Instrumental Variable Estimation in a Data Rich Environment

Jushan Bai* Serena Ng †

August 16, 2007

Abstract

We consider estimation of parameters in a regression model with endogenous regressors. The endogenous regressors along with a large number of other endogenous variables are driven by a small number of unobservable exogenous common factors. We show that the estimated common factors can be used as instrumental variables. These are not only valid instruments, they are more efficient than the observed variables in our framework. Consistency and asymptotic normality of the single equation factor instrumental variable estimator (FIV) is established. We also consider estimating panel data models in which all regressors are endogenous. We show that valid instruments can be constructed from the endogenous regressors which are themselves invalid instruments.

Keywords: high-dimensional factor models, efficient instruments.
JEL classification: C1, C2, C3, C4

*Department of Economics, NYU, 269 Mercer St, New York, NY 10012, Email: Jushan.Bai@nyu.edu.

†Department of Economics, University of Michigan and Columbia University, 420 W 118 St., IAB 1019, New York, NY 10027 Email: Serena.Ng@columbia.edu.

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

1 Introduction

The primary purpose of structural econometric modeling is to explain how endogenous variables evolve according to fundamental processes such as taste shocks, policy, and productivity variables. To completely characterize the behavior and the evolution of a particular endogenous variable in a data consistent manner, the economist needs to estimate the structural parameters of the model. It is well known that because these parameters are often coefficients attached to endogenous variables, endogeneity bias invalidates least squares estimation. There is a long history and continuing interest in estimation by instrumental variables, especially when the instruments are weak; see, for example, Andrews et al. (2006) and the references therein. This paper, however, does not consider the problem of weak instruments. Instead, we suggest a new way of constructing instrumental variables that can lead to more efficient estimates.

We show that if we have a large panel of instruments and that these variables and the endogenous regressors share some common factors, the factors estimated from the panel are valid and efficient instruments for the endogenous regressors. We provide the asymptotic theory for single equation estimation, and for systems of equations including panel data models. In the single equation case, we show that the estimated factors can be used as though they are the ideal but latent instruments. In the case of a large panel, we show that consistent estimates can be obtained by constructing valid instruments from variables that are themselves invalid instruments in a conventional sense. High dimensional factor analysis is a topic of much research in recent years especially in the context of forecasting; see, for example, Stock and Watson (2002) and Forni et al. (2005). Our analysis provides a new way of using the estimated factors not previously considered in either the factor analysis or the instrumental variables literature.

There are two reasons why the common factors can be valid instruments. In economic analysis, firms and households are assumed to make decisions given a set of primitive conditions. Some of these primitives are common to households and firms, while others are not. For example, an individual's consumption depends on cash-on-hand, which will likely be high when the economy is strong, but it may also vary according to the individual's status. Firms' decisions, on the other hand, are affected by the conditions of the aggregate economy, as well as specific conditions such as productivity. The (linearized) solution to a dynamic stochastic general equilibrium (DSGE) model is almost always a system of linear (expectational) stochastic difference equations in which the endogenous variables are expressed as a function of a small number of fundamental variables. It follows that the realized endogenous variables are functions of these fundamental variables, which are common across endogenous variables, plus expectational errors, which are specific to the endogenous variable in question. In these examples, the fundamental variables, if they were observed, would have been perfect instruments because they are correlated with the included endogenous regressors, but are uncorrelated with the equation-specific error. Our main premise is that even though the common fundamental variables are not observed, we can estimate them consistently.

An alternative view can also be developed by noting that the variables as defined in an economic model may not coincide exactly with how the measured data are defined. For example, non-durable consumption is often used to estimate preference parameters, but non-durable consumption ignores service flows, which the model's notion of consumption includes. As is well known, measurement error in the regressors will invalidate least squares estimation, but estimation by instrumental variables will yield consistent estimators. The question is just how to find these instruments. In this view, our proposed estimator works if there are many indicators of the variable that is observed with error.

