigate the effects of conditioning regarding realized random variables on the limiting distribution of an estimator for which the estimation error is given by a ratio of scalar sample aggregates. All proofs are collected in Appendix B.

Lemma (scalar conditional limiting distribution). *Let the estimation error* $\hat{\beta} - \beta = N/D$ *be a ratio of scalar random variables, where the numerator* $N = \sum_{i=1}^n N_i$ *and the denominator* $D = \sum_{i=1}^n D_i$ *are both sample aggregates for a sample of size* n*, whereas both* N *and* D *can be decomposed, employing some set of conditioning variables* $\mathcal{C}$*, such that* $N = \bar{N} + \tilde{N}$ *and* $D = \bar{D} + \tilde{D}$*, where* $\bar{N} \equiv E(N \mid \mathcal{C}) = O(n)$ *and* $\bar{D} \equiv E(D \mid \mathcal{C}) = O(n)$*, while* $\tilde{N} \mid \mathcal{C} = O_p(n^{1/2})$ *and* $\tilde{D} \mid \mathcal{C} = O_p(n^{1/2})$*. Then, provided* $\bar{D} \neq 0$, $n^{1/2}(\hat{\beta} - \beta - \bar{N}/\bar{D}) \mid \mathcal{C} \xrightarrow{d} \mathcal{N}(0, V)$*, where* $V = (\lim n^{-1}\bar{D})^{-2} V_0$ *with* V_0 *such that* $n^{-1/2}(\tilde{N} - \bar{N}\tilde{D}/\bar{D}) \mid \mathcal{C} \xrightarrow{d} \mathcal{N}(0, V_0)$.

This lemma can be applied to both IV and OLS under the three sampling schemes distinguished above. We find that the limiting distribution of the consistent IV estimator is not affected by these three sampling schemes, whereas that of the inconsistent OLS estimator is.

Theorem 1: *Let Assumption 1 hold. Under the sampling schemes set out in (2.14) the limiting distributions of simple IV and OLS estimators are given by:*

$$(a) \qquad n^{1/2}(\hat{\beta}_{IV} - \beta) \mid \mathcal{C}_j \xrightarrow{d} \mathcal{N}\left(0, \tfrac{\sigma_\varepsilon^2}{\sigma_x^2}\tfrac{1}{\rho_{xz}^2}\right) \text{ for } j \in \{\varnothing, z, x^*\},$$

$$(b1) \qquad n^{1/2}\left(\hat{\beta}_{OLS} - \beta - \rho_{x\varepsilon}\tfrac{\sigma_\varepsilon}{\sigma_x}\right) \mid \mathcal{C}_\varnothing \xrightarrow{d} \mathcal{N}\left(0, \tfrac{\sigma_\varepsilon^2}{\sigma_x^2}(1-\rho_{x\varepsilon}^2)\right),$$

$$(b2) \qquad n^{1/2}\left(\hat{\beta}_{OLS} - \beta - \rho_{x\varepsilon}\tfrac{\sigma_\varepsilon}{\sigma_x}[1-\rho_{xz}^2(1-\tfrac{1}{n\sigma_z^2}\textstyle\sum_{i=1}^n z_i^2)]^{-1}\right)\Big| \mathcal{C}_z \xrightarrow{d} \mathcal{N}\left(0, \tfrac{\sigma_\varepsilon^2}{\sigma_x^2}[1-\rho_{x\varepsilon}^2(1+2\rho_{xz}^4)]\right),$$

$$(b3) \qquad n^{1/2}\left(\hat{\beta}_{OLS} - \beta - \rho_{x\varepsilon}\tfrac{\sigma_\varepsilon}{\sigma_x}[\rho_{x\varepsilon}^2 + (1-\rho_{x\varepsilon}^2)\tfrac{1}{n\sigma_{x^*}^2}\textstyle\sum_{i=1}^n x_i^{*2}]^{-1}\right)\Big| \mathcal{C}_{x^*} \xrightarrow{d} \mathcal{N}\left(0, \tfrac{\sigma_\varepsilon^2}{\sigma_x^2}(1-\rho_{x\varepsilon}^2)[1-2(1-\rho_{x\varepsilon}^2)\rho_{x\varepsilon}^2]\right).$$

Additional requirements for result (b1) are Assumptions 2 and 4, for (b2) Assumptions 2 and 3, and for result (b3) Assumption 2.

The underlying reasons for the equivalence of the IV results under conditioning are that for the numerator we have $E(N \mid \mathcal{C}_z) = E(N \mid \mathcal{C}_{x^*}) = 0$ and (therefore) also $E(N) = 0$ and the limiting distribution will always be centered around β. It can be obtained simply from $(\bar{D}/n)^{-1}n^{-1/2}\tilde{N}$, where $\lim \bar{D}/n = \sigma_{zx}$ and $n^{-1/2}\tilde{N} \xrightarrow{d} \mathcal{N}(0, \sigma_\varepsilon^2\sigma_z^2)$, irrespective of the conditioning used. The situation is different for the inconsistent estimator. Here $\bar{N} = n\rho_{x\varepsilon}\sigma_x\sigma_\varepsilon$ differs from zero under simultaneity, so the center of the limiting distribution is not at β. Its shift $\bar{N}/\bar{D}$ differs for different conditioning variables, because $\bar{D}$ is an expression involving these conditioning variables. This also affects the second term in $\tilde{N} - \bar{N}\tilde{D}/\bar{D}$, which leads to different expressions for V_0 and thus for V. Note that $\bar{N}/\bar{D}$ is (or converges to) the inconsistency $\rho_{x\varepsilon}\sigma_\varepsilon/\sigma_x$, which is bounded, because $\rho_{x\varepsilon}$ and $\sigma_\varepsilon^2/\sigma_x^2$ are too.

Result (b1) can already be found in Goldberger (1964, p.359), who considers a model with more possibly endogenous regressors. However, he does not mention that it only holds when the IID observations on all the variables have a 4th moment corresponding to that of the Normal distribution. Result (b2) corresponds to a (in our opinion less transparent) formula given in Hausman (1978, p.1257) and Hahn and Hausman (2003, p.124), who refer for its derivation to Rothenberg (1972, p.9). Result (b3) can be found in Kiviet and Niemczyk (2007, 2012), where the issue of conditioning is not addressed explicitly. Similar distinct limiting distributions due to using alternative conditioning sets have been obtained in the context of errors in variables models by Schneeweiss (1980).

Result (a) shows that the asymptotic variance of IV is inversely related to the strength of the instrument, and the asymptotic distribution is invariant with respect to: (i) the degree of simultaneity $\rho_{x\varepsilon}$, (ii) regarding the value of $\sigma_z^2 > 0$ and (iii) with respect to the actual distribution of the disturbances, regressor and instrument.

When there is no simultaneity ($\rho_{x\varepsilon} = 0$) the asymptotic variance of OLS (both conditionally and unconditionally) is $\sigma_\varepsilon^2/\sigma_x^2$. Irrespective of the simultaneity, the asymptotic variance of IV using instrument z, is equal to this ratio $\sigma_\varepsilon^2/\sigma_x^2$ multiplied by ρ_{xz}^{-2}. Since $|\rho_{xz}| \leq 1$, this factor is never smaller than 1. Applying OLS under simultaneity has the serious consequence of inconsistency, but also the advantage that it yields a smaller asymptotic variance, because this is given by the ratio $\sigma_\varepsilon^2/\sigma_x^2$ multiplied by a factor smaller than unity. In fact, when comparing the asymptotic variances of IV and OLS we find:

Corollary 1. *Under the conditions of Theorem 1 we have*

$$AsyVar(\hat{\beta}_{IV}) \geq AsyVar(\hat{\beta}_{OLS} \mid \mathcal{C}_{\varnothing}) \geq AsyVar(\hat{\beta}_{OLS} \mid \mathcal{C}_z) \geq AsyVar(\hat{\beta}_{OLS} \mid \mathcal{C}_{x^*}).$$

So, apart from the inconsistency, the OLS asymptotic distribution seems always more attractive than that of IV, whereas the asymptotic variance of OLS gets smaller by extending the conditioning set. Conditioning OLS on instrument z_i (which it does not exploit) does reduce the asymptotic variance more the stronger the instrument is, and extending the conditioning set from z_i to $x_i^* = \pi z_i + v_i^*$ decreases the asymptotic variance even more, provided $\sigma_{v^*}^2 > 0$.

The magnitudes of the four different multiplying factors of $\sigma_\varepsilon^2/\sigma_x^2$ are depicted in Diagram 1, in separate panels for six increasing values of the strength ρ_{xz} of the instrument, and with the degree of simultaneity $\rho_{x\varepsilon}$ on the horizontal axis. Note that these six panels have different scales on their horizontal axis, because its range of admissible values is restricted by $0 \leq \rho_{x\varepsilon} \leq \sqrt{1-\rho_{xz}^2}$. All depicted curves are symmetric around zero in both $\rho_{x\varepsilon}$ and ρ_{xz}, so we may restrict ourselves to examining just positive values. Diagram 1 shows that conditioning on instrument z has a noticeable effect on the OLS variance only when the instrument is sufficiently strong, say $|\rho_{xz}| > 0.5$. Adding v_i^* to the conditioning set already consisting of z_i has little effect when the instrument is very strong and $\sigma_{v^*}^2$ is correspondingly small. But for $|\rho_{xz}| < 0.8$ it has a noteworthy effect, especially for intermediate values of the simultaneity $\rho_{x\varepsilon}$ (then the discrepancy between the circles and the crosses is substantial). When the strength parameter is below 0.8 the factor $1/\rho_{xz}^2$ for IV is so much larger than unity that it is difficult to combine it with the OLS results in the same plot. Therefore, the IV results are shown only in the bottom two panels referring to a very strong instrument.

Due to the inconsistency of OLS we should not make a comparison with IV purely based on just asymptotic variance. Therefore, in Diagram 2, we compare the asymptotic root mean squared errors (ARMSE). For IV this is simply $\sigma_\varepsilon/(\sigma_x \rho_{xz})$. For OLS we have to take into account also the squared inconsistency multiplied by n, because we confront it with the asymptotic variance. Hence, for unconditional OLS this leads to

$$ARMSE(\hat{\beta}_{OLS} \mid \mathcal{C}_{\varnothing}) = \frac{\sigma_\varepsilon}{\sigma_x}[(1-\rho_{x\varepsilon}^2) + n\rho_{x\varepsilon}^2]^{1/2},$$

and similarly for the conditional OLS results. In Diagram 2 we present the ratio (hence, $\sigma_\varepsilon/\sigma_x$ drops out) with the IV result in the denominator, giving the ARMSE of OLS as a fraction of the IV figure. There is symmetry again regarding $\rho_{x\varepsilon}$ and ρ_{xz}. The only further determinant is n. When this is not extremely small the effect of the inconsistency is so prominent that the differences in variance due to the conditioning setting become trifling and the curves for the three (un)conditional OLS results almost coincide. OLS beats IV according to its asymptotic distribution when the curve is below the dashed line at unity. Hence, we note that for $n = 100$ IV seems preferable when $|\rho_{xz}| > .5$ and $|\rho_{x\varepsilon}| > .2$. However, when $|\rho_{xz}| < .5$ and $n = 30$ the ARMSE of OLS could be half or just one fifth of that of IV. Whether this is useful information for samples that actually have size 30 depends on the accuracy of these asymptotic approximations, which we will check in the next section. Especially when the instrument is weak special weak-instrument asymptotic results are available, see Andrews and Stock (2007a), which might be more appropriate for making comparisons with the ARMSE of OLS.

3.2. Simulated distributions

We obtained the empirical densities of OLS and IV by generating the estimators $R \geq 50,000$ times and distinguishing between the sampling schemes; all series (2.1) were drawn from the

Normal distribution and we took – without loss of generality – $\beta = 0$ and $\sigma_\varepsilon = \sigma_z = 1$. In all simulations we chose $\sigma_x = 3$. This does not lead to loss of generality either, because all depicted densities are multiples of σ_x, which means that the form and relative positions of the densities would not change if we changed σ_x; the only effect would be that the scale on the horizontal axis would be different. We examined the cases $\rho_{x\varepsilon} = 0, .3, .5, .7$. This gives 4 panels of densities. Most diagrams contain results for both $\rho_{xz} = .4$ and $\rho_{xz} = .1$; then the top four panels in a diagram are those for the not so weak instrument, and the bottom four those for the much weaker instrument. Most densities are for sample size $n = 100$, but in some we took $n = 1000$. All the relevant design parameter values and their consequences for PCP are given in the diagrams.

Diagram 3 considers the unconditional case, so in all replications new vectors for the three series (2.1) have been drawn. The IV density is dashed (red) and the OLS density is solid (blue). The panels also show the standard asymptotic approximations for IV (dotted, red) and for OLS (dotted, blue). By not fully displaying the mode of all distributions we get a clearer picture of the tails. Even at $\rho_{xz} = .4$, where $PCP = 19$, the asymptotic approximation for IV is not very accurate, especially for $\rho_{x\varepsilon}$ high. Although the limiting distribution is invariant for $\rho_{x\varepsilon}$ this is clearly not the case for the finite sample distribution of IV. The asymptotic approximation for the OLS distribution is so accurate that in all these drawings it collapses with the empirical distribution. For $\rho_{xz} = .1$ the standard IV approximation starts to get really bad, but it is much worse for even smaller values of ρ_{xz} (not depicted). For $\rho_{x\varepsilon}$ away from zero OLS is clearly biased, but so is IV when the instrument is weak. In most of the panels OLS seems preferable to IV, because most of its actual probability mass is much closer to the true value of zero than is the case for IV.

Diagram 4 presents results in which the z_i series has been kept fixed. Both IV and OLS are plotted for 6 different arbitrary realizations of z_i (in the diagram indicated by 6×raw). For OLS the densities almost coincide, except for the stronger instrument when simultaneity is serious. For IV the effects of conditioning are more pronounced, especially for the stronger instrument. In Diagram 5 the same 6 realizations of the z_i series have been used after stylizing them by normalizing them such that their first two sample moments correspond to the population moments (indicated by 6×stylized). Now all 6 curves almost coincide, and apparently the effects of higher order sample moments not fully corresponding with their population counterpart has hardly any effects.

In Diagram 6 we conditioned on both z_i and v_i^*, again using 6 arbitrary realizations. Note that although $E(z_i v_i^*) = 0$ the sample correlation coefficient r_{zv^*} may deviate from zero. The effects of this type of conditioning are much more pronounced now, just a little for OLS, but extremely so for IV in the weak instrument case. Then, the empirical distribution of IV demonstrates characteristics completely different from its standard asymptotic approximation; especially for low ρ_{xz} and high $\rho_{x\varepsilon}$ bimodality is manifest with also a region of zero density between the two modes. Stylizing the z_i and v_i^* series (also making sure that their sample covariance is zero) makes the curves coincide again in Diagram 7, but shows bimodality again in the case where the instrument is weak and the simultaneity serious.

