ISSN 1440-771X

MONASH University

Australia

Department of Econometrics
and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Approximating the Distribution of the Instrumental Variables Estimator when the Concentration Parameter is Small

D. S. Poskitt and C.L. Skeels

October 2004

Working Paper 19/04

Approximating the Distribution of the Instrumental Variables Estimator when the Concentration Parameter is Small.*

D S Poskitt
Department of Econometrics
and Business Statistics
Monash University
Vic 3800, Australia

C L Skeels
Department of Economics
The University of Melbourne
Victoria 3010, Australia

Abstract

This paper presents a new approximation to the exact sampling distribution of the instrumental variables estimator in simultaneous equations models. It differs from many of the approximations currently available, Edgeworth expansions for example, in that it is specifically designed to work well when the concentration parameter is small. The approximation is remarkable for the simplicity of its final form, for its accuracy and for its ability to capture those stylized facts that characterize lack of identification and weak instrument scenarios. The development leading to the approximation is also novel in that it introduces techniques of some independent interest not seen in this literature hitherto. (JEL CLASSIFICATION C16, C30)

Keywords: concentration parameter, IV estimator, simultaneous equations model, t approximation, weak instruments.

*We are grateful to Giovanni Forchini, Grant Hillier, Sophocles Mavroeidis, Murray Smith, James Stock and Andrew Tremayne for helpful, thought provoking comments on earlier drafts of the paper. In addition we would like to thank seminar participants at the Tinbergen Institute (Amsterdam), the universities of Sydney and York, and delegates at ESAM2003 and at the 14th EC^2 Conference. The usual caveat applies.

1 Introduction

In this paper we present a new approximation to the exact sampling distribution of the instrumental variables (IV) estimator of the coefficients on the endogenous regressors in a single equation from a linear system of simultaneous equations. More specifically we examine the properties of the two-stage least squares estimator and, as will be seen, the approximation we obtain is remarkable both for its accuracy and for its ability to capture many of the stylized facts that constitute the current state of knowledge. Manipulation of our results provides simple demonstrations of many of the qualitative characteristics that have been obtained under the different paradigms used to analyze weak identification, the related issue of weak instruments, and simultaneous equations models more generally.

Recent years have seen much exploration of the consequences of weak identification and weak instruments for estimation and inference in simultaneous equations models. The literature exploring this model reveals a variety of perspectives from which the problem has been considered, ranging from exact finite sample theory for totally or partially unidentified models (Choi & Phillips, 1992, Nelson & Startz, 1990a,b, Phillips, 1989), to local-to-zero asymptotics for identified (but asymptotically unidentified) models (Staiger & Stock, 1997, Startz, Nelson, & Zivot, 2000, Wang & Zivot, 1998, Zivot, Startz, & Nelson, 1998), through to the many-instrument asymptotics of (Bekker, 1994).[1] More recently a body of literature has developed that seeks to combine the many-instrument asymptotics of Bekker (1994) with the local-to-zero asymptotics of Staiger & Stock (1997), resulting in many-weak-instruments asymptotics; see, for example, Chao & Swanson (2002, 2003) and Stock & Yogo (2003). These asymptotic approaches differ essentially in the structure of the sequence in which they nest the model of interest and, although the exact details may differ with the approach, certain stylized facts emerge from these studies as characterizing the sampling behaviour of IV estimators; including (i) sampling distributions that are complicated mixtures of Normal distributions, typically asymmetric about the parameter of inter-

[1] Given the close relationship between weak instruments and a lack of identification, this literature can be traced back through to the work of *inter alia* Sargan (1983), Sims (1980) and Basmann (1963). For a more comprehensive treatment of the literature in this area see the survey by Stock, Wright, & Yogo (2002).

est, and (ii) non-standard asymptotic results with non-degenerate limiting distributions.

At the risk of getting ahead of ourselves, we find that certain functions of the IV estimator can be approximated by various t-distributions.[2] Our approximation provides a framework that goes some way towards unifying the qualitatively similar but technically distinct results of Staiger & Stock (1997), Wang & Zivot (1998), Zivot et al. (1998) and Startz et al. (2000), on the one hand, and Nelson & Startz (1990b), Phillips (1989) and Choi & Phillips (1992) on the other. For example, t-distributions can be thought of as mixed-Normal distributions, a feature of many existing results. Similarly, the asymptotic normality implied by the many-instrument asymptotics of Bekker (1994) can also be obtained as a special case. Quite apart from its simplicity and its explanatory power, the approximation is of independent interest in view of the novelty of its development which, as far as we are aware, has not appeared in the econometrics literature heretofore.

The remainder of the paper has the following structure. In the next section we will introduce the model and establish notation whilst considering a canonical transformation. Section 3 presents the main theoretical developments of the paper. In that section we explore a skewed approximation to the non-central Wishart distribution that is based on the central Wishart distribution. This skewed approximation then forms the basis of our approximation to the distribution of the IV estimator. In Section 4 we relate the magnitude of the concentration parameter to the notion of instrument relevance or, conversely, weakness and examine how our results are related to various features that have been observed under different scenarios used to analyze weak identification/instruments. The quality of the approximations is then explored in Section 5. Section 6 presents some brief concluding remarks and discusses the practical implementation of the approximating distribution.

[2]It has been known for some time that the distribution of the IV estimator is approximately multivariate - t; see, for example, Phillips (1980, p.870). However, the approximations presented here involve different parameterizations and, as we show below, they only reduce to existing results in certain special cases.

2 Background

Consider the structural model

$$\mathbf{y} = \mathbf{Y}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma} + \mathbf{u}\,, \qquad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_T) \tag{1}$$

where the endogenous matrix variables $\mathbf{y}$ and $\mathbf{Y}$ are $T \times 1$ and $T \times n$, respectively, the matrix of explanatory variables $\mathbf{Z}_1$ is $T \times K_1$, and $\mathbf{u}$ denotes the $T \times 1$ vector of structural disturbances.[3] The vectors of structural coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $n \times 1$ and $K_1 \times 1$, respectively. The corresponding reduced form model is

$$[\mathbf{y}, \mathbf{Y}] = [\mathbf{Z}_1, \mathbf{Z}_2] \begin{bmatrix} \boldsymbol{\pi}_1 & \boldsymbol{\Pi}_1 \\ \boldsymbol{\pi}_2 & \boldsymbol{\Pi}_2 \end{bmatrix} + [\mathbf{v}, \mathbf{V}]\,. \tag{2}$$

Here the rows of the $T \times (n+1)$ matrix $[\mathbf{v}, \mathbf{V}]$ are independent normal vectors with zero mean and common $(n+1) \times (n+1)$ covariance matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}, \tag{3}$$

ω_{11} scalar, so that $[\mathbf{v}, \mathbf{V}] \sim N(\mathbf{0}, \boldsymbol{\Omega} \otimes \mathbf{I}_T)$,[4] where $[\mathbf{v}, \mathbf{V}]$ is partitioned conformably with $[\mathbf{y}, \mathbf{Y}]$. Note that, by implication, the structural variance $\sigma_u^2 = [1, -\boldsymbol{\beta}'] \boldsymbol{\Omega} [1, -\boldsymbol{\beta}']'$. Defining the $T \times K$ matrix $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$ to be of full column rank, where $\mathbf{Z}_2$ denotes the matrix of K_2 exogenous regressors excluded from equation (1) and where $K = K_1 + K_2$, it follows that $[\mathbf{y}, \mathbf{Y}] \sim N(\mathbf{Z}\boldsymbol{\Pi}, \boldsymbol{\Omega} \otimes \mathbf{I}_T)$. The components of the reduced form coefficient matrix $\boldsymbol{\Pi}$ — namely $\boldsymbol{\pi}_1$, $\boldsymbol{\Pi}_1$, $\boldsymbol{\pi}_2$ and $\boldsymbol{\Pi}_2$ — are of dimension $K_1 \times 1$, $K_1 \times n$, $K_2 \times 1$ and $K_2 \times n$, respectively.

We are interested in the IV estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{Y}'(\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_1})\mathbf{Y})^{-1} \mathbf{Y}'(\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_1})\mathbf{y}, \tag{4}$$

where for any $p \times q$ matrix $\mathbf{A}$ of full column rank the notation $\mathbf{P}_{\mathbf{A}}$ and $\mathbf{R}_{\mathbf{A}}$ denotes the symmetric, idempotent matrices $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ and $\mathbf{R}_{\mathbf{A}} = \mathbf{I}_p - \mathbf{P}_{\mathbf{A}}$, of rank q and $p-q$, respectively. The $T \times T$ matrix $\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_1} = \mathbf{R}_{\mathbf{Z}_1} - \mathbf{R}_{\mathbf{Z}}$ has rank $\nu = K_2 \geq n$ and a spectral decomposition implies that there exists

[3] As we shall discuss below, although convenient for expository purposes, the normality assumption is not critical to the development of subsequent results.