It is well recognized that use of all potentially relevant instruments in the first stage of two-stage least squares estimation will lead to a degrees of freedom problem. This motivates Kloek and Mennes (1960) to construct a small number of principal components from the predetermined variables as instruments. Our methodology is similar in some ways, but we put more structure on the predetermined variables. Our point of departure is that if the variables in the system are driven by common sources of variations, then the ideal instruments for the endogenous variables in the system are their common components. Thus, while we have many valid instruments, each is merely a noisy indicator of the ideal instruments that we do not observe. However, we can extract the ideal instruments from the valid set. We use a factor approach to estimate the feasible instruments space from the space spanned by the observed instruments. The resulting factor-based instrumental variable estimator is denoted FIV. In the terminology of Bernanke and Boivin (2003), what we propose is a way to construct instrumental variables in a 'data rich environment'. Favero and Marcellino (2001) used estimated factors as instruments to estimate forward looking Taylor rules with the motivation that the factors contain more information than a small number of series. Here, we provide a formal analysis and show that the estimated factors are more efficient instruments than the observed variables. As far as we are aware, Kapetanios and Marcellino (2006) is the only other paper that considers using estimated factors as instruments. Their framework assumes that there are many observed weak instruments having a weak factor structure. In contrast, we assume that there are many observed instruments with an identifiable factor structure. As such, we adopt standard instead of weak instrument asymptotics. A further point of departure is that we consider high dimensional simultaneous equations system in which there exist no valid instruments in the conventional sense.

The rest of this paper is organized as follows. Section 2 presents the framework for estimation using the feasible instrument set. Section 3 studies instrumental variables estimation for panel data models without observable valid instrument variables. Simulations are given in Section 4. Our analysis is confined to cases in which the model is linear in the endogenous regressors, though we permit non-linear instrumental variable estimation when the non-linearity is induced by parameter restrictions. Non-linear instrumental variable estimation is a more involved problem even when the instruments are observed, and this issue is not dealt with in our analysis.

2 The Econometric Framework

We begin with the case of a single equation. For $t = 1, \dots T$, the endogenous variable y_t is specified as a function of a $K \times 1$ vector of regressors x_t:

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

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

$$x_{2t} = \Psi' F_t + u_t \tag{2}$$

where Ψ' is a $K_2 \times r$ matrix, F_t is a $r \times 1$ vector of fundamental variables, and $r \geq K_2$ is a small number. Endogeneity arises when $E(F_t\varepsilon_t) = 0$ but $E(u_t\varepsilon_t) \neq 0$. This induces a non-zero correlation between x_{2t} and ε_t. If F_t were observed, $\beta = (\beta_1', \beta_2')'$ could be estimated, for example, by using F_t to instrument x_{2t}. This paper assumes that the ideal instrument vector F_t is not observed.

We assume that there is a 'large' panel of data, $z_{1t}, \dots z_{Nt}$ that are weakly exogenous for β and generated as follows:

$$z_{it} = \lambda_i' F_t + e_{it}. \tag{3}$$

The $r \times 1$ vector F_t above is a set of common factors, λ_i is the factor loadings, $\lambda_i' F_t$ is referred to as the common component of z_{it}, e_{it} is an idiosyncratic error that is uncorrelated with x_{2t} and uncorrelated with ε_t. Neither e_{it} nor F_t is observed. Viewed from the factor model perspective, x_{2t} is just K_2 of the many other variables in the economic system that has a common component and an idiosyncratic component. This assumption underlies the co-movement observed for economic time series.

Although z_t, like x_{2t}, is driven by F_t, we assume e_{it} is uncorrelated with ε_t, and z_{it} is correlated with x_{2t} through F_t. Thus, z_{it} is weakly exogenous for β, and $\{z_{it}\}$ constitutes a large panel of valid instruments. While valid, z_{it} is a 'noisy' instrument for each x_{2t} because the ideal instrument for x_{2t} is F_t. When the context is clear, we will simply refer to F_t as instruments instead of 'factor-based instruments'. We cannot use F_t only because it is not observed. The idea is to use estimated F_t as instrument.

2.1 Assumptions and Estimation of F_t

We estimate the factors from a panel of instruments z_{it}, $i = 1, \ldots N, t = 1, \ldots T$, by the method of principal components. Let $z_t = (z_{1t}, z_{2t}, ..., z_{Nt})'$ be the $N \times 1$ vector of the instrumental variables, and let $Z = (z_1, z_2, ..., z_T)$, which is $N \times T$. We define $F = (F_1, ..., F_T)'$ to be the $T \times r$ factor matrix, and $\Lambda = (\lambda_1, ..., \lambda_N)'$ to be the $N \times r$ factor loading matrix. The estimated factors, denoted $\widetilde{F} = (\widetilde{F}_1, ..., \widetilde{F}_T)'$, is a $T \times r$ matrix consisting of r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the matrix $Z'Z/(TN)$ in decreasing order. Then $\widetilde{\Lambda} = (\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_N)' = Z\widetilde{F}/T$, which is $N \times r$, is an estimate for the factor loading matrix Λ. Let $\widetilde{e} = Z - \widetilde{\Lambda}\widetilde{F}'$ be the residual matrix $(N \times T)$. Also let $\widetilde{V}$ be the $r \times r$ diagonal matrix consisting of the r largest eigenvalues of $Z'Z/(TN)$. Hereafter, variables denoted with a 'tilde' are (based on) principal component estimates associated with the factor model (3), while 'hatted' variables are estimated from the regression model. The following assumption is concerned with the factor model (3).