From all these densities it seems obvious that the distribution of IV gets more problematic under conditioning, and that it seems worthwhile to develop a method such that OLS could be employed for inference. That this conclusion does not just pertain to very small samples can be learned from Diagram 8, where $n = 1000$ and $\rho_{xz} = .1$, giving rise to PCP just above 10. The top four diagrams are unconditional, whereas the bottom four diagrams present results conditional on x_i^*. In most cases all realizations of the OLS estimator are much closer to the

true value than most of those obtained by IV. Especially if OLS could be corrected for bias (without a major increase in variance) it would provide a welcome alternative to IV, not only when the sample is small, and even when the instrument is not all that weak.

4. Robust OLS inference under simultaneity

In Kiviet and Niemczyk (2012) an attempt was made to design inference on β based on a bias corrected OLS estimator and an assumption regarding $\rho_{x\varepsilon}$. An infeasible version, which used also the true values of σ_x and σ_ε worked well, which is promising. However, after replacing σ_x and σ_ε by sample equivalents it became clear that such a substitution at the same time alters the limiting distribution of the bias corrected estimator, so that further refinements of the asymptotic approximations are called for.

4.1. The limiting distribution of bias corrected OLS

We shall derive the appropriate limiting distribution in the hope that it will lead to an asymptotic approximation that is still reasonably accurate in finite samples. We will focus on the unrestrained sampling scheme, because we found that the variance of OLS has a tendency to decrease by conditioning on exogenous regressors. Therefore, any resulting unconditional (regarding the exogenous variables) confidence sets are expected to have larger coverage probability (thus will be conservative) under conditioning. Moreover, when we will compare unconditional robust OLS inference with unconditional IV inference and the former is found to provide a useful alternative, then we would know for sure that this will certainly be the case under a restrained sampling scheme, because we have seen that conditioning worsens the situation for IV and improves it for OLS.

So, our starting point is result (b1) of Theorem 1. This suggests the unfeasible bias corrected OLS-based estimator $\hat{\beta}_{UOLS}(\rho_{x\varepsilon}, \sigma_\varepsilon/\sigma_x) \equiv \hat{\beta}_{OLS} - \xi\sigma_\varepsilon^2/\sigma_x^2 = \hat{\beta}_{OLS} - \rho_{x\varepsilon}\sigma_\varepsilon/\sigma_x$ for β, with limiting distribution

$$n^{1/2}\left(\hat{\beta}_{UOLS}(\rho_{x\varepsilon}, \sigma_\varepsilon/\sigma_x) - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\sigma_x^2}(1-\rho_{x\varepsilon}^2)\right). \tag{4.1}$$

This estimator is unfeasible, because it requires the values of $\rho_{x\varepsilon}$ and $\sigma_\varepsilon/\sigma_x$. Below we shall show that the noise-signal related quantity $\sigma_\varepsilon/\sigma_x$ can be estimated consistently, but this estimator is unfeasible as well, although it just requires knowledge of the value of $\rho_{x\varepsilon}$ or a consistent estimator of it. The latter would require exploiting at least one valid and relevant instrument z_i, which in its turn is a feasible procedure only when adopting the value zero for $\rho_{z\varepsilon}$. Since testing the hypothesis $\rho_{z\varepsilon} = 0$ is only possible by exploiting at least one other valid and relevant instrument (which should be linearly independent from z_i), we find ourselves in a recursion. This highlights that making an untestable assumption is unavoidable. One either needs to adopt validity of at least one instrument, or one should adopt a particular value for $\rho_{x\varepsilon}$, provided that this would lead to an accurate and practicable inference procedure. We set out to examine whether this is the case.

Consider the estimator for β

$$\hat{\beta}_{KLS}(\rho_{x\varepsilon}) \equiv \hat{\beta}_{OLS} - \rho_{x\varepsilon}\left(\frac{\hat{\sigma}_\varepsilon^2/(1-\rho_{x\varepsilon}^2)}{n^{-1}\sum_{i=1}^n x_i^2}\right)^{1/2}, \tag{4.2}$$

which is a function of the OLS estimator for β, the (ML) estimator for σ^2 given by $\hat{\sigma}_\varepsilon^2 \equiv n^{-1}\sum_{i=1}^n (y_i - x_i\hat{\beta}_{OLS})^2$ and of $\rho_{x\varepsilon}$. We nickname this estimator KLS (kinky least-squares) for which we find the following.

Theorem 2 (unconditional limiting distribution of KLS). *Under Assumptions 1, 2 and 4 we have* $n^{1/2}(\hat{\beta}_{KLS} - \beta) \to \mathcal{N}(0, \sigma_\varepsilon^2/\sigma_x^2)$.

Hence, remarkably, the standard OLS estimator, which has asymptotic variance $\sigma_\varepsilon^2/\sigma_x^2$ when $\rho_{x\varepsilon} = 0$, and when $\rho_{x\varepsilon} \neq 0$ has inconsistency $\rho_{x\varepsilon}\sigma_\varepsilon/\sigma_x$ and – provided both ε and x have fourth moments as if they were Gaussian – the smaller unconditional asymptotic variance $\sigma_\varepsilon^2(1-\rho_{x\varepsilon}^2)/\sigma_x^2$, has again the original asymptotic variance $\sigma_\varepsilon^2/\sigma_x^2$ after one has subtracted an OLS-based estimate of its inconsistency based on the true value of $\rho_{x\varepsilon}$. To estimate the variance of $\hat{\beta}_{KLS}$ one should not use the standard expression $\widehat{Var}(\hat{\beta}_{OLS}) = \hat{\sigma}_\varepsilon^2/\sum_{i=1}^n x_i^2$, but

$$\widehat{Var}(\hat{\beta}_{KLS}) \equiv \frac{\hat{\sigma}_\varepsilon^2}{(1-\rho_{x\varepsilon}^2)\sum_{i=1}^n x_i^2} = \frac{1}{1-\rho_{x\varepsilon}^2}\widehat{Var}(\hat{\beta}_{OLS}), \tag{4.3}$$

because that yields $\operatorname{plim} n\widehat{Var}(\hat{\beta}_{KLS}) = \sigma_\varepsilon^2/\sigma_x^2$, as it should.

4.2. Simulation results on KLS-based inference

In a Monte Carlo study based on 100,000 replications we have examined the bias of both $\hat{\beta}_{OLS}$ and $\hat{\beta}_{KLS}$, the Monte Carlo estimate of $Var(\hat{\beta}_{KLS})$ and the estimated expectation of its empirical estimate $\widehat{Var}(\hat{\beta}_{KLS})$. Also we have examined the coverage probability of an asymptotic KLS based confidence interval with nominal confidence coefficient of 95%. Hence, we analyzed the frequency over the Monte Carlo replications by which the true value of β was covered by the interval

$$[\hat{\beta}_{KLS} - 1.96 \times \widehat{SD}(\hat{\beta}_{KLS}), \hat{\beta}_{KLS} + 1.96 \times \widehat{SD}(\hat{\beta}_{KLS})], \tag{4.4}$$

where both $\hat{\beta}_{KLS}$ and $\widehat{SD}(\hat{\beta}_{KLS}) \equiv (\widehat{Var}(\hat{\beta}_{KLS}))^{1/2} = (1-\rho_{x\varepsilon}^2)^{-1/2}\widehat{SD}(\hat{\beta}_{OLS})$ are calculated on the basis of the true value of $\rho_{x\varepsilon}$.

Since the simulation results in Table 1 are purely OLS based, they do not require the availability of an instrumental variable, and thus instrument strength is irrelevant. No conditioning on exogenous variables occurred and all results are invariant regarding the true values of β and σ_ε. Just n, $\rho_{x\varepsilon}$ and $\sigma_x/\sigma_\varepsilon$ matter, but the latter is found (and can be shown) not to affect the coverage probability of KLS confidence intervals. The bias of OLS is found to be (almost) invariant with respect to n, and to be extremely close to the inconsistency, which predicts it to increase with (the absolute value of) $\rho_{x\varepsilon}$ and to be inversely related to $\sigma_x/\sigma_\varepsilon$. The KLS coefficient estimator, which is based on the true value of $\rho_{x\varepsilon}$, is almost unbiased, and variance estimator (4.3) is found to have a minor negative bias for the true variance. The actual coverage probability of the KLS confidence intervals is remarkably close to the nominal value of 95%. These estimates have a Monte Carlo standard error of about .0007, hence we establish a significant slight under-coverage over all the design parameter values examined, also when $\rho_{x\varepsilon} = 0$. The latter is due to using a critical value from the normal and not from the Student distribution. As far as the slight under-coverage is due to inaccuracies for n small in the OLS bias approximations, this could be repaired possibly by employing higher-order approximations as in Kiviet and Phillips (1996). The appearance of the factor $1-\rho_{x\varepsilon}^2$ in the denominator of (4.3) indicates that $\rho_{x\varepsilon}$ values really close to unity will be problematic.

Table 1 Qualities of KLS-based inference

$\sigma_x/\sigma_\varepsilon$	n	$\rho_{x\varepsilon}$	$E(\hat{\beta}_{OLS})-\beta$	$E(\hat{\beta}_{KLS})-\beta$	$Var(\hat{\beta}_{KLS})$	$E[\widehat{Var}(\hat{\beta}_{KLS})]$	Cov.Prob.
3	30	.0	-.000	-.000	.00395	.00397	.9405
		.3	.100	-.002	.00396	.00397	.9409
		.6	.200	-.003	.00401	.00397	.9410
		.9	.300	-.005	.00408	.00397	.9401
	300	.0	.000	.000	.000372	.000373	.9487
		.3	.100	-.000	.000374	.000373	.9481
		.6	.200	-.000	.000375	.000373	.9479
		.9	.300	-.000	.000377	.000373	.9483
10	30	.3	.030	-.001	.000356	.000358	.9409
		.9	.090	-.002	.000367	.000357	.9401
	300	.3	.030	-.000	.000034	.000034	.9481
		.9	.090	-.000	.000034	.000034	.9483

Interval (4.4) is based on standard OLS statistics only. When $\zeta_{\alpha/2}$ expresses the $\alpha/2$ quantile of the standard Normal distribution an interval with confidence coefficient $1-\alpha$ is obtained by

$$\hat{\beta}_{OLS} - [(n^{1/2}\rho_{x\varepsilon} \pm \zeta_{\alpha/2})/(1-\rho_{x\varepsilon}^2)^{1/2}]\widehat{SD}(\hat{\beta}_{OLS}). \tag{4.5}$$

It can deviate substantially from a standard OLS interval when $\rho_{x\varepsilon}$ is far from zero. Its location with respect to the standard OLS interval shifts, due to the bias correction. Moreover, its width is multiplied by the factor $(1-\rho_{x\varepsilon}^2)^{-1/2}$ and hence bulges up for $\rho_{x\varepsilon}^2$ close to unity. If one wants to base inference on an interval assumption, say $\rho_{x\varepsilon} \in [\rho_{x\varepsilon}^L, \rho_{x\varepsilon}^R]$, then a conservative $(1-\alpha)100\%$ confidence interval is obtained by taking the union of all intervals for $\rho_{x\varepsilon}$ in that interval.

Diagrams 9 and 10 provide an impression of what such intervals might achieve in comparison to those provided by IV under ideal circumstances, namely based on the actual finite sample distribution of IV, which in practice is very hard to approximate, whereas we have seen that this works fine for the KLS estimator. Diagram 9 replicates the results of Diagram 3 for the finite sample distributions of IV and OLS, but in addition it shows the densities of KLS when based on an (in)correct assumption regarding $\rho_{x\varepsilon}$. The blue curve uses $\rho_{x\varepsilon} = 0$, so here KLS simplifies to OLS. The other densities use the assumptions $\rho_{x\varepsilon} = -.9$ (red), $-.6$ (green) or $-.3$ (cyan), .3 (magenta), .6 (yellow), .9 (black). Hence, in each panel at most one KLS curve uses a $\rho_{x\varepsilon}$ value very close to its true value (which is mentioned in the panel). The superiority of KLS when employing a $\rho_{x\varepsilon}$ value that is wrong by about a ± 0.3 margin is apparent, because it yields a probability mass much closer to the true value of zero than IV, especially when the instrument is weak. In Diagram 10 densities of the same estimators are presented but now for $n = 1000$. For a not very weak instrument and a large sample IV may do better, but for cases where PCP is smaller than or around 10 the instrument-free KLS inference seems a very welcome alternative. This will be even more the case in practice, when the available (weak) instrument could actually be endogenous.

4.3. Comparison of the qualities of alternative confidence sets

As already argued above, the reputation of any inference technique should be based on criteria such as: is it based on credible assumptions; does it fulfill (also in finite samples) its promises regarding its claimed accuracy; can it boast other credentials, such as robustness; and is its inference more efficient than that produced by competing techniques. As a rule, in most contexts there does not exist a single technique which uniformly dominates all others under all relevant circumstances.

For the DGP examined above we will now compare the properties of various confidence sets on the scalar parameter β. We will examine KLS-based confidence sets with IV-based confidence sets, where the latter are constructed from either the standard asymptotic Wald (W) test statistic or by inversion of the Anderson-Rubin (AR) statistic. The W procedure is known to be defective when instruments are weak, whereas the AR procedure is known to be robust under the null when instruments are weak. Mikusheva (2010) claims that it performs reasonably well amongst other techniques which focus on invariance of the asymptotic null distribution with respect to weakness of instruments. However, in the present just identified case these other (LM and LR based) techniques simplify to the AR procedure.

In the present context, the Wald statistic for $H_0 : \beta = \beta_0$ is given by

$$W(\beta_0) = \left(\frac{\hat{\beta}_{IV} - \beta_0}{\widehat{SD}(\hat{\beta}_{IV})} \right)^2, \text{ with } \widehat{SD}(\hat{\beta}_{IV}) = \left(\frac{(y - x\hat{\beta}_{IV})'(y - x\hat{\beta}_{IV})/(n-1)}{x'z(z'z)^{-1}z'x} \right)^{1/2}, \tag{4.6}$$

where y, x and z are $n \times 1$ vectors containing all the sample data. For nominal $(1-\alpha) \times 100\%$ confidence coefficient, it yields confidence interval

$$\left(\hat{\beta}_{IV} + \zeta_{\alpha/2}\widehat{SD}(\hat{\beta}_{IV}), \hat{\beta}_{IV} + \zeta_{1-\alpha/2}\widehat{SD}(\hat{\beta}_{IV}) \right). \tag{4.7}$$

The AR test statistic for the same hypothesis is performed by substituting the reduced form equation in the structural form relationship and next testing the significance of the slope in the equation $y - x\beta_0 = z\pi(\beta - \beta_0) + \varepsilon + (\beta - \beta_0)v$. Under the null the tests statistic

$$AR(\beta_0) = (n-1)\frac{(y - x\beta_0)'z(z'z)^{-1}z'(y - x\beta_0)}{(y - x\beta_0)'[I - z(z'z)^{-1}z'](y - x\beta_0)} \tag{4.8}$$

is asymptotically distributed as $\chi^2(1)$ and under normality exactly as $F(1, n-1)$. It implies the confidence interval

$$C_{AR}(\alpha) = \{\beta_0 : AR(\beta_0) < F_\alpha(1, n-1)\}. \tag{4.9}$$

This can be established by solving $a\beta^2 + b\beta + c \geq 0$, where $a = x'Ax$, $b = -2x'Ay$ and $c = y'Ay$, with $A = I - dz(z'z)^{-1}z'$ and $d = 1 + (n-1)/F_\alpha(1, n-1)$. If $\Delta = b^2 - 4ac \geq 0$ then an interval follows easily provided $a \leq 0$; when $a > 0$, though, the interval consists of the whole real line, except for a finite interval, and thus has infinite length. When $\Delta < 0$ the interval is empty for $a < 0$ and equals the full real line when $a > 0$ (note that $\Delta < 0$ and $a = 0$ cannot occur). These anomalies, if they really do occur, would be the result of uncomfortable behavior of the AR statistic under the alternative hypothesis, implying either unit rejection probability for all alternatives (CI empty), or zero rejection probability for either all alternatives (CI is the full real line) or for just a closed set of alternatives (CI is the real line, except for a particular interval). In the simulations to follow, we will monitor the occurrence of these anomalies.