[4] The notation $\mathbf{X} \sim N(\mathbf{M}, \boldsymbol{\Omega})$ should be read as $\text{vec}(\mathbf{X}) \sim N(\text{vec}(\mathbf{M}), \boldsymbol{\Omega})$.

a $T \times \nu$ matrix $\mathbf{C} = \mathbf{R}_{\mathbf{Z}_1}\mathbf{Z}_2'(\mathbf{Z}_2'\mathbf{R}_{\mathbf{Z}_1}\mathbf{Z}_2)^{-1/2}$ such that $\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_1} = \mathbf{C}\mathbf{C}'$, where $\mathbf{C}'\mathbf{C} = \mathbf{I}_\nu$ and $\mathbf{C}'\mathbf{Z}_1 = \mathbf{0}$. If we pre-multiply by $\mathbf{C}'$, so that $[\tilde{\mathbf{y}}, \widetilde{\mathbf{Y}}] = \mathbf{C}'[\mathbf{y}, \mathbf{Y}]$ and $\widetilde{\mathbf{Z}}_2 = \mathbf{C}'\mathbf{Z}_2$, then the model becomes

$$\begin{aligned} \tilde{\mathbf{y}} &= \widetilde{\mathbf{Y}}\boldsymbol{\beta} + \tilde{\mathbf{u}}\,, \qquad \tilde{\mathbf{u}} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I}_\nu) \\ [\tilde{\mathbf{y}}, \widetilde{\mathbf{Y}}] &= \widetilde{\mathbf{Z}}_2[\boldsymbol{\pi}_2, \boldsymbol{\Pi}_2] + [\tilde{\mathbf{v}}, \widetilde{\mathbf{V}}]\,, \qquad [\tilde{\mathbf{v}}, \widetilde{\mathbf{V}}] \ \sim N(\mathbf{0}, \boldsymbol{\Omega} \otimes \mathbf{I}_\nu) \end{aligned}$$

where $\tilde{\mathbf{u}} = \mathbf{C}'\mathbf{u}$ and $[\tilde{\mathbf{v}}, \widetilde{\mathbf{V}}] = \mathbf{C}'[\mathbf{v}, \mathbf{V}]$.

Henceforth we will discuss the transformed system, unless explicitly stated otherwise, and for notational convenience we will drop the tilde and revert to generic symbolism for the variables. Thus the IV estimator is now

$$\widehat{\boldsymbol{\beta}} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{y}\,, \tag{4a}$$

where $[\mathbf{y}, \mathbf{Y}] \sim N([\mathbf{m}, \mathbf{M}], \boldsymbol{\Omega} \otimes \mathbf{I}_\nu)$ with $[\mathbf{m}, \mathbf{M}] = \mathbf{Z}_2[\boldsymbol{\pi}_2, \boldsymbol{\Pi}_2]$ and

$$\mathbf{S} = [\mathbf{y}, \mathbf{Y}]'[\mathbf{y}, \mathbf{Y}] \sim \mathrm{W}_{n+1}(\nu, \boldsymbol{\Omega}, \nu\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}})\,, \tag{5}$$

where $\boldsymbol{\Delta} = \nu^{-1}[\mathbf{m}, \mathbf{M}]'[\mathbf{m}, \mathbf{M}]$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{\Omega}^{\frac{1}{2}}$, with $\boldsymbol{\Omega}^{\frac{1}{2}}$ the symmetric square root of $\boldsymbol{\Omega}$.[5] That is, $\mathbf{S}$ has a non-central Wishart distribution with ν degrees of freedom, covariance matrix $\boldsymbol{\Omega}$ and non-centrality parameter $\nu\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}}$.[6] Further, we have the usual compatibility condition

$$\boldsymbol{\pi}_2 = \boldsymbol{\Pi}_2\boldsymbol{\beta}\,,$$

and so

$$\boldsymbol{\Delta} = \begin{bmatrix} \delta_{11} & \boldsymbol{\delta}_{12} \\ \boldsymbol{\delta}_{21} & \boldsymbol{\Delta}_{22} \end{bmatrix} = \nu^{-1}[\boldsymbol{\beta}, \mathbf{I}_n]'\boldsymbol{\Pi}_2'\mathbf{Z}_2'\mathbf{Z}_2\boldsymbol{\Pi}_2[\boldsymbol{\beta}, \mathbf{I}_n]\,,$$

[5] If the spectral decomposition of $\boldsymbol{\Omega}$ is $\mathbf{H}'\boldsymbol{\Omega}\mathbf{H} = \mathbf{D}$, where $\mathbf{H}$ is an orthogonal matrix of characteristic vectors of $\boldsymbol{\Omega}$ and $\mathbf{D} = \mathrm{diag}[\lambda_1(\boldsymbol{\Omega}), \ldots, \lambda_{n+1}(\boldsymbol{\Omega})]$ is the diagonal matrix of characteristic roots, then $\boldsymbol{\Omega}^{\frac{1}{2}} = \mathbf{H}\mathbf{D}^{\frac{1}{2}}\mathbf{H}'$ where $\mathbf{D}^{\frac{1}{2}} = \mathrm{diag}[\lambda_1(\boldsymbol{\Omega})^{\frac{1}{2}}, \ldots, \lambda_{n+1}(\boldsymbol{\Omega})^{\frac{1}{2}}]$; see, for example, Searle (1982, Section 11.6).

[6] In Footnote 3 we made comment about the normality assumption not being crucial to subsequent developments. We address this point here. First, from an exact distribution perspective, note that the normality assumption can be relaxed because $\mathbf{S}$ will have a Wishart distribution for any elliptically symmetric distribution on $[\mathbf{y}\ \mathbf{Y}]$. Second, taking a different perspective, if we briefly revert to $[\tilde{\mathbf{y}}, \widetilde{\mathbf{Y}}]$ to denote the variables in our transformed space, observe that each of their elements are linear combinations of the original variables $[\mathbf{y}\ \mathbf{Y}]$. Under reasonably general conditions it follows that the elements of $[\tilde{\mathbf{y}}, \widetilde{\mathbf{Y}}]$ will be approximately normally distributed and so $\mathbf{S}$ will be approximately Wishart. Provided that this latter approximation is not too coarse, our results will carry over without change.

where the partition of $\mathbf{\Delta}$ occurs after the first row and column. All subsequent partitions of matrices will be conformable with that of $\mathbf{\Omega}$ and $\mathbf{\Delta}$ unless stated otherwise.

Following standard practice we will use $\mathbf{\Gamma}_{22} = \nu\mathbf{\Omega}_{22}^{-\frac{1}{2}}\mathbf{\Delta}_{22}(\mathbf{\Omega}_{22}^{-\frac{1}{2}})'$ to denote the concentration parameter. We will refer to

$$\mu^2 = \text{tr}\{\mathbf{\Gamma}_{22}\} = \nu \times \text{tr}\{\mathbf{\Omega}_{22}^{-1}\mathbf{\Delta}_{22}\}$$

as the concentration coefficient. Noting that the Euclidean norm of $\mathbf{A}$ is $\|\mathbf{A}\| = \sqrt{\text{tr}\{\mathbf{A}'\mathbf{A}\}}$, we see that the concentration coefficient is simply a natural measure of the magnitude of the concentration parameter. The importance of the magnitude of the concentration parameter for the sampling behaviour of $\widehat{\boldsymbol{\beta}}$ has been well documented in the literature, see, *inter alia*, Mariano (1982, Sections 3 and 4) and Phillips (1983, Section 3.6). Rothenberg (1984) discusses Edgeworth type expansions of the distribution of the IV estimator, as in Sargan & Mikhail (1971) and Anderson & Sawa (1973, 1979), and points out that the resulting approximations can be poor if the concentration parameter is not large. One of the main contributions of this paper is to provide an approximation to the distribution of $\widehat{\boldsymbol{\beta}}$ that is designed to work well when $\mathbf{\Gamma}_{22}$ is small.

A major stumbling block in the development of exact distribution theory for the IV estimator in (4), or equivalently (4a), is the implied non-centrality in the distribution of $\mathbf{S}$. Our approximation, which is developed in a series of results presented in the following section, exploits a technique used by Steyn & Roux (1972) to approximate the non-central Wishart distribution by a central Wishart distribution when the non-centrality parameter $\mathbf{\Delta}$ is small. Once in the central case it proves to be relatively straight-forward to derive a corresponding approximation to the sampling distribution of $\widehat{\boldsymbol{\beta}}$. To relate the magnitude of $\mathbf{\Delta}$ to the concentration parameter note that $\|\mathbf{\Delta}\| \leq (\|\boldsymbol{\beta}\|^2 + n)\|\mathbf{\Delta}_{22}\|$ and $\|\mathbf{\Delta}_{22}\| \leq \nu^{-1}\|\mathbf{\Omega}_{22}^{\frac{1}{2}}\|^2\|\mathbf{Z}_2\mathbf{\Pi}_2(\mathbf{\Omega}_{22}^{-\frac{1}{2}})'\|^2 = \nu^{-1}\|\mathbf{\Omega}_{22}^{\frac{1}{2}}\|^2\mu^2$. If $0 < \mathbf{\Omega} < \infty$, meaning that the characteristic values of $\mathbf{\Omega}$ satisfy the inequalities $0 < \lambda_{\min}(\mathbf{\Omega}) \leq \lambda_{\max}(\mathbf{\Omega}) < \infty$, which we can suppose without loss of generality, then $0 < \|\mathbf{\Omega}_{22}^{\frac{1}{2}}\|^2 < \infty$ and small values of the concentration coefficient μ^2 imply that $\mathbf{\Delta}$ must also be small. Hence our approximation is applicable under circumstances that differ significantly from those for which standard approximations are designed and is complementary to them.

3 The Approximation

In essence our approximation is obtained by perturbing the covariance matrix of a central Wishart variate to match moments (to some order of accuracy) with the non-central Wishart distribution of interest. In order to make these ideas concrete we begin by presenting a differential equation for the characteristic function of $\mathbf{S}$. Steyn & Roux (1972) originally gave an equivalent result in terms of the moment generating function of a non-central Wishart variate; they established the result by using the representation of $\mathbf{S}$ in terms of Normal vectors.