Assumption A:

a. $E\|F_t\|^4 \le M$ and $\frac{1}{T}\sum_{t=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F > 0$, is a $r \times r$ non-random matrix.

b. λ_i is either deterministic such that $\|\lambda_i\| \le M$, or it is stochastic such that $E\|\lambda_i\|^4 \le M$. In either case, $N^{-1}\Lambda'\Lambda \xrightarrow{p} \Sigma_\Lambda > 0$, a $r \times r$ non-random matrix, as $N \to \infty$.

c.i $E(e_{it}) = 0$, $E|e_{it}|^8 \le M$.

c.ii $E(e_{it}e_{js}) = \sigma_{ij,ts}$, $|\sigma_{ij,ts}| \le \bar{\sigma}_{ij}$ for all (t,s) and $|\sigma_{ij,ts}| \le \tau_{ts}$ for all (i,j) such that $\frac{1}{N}\sum_{i,j=1}^{N} \bar{\sigma}_{ij} \le M$, $\frac{1}{T}\sum_{t,s=1}^{T} \tau_{ts} \le M$, and $\frac{1}{NT}\sum_{i,j,t,s=1} |\sigma_{ij,ts}| \le M$.

c.iii For every (t,s), $E|N^{-1/2}\sum_{i=1}^{N}\Big[e_{is}e_{it} - E(e_{is}e_{it})\Big]|^4 \le M$.

d. $\{\lambda_i\}$, $\{F_t\}$, and $\{e_{it}\}$, are three mutually independent groups. Dependence within each group is allowed.

Assumption A was used in Bai and Ng (2002) and Bai (2003) to obtain properties of $\widetilde{F}$ and $\widetilde{\Lambda}$ as estimators for $F = (F_1, ..., F_T)'$ and $\Lambda = (\lambda_1, ..., \lambda_N)'$, respectively. Assumptions (A.a) and (A.b) imply existence of r factors. The idiosyncratic errors e_{it} are allowed to be cross-sectionally and serially correlated, but only weakly as stated under condition (A.c). If e_{it} are iid, then A.c(ii) and A.c(iii) are satisfied. For Assumption (A.d), within group dependence means that F_t can be serially correlated, λ_i can be correlated over i, and e_{it} can have serial and cross-sectional correlations. All these correlations cannot be too strong so that (A.a)-(A.c) hold. However, we assume no dependence between the factor loadings and the factors, or between the factors and the idiosyncratic errors, etc, which is the meaning of mutual independence between groups.

The variable x_{1t} serves as its own instrument because it is predetermined. Let $F_t^+ = (x_{1t}', F_t')'$, the vector of ideal instruments with dimension $K_1 + r$. Let β^0 denote the true value of β. Introduce $\varepsilon_t(\beta) = y_t - x_t'\beta$ and thus $\varepsilon_t = \varepsilon_t(\beta^0)$.

Assumption B

a. $E(\varepsilon_t) = 0$, $E|\varepsilon_t|^{4+\delta} < \infty$ for some $\delta > 0$. The vector process $g_t(\beta^0) = F_t^+\varepsilon_t$ satisfies $E[g_t(\beta^0)] = 0$ with $E(g_t(\beta)) \ne 0$ when $\beta \ne \beta^0$. Let $\bar{g}^0 = \frac{1}{T}\sum_{t=1}^{T} F_t^+\varepsilon_t$, and $\sqrt{T}\bar{g}^0 = T^{-1/2}\sum_{t=1}^{T} F_t^+\varepsilon_t \xrightarrow{d} N(0, S^0)$ for some $S^0 > 0$.

b. $x_{2t} = \Psi' F_t + u_t$ with $\Psi'\Psi > 0$, $E(F_t u_t) = 0$, and $E(u_t\varepsilon_t) \ne 0$.

c. For all i and t, $E(e_{it}u_t) = 0$, and $E(e_{it}\varepsilon_t) = 0$.