Since we will focus on the width of the interval, we will examine the $\Delta \geq 0$ and $a \leq 0$ cases exclusively, in order to exclude non-existing intervals and intervals of infinite length. For those with finite length we will report the median width, because the occurrence of outliers clearly indicates that for both W and AR the random variable width has no finite moments.

Regarding the KLS-based intervals we will present results for three intervals $[\rho^L_{x\varepsilon}, \rho^U_{x\varepsilon}]$, namely: (A) $\rho^L_{x\varepsilon} = \rho^U_{x\varepsilon} = \rho_{x\varepsilon}$; (B) $\rho^L_{x\varepsilon} = \rho_{x\varepsilon} - 0.2, \rho^U_{x\varepsilon} = \rho_{x\varepsilon} + 0.2$; and (C) $\rho^L_{x\varepsilon} = 0$, $\rho^U_{x\varepsilon} = 0.5$. Of course, in practice it will be unrealistic that one could ever attain (A), and although (B) is more realistic, in practice one might specify occasionally an interval that excludes the true value. This is possible with interval (C), which simply states that any simultaneity will be non-negative and not exceeding 0.5. We ran 100,000 replications for a few combinations of $\rho_{x\varepsilon}$ and ρ_{xz} values for $n = 100$, giving rise to PCP values equal to 56.25, 4.16 and 1.01 respectively.

Table 2 Actual properties of various confidence intervals; $\sigma_x/\sigma_\varepsilon = 10$, $n = 100$, $\alpha = 0.05$
CP = coverage prob.; MW = median width; PF = prob. of finite width interval

$\rho_{x\varepsilon}$	ρ_{xz}	W		AR			KLS(A)		KLS(B)		KLS(C)	
		CP	MW	PF	CP	MW	CP	MW	CP	MW	CP	MW
.0	.6	.954	.065	1.00	.950	.069	.948	.039	1.000	.081	.973	.100
	.2	.993	.188	.52	.949	.246	.947	.039	1.000	.081	.973	.100
	.1	.998	.305	.17	.949	.349	.947	.039	1.000	.081	.973	.100
.2	.6	.953	.065	1.00	.950	.069	.947	.039	1.000	.083	1.000	.098
	.2	.985	.186	.52	.950	.244	.947	.039	1.000	.083	1.000	.098
	.1	.994	.300	.17	.944	.339	.947	.039	1.000	.083	1.000	.098
.4	.6	.949	.065	1.00	.950	.069	.947	.039	1.000	.091	.999	.091
	.2	.958	.179	.52	.951	.239	.947	.039	1.000	.091	.999	.091
	.1	.971	.283	.17	.924	.323	.947	.039	1.000	.091	.999	.091
.6	.6	.941	.064	1.00	.950	.070	.947	.039	1.000	.115	.656	.080
	.2	.905	.166	.52	.952	.234	.946	.039	1.000	.115	.657	.080
	.05	.903	.252	.17	.895	.293	.946	.039	1.000	.115	.658	.080

Table 2 shows that for a weak instrument the IV-based Wald test can both be conservative (when the simultaneity is moderate) and yield under-coverage (when the simultaneity is more serious). The AR procedure is of little practical use in this model, because it does not improve on the Wald test when the latter works well, and although it does produce intervals with appropriate coverage when the instrument is weak, much too frequently it does not deliver an interval (of finite length) at all. Moreover, the few intervals that it delivers when $\rho_{xz} = .1$ have larger width on average than those produced by the Wald procedure, even when the latter are conservative. The KLS(A) procedure performs close to perfect. Its coverage is not only very close to 95%, but its performance is also invariant with respect to both $\rho_{x\varepsilon}$ and ρ_{zx}, and the width of the interval is 60% (for $\rho_{xz} = .6$) or just about 15% (for $\rho_{xz} = .1$) of the Wald interval. The more realistic KLS(B) intervals are much too conservative, but nevertheless have smaller width than the Wald intervals when the instrument is weak. The same holds for the realistic KLS(C) intervals, provided that the true value of $\rho_{x\varepsilon}$ is in the interval. If this is not the case, then the KLS procedure breaks down.

4.4. Empirical illustration

The application that has undoubtedly received most attention over the last two decades in the debate and research on the effects of weakness and validity of instruments on the coefficient estimate of an endogenous variable is Angrist and Krueger (1991), who analyzed the returns to schooling on log wage. In Donald and Newey (2001) and Flores-Lagunes (2007) many variants of IV based estimates have been obtained. Donald and Newey (2001, p.1178) present also the OLS results for the coefficient estimate for schooling, being .0673 with a standard error equal to .0003. This has been obtained from a sample of size 329500. They also indicate sample second moments of the structural and reduced form disturbances which suggest estimates of the simultaneity coefficient equal to -.127, -.192 and -.204 according to 2SLS, LIML and B2SLS respectively. However, when the IV estimates are biased or even inconsistent due to the use of invalid or weak instruments these assessments may be misleading. From the OLS estimates we can deduce for different assumptions on $\rho_{x\varepsilon}$ the KLS inferences collected in Table 3.

Table 3 KLS estimates and confidence intervals for the effect of schooling

$\rho_{x\varepsilon}$	$\hat{\beta}_{KLS}$	$\widehat{SD}(\hat{\beta}_{KLS})$	95% CI for β
-.5	.1667	.0003	(.1660, .1674)
-.4	.1425	.0003	(.1418, .1431)
-.3	.1215	.0003	(.1208, .1221)
-.2	.1025	.0003	(.1019, .1031)
-.1	.0846	.0003	(.0840, .0852)
0	.0673	.0003	(.0667, .0679)
.1	.0500	.0003	(.0494, .0506)
.2	.0321	.0003	(.0315, .0327)
.3	.0131	.0003	(.0125, .0138)
.4	-.0079	.0003	(-.0085, -.0072)
.5	-.0321	.0003	(-.0328, -.0314)

Note that unlike the IV results the KLS inferences are not affected by the quality or validity of any of the external instrumental variables. However, like the IV results, they do assume that apart from the schooling variable all the partialled out regressors are exogenous. We find that when $\rho_{x\varepsilon}$ were close to -.2 indeed, KLS infers that the effects of schooling are close to .10 and in fact much closer than IV is able to produce. The IV-based techniques examined in Flores-Lagunes (2007) yield much wider confidence intervals of about $[.8, .13]$. From KLS one can deduce that such values are plausible only if $\rho_{x\varepsilon}$ is in the range (-.35,-.08). If $\rho_{x\varepsilon}$ is in fact mildly positive (as many have argued on theoretical grounds), then the effect of schooling on wage is actually much smaller than .10 and possibly just around .03 as Table 3 shows. Then the present IV findings would be the result of their inescapable frailty when instruments are weak or even invalid.

5. Conclusions

It is well-known now that the actual distribution of IV estimators gets rather anomalous when based on weak instruments. This paper shows that another factor causing serious deviations of

the finite sample distribution of IV from its standard asymptotic approximation is the sampling scheme. Conditioning on exogenous regressors aggravates the anomalies of IV. It leads more often to the occurrence of bimodality and occasionally to an area close to the median of the distribution where the density is zero. On the other hand, the distribution of the OLS estimator in simultaneous equations is much smoother, always unimodal and much less dispersed than for IV, it is much less affected by the sampling scheme and its asymptotic approximation is remarkably accurate even in small samples. Its major problem is its inconsistency, although this is bounded. We demonstrate how the OLS estimator can be corrected to render it consistent, without using instrumental variables at all. According to the established view, however, this corrected estimator is unfeasible, because it is based on an unknown parameter, namely the degree of simultaneity. But in fact similar problems afflict the single equation IV estimator, because its consistency is also based on making assumptions on parameter values, namely the coefficient values (possibly their exclusion) of at least as many regressors as the relationship involves endogenous regressors with unknown coefficients. Only a surplus of such restrictions and corresponding orthogonality conditions can be verified by empirical-statistical methods; an initial set as large as the number of included endogenous regressors requires parameter values chosen on the basis of rhetorical arguments to be deduced from economic theory or natural experiment settings. Even when these adopted restrictions are valid we show how vulnerable IV inferences are when instruments are weak. To be able to avoid this situation we derive the limiting distribution of bias corrected OLS, which allows to produce inference on the structural coefficient for a range of possible values of the simultaneity parameter. By taking the union of a series of confidence intervals we obtain relatively efficient, accurate, robust and often more credible inferences, without the need to nominate instrumental variables. In future research we plan to generalize the method for models under less strict independence restrictions involving more than just one endogenous regressor from which exogenous regressors are not necessarily partialled out in order to construct accurate external-instrument-free inference for all structural coefficients.

References

Andrews, D.W.K., Stock, J.H., 2007a. Inference with Weak Instruments, Chapter 6 in: Blundell, R., Newey, W.K., Persson, T. (eds.), *Advances in Economics and Econometrics, Theory and Applications*, 9th Congress of the Econometric Society, Vol. 3. Cambridge, UK: Cambridge University Press.

Andrews, D.W.K., Stock, J.H., 2007b. Testing with many weak instruments. *Journal of Econometrics* 138, 24-46.

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979-1014.

Angrist, J.D., Pischke, J-S., 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24, 3-30.

Conley, T.G., Hansen, C.B., Rossi, P.E., 2012. Plausibly exogenous. *The Review of Economics and Statistics* 94, 260-272.

Donald, S.G., Newey, W., 2001. Choosing the number of instruments. *Econometrica* 69, 1161-1191.

Erickson, T., 1993. Restricting regression slopes in the errors-in-variables model by bounding the error correlation. *Econometrica* 61, 959-969.

Flores-Lagunes, A., 2007. Finite sample evidence of IV estimators under weak instruments. *Journal of Applied Econometrics* 22, 677-694.

Forchini, G., 2006. On the bimodality of the exact distribution of the TSLS estimator. *Econometric Theory* 22, 932-946.

Goldberger, A.S., 1964. *Econometric Theory.* John Wiley & Sons. New York.

Hahn, J., Hausman, J.A., 2003. Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review* 93, 181-125.

Hausman, J.A., 1978. Specification tests in econometrics. *Econometrica* 46, 1251-1271.

Hu, Y., 2006. Bounding parameters in a linear regression model with a mismeasured regressor using additional information. *Journal of Econometrics* 133, 51-70.

Hillier, G., 2006. Yet more on the exact properties of IV estimators. *Econometric Theory* 22, 913-931.

Kapteyn, A., Wansbeek, T., 1984. Errors in variables: Consistent Adjusted Least Squares (CALS) estimation. *Communications in Statistics: Theory and Methods* 13, 1811-1837.

Kiviet, J.F., Niemczyk, J., 2007. The asymptotic and finite sample distribution of OLS and simple IV in simultaneous equations. *Journal of Computational Statistics and Data Analysis* 51, 3296-3318.

Kiviet, J.F., Niemczyk, J., 2012. The asymptotic and finite sample (un)conditional distributions of OLS and simple IV in simultaneous equations. *Journal of Computational Statistics and Data Analysis* 56, 3567-3586.

Kiviet, J.F., Phillips, G.D.A., 1996. The bias of the ordinary least squares estimator in simultaneous equation models. *Economics Letters* 53, 161-167.

Kraay, A., 2012. Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach. *Journal of Applied Econometrics* 27, 108-128.

Maddala, G.S., Jeong, J., 1992. On the exact small sample distribution of the instrumental variable estimator. *Econometrica* 60, 181-183.

Manski, C.F., 2003. *Partial Identification of Probability Distributions.* New York, Springer.

Manski, C.F., 2007. *Identification for Prediction and Decision.* Cambridge, MA. Harvard University Press.

Manski, C.F., Pepper, J.V., 2000. Monotone instrumental variables: With an application to the demand for schooling. *Econometrica* 68, 997-1010.

Manski, C.F., Pepper, J.V., 2009. More on monotone instrumental variables. *Econometrics Journal* 12, S200-S216.

Mariano, R.S., 1973. Approximations to the distribution functions of the ordinary least-squares and two-stage least-squares estimators in the case of two included endogenous variables. *Econometrica* 41, 67-77.

Mikusheva, A., 2010. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics* 157, 236-247.

Nelson, C.R., Startz, R., 1990a. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58, 967-976.

Nelson, C.R., Startz, R., 1990b. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 63, S125-S140.

Nevo, A., Rosen, A.M., 2012. Identification with imperfect instruments. *The Review of Economics and Statistics* 94, 659-671.

Phillips, P.C.B., 1983. Exact small sample theory in the simultaneous equations model. In M.D. Intriligator, Z. Griliches (eds.), *Handbook of Econometrics*, Ch. 8, 449-516. North–Holland.

Phillips, P.C.B., 1989. Partially identified econometric models. *Econometric Theory* 5, 181-240.

Phillips, P.C.B., 2006. A remark on bimodality and weak instrumentation in structural equation estimation. *Econometric Theory* 22, 947–960.

Phillips, P.C.B., 2009. Exact distribution theory in structural estimation with an identity. *Econometric Theory* 25, 958-984.

Rothenberg, T.J., 1972. The asymptotic distribution of the least squares estimator in the errors in variables model. *Unpublished mimeo, University of California, Berkeley.*

Sawa, T., 1969. The exact sampling distribution of ordinary least squares and tw0-stage least squares estimators. *Journal of the American Statistical Association* 64, 923-937.

Schneeweiss, H., 1980. Different asymptotic variances for the same estimator in a regression with errors in the variables. *Methods of Operations Research* 36, 249-269.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 557-586.

Tamer, E., 2010. Partial identification in econometrics. *Annual Reviews Economics* 2, 167-195.

Woglom, G., 2001. More results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 69, 1381-1389.