Lemma 1. *Let* $\mathbf{S} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Omega}, \nu\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{\Delta}\mathbf{\Omega}^{-\frac{1}{2}})$*, and let* $\Phi_S(\mathbf{T})$ *denote the characteristic function of* $\mathbf{S}$*, so that*

$$\Phi \equiv \Phi_S(\mathbf{T}) = |\mathbf{\Psi}|^{-\nu/2} \exp\left\{\mathrm{tr}\left(i\nu\mathbf{\Omega}\mathbf{T}\mathbf{\Psi}^{-1}\mathbf{\Omega}^{-1}\mathbf{\Delta}\right)\right\},$$

where $i^2 = -1$*,* $\mathbf{\Psi} = \mathbf{I}_{n+1} - 2i\mathbf{\Omega}\mathbf{T}$*, and* $\mathbf{T} = \{\tau_{jk}\}$*, with* $\tau_{jk} = \frac{1}{2}(1+\delta_{jk})\eta_{jk}$*,* $\eta_{jk} = \eta_{kj}$*,* $j, k = 1, \ldots, n+1$*, and* δ_{jk} *is Kronecker's delta,*

$$\delta_{jk} = \begin{cases} 1, & \text{if } j = k, \\ 0 & \text{otherwise.} \end{cases}$$

Then Φ *satisfies the differential equation*

$$\frac{\partial\Phi}{\partial\mathbf{H}} = i\nu\left[\mathbf{\Psi}^{-1}\mathbf{\Omega} + \mathbf{\Psi}^{-1}\mathbf{\Delta}(\mathbf{\Psi}^{-1})'\right]\Phi, \tag{6}$$

where $\partial\Phi/\partial\mathbf{H} = \{\partial\Phi/\partial\eta_{jk}\}$*,* $j, k = 1, \ldots, n+1$*.*

Proof. The characteristic function Φ is a rearrangement of that given in Gupta & Nagar (2000, Theorem 3.5.3). The differential equation (6) is given in Gupta & Nagar (2000, proof of Theorem 3.5.4). □

Corollary 1. *If* $\mathbf{S} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Omega})$ *then equation* (6) *reduces to*

$$\frac{\partial\Phi}{\partial\mathbf{H}} = i\nu(\mathbf{I}_{n+1} - 2i\mathbf{\Omega}\mathbf{T})^{-1}\mathbf{\Omega}\Phi. \tag{7}$$

We now have the following approximation to the non-central Wishart distribution.

Theorem 1. *Let* $\mathbf{W} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Omega}, \nu\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{\Delta}\mathbf{\Omega}^{-\frac{1}{2}})$*, where* $\|\mathbf{\Omega}\| = O(1)$ *and* $\|\mathbf{\Delta}\| = o(1)$*, then* $f(\mathbf{W}) = f(\widetilde{\mathbf{W}}) + O(\|\mathbf{\Delta}\|)$*, where* $\widetilde{\mathbf{W}} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Sigma})$ *with* $\mathbf{\Sigma} = \mathbf{\Omega} + \mathbf{\Delta}$ *and* $f(\cdot)$ *generically denoting the relevant density functions.*

Proof. Our proof parallels that of Steyn & Roux (1972, Section 4) — see also Gupta & Nagar (2000, pp.125–6) — although we explicitly control the error of the approximation by reference to the order of magnitude of $\|\mathbf{\Delta}\|$. If $\mathbf{W} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Omega}, \nu\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{\Delta}\mathbf{\Omega}^{-\frac{1}{2}})$ then $\mathrm{E}[\mathbf{W}] = \nu(\mathbf{\Omega} + \mathbf{\Delta})$. Similarly, if $\widetilde{\mathbf{W}} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Sigma})$ then $\mathrm{E}[\widetilde{\mathbf{W}}] = \nu\mathbf{\Sigma}$. So a method of moments approximation suggests choosing $\mathbf{\Sigma} = \mathbf{\Omega}+\mathbf{\Delta}$. With this motivation, suppose that we replace $\mathbf{\Omega}$ in equation (7) by $\mathbf{\Sigma}$. This yields

$$\begin{aligned}
\frac{\partial \Phi}{\partial \mathbf{H}} &= i\nu \left[\mathbf{I}_{n+1} - 2i(\mathbf{\Omega} + \mathbf{\Delta})\mathbf{T}\right]^{-1} (\mathbf{\Omega} + \mathbf{\Delta})\Phi \\
&= i\nu[\mathbf{I}_{n+1} + 2i(\mathbf{\Omega} + \mathbf{\Delta})\mathbf{T} + \{2i(\mathbf{\Omega} + \mathbf{\Delta})\mathbf{T}\}^2 + \dots\ ](\mathbf{\Omega} + \mathbf{\Delta})\Phi \\
&= i\nu[\mathbf{\Psi}^{-1} + \mathbf{\Psi}^{-1}(2i\mathbf{\Delta}\mathbf{T})\mathbf{\Psi}^{-1} + O(\|\mathbf{\Delta}\|^2)](\mathbf{\Omega} + \mathbf{\Delta})\Phi \\
&= i\nu[\mathbf{\Psi}^{-1}\mathbf{\Omega} + \mathbf{\Psi}^{-1}\mathbf{\Delta}\left\{\mathbf{I}_{n+1} + 2i\mathbf{T}\mathbf{\Psi}^{-1}\mathbf{\Omega}\right\} + O(\|\mathbf{\Delta}\|^2)]\Phi \\
&= i\nu[\mathbf{\Psi}^{-1}\mathbf{\Omega} + \mathbf{\Psi}^{-1}\mathbf{\Delta}(\mathbf{\Psi}^{-1})']\Phi + O(\|\mathbf{\Delta}\|^2)
\end{aligned}$$

which is the same as equation (6) except for terms of order $O(\|\mathbf{\Delta}\|^2)$. The latter implies, via inversion of the characteristic function, that the approximating distribution is accurate to terms of order $O(\|\mathbf{\Delta}\|)$. □

The significance of Theorem 1, for our purposes, is that it provides guidance on the construction of a rather stronger result. Set $\mathbf{G} = [\mathbf{m}, \mathbf{M}]$ and suppose that the $\nu \times (n+1)$ matrix $\mathbf{N} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{\nu(n+1)})$. Then, by definition,

$$\begin{aligned}
\mathbf{W} &= \mathbf{\Omega}^{1/2}(\mathbf{N} + \mathbf{G}\mathbf{\Omega}^{-1})'(\mathbf{N} + \mathbf{G}\mathbf{\Omega}^{-1})\mathbf{\Omega}^{1/2} \\
&\sim \mathrm{W}_{n+1}(\nu, \mathbf{\Omega}, \nu\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{\Delta}\mathbf{\Omega}^{-\frac{1}{2}})\,.
\end{aligned}$$

Similarly, if $\mathbf{\Sigma} = \mathbf{\Omega} + \mathbf{\Delta}$ then

$$\widetilde{\mathbf{W}} = \mathbf{\Sigma}^{1/2}\mathbf{N}'\mathbf{N}\mathbf{\Sigma}^{1/2} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Sigma})\,.$$

From the definitions of $\mathbf{W}$ and $\widetilde{\mathbf{W}}$ it follows that

$$\begin{aligned}
\mathbf{W} = \mathbf{\Omega}^{1/2}\mathbf{\Sigma}^{-1/2}\widetilde{\mathbf{W}}\mathbf{\Sigma}^{-1/2}\mathbf{\Omega}^{1/2} &+ \mathbf{\Omega}^{1/2}\mathbf{N}'\mathbf{G}\mathbf{\Omega}^{-1/2} \\
&+ \mathbf{\Omega}^{-1/2}\mathbf{G}'\mathbf{N}\mathbf{\Omega}^{1/2} + \nu\mathbf{\Omega}^{-1/2}\mathbf{\Delta}\mathbf{\Omega}^{-1/2}\,.
\end{aligned} \tag{8}$$

The middle two terms on the right hand side of equation (8) are of order $O_p(||\mathbf{G}||)$, since $||\mathbf{AB}|| \leq ||\mathbf{A}|| \cdot ||\mathbf{B}||$, and the final term is of order $O(||\mathbf{\Delta}||)$. This observation leads to the following result.

Theorem 2. *If the matrix* $\mathbf{W} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Omega}, \nu\mathbf{\Omega}^{-1/2}\mathbf{\Delta}\mathbf{\Omega}^{-1/2})$*, where* $0 < \mathbf{\Omega} < \infty$ *and* $\mathbf{\Delta} = \nu^{-1}[\mathbf{m}, \mathbf{M}]'[\mathbf{m}, \mathbf{M}]$ *with* $[\mathbf{m}, \mathbf{M}] = o(1)$*, then there exists a random matrix* $\widetilde{\mathbf{W}}$*, defined on the same probability space as* $\mathbf{W}$*, such that*

$$\mathbf{W} = \widetilde{\mathbf{W}} + O_p(|||[\mathbf{m}, \mathbf{M}]|||)\,, \tag{9}$$

where $\widetilde{\mathbf{W}} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Sigma})$ *with* $\mathbf{\Sigma} = \mathbf{\Omega} + \mathbf{\Delta}$*.*

Proof. Since $||\mathbf{\Delta}|| \leq \nu^{-1}||\mathbf{G}||^2$ we can deduce from equation (8) that

$$\mathbf{W} = \mathbf{\Omega}^{1/2}\mathbf{\Sigma}^{-1/2}\widetilde{\mathbf{W}}\mathbf{\Sigma}^{-1/2}\mathbf{\Omega}^{1/2} + O_p(||\mathbf{G}||) \tag{10}$$

where, we recall, $\mathbf{G} = [\mathbf{m}, \mathbf{M}]$. To complete the proof note that a consequence of Lemma (A.1) is that $\mathbf{\Omega}^{1/2} = \mathbf{\Sigma}^{1/2} + O(||\mathbf{\Delta}||^{\frac{1}{2}})$ and therefore

$$\mathbf{\Omega}^{1/2}\mathbf{\Sigma}^{-1/2} = \mathbf{I}_{n+1} + O(||\mathbf{G}||)\,, \tag{11}$$

which when substituted into equation (10) yields, as required,

$$\begin{aligned}\mathbf{W} &= [\mathbf{I}_{n+1} + O(||\mathbf{G}||)]\widetilde{\mathbf{W}}[\mathbf{I}_{n+1} + O(||\mathbf{G}||)] + O_p(||\mathbf{G}||) \\ &= \widetilde{\mathbf{W}} + O_p(||\mathbf{G}||)\,.\end{aligned}$$

□

Having reduced the problem to one involving a central Wishart distribution we are in a position to exploit the following result:

Lemma 2. *Suppose that* $\mathbf{S} \sim \mathrm{W}_{n+1}(\nu, \mathbf{\Sigma})$*. Partition*

$$\mathbf{S} = \begin{bmatrix}\mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22}\end{bmatrix} \qquad \textit{and} \qquad \mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{bmatrix}$$

conformably, with $\mathbf{S}_{22}$ $p \times p$*. If* $\mathbf{B} = \mathbf{S}_{22}^{-1}\mathbf{S}_{21}$*, then the density function of* $\mathbf{B}$ *is*

$$\begin{aligned}f(\mathbf{B}) = \frac{\Gamma_p\left[\frac{\nu+n-p+1}{2}\right]}{\pi^{p(n-p+1)/2}\Gamma_p\left[\frac{\nu}{2}\right]} &|\mathbf{\Sigma}_{22}|^{-v/2}\,|\mathbf{\Sigma}_{11\cdot 2}|^{-p/2} \\ &\times |\mathbf{\Sigma}_{22}^{-1} + (\mathbf{B} - \boldsymbol{\theta})\mathbf{\Sigma}_{11\cdot 2}^{-1}(\mathbf{B} - \boldsymbol{\theta})'|^{-(\nu+n-p+1)/2},\end{aligned} \tag{12}$$

where $\boldsymbol{\theta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ *and* $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. *That is,* $\mathbf{B}$ *has a matrix variate t-distribution with* $\nu - p + 1$ *degrees of freedom and parameters* $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}_{22}$ *and* $\boldsymbol{\Sigma}_{11\cdot 2}$.

Proof. See Kshirsagar (1961, Section 4).[7] □

We now explore the sampling behaviour of the IV estimator $\widehat{\boldsymbol{\beta}}$ defined in equation (4), or equivalently equation (4a).

Theorem 3. *The estimator* $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} + o_p(1)$ *if* $[\mathbf{m}, \mathbf{M}] = o(1)$ *and, by implication,* $\widehat{\boldsymbol{\beta}}$ *converges in distribution to* $\widetilde{\boldsymbol{\beta}}$ *as* $\|\boldsymbol{\Delta}\| \to 0$ *where*

$$f(\widetilde{\boldsymbol{\beta}}) = \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\pi^{n/2}\Gamma\left[\frac{\nu-n+1}{2}\right]} |\boldsymbol{\Theta}|^{1/2} \left[1 + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\theta})'\boldsymbol{\Theta}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\theta})\right]^{-(\nu+1)/2}, \tag{13}$$

with $\boldsymbol{\theta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$ *and* $\boldsymbol{\Theta} = (\sigma_{11} - \boldsymbol{\sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21})^{-1}\boldsymbol{\Sigma}_{22}$. *That is, the distribution of* $\widehat{\boldsymbol{\beta}}$ *is approximately multivariate t with* $\nu - n + 1$ *degrees of freedom, location parameter* $\boldsymbol{\theta}$ *and scale parameter* $\boldsymbol{\Theta}^{-1}$. *We shall write*

$$\widehat{\boldsymbol{\beta}} \underset{a}{\sim} \mathbf{t}_n(\nu - n + 1, \boldsymbol{\theta}, \boldsymbol{\Theta}).$$

Proof. Let $\mathbf{S} \sim \mathrm{W}_{n+1}(\nu, \boldsymbol{\Omega}, \nu\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}})$ be as defined in equation (5) and, as previously, let $\boldsymbol{\Sigma} = \boldsymbol{\Omega} + \boldsymbol{\Delta}$ and $\mathbf{G} = [\mathbf{m}, \mathbf{M}]$. Partition

$$\mathbf{S} = \begin{bmatrix} s_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{21} & \mathbf{S}_{22} \end{bmatrix}$$

such that s_{11} is scalar and $\mathbf{S}_{22}$ is $n \times n$. Then $\widehat{\boldsymbol{\beta}} = \mathbf{S}_{22}^{-1}\mathbf{s}_{21}$. Now, from Theorem 2 we know that there exists a $\widetilde{\mathbf{S}} \sim \mathrm{W}_{n+1}(\nu, \boldsymbol{\Sigma})$ such that $\mathbf{S} = \widetilde{\mathbf{S}} + O_p(||\mathbf{G}||)$. Partitioning $\widetilde{\mathbf{S}}$ conformably with $\mathbf{S}$, it follows that

$$\left(\widetilde{\mathbf{S}}_{0,22} + O_p(||\mathbf{G}||)\right)\widehat{\boldsymbol{\beta}} = \widetilde{\mathbf{s}}_{0,21} + O_p(||\mathbf{G}||). \tag{14}$$

From (14) we can deduce that $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} + O_p(||\mathbf{G}||)$ where $\widetilde{\boldsymbol{\beta}} = \widetilde{\mathbf{S}}_{0,22}^{-1}\widetilde{\mathbf{s}}_{0,21}$. The proof is completed by applying Lemma 2 (with $p = n$) and rearranging terms, noting that $\Gamma_n\left(\frac{\nu+1}{2}\right)\Gamma\left(\frac{\nu-n+1}{2}\right) = \Gamma\left(\frac{\nu+1}{2}\right)\Gamma_n\left(\frac{\nu}{2}\right)$. □

[7] A detailed description of the matrix variate t-distribution can be found in Gupta & Nagar (2000, Chapter 4). An early study of this distribution appears in Dickey (1967).

Theorem 3 is the key result of the paper and our subsequent results follow directly from it. It is worth remarking that a t distribution can be thought of as a mixed Gaussian distribution. Hence, although considerably simpler in final form than the results of either Choi & Phillips (1992) or Staiger & Stock (1997), Theorem 3 is qualitatively similar to both in that the distribution of the IV estimator is approximated by a mixed Gaussian distribution. Some simplification of both Theorem 3 and subsequent results is available if we transform the coordinate space.

Corollary 2. *Let*

$$\mathbf{r} = \boldsymbol{\Theta}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}) \tag{15}$$

then we have the following approximation to $f(\mathbf{r})$, the density function of $\mathbf{r}$:

$$f(\mathbf{r}) \approx \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\pi^{n/2}\Gamma\left[\frac{\nu-n+1}{2}\right]} \left(1 + \mathbf{r}'\mathbf{r}\right)^{-(\nu+1)/2} .$$

Proof. In equation (13) make the transformation (15), for which the Jacobian of transformation is $|\boldsymbol{\Theta}|^{1/2}$. The result follows immediately. □

On the basis of Theorem 3 we can approximate the sampling distribution of a fixed linear combination of the elements of $\widehat{\boldsymbol{\beta}}$.

Theorem 4. *For any fixed vector $\boldsymbol{\alpha}$*

$$\boldsymbol{\alpha}'\widehat{\boldsymbol{\beta}} \underset{a}{\sim} t_1(\nu - n + 1, \xi, \kappa^2) ,$$

where $\xi = \boldsymbol{\alpha}'\boldsymbol{\theta}$ and $\kappa = (\boldsymbol{\alpha}'\boldsymbol{\Theta}^{-1}\boldsymbol{\alpha})^{-1/2} > 0$. That is, writing $\eta = \boldsymbol{\alpha}'\widehat{\boldsymbol{\beta}}$, we have the following approximation to the density of η:

$$f(\eta) \approx \frac{\Gamma\left[\frac{\nu-n+2}{2}\right]}{\pi^{1/2}\Gamma\left[\frac{\nu-n+1}{2}\right]} \kappa \left[1 + \kappa^2(\eta - \xi)^2\right]^{-(\nu-n+2)/2} .$$

Proof. The result follows directly from Theorem 3 using Slutsky's theorem and, for example, Gupta & Nagar (2000, Theorem 4.3.7). □

Corollary 3. *$t_\alpha = \kappa\boldsymbol{\alpha}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta})$ has (approximately) a t-distribution with $\nu - n + 1$ degrees of freedom and approximate density function*

$$f(t_\alpha) \approx \frac{\Gamma\left[\frac{\nu-n+2}{2}\right]}{\pi^{1/2}\Gamma\left[\frac{\nu-n+1}{2}\right]} \left(1 + t_\alpha^2\right)^{-(\nu-n+2)/2} .$$

Proof. Follows directly from Theorem 4 on making the simple transformation $t_\alpha = \kappa(\eta - \xi)$. □

We have used $\|\mathbf{\Delta}\|$ to control our approximation error and it is clear that the accuracy of the approximations is contingent on the closeness of $\mathbf{\Delta}$ to $\mathbf{0}$. Indeed, if $\mathbf{\Delta} = \mathbf{0}$, for whatever reason, then our approximations are exact results. In the next section we explore various circumstances where we might expect $\mathbf{\Delta}$ to be small and examine the consequences for the sampling behaviour of $\widehat{\boldsymbol{\beta}}$. Note that, by assumption, $T \geq \nu$ and so there is no conflict inherent in allowing both these quantities to diverge, as is sometimes done in what follows. Further, there need be no relationship between the rates of divergence of ν and T, beyond the fact that the former is bounded from above by the latter. Moreover, we can also allow for the possibility of T and ν being small because such events obviously do not preclude $\mathbf{\Delta}$ being small.

4 Low Concentration — Weak Instrument — Scenarios

In what follows we will classify any situation where μ^2 is small as being one involving the use of weak instruments. Since the literature lacks unanimity on the appropriate paradigm for weak instruments we should, perhaps, motivate this nomenclature, even though the debate is not germane to the contributions of this paper. Consider the multivariate OLS regression of the endogenous regressor $\mathbf{Y}$ on the instruments in $\mathbf{Z}_2$ and, following Hooper (1959), let

$$r^2 = \nu^{-1}\mathrm{tr}\{(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{P}_{\mathbf{Z}_2}\mathbf{Y}\}\,.$$

The statistic $r^2 = 0$ if $\mathbf{Y}$ and $\mathbf{Z}_2$ are orthogonal, $r^2 = 1$ if there exists a coefficient $\mathbf{B}$ such that $\mathbf{Y} = \mathbf{Z}_2\mathbf{B}$, and more generally r^2

> "can be naturally interpreted as that part of the total variance of the jointly dependent variables that is accounted for by the systematic part of the reduced form" (Hooper, *op. cit.*, p. 250.)