Appendices

A. The relevance of conditioning on unobserved variables

The DGP introduced in section 2.1 with just three fundamental parameters can of course easily be generalized in various directions. We will indicate here one particular type of possible extension, just because that supports the interpretation of the component v_i^* of the reduced form disturbance v_i and provides arguments to examine in Monte Carlo experiments both cases where v_i^* is either random or fixed. When allowing for overidentification, this requires to replace $z_i \sim IID(0,1)$ by the $L > 1$ series $z_i^{(l)} \sim IID(0,1)$, $l = 1, ..., L$. Then the reduced form equation generalizes to

$$x_i = \sum\nolimits_{l=1}^{L} \pi_l z_i^{(l)} + v_i, \tag{A.1}$$

where again $v_i = v_i^* + \xi\varepsilon_i$. Without loss of generality we may assume that the L instruments (possibly after a transformation) are all mutually independent. Then it follows that

$$\sigma_x^2 = \sum\nolimits_{l=1}^{L} \pi_l^2 + \xi^2 + \sigma_{v^*}^2. \tag{A.2}$$

With $E(x_i z_i^{(l)}) = \rho_{xz^{(l)}}\sigma_x = \pi_l$ and (2.10) one finds $\sigma_{v^*}^2 = \sigma_x^2(1 - \rho_{x\varepsilon}^2 - \sum\nolimits_{l=1}^{L} \rho_{xz^{(l)}}^2)$. This again implies admissibility restrictions, which here involve an $(L+1)$-dimensional unit sphere, namely

$$\rho_{x\varepsilon}^2 + \rho_{xz^{(1)}}^2 + ... + \rho_{xz^{(L)}}^2 \leq 1. \tag{A.3}$$

Now imagine that a researcher has available (or is aware of) only $L^{\#}$ of the instrumental variables, namely $z_i^{(l)}$ for $l = 1, ..., L^{\#}$, where $0 < L^{\#} < L$. Then $x_i = \pi_1 z_i^{(1)} + ... + \pi_{L^{\#}} z_i^{(L^{\#})} + v_i^{\#*} + v_i^* + \xi\varepsilon_i$, with $v_i^{\#*} = \pi_{L^{\#}+1} z_i^{(L^{\#}+1)} + ... + \pi_L z_i^{(L)}$. Now $v_i^{\#*} + v_i^*$ plays the role of the earlier v_i^*. Actual inference can now only be conditioned on the $L^{\#}$ available instruments. Thus, in a Monte Carlo analysis of that situation one may choose to keep only these $L^{\#}$ instruments fixed. However, it does make sense too to keep $v_i^{\#*}$ fixed as well, simply if one is of the opinion that this mimics the actual population better. Keeping both $v_i^{\#*}$ and v_i^* fixed simply represents the extreme case in which all the contributions to x_i which are exogenous with respect to ε_i but are not explicitly taken into account by the available instruments, are supposed to represent the actual population of interest best. For our simple model where $L = 1$ there are no such intermediary cases for conditioning on just $L^{\#}$ instruments, but only the two extremes $\mathcal{C}_z$ and $\mathcal{C}_{x^*}$.

B. Proofs

Lemma: From $D^{-1} = [\bar{D} + (D - \bar{D})]^{-1} = \bar{D}^{-1}(1 + \bar{D}^{-1}\tilde{D})^{-1}$, where $\bar{D}^{-1}\tilde{D} = O_p(n^{-1/2})$, the Taylor expansion $(1 + \bar{D}^{-1}\tilde{D})^{-1} = 1 - \bar{D}^{-1}\tilde{D} + o_p(n^{-1/2})$ yields $D^{-1} = \bar{D}^{-1}(1 - \bar{D}^{-1}\tilde{D}) + o_p(n^{-3/2})$, giving $R = N/D = (\bar{N} + \tilde{N})\bar{D}^{-1}(1 - \bar{D}^{-1}\tilde{D}) + o_p(n^{-1/2})$. Note that $\tilde{N}\bar{D}^{-2}\tilde{D} = O_p(n^{-1})$. So, conditional on $\mathcal{C}$, we have

$$n^{1/2}(R - \bar{N}/\bar{D}) = (\bar{D}/n)^{-1} n^{-1/2}[\tilde{N} - (\bar{N}/\bar{D})\tilde{D}] + o_p(1).$$

Now the result of the lemma follows directly by omitting the remainder term and invoking standard results attributed to Slutsky and Cramér to the right-hand term.

Theorem 1:

For obtaining the unconditional limiting distribution of IV we do not need the Lemma, and can just follow the standard textbook approach. This exploits that according to the Law of Large Numbers (LLN) $\operatorname{plim} n^{-1}\sum_{i=1}^n z_i x_i = \sigma_{xz} \neq 0$, while a standard Central Limit Theorem (CLT) yields $n^{-1/2}\sum_{i=1}^n z_i\varepsilon_i \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2\sigma_z^2\right)$, since $E(z_i\varepsilon_i) = 0$, $E(z_i z_j \varepsilon_i\varepsilon_j) = 0$ for $i \neq j$ and $E(z_i^2\varepsilon_i^2) = \sigma_\varepsilon^2\sigma_z^2$, because of the the independence of z_i and ε_i. Then it follows from (3.1) that $n^{1/2}(\hat\beta_{IV} - \beta) = \left(n^{-1}\sum_{i=1}^n z_i x_i\right)^{-1} n^{-1/2}\sum_{i=1}^n z_i\varepsilon_i \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2\sigma_z^2/\sigma_{xz}^2\right)$, where $\sigma_z^2/\sigma_{xz}^2 = 1/(\rho_{xz}^2\sigma_x^2)$.

When conditioning $\hat\beta_{IV}$ on the z_i we employ the Lemma with $N = \sum_{i=1}^n z_i\varepsilon_i$ and $D = \sum_{i=1}^n z_i x_i$ and find

$$\begin{aligned}
\bar N &\equiv E(N \mid \mathcal{C}_z) = E(\textstyle\sum_{i=1}^n z_i\varepsilon_i \mid \mathcal{C}_z) = \sum_{i=1}^n z_i E(\varepsilon_i \mid \mathcal{C}_z) = 0,\\
\tilde N &\equiv N - \bar N = \textstyle\sum_{i=1}^n z_i\varepsilon_i, \text{ with } \tilde N \mid \mathcal{C}_z = O_p(n^{1/2}),\\
\bar D &\equiv E(D \mid \mathcal{C}_z) = E(\textstyle\sum_{i=1}^n z_i x_i \mid \mathcal{C}_z) = \sum_{i=1}^n z_i E(x_i \mid \mathcal{C}_z) = \pi\sum_{i=1}^n z_i^2 = O(n),\\
\tilde D &\equiv D - \bar D = \textstyle\sum_{i=1}^n z_i(x_i - \pi z_i) = \sum_{i=1}^n z_i v_i, \text{ with } \tilde D \mid \mathcal{C}_z = O_p(n^{1/2}).
\end{aligned}$$

We conclude that $\tilde N \mid \mathcal{C}_z = O_p(n^{1/2})$, because $Var\left(\sum_{i=1}^n z_i\varepsilon_i \mid \mathcal{C}_z\right) = \sigma_\varepsilon^2\sum_{i=1}^n z_i^2 = O(n)$. The result for $\bar D$ follows immediately from (2.3), whereas $\tilde D \mid \mathcal{C}_z = O_p(n^{1/2})$, because we find $Var\left(\sum_{i=1}^n z_i v_i \mid \mathcal{C}_z\right) = \sigma_v^2\sum_{i=1}^n z_i^2 = O(n)$. Hence, the conditions of the Lemma are satisfied, although $\bar N$ is actually much smaller than $O(n)$. Moreover, $\tilde N - \bar N\tilde D/\bar D = \tilde N = \sum_{i=1}^n z_i\varepsilon_i$, thus $n^{-1/2}(\tilde N - \bar N\tilde D/\bar D) = n^{-1/2}\sum_{i=1}^n z_i\varepsilon_i$, for which we have already obtained that it has conditional expectation zero and conditional variance $\sigma_\varepsilon^2\frac{1}{n}\sum_{i=1}^n z_i^2$, thus a standard CLT yields $V_0 = \lim \sigma_\varepsilon^2\frac{1}{n}\sum_{i=1}^n z_i^2 = \sigma_\varepsilon^2\sigma_z^2$. Since $\lim \bar D/n = \pi\sigma_z^2 = \rho_{xz}\sigma_z\sigma_x$ we find that conditioning on the instruments has no effect. This is because $\bar N = 0$ and due to $\lim n^{-1}\bar D = \lim \frac{1}{n}\sum_{i=1}^n E(z_i x_i \mid z_i) = \pi \lim \frac{1}{n}\sum_{i=1}^n z_i^2 = \pi\sigma_z^2 = \operatorname{plim}\frac{1}{n}\sum_{i=1}^n z_i x_i$.

Conditioning IV on x_i^* and exploiting $E(x_i^*\varepsilon_i) = 0$ yields

$$\begin{aligned}
\bar N &\equiv E(N \mid \mathcal{C}_{x^*}) = E(\textstyle\sum_{i=1}^n z_i\varepsilon_i \mid \mathcal{C}_{x^*}) = \sum_{i=1}^n z_i E(\varepsilon_i \mid \mathcal{C}_{x^*}) = 0,\\
\tilde N &\equiv N - \bar N = \textstyle\sum_{i=1}^n z_i\varepsilon_i, \text{ with } \tilde N \mid \mathcal{C}_{x^*} = O_p(n^{1/2}),\\
\bar D &\equiv E(D \mid \mathcal{C}_{x^*}) = E[\textstyle\sum_{i=1}^n z_i(x_i^* + \xi\varepsilon_i) \mid \mathcal{C}_{x^*}] = \sum_{i=1}^n z_i x_i^* = \pi\sum_{i=1}^n z_i^2 + \sum_{i=1}^n z_i v_i^* = O(n),\\
\tilde D &\equiv D - \bar D = \textstyle\sum_{i=1}^n z_i(x_i - \pi z_i - v_i^*) = \xi\sum_{i=1}^n z_i\varepsilon_i, \text{ with } \tilde D \mid \mathcal{C}_{x^*} = O_p(n^{1/2}).
\end{aligned}$$

That $\tilde N \mid \mathcal{C}_{x^*} = O_p(n^{1/2})$ is now self-evident. Because $Var\left(\xi\sum_{i=1}^n z_i\varepsilon_i \mid \mathcal{C}_{x^*}\right) = \sigma_\varepsilon^2\sum_{i=1}^n z_i^2 = O(n)$ we find $\tilde D \mid \mathcal{C}_{x^*} = O_p(n^{1/2})$. Again $\bar N = 0$, giving $n^{-1/2}(\tilde N - \bar N\tilde D/\bar D) = n^{-1/2}\sum_{i=1}^n z_i\varepsilon_i$ and $V_0 = \lim \sigma_\varepsilon^2\frac{1}{n}\sum_{i=1}^n z_i^2 = \sigma_\varepsilon^2\sigma_z^2$. Because $\lim \bar D/n = \pi\sigma_z^2 + \lim \sigma_\varepsilon^2\frac{1}{n}\sum_{i=1}^n E(z_i v_i^*) = \pi\sigma_z^2 = \rho_{xz}\sigma_z\sigma_x$, we again find the same limiting distribution, and hence have established result (a) of the Theorem.

For simple OLS we have $N = \sum_{i=1}^n x_i\varepsilon_i$ and $D = \sum_{i=1}^n x_i^2$. For $\mathcal{C}$ empty and using decomposition $x_i = \pi z_i + v_i^* + \xi\varepsilon_i = x_i^* + \xi\varepsilon_i$ we easily obtain

$$\begin{aligned}
\bar N &= E(\textstyle\sum_{i=1}^n x_i\varepsilon_i) = n\xi\sigma_\varepsilon^2 = O(n),\\
\tilde N &= \textstyle\sum_{i=1}^n(x_i\varepsilon_i - \xi\sigma_\varepsilon^2) = \sum_{i=1}^n[x_i^*\varepsilon_i + \xi(\varepsilon_i^2 - \sigma_\varepsilon^2)] = O_p(n^{1/2}),\\
\bar D &= E(\textstyle\sum_{i=1}^n x_i^2) = n\sigma_x^2 = O(n),\\
\tilde D &= \textstyle\sum_{i=1}^n(x_i^2 - \sigma_x^2) = O_p(n^{1/2}),
\end{aligned}$$

where the orders of probability of $\tilde N$ and $\tilde D$ follow because their expectation is zero and their variance is $O(n)$, provided x_i^* and ε_i have finite 4th moments, which is the case under

Assumptions 2 and 4. We may write $\tilde{N} - (\bar{N}/\bar{D})\tilde{D}$ as $\sum_{i=1}^n u_i$. Here, using $\bar{N}/\bar{D} = \xi\sigma_\varepsilon^2/\sigma_x^2$, we have

$$u_i = x_i^*\varepsilon_i + \xi(\varepsilon_i^2 - \sigma_\varepsilon^2) - \xi\frac{\sigma_\varepsilon^2}{\sigma_x^2}[(x_i^{*2} - \sigma_{x^*}^2) + 2\xi x_i^*\varepsilon_i + \xi^2(\varepsilon_i^2 - \sigma_\varepsilon^2)], \tag{B.1}$$

with $E(u_i) = 0$, $E(u_iu_j) = 0$ for $i \neq j$. Under Assumptions 2 and 4 we find that

$$\begin{aligned}
E(u_i^2) &= \sigma_\varepsilon^2\sigma_{x^*}^2 + 2\xi^2\sigma_\varepsilon^4 + \xi^2\frac{\sigma_\varepsilon^4}{\sigma_x^4}[2\sigma_{x^*}^4 + 4\xi^2\sigma_{x^*}^2\sigma_\varepsilon^2 + 2\xi^4\sigma_\varepsilon^4] - 4\xi^2\sigma_{x^*}^2\frac{\sigma_\varepsilon^4}{\sigma_x^2} - 4\xi^4\frac{\sigma_\varepsilon^6}{\sigma_x^2} \\
&= \sigma_\varepsilon^2\sigma_x^2 - \xi^2\sigma_\varepsilon^4 + 2\xi^2\sigma_\varepsilon^4 + \xi^2\frac{\sigma_\varepsilon^4}{\sigma_x^4}[2(\sigma_x^4 - 2\xi^2\sigma_x^2\sigma_\varepsilon^2 + \xi^4\sigma_\varepsilon^4) + 4\xi^2\sigma_x^2\sigma_\varepsilon^2 - 4\xi^4\sigma_\varepsilon^4 + 2\xi^4\sigma_\varepsilon^4] \\
&\quad -4\xi^2\sigma_x^2\frac{\sigma_\varepsilon^4}{\sigma_x^2} + 4\xi^4\frac{\sigma_\varepsilon^6}{\sigma_x^2} - 4\xi^4\frac{\sigma_\varepsilon^6}{\sigma_x^2} \\
&= \sigma_\varepsilon^2\sigma_x^2(1 - \rho_{x\varepsilon}^2).
\end{aligned}$$

is finite and constant. So $n^{-1/2}\sum_{i=1}^n u_i \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2\sigma_x^2(1-\rho_{x\varepsilon}^2)\right)$ according to the CLT and, making use of $n^{-1}\bar{D} = \sigma_x^2$, we find from the Lemma the unconditional limiting distribution indicated as (b1) in the Theorem.