Now, by way of analogy we have the result that

$$\mathsf{E}[\|\mathbf{Y}(\mathbf{\Omega}_{22}^{-\frac{1}{2}})'\|^2] = \nu(\mathrm{tr}\{\mathbf{\Omega}_{22}^{-\frac{1}{2}}\mathbf{\Delta}_{22}(\mathbf{\Omega}_{22}^{-\frac{1}{2}})'\} + n) \;= \mu^2 + \nu n\,.$$

Thus μ^2 is proportional to the regression mean square in the regression of $\mathbf{Y}(\mathbf{\Omega}_{22}^{-\frac{1}{2}})'$ on $\mathbf{Z}_2(\mathbf{\Omega}_{22}^{-\frac{1}{2}})'$ and is the population counterpart of the explained sum of squares in the definition of r^2. Consequently μ^2, as with $r^2/(1-r^2)$, may be interpreted as providing a measure of the signal-to-noise ratio in the reduced form. Hence μ^2 provides a natural measure of the strength of the instruments and μ^2 being small clearly delineates situations where the instruments can be thought of as weak.

In order to further investigate different low concentration (small μ^2) - weak instrument scenarios we will assume that the original (untransformed) exogenous variables satisfy the following conditions:

1. The matrix $\mathbf{D}_Z^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{D}_Z^{-1}$ is nonsingular for all $T > K$ where $\mathbf{D}_Z^2 = \text{diag}\{\mathbf{Z}'\mathbf{Z}\}$, so $\mathbf{D}_Z$ is a diagonal matrix whose nonzero elements equal the square roots of the diagonal elements of $\mathbf{Z}'\mathbf{Z}$. Moreover, as T increases the $\liminf_{T\to\infty} \lambda_{\min}\{\mathbf{D}_Z^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{D}_Z\} > 0$ a.s.

2. Let $z_{t\,i}$ denote the ti'th element of $\mathbf{Z}$, $t = 1, \ldots, T$, $i = 1, \ldots, K$. Then

$$\lim_{T\to\infty} \sum_{s=1}^{T} z_{s\,i}^2 = \infty \quad \text{and} \quad \lim_{T\to\infty} (\sum_{s=1}^{T} z_{s\,i}^2)^{-1} z_{t\,i}^2 = 0$$

   a.s. for all $i = 1, \ldots, K$ and $t = 1, \ldots, T$.

The first condition guarantees that the exogenous regressors are linearly independent. The second condition implies that the informational content of each exogenous variable increases unboundedly as T increases whilst no single observation can exert an undue influence on the overall sum of squares. The significance of these conditions is that they ensure that any instrument weakness associated with low concentration cannot be due to the use of redundant instruments.

If we let $\mathbf{Z}_D = \mathbf{Z}\mathbf{D}_Z^{-1}$ then it is relatively straightforward to show that $\mathbf{P} = \mathbf{P}_\mathbf{Z} - \mathbf{P}_{\mathbf{Z}_1} = \mathbf{P}_{\mathbf{Z}_D} - \mathbf{P}_{\mathbf{Z}_{1D}}$ where $\mathbf{Z}_{1D} = \mathbf{Z}_1\mathbf{D}_{Z_1}^{-1}$, $\mathbf{D}_{Z_1}^2 = \text{diag}\{\mathbf{Z}_1'\mathbf{Z}_1\}$, and that, in an obvious notation,

$$\mathbf{\Delta}_{22} = \nu^{-1}\mathbf{\Pi}_2'\mathbf{D}_{Z_2}\left(\mathbf{Z}_{2D}'\mathbf{P}\mathbf{Z}_{2D}\right)\mathbf{D}_{Z_2}\mathbf{\Pi}_2\,.$$

Given that $\mathbf{Z}_{2D}'\mathbf{P}\mathbf{Z}_{2D}$ is positive definite for all T it is plain that the proximity of μ^2, and therefore $\mathbf{\Delta}$, to zero depends on the size of ν and the re-scaled

reduced form coefficients $\mathbf{D}_{Z_2}\mathbf{\Pi}_2$.

Suppose that $\mathbf{\Pi}_2 = \mathbf{0}$, the model is commonly said to be completely unidentified. Then $\mu^2 = 0$, $\mathbf{\Delta}$ is null, and the above results are exact. In this case the standardizing transformation of Corollary 2 reduces to

$$\mathbf{r} = \mathbf{\Omega}^{1/2}(\widehat{\boldsymbol{\beta}} - \mathbf{\Omega}_{22}^{-1}\boldsymbol{\omega}_{21})/\omega, \tag{16}$$

where $\omega^2 = \omega_{11} - \boldsymbol{\omega}_{12}\mathbf{\Omega}_{22}^{-1}\boldsymbol{\omega}_{21}$, which is exactly the transformation adopted in the exact finite sample literature; see Phillips (1983, Section 3.3). In the special case of $n = 1$, $f(\cdot)$ is the same density as that given by Phillips (1983, equation 3.38).[8]

Examination of equation (15) when $\mathbf{\Pi}_2 \neq \mathbf{0}$ makes it clear that the size of $\mathbf{\Delta}$ impinges upon both the scale and location of the approximating distribution. Looking first at the location of the approximation, equation (13) implies that, to the order of our approximation, the distribution of $\widehat{\boldsymbol{\beta}}$ is symmetric about $\boldsymbol{\theta}$. It can be shown that

$$\boldsymbol{\theta} = \mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} = \boldsymbol{\beta} + (\mathbf{\Omega}_{22} + \mathbf{\Delta}_{22})^{-1}(\boldsymbol{\omega}_{21} - \mathbf{\Omega}_{22}\boldsymbol{\beta}). \tag{17}$$

Noting that the covariance between the structural and reduced form disturbances can be written

$$\operatorname{cov}(\operatorname{vec}[\widetilde{\mathbf{V}}], \tilde{\mathbf{u}}) = \boldsymbol{\gamma} \otimes \mathbf{I}_T\,, \quad \text{where} \quad \boldsymbol{\gamma} = (\boldsymbol{\omega}_{21} - \mathbf{\Omega}_{22}\boldsymbol{\beta})\,,$$

we see that the correlation parameter $\boldsymbol{\rho} = \mathbf{\Omega}_{22}^{-\frac{1}{2}}\boldsymbol{\gamma}/\sigma_u$ is a measure of the extent of simultaneity in the equation of interest. It is well-known that $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$ in equation (1) if $\boldsymbol{\rho} = \mathbf{0}$; see Hillier, Kinal, & Srivastava (1984, p.190) or Mariano (1977, p.493). When $\boldsymbol{\rho} \neq \mathbf{0}$, $\boldsymbol{\theta} \neq \boldsymbol{\beta}$, and so the (approximate) distribution of $\widehat{\boldsymbol{\beta}}$ is centred at and symmetric about some point other than $\boldsymbol{\beta}$. That is, our approximation reflects the well known fact that $\widehat{\boldsymbol{\beta}}$ is asymmetrically distributed about $\boldsymbol{\beta}$.

Rewriting equation (17) as

$$\boldsymbol{\theta} = \boldsymbol{\beta} + (\mathbf{\Omega}_{22}^{-\frac{1}{2}})'(\mathbf{I}_n + \nu^{-1}\mathbf{\Gamma}_{22})^{-1}\boldsymbol{\rho}\sigma_u\,,$$

[8] As observed by Phillips (1982, Footnote 9, Section 3), $\boldsymbol{\beta}$ is not separately identified when $\mathbf{\Pi}_2 = \mathbf{0}$, hence it is unnecessary to impose the additional restriction of $\boldsymbol{\beta} = \mathbf{0}$ to obtain this distribution, as was done in Phillips (1983). Throughout we implicitly assume that $\|\boldsymbol{\beta}\| > 0$.

it can also be seen that (i) the direction of the bias in $\widehat{\boldsymbol{\beta}}$ is the same as that of $\boldsymbol{\rho}$ and that (ii) the extent to which $\boldsymbol{\Delta}_{22} \neq \mathbf{0}$ will clearly have a substantial effect on the magnitude of the bias, with bias a decreasing function of μ^2 (see Mariano, 1977, p.494).