Conditioning on the instrument we find

$$\begin{aligned}
\bar{N} &= E(\textstyle\sum_{i=1}^n x_i\varepsilon_i \mid \mathcal{C}_z) = \sum_{i=1}^n E[(\pi z_i + v_i^* + \xi\varepsilon_i)\varepsilon_i \mid \mathcal{C}_z] = n\xi\sigma_\varepsilon^2 = O(n), \\
\tilde{N} &= \textstyle\sum_{i=1}^n[\pi z_i\varepsilon_i + v_i^*\varepsilon_i + \xi(\varepsilon_i^2 - \sigma_\varepsilon^2)] \text{ and } \tilde{N} \mid \mathcal{C}_z = O_p(n^{1/2}), \\
\bar{D} &= E(\textstyle\sum_{i=1}^n x_i^2 \mid \mathcal{C}_z) = \pi^2\sum_{i=1}^n z_i^2 + n\sigma_v^2 = O(n), \\
\tilde{D} &= \textstyle\sum_{i=1}^n[2\pi z_iv_i + (v_i^2 - \sigma_v^2)] \text{ and } \tilde{D} \mid \mathcal{C}_z = O_p(n^{1/2}),
\end{aligned}$$

where the orders of probability of $\tilde{N}$ and $\tilde{D}$ again follow upon noting that their conditional expectation is zero and their conditional variance is $O(n)$, provided ε_i and v_i have finite fourth moment. Here we find for $\tilde{N} - (\bar{N}/\bar{D})\tilde{D} = \sum_{i=1}^n u_i$ that it has

$$\begin{aligned}
u_i &= \pi z_i\varepsilon_i + v_i^*\varepsilon_i + \xi(\varepsilon_i^2 - \sigma_\varepsilon^2) \\
&\quad -\frac{\bar{N}}{\bar{D}}[2\pi z_i(v_i^* + \xi\varepsilon_i) + (v_i^{*2} - \sigma_{v^*}^2) + 2\xi v_i^*\varepsilon_i + \xi^2(\varepsilon_i^2 - \sigma_\varepsilon^2)],
\end{aligned} \tag{B.2}$$

with $E(u_i \mid \mathcal{C}_z) = 0$ and $E(u_iu_j \mid \mathcal{C}_z) = 0$ for $i \neq j$. Substituting $\bar{N} = n\xi\sigma_\varepsilon^2$, and using Assumptions 2 and 3, we obtain

$$\begin{aligned}
E(u_i^2 \mid \mathcal{C}_z) = \sigma_\varepsilon^2\{&\pi^2z_i^2 + \sigma_{v^*}^2 + 2\xi^2\sigma_\varepsilon^2 + \bar{D}^{-2}n^2\xi^2\sigma_\varepsilon^2[4\pi^2\sigma_v^2z_i^2 + 2\sigma_v^4] \\
&-4\bar{D}^{-1}n\xi^2[\pi^2z_i^2\sigma_\varepsilon^2 + \sigma_{v^*}^2\sigma_\varepsilon^2 + \xi^2\sigma_\varepsilon^4]\},
\end{aligned}$$

which is finite. Therefore, a standard CLT applies, giving

$$n^{-1/2}\textstyle\sum_{i=1}^n u_i \mid \mathcal{C}_z \xrightarrow{d} \mathcal{N}\left(0, \lim n^{-1}\sum_{i=1}^n E(u_i^2 \mid \mathcal{C})\right).$$

Making use of $\lim n^{-1}\bar{D} = \lim n^{-1}\left(\pi^2\sum_{i=1}^n z_i^2 + n\sigma_v^2\right) = \pi^2\sigma_z^2 + \sigma_v^2 = \sigma_x^2$ we find

$$\lim n^{-1}\textstyle\sum_{i=1}^n E(u_i^2 \mid \mathcal{C}_z) = \sigma_\varepsilon^2\sigma_x^2[1 - \rho_{x\varepsilon}^2(1 + 2\rho_{xz}^4)]$$

from which the conditional limiting distribution (2b) is found.

When conditioning on x^* we find

$$\begin{aligned}
\bar{N} &= E(\textstyle\sum_{i=1}^n x_i\varepsilon_i \mid \mathcal{C}_{x^*}) = \sum_{i=1}^n E[(x_i^* + \xi\varepsilon_i)\varepsilon_i \mid \mathcal{C}_{x^*}] = n\xi\sigma_\varepsilon^2 = O(n), \\
\tilde{N} &= \textstyle\sum_{i=1}^n [x_i^*\varepsilon_i + \xi(\varepsilon_i^2 - \sigma_\varepsilon^2)] \text{ and } \tilde{N} \mid \mathcal{C}_{x^*} = O_p(n^{1/2}), \\
\bar{D} &= E(\textstyle\sum_{i=1}^n x_i^2 \mid \mathcal{C}_{x^*}) = \sum_{i=1}^n x_i^{*2} + n\xi^2\sigma_\varepsilon^2 = O(n), \\
\tilde{D} &= \textstyle\sum_{i=1}^n [2\xi x_i^*\varepsilon_i + \xi^2(\varepsilon_i^2 - \sigma_\varepsilon^2)] \text{ and } \tilde{D} \mid \mathcal{C}_{x^*} = O_p(n^{1/2}),
\end{aligned}$$

where the orders of probability of $\tilde{N}$ and $\tilde{D}$ again follow provided ε_i has finite 4th moment. Here $\tilde{N} - (\bar{N}/\bar{D})\tilde{D} = \sum_{i=1}^n u_i$ has

$$u_i = x_i^*\varepsilon_i + \xi(\varepsilon_i^2 - \sigma_\varepsilon^2) - \frac{\bar{N}}{\bar{D}}[2\xi x_i^*\varepsilon_i + \xi^2(\varepsilon_i^2 - \sigma_\varepsilon^2)], \tag{B.3}$$

with $E(u_i \mid \mathcal{C}_{x^*}) = 0$, $E(u_iu_j \mid \mathcal{C}_{x^*}) = 0$ for $i \neq j$ and under Assumption 2

$$E(u_i^2 \mid \mathcal{C}_{x^*}) = \sigma_\varepsilon^2\{x_i^{*2} + 2\xi^2\sigma_\varepsilon^2 + \bar{D}^{-2}n^2\xi^2\sigma_\varepsilon^2[4\xi^2x_i^{*2}\sigma_\varepsilon^2 + 2\xi^4\sigma_\varepsilon^4] - 4\bar{D}^{-1}n\xi^2[x_i^{*2}\sigma_\varepsilon^2 + \xi^2\sigma_\varepsilon^4]\},$$

which is finite again. Making use of $\lim n^{-1}\bar{D} = \lim n^{-1}\left(\sum_{i=1}^n x_i^{*2} + n\xi^2\sigma_\varepsilon^2\right) = \sigma_{x^*}^2 + \xi^2\sigma_\varepsilon^2 = \sigma_x^2$ we obtain

$$\begin{aligned}
\lim n^{-1}\textstyle\sum_{i=1}^n E(u_i^2 \quad &| \quad \mathcal{C}_{x^*}) = \sigma_\varepsilon^2\{\sigma_x^2 + \rho_{x\varepsilon}^2\sigma_x^2 + \frac{\rho_{x\varepsilon}^2}{\sigma_x^2}[4\rho_{x\varepsilon}^2\sigma_x^2\sigma_{x^*}^2 + 2\rho_{x\varepsilon}^4\sigma_x^4] - 4[\rho_{x\varepsilon}^2\sigma_{x^*}^2 + \rho_{x\varepsilon}^4\sigma_x^2]\} \\
&= \sigma_\varepsilon^2\sigma_x^2\left\{1 + \rho_{x\varepsilon}^2 + \rho_{x\varepsilon}^2[4\rho_{x\varepsilon}^2(1-\rho_{x\varepsilon}^2) + 2\rho_{x\varepsilon}^4] - 4[\rho_{x\varepsilon}^2(1-\rho_{x\varepsilon}^2) + \rho_{x\varepsilon}^4]\right\} \\
&= \sigma_\varepsilon^2\sigma_x^2(1-\rho_{x\varepsilon}^2)[1 - 2(1-\rho_{x\varepsilon}^2)\rho_{x\varepsilon}^2]
\end{aligned}$$

from which the limiting distribution of $\hat{\beta}_{OLS}$ conditional on $\mathcal{C}_{x^*}$, indicated as result (b3), follows.

Corollary 1:
It is obvious that $1 \geq (1-\rho_{x\varepsilon}^2) \geq [1 - \rho_{x\varepsilon}^2(1 + 2\rho_{xz}^4)]$, because $\rho_{x\varepsilon}^2 \geq 0$ and $1 + 2\rho_{xz}^4 \geq 1$. From $0 \leq \rho_{xz}^2 + \rho_{x\varepsilon}^2 \leq 1$ it follows that $1 + 2\rho_{xz}^4 \leq 1 + 2(1-\rho_{x\varepsilon}^2)^2$, thus

$$1 - \rho_{x\varepsilon}^2(1 + 2\rho_{xz}^4) \geq 1 - \rho_{x\varepsilon}^2[1 + 2(1-\rho_{x\varepsilon}^2)^2] = (1-\rho_{x\varepsilon}^2)[1 - 2(1-\rho_{x\varepsilon}^2)\rho_{x\varepsilon}^2].$$

Theorem 2:
Substituting (2.4) and (3.2) in the estimator for $\hat{\sigma}_\varepsilon^2$ we find

$$\hat{\sigma}_\varepsilon^2 \equiv n^{-1}\sum_{i=1}^n (y_i - x_i\hat{\beta}_{OLS})^2 = n^{-1}\sum_{i=1}^n \varepsilon_i^2 - n^{-1}\left(\sum_{i=1}^n x_i\varepsilon_i\right)^2 / \sum_{i=1}^n x_i^2 \tag{B.4}$$

for which

$$\begin{aligned}
\operatorname{plim}\hat{\sigma}_\varepsilon^2 &= \lim n^{-1}\sum_{i=1}^n E(\varepsilon_i^2) - \left(\lim n^{-1}\sum_{i=1}^n E(x_i\varepsilon_i)\right)^2 / \lim n^{-1}\sum_{i=1}^n E(x_i^2) \\
&= \sigma_\varepsilon^2 - (\xi\sigma_\varepsilon^2)^2/\sigma_x^2 = \sigma_\varepsilon^2(1-\rho_{x\varepsilon}^2).
\end{aligned}$$

So, if $\rho_{x\varepsilon}$ were known, $\hat{\sigma}_\varepsilon^2/(1-\rho_{x\varepsilon}^2)$ establishes a consistent estimator of σ_ε^2 and likewise $\hat{\beta}_{KLS}(\rho_{x\varepsilon})$ for β. The limiting distribution of $\hat{\beta}_{KLS}(\rho_{x\varepsilon})$ does not simply coincide with that of $\hat{\beta}_{OLS}(\rho_{x\varepsilon}, \sigma_\varepsilon/\sigma_x)$ given in (4.1), due to the randomness of the correction term. To derive it, we substitute (B.4) in (3.2) to obtain (omitting from now on in the notation the dependence of $\hat{\beta}_{KLS}$ on $\rho_{x\varepsilon}$)

$$\hat{\beta}_{KLS} - \beta = \frac{\Sigma_i x_i\varepsilon_i}{\Sigma_i x_i^2} - \frac{\rho_{x\varepsilon}}{(1-\rho_{x\varepsilon}^2)^{1/2}}\left(\frac{\Sigma_i\varepsilon_i^2}{\Sigma_i x_i^2} - \frac{(\Sigma_i x_i\varepsilon_i)^2}{(\Sigma_i x_i^2)^2}\right)^{1/2}. \tag{B.5}$$

This has now to be linearized and scaled in order that we can invoke a CLT to its leading terms. We consider first the major components of (B.5) separately. When deriving result (b1) from Theorem 1 we obtained

$$\frac{\Sigma_i x_i \varepsilon_i}{\Sigma_i x_i^2} = \frac{\xi \sigma_\varepsilon^2}{\sigma_x^2} + \frac{1}{n\sigma_x^2}\Sigma_i u_i + o_p(n^{-1/2}), \tag{B.6}$$

with u_i as defined in (B.1). In similar way, employing the Lemma, we obtain

$$\frac{\Sigma_i \varepsilon_i^2}{\Sigma_i x_i^2} = \frac{\sigma_\varepsilon^2}{\sigma_x^2} + \frac{1}{n\sigma_x^2}[\Sigma_i\left(\varepsilon_i^2 - \sigma_\varepsilon^2\right) - \frac{\sigma_\varepsilon^2}{\sigma_x^2}\Sigma_i\left(x_i^2 - \sigma_x^2\right)] + o_p(n^{-1/2}).$$

From (B.6) we find

$$\frac{\left(\Sigma_i x_i \varepsilon_i\right)^2}{\left(\Sigma_i x_i^2\right)^2} = \frac{\xi^2\sigma_\varepsilon^4}{\sigma_x^4} + \frac{2\xi\sigma_\varepsilon^2}{n\sigma_x^4}\Sigma_i u_i + o_p(n^{-1/2}).$$

Combining the latter two results yields

$$\begin{aligned}\frac{\Sigma_i \varepsilon_i^2}{\Sigma_i x_i^2} - \frac{\left(\Sigma_i x_i \varepsilon_i\right)^2}{\left(\Sigma_i x_i^2\right)^2} &= \left(\frac{\sigma_\varepsilon^2}{\sigma_x^2} - \frac{\xi^2\sigma_\varepsilon^4}{\sigma_x^4}\right) + \frac{1}{n\sigma_x^2}\Sigma_i\left(\varepsilon_i^2 - \sigma_\varepsilon^2\right) \\ &\quad - \frac{\sigma_\varepsilon^2}{n\sigma_x^4}\Sigma_i\left(x_i^2 - \sigma_x^2\right) - \frac{2\xi\sigma_\varepsilon^2}{n\sigma_x^4}\Sigma_i u_i + o_p(n^{-1/2}).\end{aligned}$$

This has the form $v = \theta + w + o_p(n^{-1/2})$, where $\theta = O(1)$ with w containing three terms and all are $O_p(n^{-1/2})$. One can easily obtain by a Taylor expansion (and check by taking the square) that $v^{1/2} = \theta^{1/2} + \frac{1}{2}\theta^{-1/2}w + o_p(n^{-1/2})$. Hence,

$$\begin{aligned}\left(\frac{\Sigma_i \varepsilon_i^2}{\Sigma_i x_i^2} - \frac{\left(\Sigma_i x_i \varepsilon_i\right)^2}{\left(\Sigma_i x_i^2\right)^2}\right)^{1/2} &= \frac{\sigma_\varepsilon(1-\rho_{x\varepsilon}^2)^{1/2}}{\sigma_x} + \frac{1}{2n\sigma_\varepsilon\sigma_x(1-\rho_{x\varepsilon}^2)^{1/2}} \times \\ &\quad \Sigma_i[\left(\varepsilon_i^2 - \sigma_\varepsilon^2\right) - \frac{\sigma_\varepsilon^2}{\sigma_x^2}\left(x_i^2 - \sigma_x^2\right) - \frac{2\rho_{x\varepsilon}\sigma_\varepsilon}{\sigma_x}u_i] + o_p(n^{-1/2}).\end{aligned}$$

Collecting the various components of (B.5) now yields

$$\begin{aligned}n^{1/2}(\hat{\beta}_{KLS} - \beta) &= \frac{1}{n^{1/2}\sigma_x^2}\Sigma_i u_i \\ &\quad - \frac{\rho_{x\varepsilon}}{2n^{1/2}\sigma_\varepsilon\sigma_x(1-\rho_{x\varepsilon}^2)}\Sigma_i[\left(\varepsilon_i^2 - \sigma_\varepsilon^2\right) - \frac{\sigma_\varepsilon^2}{\sigma_x^2}\left(x_i^2 - \sigma_x^2\right) - \frac{2\rho_{x\varepsilon}\sigma_\varepsilon}{\sigma_x}u_i] + o_p(1) \\ &= \frac{1}{\sigma_x^2(1-\rho_{x\varepsilon}^2)}\frac{1}{n^{1/2}}\Sigma_i[u_i - \frac{\sigma_x\rho_{x\varepsilon}}{2\sigma_\varepsilon}\left(\varepsilon_i^2 - \sigma_\varepsilon^2\right) + \frac{\sigma_\varepsilon\rho_{x\varepsilon}}{2\sigma_x}\left(x_i^2 - \sigma_x^2\right)] + o_p(1),\end{aligned}$$