The scale factor $\boldsymbol{\Theta} = (\sigma_{11} - \boldsymbol{\sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21})^{-1}\boldsymbol{\Sigma}_{22}$ is somewhat more difficult to interpret. Note however from (17) that $\sigma_{11} - \boldsymbol{\sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$ equals

$$\begin{aligned} \sigma_{11} - \boldsymbol{\theta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\theta} &= \sigma_{11} - \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\boldsymbol{\omega}_{21} - \boldsymbol{\Omega}_{22}\boldsymbol{\beta}) \\ &\quad -(\boldsymbol{\omega}_{21} - \boldsymbol{\Omega}_{22}\boldsymbol{\beta})'\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\omega}_{21} - \boldsymbol{\Omega}_{22}\boldsymbol{\beta}) \end{aligned} \tag{18}$$

and substituting from the relationship $\boldsymbol{\Sigma} = \boldsymbol{\Omega} + \boldsymbol{\Delta}$ into σ_{11} and $\boldsymbol{\Sigma}_{22}$ and simplifying, recognizing that $\delta_{11} = \boldsymbol{\beta}'\boldsymbol{\Delta}_{22}\boldsymbol{\beta}$, we find that the first three terms on the right hand side of (18) equal σ_u^2. Similarly, some simple if somewhat tedious algebra shows that the last term equals $-\sigma_u^2\boldsymbol{\rho}'(\mathbf{I}_n + \nu^{-1}\boldsymbol{\Gamma}_{22})^{-1}\boldsymbol{\rho}$ and hence that

$$\boldsymbol{\Theta} = \frac{\boldsymbol{\Omega}_{22}^{\frac{1}{2}}(\mathbf{I}_n + \nu^{-1}\boldsymbol{\Gamma}_{22})(\boldsymbol{\Omega}_{22}^{\frac{1}{2}})'}{\sigma_u^2(1 - \boldsymbol{\rho}'(\mathbf{I}_n + \nu^{-1}\boldsymbol{\Gamma}_{22})^{-1}\boldsymbol{\rho})} \ .$$

Thus $\boldsymbol{\Theta}$ is an increasing function of $\boldsymbol{\Delta}_{22}$ and the greater is $\boldsymbol{\Theta}$ the more concentrated will be the approximating distribution of $\widehat{\boldsymbol{\beta}}$ about $\boldsymbol{\theta}$, because the variance of the distribution in Theorem 3 is $(\nu - n - 1)^{-1}\boldsymbol{\Theta}^{-1}$. Thus we find that the bias and dispersion of $\widehat{\boldsymbol{\beta}}$ are decreasing functions of our measure μ^2.

Now suppose that $\boldsymbol{\Pi}_2 \neq \mathbf{0}$ but is local to zero as T increases, which for ν given we define to mean that $\|\mathbf{D}_{Z_2}\boldsymbol{\Pi}_2\| \to 0$ as $T \to \infty$. Then our limiting distribution leads to the conclusion that the IV estimator is inconsistent, in accord with the results of Staiger & Stock (1997). The restriction $\|\mathbf{D}_{Z_2}\boldsymbol{\Pi}_2\| \to 0$ as $T \to \infty$ implies that the reduced form coefficients $\boldsymbol{\Pi}_2$ must decline to zero faster than the growth rate in the instruments. If we also allow the number of instruments ν to grow with sample size then, noting that $(\nu - n - 1)^{-1}\boldsymbol{\Theta}^{-1} \to \mathbf{0}$ as $\nu \to \infty$, we see from equation (17) that

$$\text{plim}_{T\to\infty}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \lim_{T\to\infty}(\boldsymbol{\Omega}_{22}^{-\frac{1}{2}})'(\mathbf{I}_n + \nu^{-1}\boldsymbol{\Gamma}_{22})^{-1}\boldsymbol{\rho}\sigma_u = (\boldsymbol{\Omega}_{22}^{-\frac{1}{2}})'\boldsymbol{\rho}\sigma_u \, . \tag{19}$$

Thus the IV estimator now converges to a non-random limit. Notice that the right hand side of (19) equals the probability limit of the OLS estimation error in the totally unidentified case, *cf.* Zivot et al. (1998). The IV estimator

is therefore still inconsistent. These latter results are in line with the findings of Chao & Swanson (2002, 2003).

Consideration of the case $\nu \to \infty$ corresponds to the many-instrument asymptotics of Bekker (1994). Bekker's arguments yield Gaussian approximations to the sampling distribution of $\widehat{\boldsymbol{\beta}}$, as do the developments in Chao & Swanson (2002, 2003), and Gaussianity is of course compatible with our results since as $\nu \to \infty$ our approximating t-distribution tends to a Normal distribution. In contrast to the situation considered here, however, within the Bekker (1994) framework the reduced form model (2) is allowed to expand at the same rate as the sample size while holding the structural form model (1) fixed, so $(T-n)^{-1}\boldsymbol{\Pi}_2'\mathbf{Z}_2'\mathbf{P}\mathbf{Z}_2\boldsymbol{\Pi}_2$ is held constant whilst ν and T tend to infinity such that $\nu/T \to \alpha$ where $0 \leq \alpha < 1$. Thus, although Bekker (1994) finds that the IV estimator will be consistent whenever ν grows at a slower rate than T, we have a situation where $\mu^2 \to 0$ as $\nu \to \infty$ and the weakness of the instruments leads to the IV estimator being inconsistent.

At the other extreme, suppose that ν is small. It is well known that if $\tau \sim t_k$ then τ possesses only $k-1$ moments. In our case $k = \nu - n + 1$ and so Corollary 3 implies that, to our order of approximation, the first two moments will not exist unless the degree of over-identification of the model $\nu - n \geq 2$. Essentially this same result appears in Mariano (1982, equation 4.29). See also Kinal (1980) for more general results on the existence of moments.

5 Illustrations

We will now illustrate the behaviour of our approximation. To provide a basis for comparison we have graphed our approximation against the exact finite sample distribution and we have also plotted the standard asymptotic normal approximation. We have also selected our basic parameter values to correspond to those used by Woglom (2001) to illustrate the exact small sample properties of the IV estimator in the simplest case where $n = 1$ and the model is exactly identified. Note that in this case the concentration parameter and the concentration coefficient are identically equal.

Figure 1 indicates quite clearly that when the sample size T is small and the concentration coefficient is small relative to T then the t approximation

can be extremely close to the exact finite sample distribution. Indeed, the two distributions can be virtually indistinguishable, even when the degree of endogeneity is quite high, $\rho^2 = 0.5$, as is made plain by the first three panels of Figure 1 where the exact distribution is almost totally obscured by our approximation. Increasing the degree of endogoneity (to the extreme) by setting $\rho^2 = 0.99$, as in Figure 2, introduces bi-modality into the exact finite sample distribution and one would not expect any symmetric uni-modal approximation to capture such small sample properties well.[9] Nevertheless, as long as the signal-to-noise ratio in the first stage regression is not too large, the t approximation appears to mimic the tail behaviour of the exact finite sample distribution reasonably well.

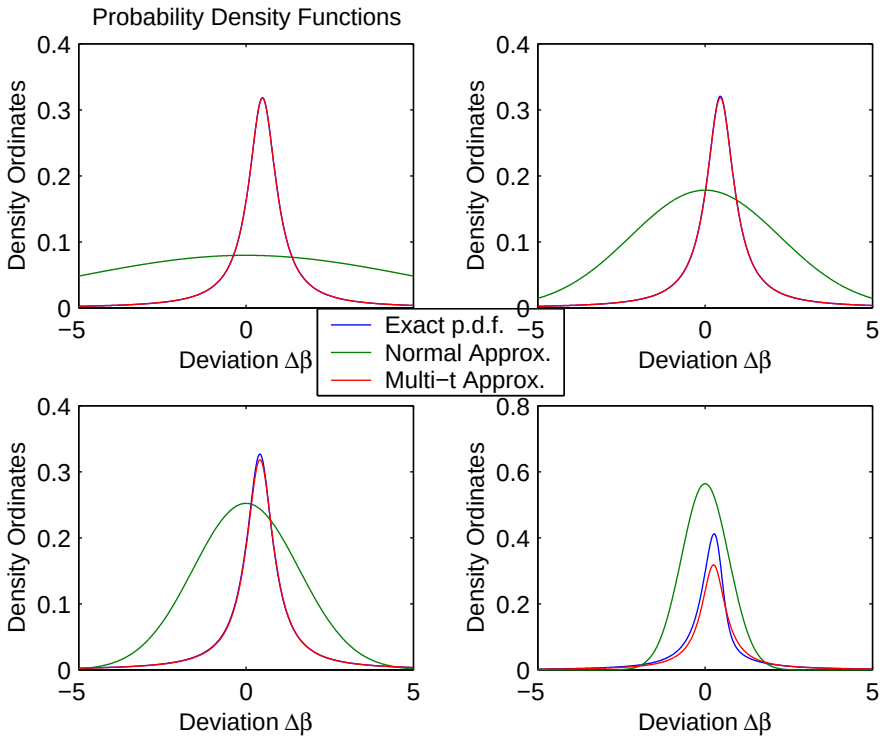

Figure 1: Density Functions of $\Delta\beta = \hat{\beta} - \beta$ with $T = 20$, $\rho^2 = 0.5$, $\sigma_u^2 = 1$ and $\mathbf{\Omega}_{22} = \sigma_V^2 = 2$. Concentration coefficient (starting in top left hand panel and proceeding linearly) $\mu^2 = 0.001T, 0.005T, 0.01T, 0.05T$

It is clear from Figures 1 and 2 that the normal approximation is not

[9]See Woglom (2001) for a detailed discussion of the differing circumstances giving rise to such finite sample properties.

working very well whatever the current circumstances. It may be felt that this is due to the sample size T being small. T only enters the distributions via its influence on the magnitude of the concentration parameter, however, and if we imagine a local to zero scenario of the type considered above then both Figures 1 and 2 will be applicable for any value of T provided that $\mu^2 = 0.02/T, 0.1/T, 0.2/T, 1.0/T$.