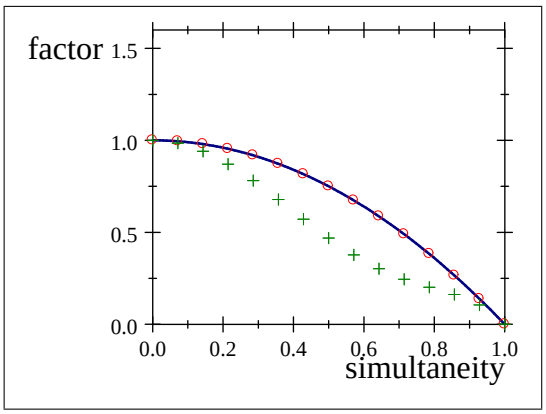
from which we should derive the variance of its $O_p(1)$ terms. Substituting (B.1) and employing Assumptions 2 and 4 we find

$$\begin{aligned}&Var[u_i - \frac{\sigma_x\rho_{x\varepsilon}}{2\sigma_\varepsilon}\left(\varepsilon_i^2 - \sigma_\varepsilon^2\right) + \frac{\sigma_\varepsilon\rho_{x\varepsilon}}{2\sigma_x}\left(x_i^2 - \sigma_x^2\right)] \\ = \; &Var\{x_i^*\varepsilon_i + \xi(\varepsilon_i^2 - \sigma_\varepsilon^2) - \xi\frac{\sigma_\varepsilon^2}{\sigma_x^2}[(x_i^{*2} - \sigma_{x^*}^2) + 2\xi x_i^*\varepsilon_i + \xi^2(\varepsilon_i^2 - \sigma_\varepsilon^2)] \\ &-\frac{1}{2}\xi\left(\varepsilon_i^2 - \sigma_\varepsilon^2\right) + \frac{\sigma_\varepsilon^2}{2\sigma_x^2}\xi[(x_i^{*2} - \sigma_{x^*}^2) + 2\xi x_i^*\varepsilon_i + \xi^2(\varepsilon_i^2 - \sigma_\varepsilon^2)]\}\end{aligned}$$

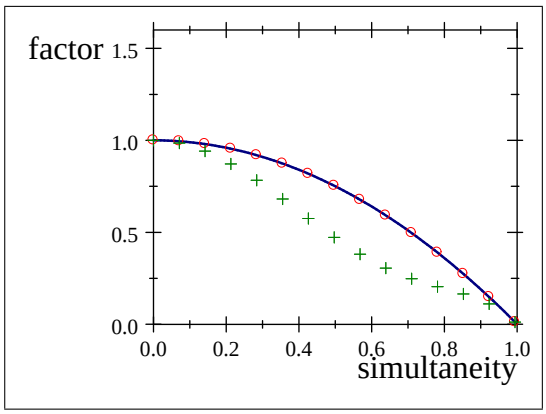
$$= Var\{(1-\xi^2\frac{\sigma_\varepsilon^2}{\sigma_x^2})x_i^*\varepsilon_i + \frac{1}{2}(\xi - \xi^3\frac{\sigma_\varepsilon^2}{\sigma_x^2})(\varepsilon_i^2 - \sigma_\varepsilon^2) - \frac{1}{2}\xi\frac{\sigma_\varepsilon^2}{\sigma_x^2}(x_i^{*2} - \sigma_{x^*}^2)\}$$

$$\begin{aligned}
&= (1-\xi^2\frac{\sigma_\varepsilon^2}{\sigma_x^2})^2\sigma_{x^*}^2\sigma_\varepsilon^2 + \frac{1}{2}(\xi - \xi^3\frac{\sigma_\varepsilon^2}{\sigma_x^2})^2\sigma_\varepsilon^4 + \frac{1}{2}\xi^2\frac{\sigma_\varepsilon^4}{\sigma_x^4}\sigma_{x^*}^4 \\
&= (1-\rho_{x\varepsilon}^2)^3\sigma_x^2\sigma_\varepsilon^2 + \frac{1}{2}\rho_{x\varepsilon}^2(1-\rho_{x\varepsilon}^2)^2\sigma_x^2\sigma_\varepsilon^2 + \frac{1}{2}\rho_{x\varepsilon}^2(1-\rho_{x\varepsilon}^2)^2\sigma_x^2\sigma_\varepsilon^2 \\
&= \sigma_x^2\sigma_\varepsilon^2(1-\rho_{x\varepsilon}^2)^2.
\end{aligned}$$

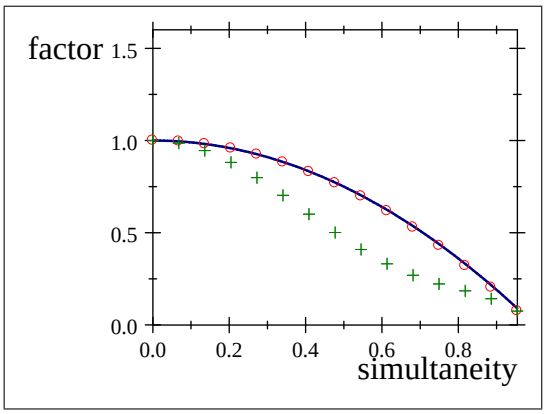
Now invoking a standard CLT yields the amazingly simple result of the Theorem.



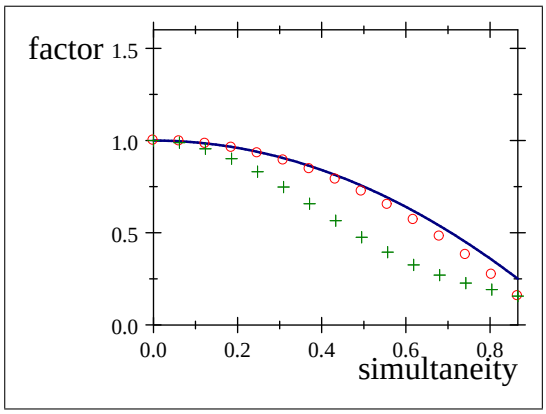
strenghth = 0.01



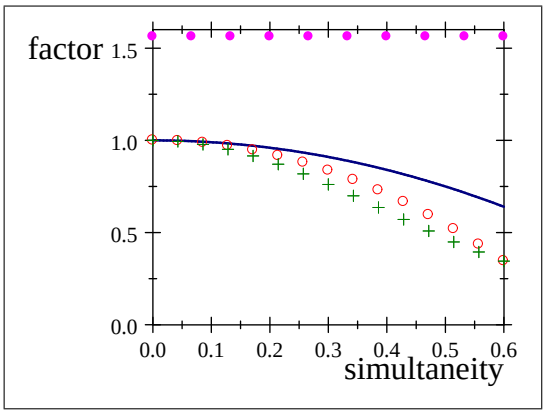
strenghth = 0.1



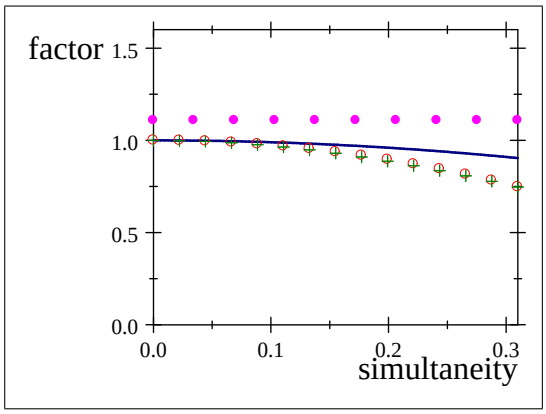
strenghth = 0.3



strenghth = 0.5

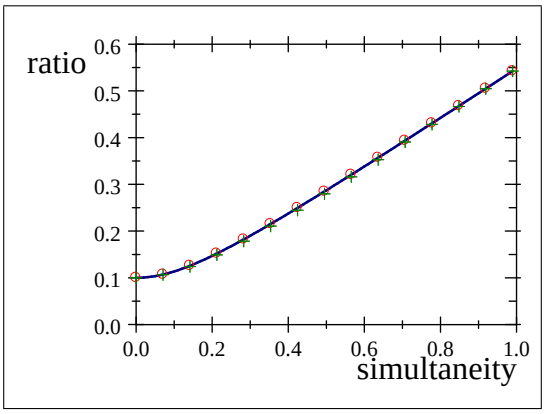


strenghth = 0.8

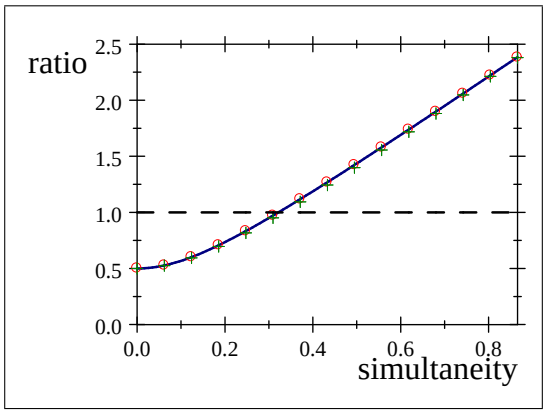


strenghth = 0.95

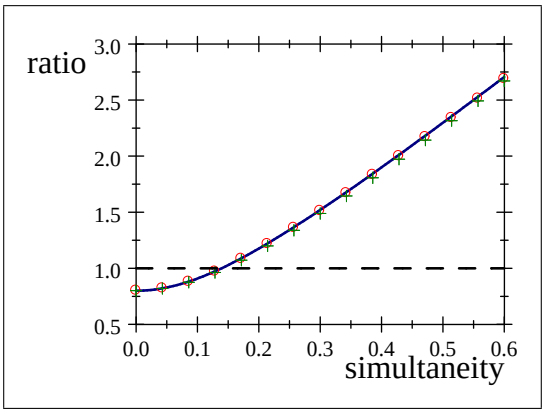
Diagram 1: Factor of $\sigma_\varepsilon^2/\sigma_x^2$ to obtain the asymptotic variance for OLS when conditioned on: $\mathcal{C}_\varnothing$ (blue, solid), $\mathcal{C}_z$ (red, circles), $\mathcal{C}_{x^*}$ (green, crosses) respectively, and for IV (magenta, dots); simultaneity = $\rho_{x\varepsilon}$; strength = ρ_{xz}.



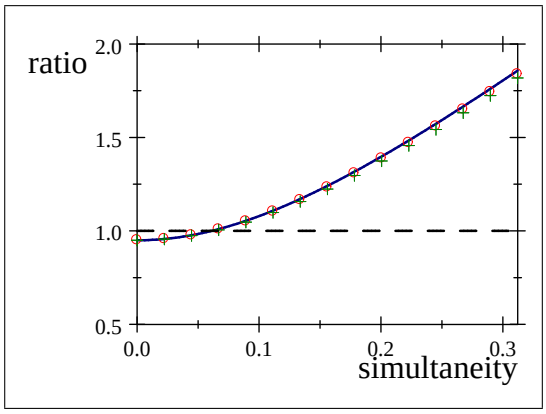
strenghth $= 0.1; n = 30;\ PCP = .3$



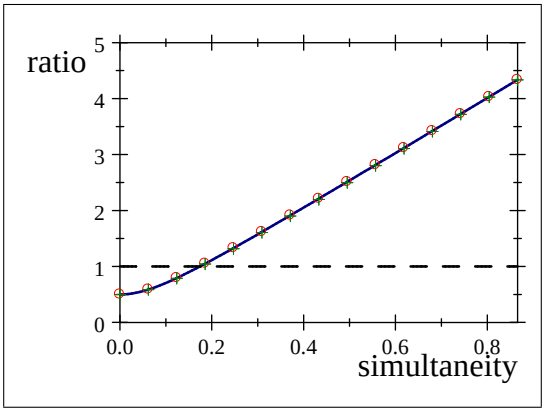
strenghth $= 0.5; n = 30;\ PCP = 10$



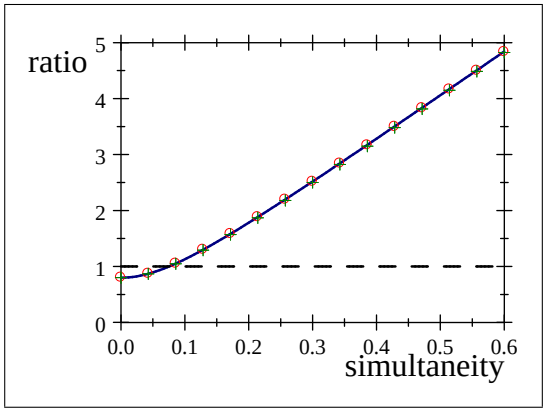
strenghth $= 0.8; n = 30;\ PCP = 53$



strenghth $= 0.95; n = 30;\ PCP = 270$



strenghth $= 0.5; n = 100;\ PCP = 33$



strenghth $= 0.8; n = 100;\ PCP = 178$

Diagram 2: Relative asymptotic precision: $ARMSE\,(OLS \mid \mathcal{C})\,/ARMSE\,(IV)$, where $\mathcal{C} = \mathcal{C}_{\varnothing}$ (blue, solid), $\mathcal{C}_z$ (red, circles), $\mathcal{C}_{x^*}$ (green, crosses).

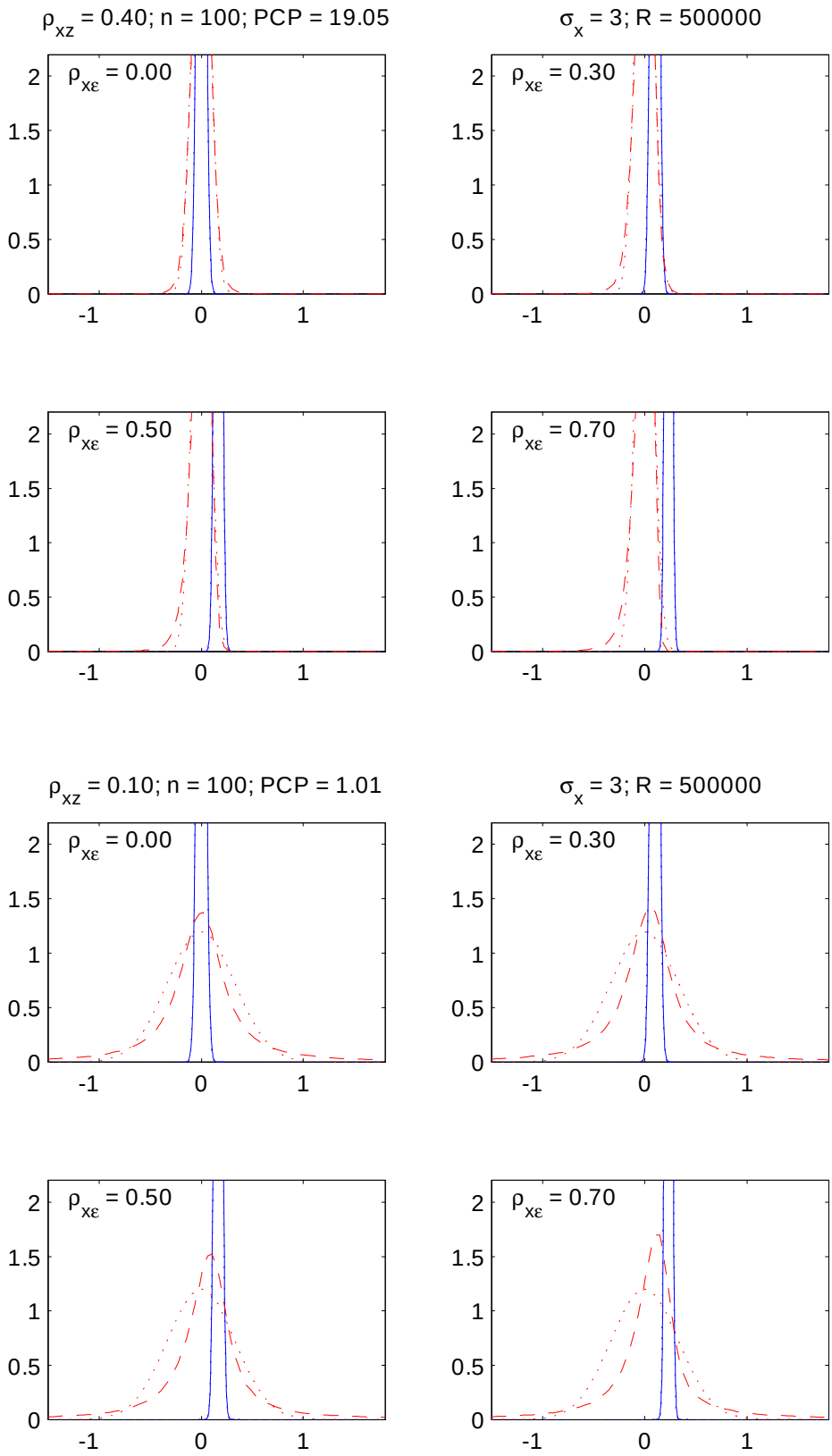


Diagram 3: Probability densities for IV (dashed, red) and OLS (solid, blue), both unconditional, supplemented by their respective asymptotic approximations (dotted)

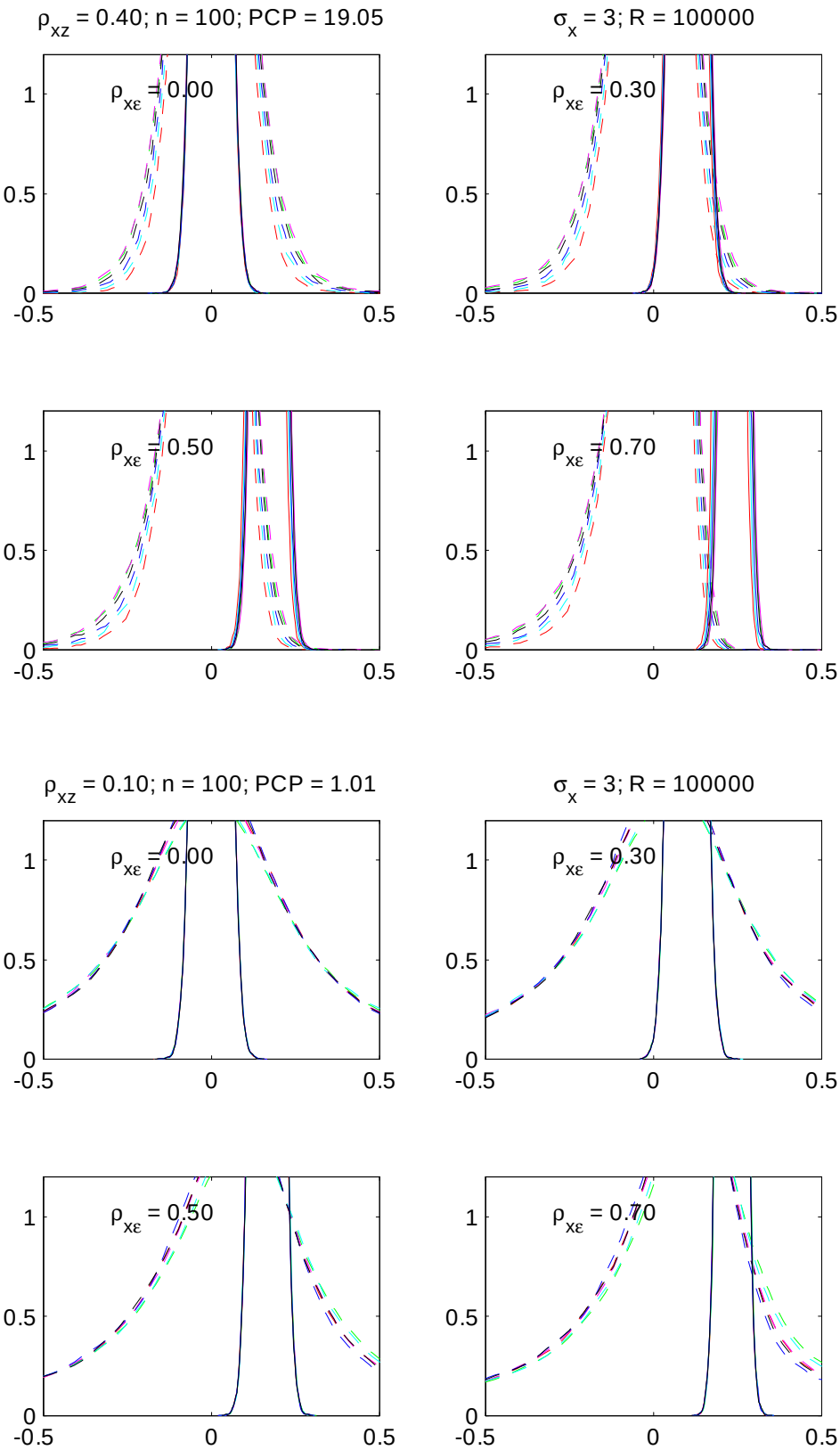


Diagram 4: IV (dashed) and OLS (solid) conditional on z (6×raw)

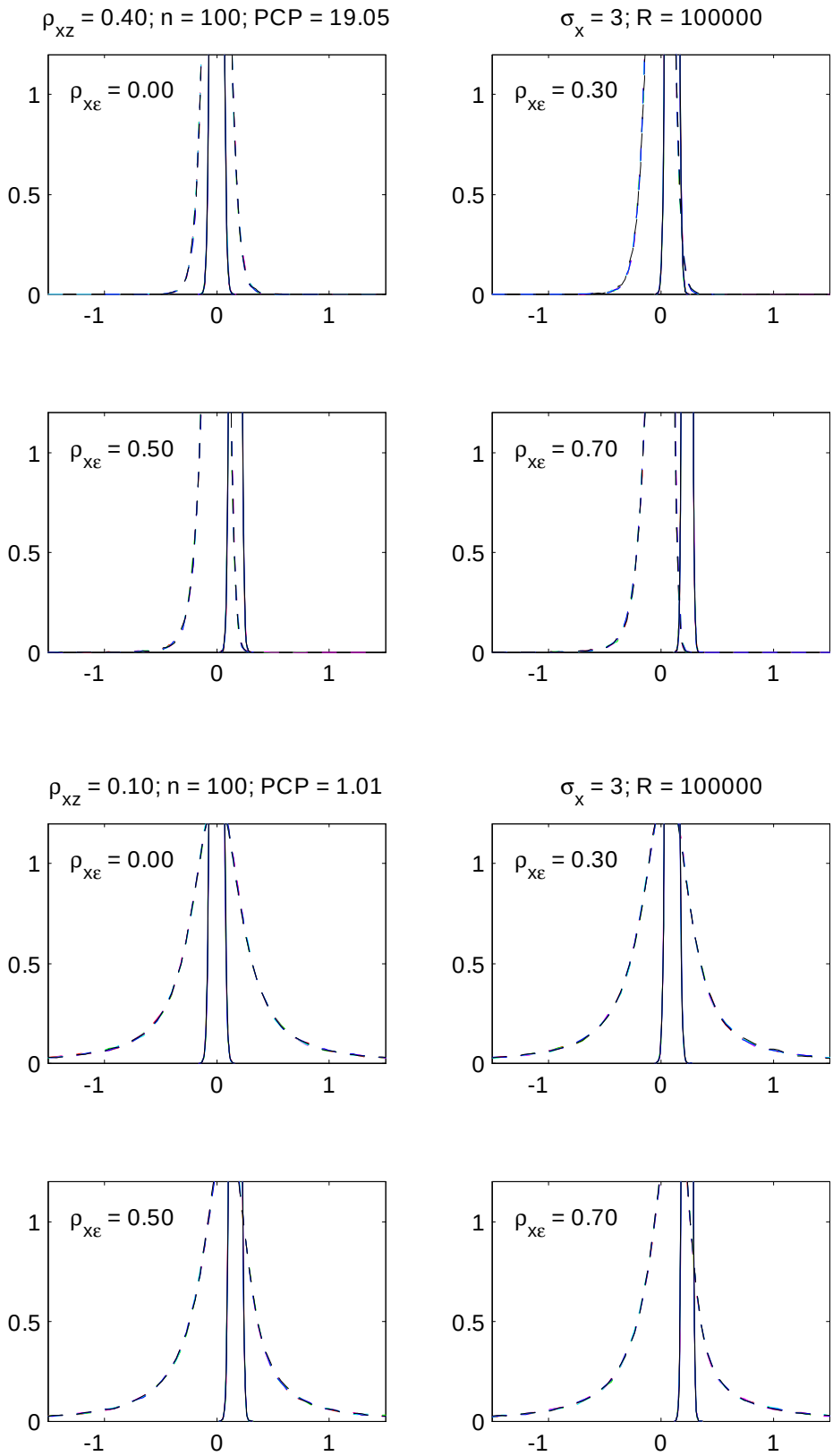


Diagram 5: IV (dashed) and OLS (solid) conditional on z (6×stylized)

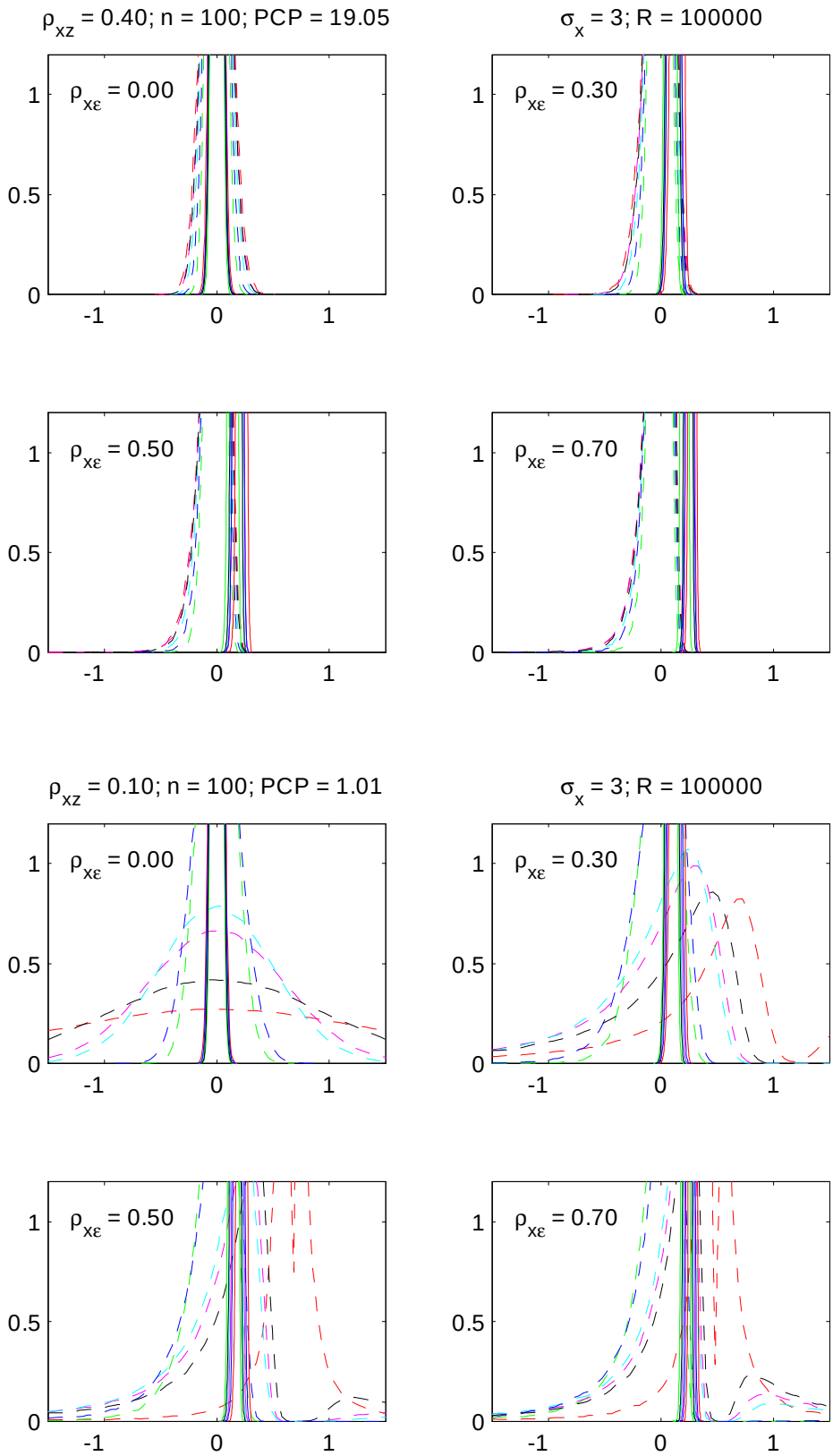


Diagram 6: IV (dashed) and OLS (solid) conditional on x^* (6×raw)
The 6 arbitrary realizations had: r_{zv^*} = -.14, .11, -.03, -.04, .08, -.05 respectively.