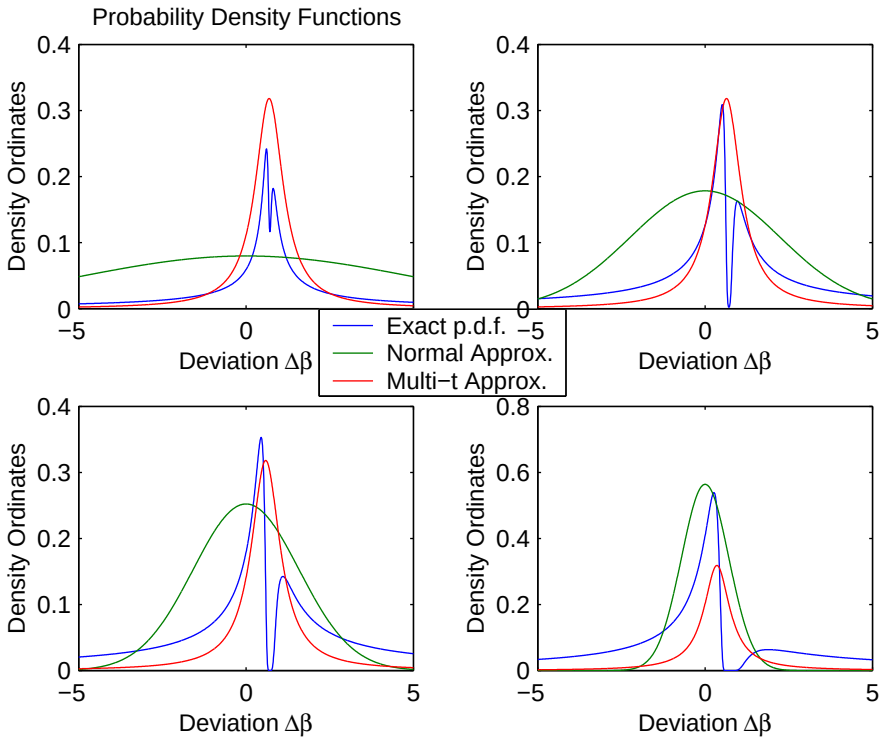

Figure 2: Density Functions of $\Delta\beta = \hat{\beta} - \beta$ with $T = 20$, $\rho^2 = 0.99$, $\sigma_u^2 = 1$ and $\mathbf{\Omega}_{22} = \sigma_V^2 = 2$. Concentration coefficient (starting in top left hand panel and proceeding linearly) $\mu^2 = 0.001T, 0.005T, 0.01T, 0.05T$

To show how increasing sample size effects the distributions we present in Figure 3 the counterpart to Figure 1, with $\mu^2 = 0.001T, 0.005T, 0.01T, 0.05T$ and $T = 250$. A reduction in asymptotic bias is readily apparent, as is the increased concentration of the distributions about their respective means. But despite the improvement in the asymptotic approximation, the t approximation to the exact density is still performing relatively well.

The preceding illustrations have all considered the exactly identified case. To indicate the impact of increasing the number of instruments Figure 4

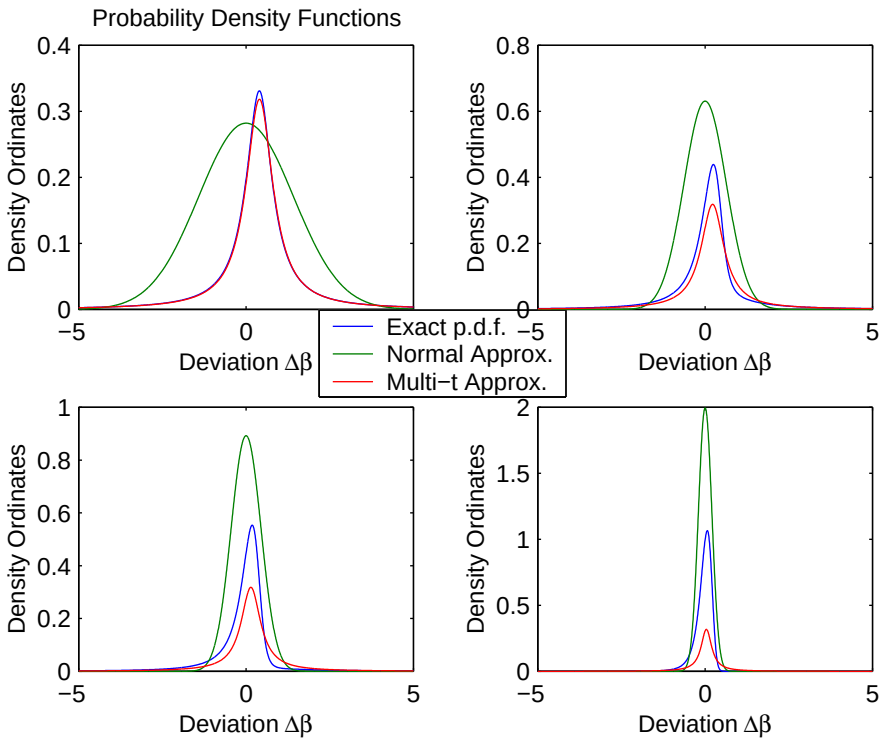

Figure 3: Density Functions of $\Delta\beta = \hat{\beta} - \beta$ with $T = 250$, $\rho^2 = 0.5$, $\sigma_u^2 = 1$ and $\mathbf{\Omega}_{22} = \sigma_V^2 = 2$. Concentration coefficient (starting in top left hand panel and proceeding linearly) $\mu^2 = 0.001T, 0.005T, 0.01T, 0.05T$

presents the three distributions in the moderately endogenous situation $\rho^2 = 0.5$ when $\mu^2 = 0.01T$, $T = 20$ and $\nu = 2, 4, 8, 16$. The relative superiority of the t approximation over the asymptotic normal approximation in these circumstances is apparent. Re-interpreting Figure 4 as representing a local-to-zero scenario, in which $\mu^2 = 0.2/T$ as $T \to \infty$, we can see that the figure clearly illustrates the inconsistency of the IV estimator discussed above.

Finally, the exact finite sample distribution of the IV estimator has been known for some time, see Phillips (1980), and in the special case of $n = 1$ can be traced back to the work of Richardson (1968). The need for an approximation is obviously therefore brought into question. When $n = 1$ the exact finite sample distribution of the IV estimator can be expressed as a convergent series in confluent hypergeometric functions and can be fairly readily evaluated, as we have done here.[10] When $n \geq 2$, however, the density involves

[10]The computations for this paper have been conducted using MATLAB.

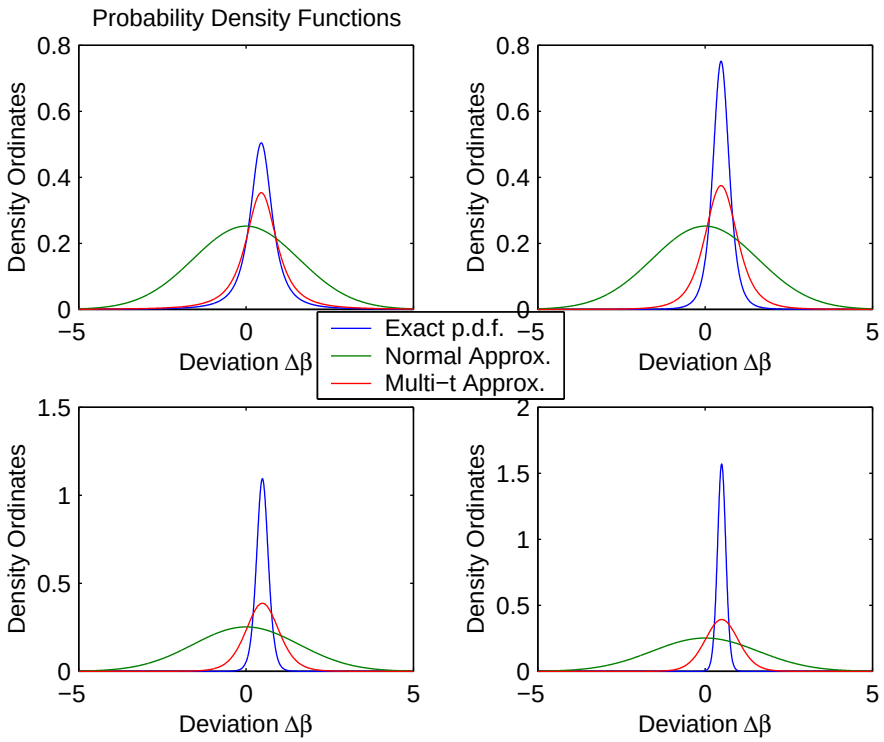

Figure 4: Density Functions of $\Delta\beta = \hat{\beta} - \beta$ with $T = 20$, $\rho^2 = 0.5$, $\sigma_u^2 = 1$, $\mathbf{\Omega}_{22} = \sigma_V^2 = 2$ and $\mu^2 = 0.01T$. Degree of over-identification (starting in top left hand panel and proceeding linearly) $\nu - n = 1, 3, 7, 15$

invariant polynomials of matrix argument and the computational burden of evaluating the exact distribution presents numerical problems which, to the best of our knowledge, are as yet unresolved. Hence the need for an approximation which (*a fortiori* - on the basis of the evidence presented thus far) we can anticipate will work well in various different situations and which can be applied using standard software.

6 Discussion

A feature of the results presented in this paper is that they are amenable to straightforward manipulation and inspection, and they provide simple demonstrations of many of the qualitative properties that have been obtained under various different paradigms used to analyze models with weak instruments. As we have seen, these range from exact finite sample theory

through to large sample and many instrument asymptotic results.

Another attraction of the results presented in this paper is their relative simplicity. This makes them easy to implement for practitioners, clearly the approximation to the distribution of the IV estimator given here is no more difficult to employ than are the Normal approximations that arise in much standard asymptotic analysis. But this raises two practical questions:

- First, how is the practitioner going to ascertain when μ^2 is small and hence when the use of the approximation developed here is appropriate?
- Second, if the use of the approximation is deemed appropriate, how are the nuisance parameters going to be estimated?

With regard to the latter, whatever the values of $\mathbf{\Pi}_2$ and $\mathbf{\Gamma}_{22}$, both $\mathbf{\Omega}$ and $\mathbf{\Delta}$ can be consistently estimated from the first stage reduced form regression. Expressed in terms of the original (untransformed) variables we have

$$\widehat{\mathbf{\Delta}} = \nu^{-1}[\mathbf{y}, \mathbf{Y}]'\mathbf{R}_{\mathbf{Z}_1}\mathbf{Z}_2(\mathbf{Z}_2'\mathbf{R}_{\mathbf{Z}_1}\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{R}_{\mathbf{Z}_1}[\mathbf{y}, \mathbf{Y}]$$

and

$$\widehat{\mathbf{\Omega}} = (T-\nu)^{-1}([\mathbf{y}, \mathbf{Y}]'(\mathbf{R}_{\mathbf{Z}_1} - \mathbf{R}_{\mathbf{Z}_1}\mathbf{Z}_2(\mathbf{Z}_2'\mathbf{R}_{\mathbf{Z}_1}\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{R}_{\mathbf{Z}_1})[\mathbf{y}, \mathbf{Y}]$$

where $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$. The estimates $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{\Delta}}$ can clearly be used to construct "plug in" values for the nuisance parameters that appear in $\boldsymbol{\theta}$ and $\mathbf{\Theta}$.