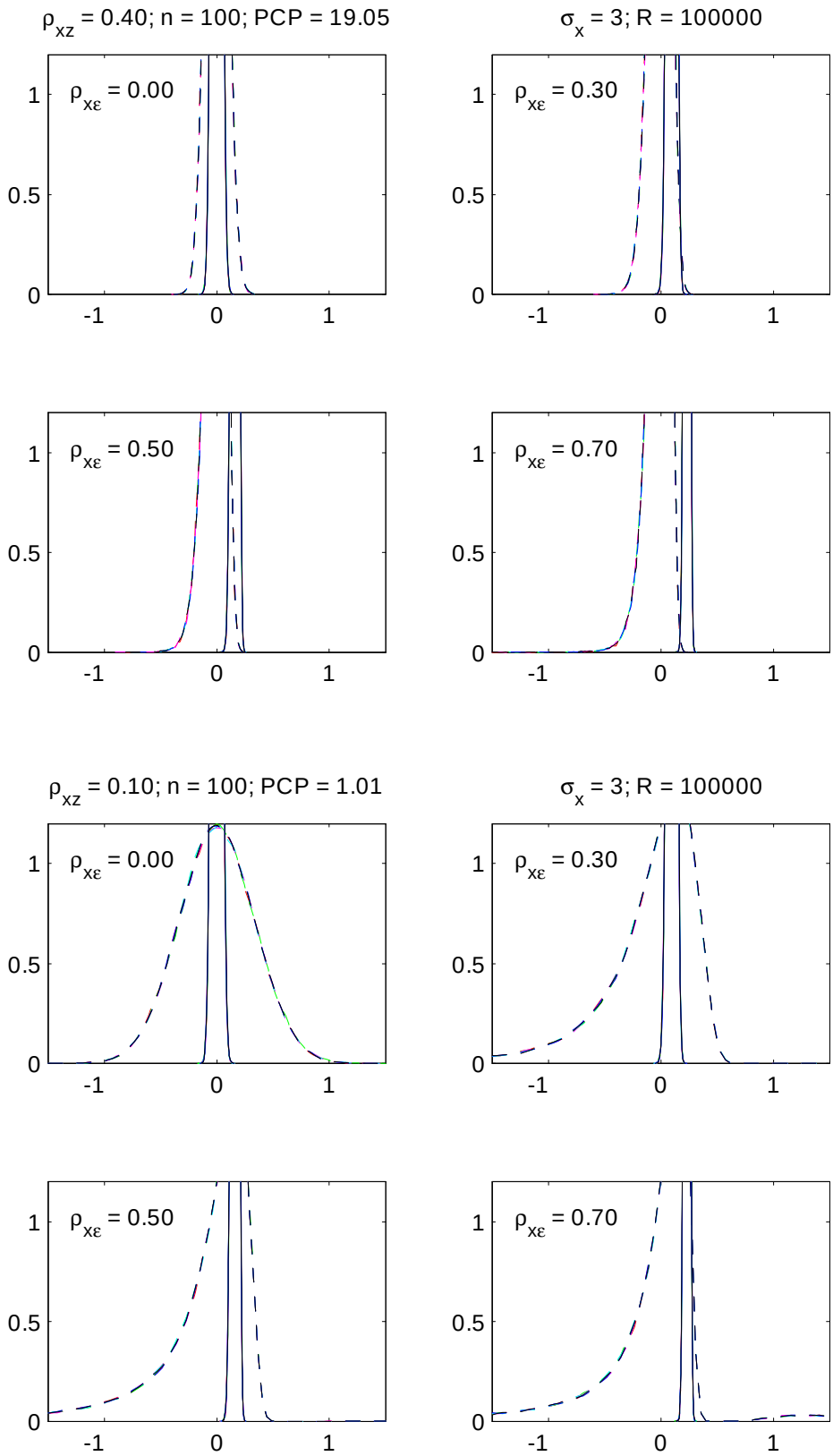


Diagram 7: IV (dashed) and OLS (solid) conditional on x^* (6×stylized)

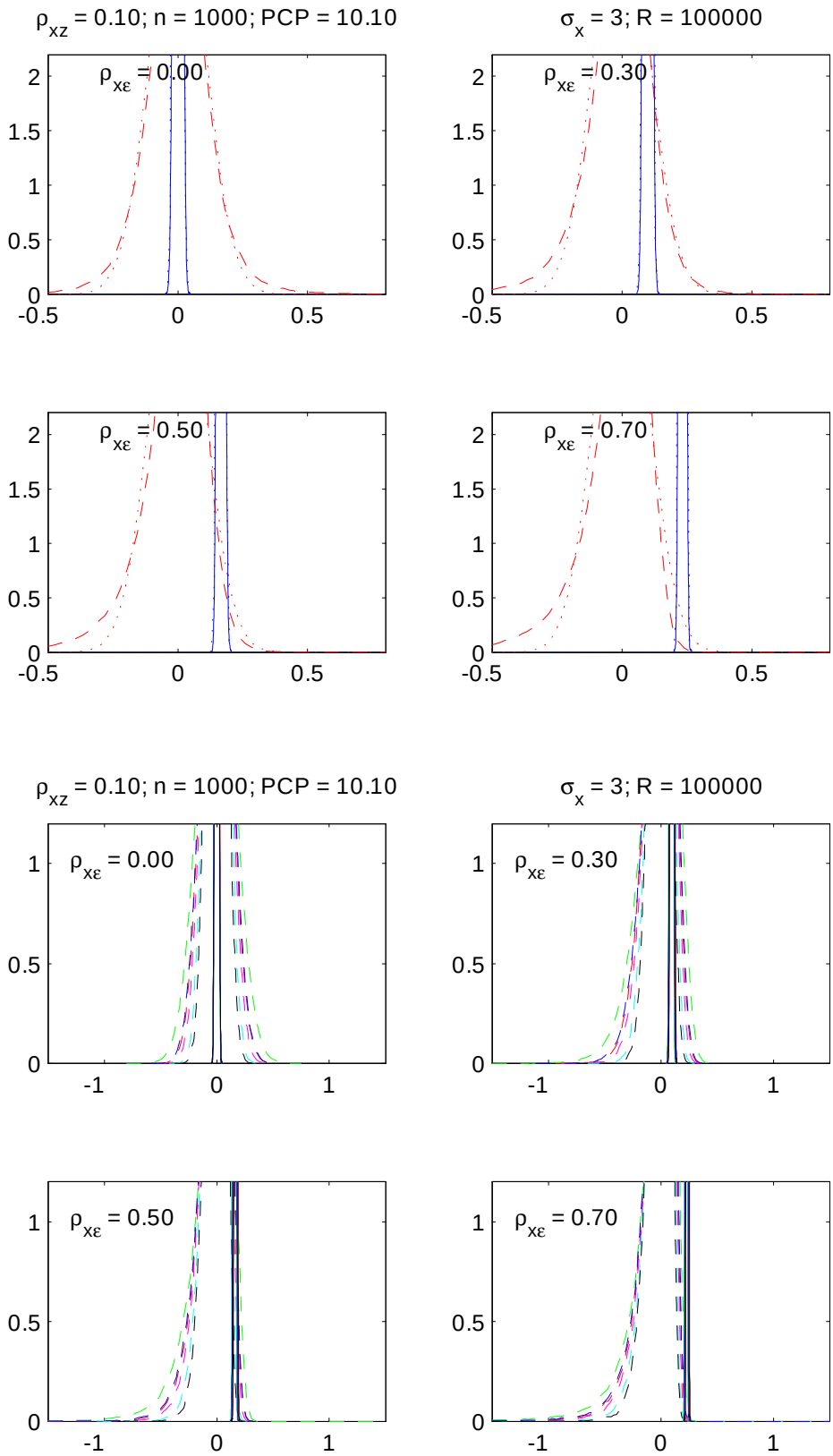


Diagram 8: IV (dashed) and OLS (solid) for $n = 1000$ and $\rho_{xz} = .1$; unconditional in the top panel and conditional on x^* (6×raw) in the bottom panel

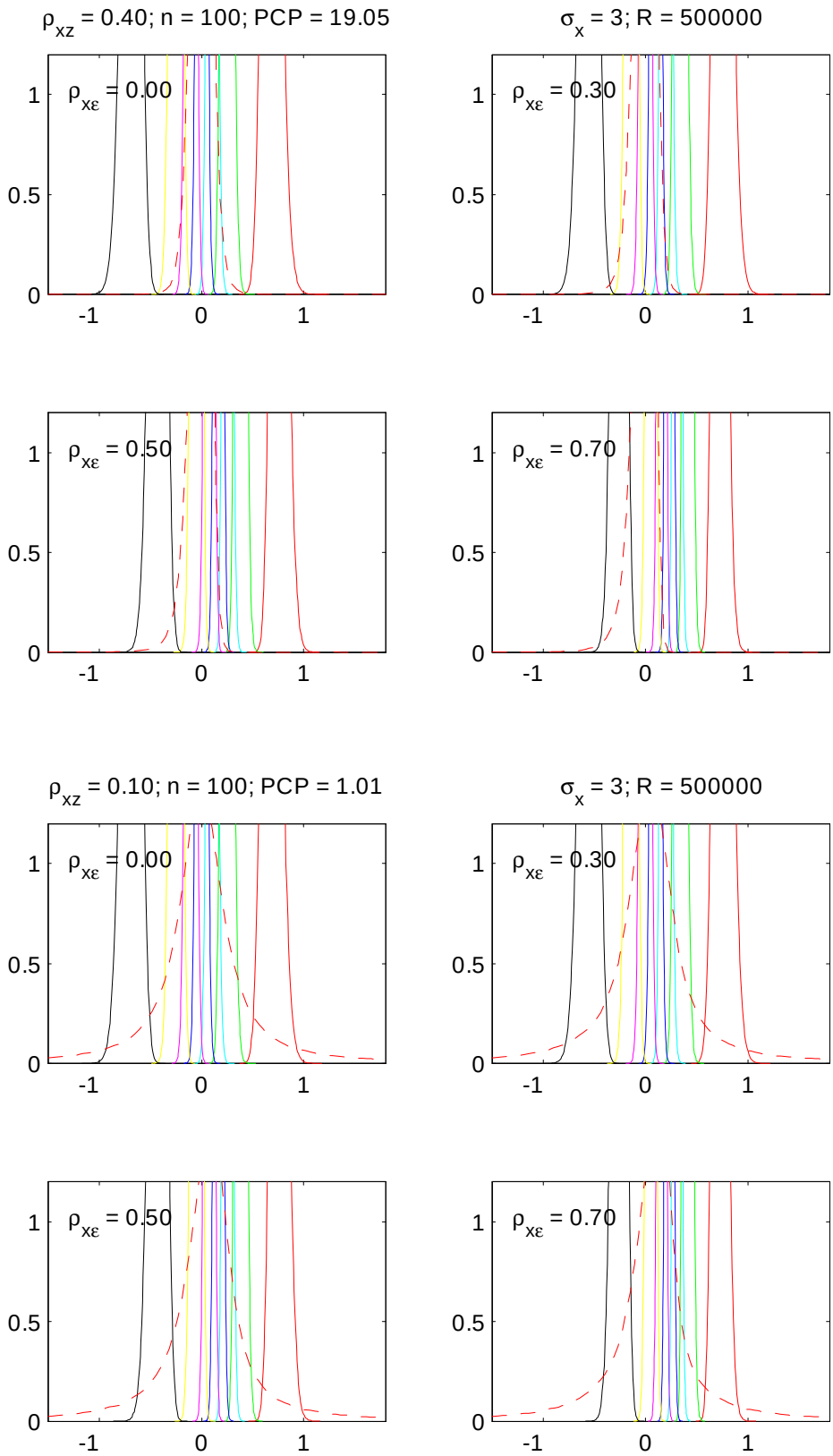


Diagram 9: Unconditional IV (dashed) and KLS for $n = 100$ where KLS uses: $\rho_{x\varepsilon} = -.9$ (red), $-.6$ (green) or $-.3$ (cyan), 0 (blue), .3 (magenta), .6 (yellow), .9 (black)

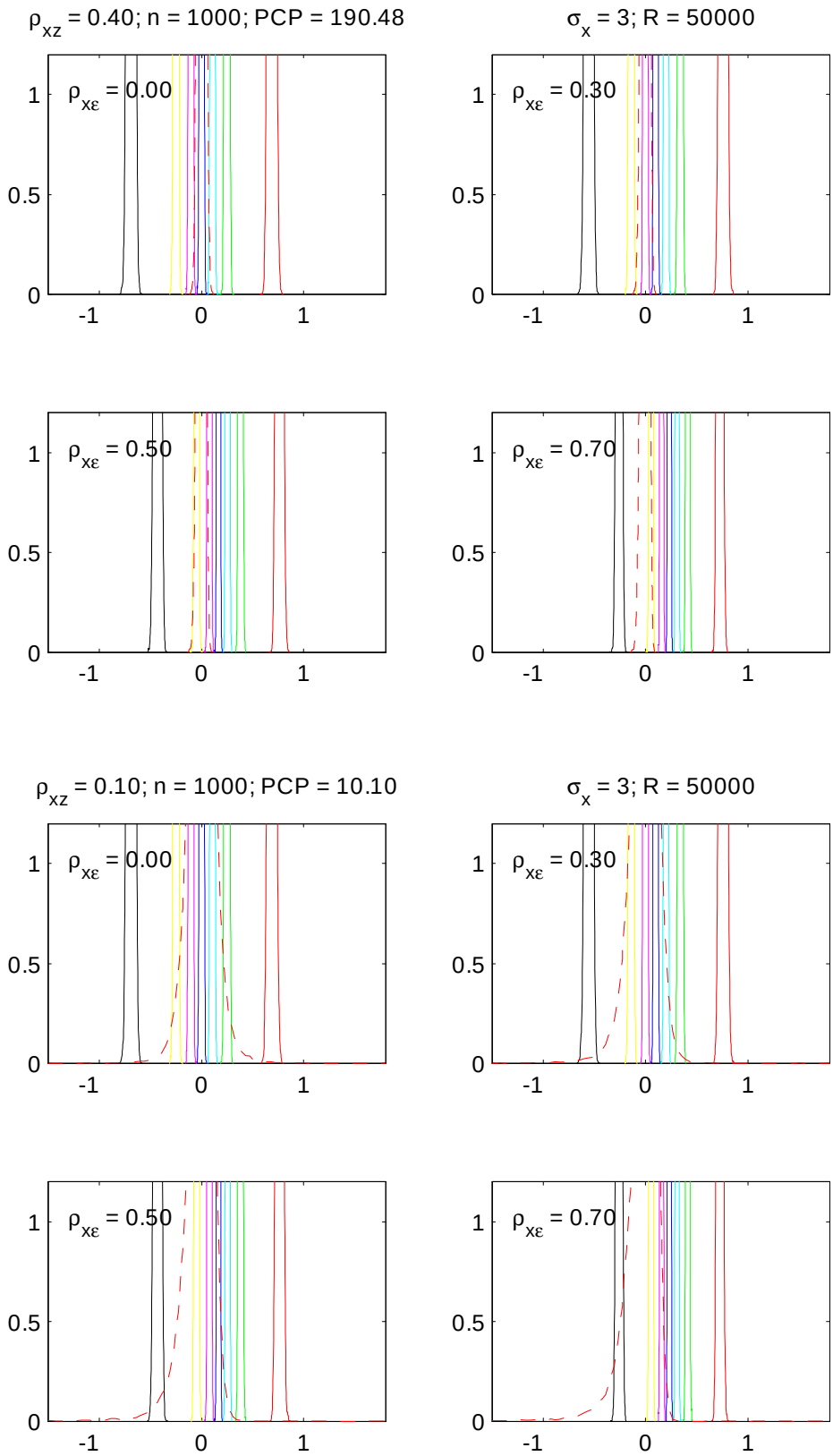


Diagram 10: Unconditional IV (dashed) and KLS for $n = 1000$ and where KLS uses: $\rho_{x\varepsilon} = -.9$ (red), $-.6$ (green) or $-.3$ (cyan), 0 (blue), .3 (magenta), .6 (yellow), .9 (black)