As for the first question, Poskitt & Skeels (2002) have recently developed a multivariate measure of the magnitude of the concentration parameter ideally suited to this task. Their statistic is calculated from the first stage reduced form regression and uses Wilks'-Λ distribution to construct a probabilistic calibration of $\mathbf{\Gamma}_{22}$. The statistic can be interpreted as providing a likelihood ratio test of the null-hypothesis that the endogenous regressors and the instruments are orthogonal and significantly large values of the statistic are associated with large values of the parameter $\nu\mathbf{\Delta}_{22}$. Hence the Poskitt & Skeels (2002) statistic can be used to screen out situations where the concentration coefficient μ^2 appears to be large and use of the approximation inappropriate, thereby designating situations where the approximation is likely to work well.

As a final remark, we recognize that approximating the sampling distribution of the statistic $\hat{\boldsymbol{\beta}}$ is, of itself, of secondary importance to the problem of making inferences about the parameter vector $\boldsymbol{\beta}$. The use of our t-approximation as an inferential tool will be addressed in detail in a companion paper. At this point it is, perhaps, worth pointing out that although our approximation is based on a t-distribution, it does not follow that inferential procedures based upon it will automatically suffer from the problems described by Dufour (1997). Consider, for example, constructing a confidence set for $\boldsymbol{\beta}$ using quantile points determined from the standard t-distribution and inverting (15). Since both $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ are functions of $\boldsymbol{\beta}_0$, the relationship between $\mathbf{r}$ and $\boldsymbol{\beta}_0$ is non-linear. Consequently, although $\mathbf{r}$ has a spherically symmetric distribution, the confidence sets so formed are neither likely to be, for example, elliptical, nor need they always be bounded. As such, confidence regions derived from $\mathbf{r}$ are not conventional Wald-type intervals. Indeed, application of our t-distribution, in conjunction with the estimated values of $\boldsymbol{\Omega}$ and $\boldsymbol{\Delta}$, as given above, can be viewed as an application of the conditioning principle for obtaining similar tests based on non-pivotal statistics described by Moreira (2003). Such issues remain the subject of ongoing research.

References

Anderson, T. W., Sawa, T., 1973. Distributions of estimates of coefficients of a single equation in a simultaneous system and their asymptotic expansions, Econometrica 41, 683–714.

Anderson, T. W., Sawa, T., 1979. Evaluation of the distribution function of the two-stage least squares estimate, Econometrica 47, 163–182.

Basmann, R. L., 1963. Exact finite sample distributions for some econometric estimators and test statistics: A survey and appraisal, in: Intriligator, M. D., Kendrick, D. A., eds., Frontiers of Quantitative Economics, volume II, chapter 4 (North-Holland Publishing Company).

Bekker, P. A., 1994. Alternative approximations to the distributions of instrumental variable estimators, Econometrica 58, 1443–1458.

Chao, J. C., Swanson, N. R., 2002. Consistent estimation with a large number of weak instruments, manuscript, University of Maryland.

Chao, J. C., Swanson, N. R., 2003. Asymptotic normality of single-equation estimators for the case with a large number of weak instruments, manuscript, University of Maryland.

Choi, I., Phillips, P. C. B., 1992. Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified models, Journal of Econometrics 51, 113–150.

Dickey, J. M., 1967. Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution, The Annals of Mathematical Statistics 38, 511–518.

Dufour, J.-M., 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models, Econometrica 65, 1365–1388.

Gupta, A. K., Nagar, D. K., 2000. Matrix Variate Distributions (Chapman & Hall/CRC, Boca Raton).

Hillier, G. H., Kinal, T. W., Srivastava, V. K., 1984. On the moments of ordinary least squares and instrumental variables estimators in a general structural equation, Econometrica 52, 185–202.

Hooper, J. W., 1959. Simultaneous equations and canonical correlation theory, Econometrica 27, 245–256.

Kinal, T. W., 1980. The existence of moments of k-class esimators, Econometrica 49, 241–249.

Kshirsagar, A. M., 1961. Some extensions of the multivariate t distribution and the multivariate generalization of the distribution of the regression coefficients, Proceedings of the Cambridge Philosophical Society 57, 80–85.

Mariano, R. S., 1977. Finite sample properties of instrumental variable estimators of structural coefficients, Econometrica 45, 487–496.

Mariano, R. S., 1982. Analytical small-sample distribution theory in econometrics: The simultaneous-equations case, International Economic Review 23, 503–533.

Moreira, M. J., 2003. A conditional likelihood ratio test for structural models, Econometrica 71, 1027–1048.

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

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

Phillips, P. C. B., 1980. The exact distribution of instrumental variable estimators in an equation containing n+1 endogenous variables, Econometrica 48, 861–878.

Phillips, P. C. B., 1982. Small sample distribution theory in econometric models of simultaneous equations, Cowles Foundation Discussion Paper No. 617, Yale University.

Phillips, P. C. B., 1983. Exact small sample theory in the simultaneous equations model, in: Griliches, Z., Intriligator, M. D., eds., Handbook of Econometrics, volume 1, chapter 8, 449–516 (North Holland, Amsterdam).

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

Poskitt, D. S., Skeels, C. L., 2002. Assessing instrumental variable relevance: An alternative measure and some exact finite sample theory, Paper presented at the Australasian Meeting of the Econometric Society, Brisbane.

Richardson, D. H., 1968. The exact distribution of a structural coefficient estimator, Journal of the American Statistical Association 63, 1214–1226.

Rothenberg, T. J., 1984. Approximating the distributions of econometric estimators and test statistics, in: Griliches, Z., Intriligator, M. D., eds.,

Handbook of Econometrics, volume 2, chapter 15, 881–935 (North Holland, Amsterdam).

Sargan, J. D., 1983. Identification and lack of identification, Econometrica 51, 1605–1633.

Sargan, J. D., Mikhail, W. M., 1971. A general approximation to the distribution of instrumental variables estimates, Econometrica 39, 131–169.

Searle, S. R., 1982. Matrix Algebra Useful For Statistics (John Wiley & Sons, New York).

Sims, C. A., 1980. Macroeconomics and reality, Econometrica 48, 1–43.

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

Startz, R., Nelson, C. R., Zivot, E., 2000. Improved inference for the instrumental variable estimator, paper presented at the Eighth World Congress of the Econometric Society, Seattle.

Steyn, H. S., Roux, J. J. J., 1972. Approximations for the non-central Wishart distributions, South African Statistical Journal 6, 165–173.

Stock, J. H., Wright, J., Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments, Journal of Business & Economic Statistics 20, 518–529.

Stock, J. H., Yogo, M., 2003. Asymptotic distributions of instrumental variables statistics with many weak instruments, manuscript, Harvard University.

Wang, J., Zivot, E., 1998. Inference on structural parameters in instrumental variables regression with weak instruments, Econometrica 66, 1389–1404.

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

Zivot, E., Startz, R., Nelson, C. R., 1998. Valid confidence intervals and inference in the presence of weak instruments, International Economic Review 39, 1119–1144.

Appendix

Lemma A.1. *Let* $\mathbf{A}_n$ *denote a sequence of symmetric matrices such that* $\mathbf{A}_n = \mathbf{A} + \mathbf{B}_n$*, where* $\mathbf{A} = \mathbf{A}'$*,* $0 < \mathbf{A} < \infty$*, and* $\mathbf{B}_n = O(h(n))$ *with* $h(n) \to 0$ *as* $n \to \infty$*. Then* $\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}} = O(h(n)^{\frac{1}{2}})$

Proof. Taking the trace of left and right hand sides in the expression

$$(\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}})(\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}) = \mathbf{A}_n - \mathbf{A} + (\mathbf{A}^{\frac{1}{2}} - \mathbf{A}_n^{\frac{1}{2}})\mathbf{A}^{\frac{1}{2}} + \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}} - \mathbf{A}_n^{\frac{1}{2}})$$

we find that

$$||\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}||^2 = O(h(n)) + 2\mathrm{tr}\{(\mathbf{A}^{\frac{1}{2}} - \mathbf{A}_n^{\frac{1}{2}})\mathbf{A}^{\frac{1}{2}}\}\,.$$

From the Cauchy-Schwartz inequality it now follows that

$$||\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}||^2 \leq O(h(n)) + 2||\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}|| \cdot ||\mathbf{A}^{\frac{1}{2}}||\,. \tag{A.1}$$

Let $z_n = ||\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}||/||\mathbf{B}_n||^{\frac{1}{2}}$. Then (A.1) implies that

$$z_n \leq \frac{O(h(n)/||\mathbf{B}_n||^{\frac{1}{2}})}{|(||\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}|| - 2||\mathbf{A}^{\frac{1}{2}}||)|}\,,$$

from which we can conclude that $||\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}|| \leq O(h(n)^{\frac{1}{2}})$, as required, where the final inequality follows on noting first that $||\mathbf{B}_n|| = O(h(n))$ by assumption and second that $0 < |(||\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}|| - 2||\mathbf{A}^{\frac{1}{2}}||)| < \infty$ for sufficiently large n.[11]. □

[11]This final point follows from the fact that $0 < \mathbf{A} < \infty$, coupled with a continuity argument that implies $\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}} \to 0$ as $n \to \infty$ because $\mathbf{A}_n - \mathbf{A} \to 0$.