Part (a) states that the model is correctly specified and a set of orthogonality conditions hold at β^0. In general, S^0 is the limit of $T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T} E[F_t^+ F_s^{+\prime}\varepsilon_t\varepsilon_s]$. However, to focus on the main idea, we shall assume throughout $F_t^+\varepsilon_t$ to be serially uncorrelated so that S^0 is the probability limit of $T^{-1}\sum_{t=1}^{T} F_t^+ F_t^{\prime +}\varepsilon_t^2$. Heteroskedasticity of ε_t is allowed and will be reflected in the asymptotic variance, S^0. Validity of F_t as an instrument requires that F_{jt} has a non-zero loading on x_{2t} for each $j = 1, \dots r$. Thus, F_t is the ideal but infeasible instrument for x_{2t}. The requirement that $\Psi'\Psi > 0$ in part (b) is thus important for our analysis. Part (c) assumes that the correlation between the instruments and the endogenous regressor come through F_t and not e_{it}. It further implies that all the instruments are valid. This assumption is stronger than is necessary and can be relaxed, see Remark 1 below.

In certain cases, lags of F_t can also serve as instruments, though in general, lags of F_t should provide no further information about x_{2t} once conditioned on F_t. When u_t is serially uncorrelated and all the dynamics in x_{2t} are due to F_t, then lags of F_t are better instruments than lags of x_{2t}. Lags of x_{2t} can be better instruments if F_t does not contribute to the dynamics in x_{2t}.

In order to use past values of the observed variables as instruments,, we also need

Condition C: (a) $E(x_{2t}x'_{2t-j}) \neq 0$ for some $j > 1$. and (b) $E(\varepsilon_t|I_{t-1}) = 0$ where $I_{t-1} = \{x_{1t-j}, x_{2t-j}, y_{t-j}\}_{j=1}^{t-1}$.

Essentially, x_{2t} must be serially correlated and ε_t must be uncorrelated with the past observations. If lags of x_{2t} are valid instruments, they are in general better instruments than lags of y_t because the latter are correlated with x_{2t} through the correlation between x_{2t} and its past values.[1]

2.2 A Feasible Factor IV Estimator

The conventional treatment of endogeneity bias is to use lags of y_t, x_{1t} and x_{2t} as instruments for x_{2t} and invoke Condition C. Our point of departure is to note that g_t contains all the information about β. The reason why the moments g_t are not used to estimate β is that F_t is not observed. We suggest to use $\widetilde{F}_t$ in place of F_t. To fix ideas and for notational simplicity, we assume the absence of regressor x_{1t} ($K_1 = 0$) so that the instrument is $\widetilde{F}_t$. It is understood that when x_{1t} is present, the results still go through upon replacing $\widetilde{F}_t$ in the estimator below by $\widetilde{F}_t^+ = (x'_{1t}, \widetilde{F}'_t)'$.

Define $\widetilde{g}_t(\beta) = \widetilde{F}_t\varepsilon_t(\beta)$. Consider estimating β using the r moment conditions $\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^T \widetilde{F}_t\varepsilon_t(\beta)$. Let W_T be a $r \times r$ positive definite weighting matrix. Where appropriate, the dependence of $\bar{g}$ on β will be suppressed. The linear GMM estimator is defined as

$$\begin{aligned}\breve{\beta}_{FIV} &= \underset{\beta}{\text{argmin}}\ \bar{g}(\beta)'W_T\bar{g}(\beta)\\ &= (S'_{\widetilde{F}x}W_TS_{\widetilde{F}x})^{-1}S'_{\widetilde{F}x}W_TS_{\widetilde{F}y}\end{aligned}$$

[1] When x_{1t} is strongly exogenous such that $E(x_{1t}\varepsilon_s) = 0$ for all t and s, ε_t itself is allowed to be serially correlated of unknown form (this situation of course rules out x_{1t} being the lag of y_t). When ε_t is serially correlated, the lags of x_{2t} cannot be used as instruments since x_{2t-j} can be correlated with ε_{t-j}, which is correlated with ε_t.

where $S_{\widetilde{F}x} = \frac{1}{T}\sum_{t=1}^{T}\widetilde{F}_t x_t'$. Let $\breve{\varepsilon}_t = y_t - x_t'\breve{\beta}_{FIV}$ and let $\breve{S} = \frac{1}{T}\sum_{t=1}^{T}\widetilde{F}_t\widetilde{F}_t'\breve{\varepsilon}_t^2$. Then the efficient GMM estimator, which is our main focus, is to let $W_T = \breve{S}^{-1}$, giving

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

Theorem 1 *Under Assumptions A and B, as $N, T \to \infty$,*

$$\sqrt{T}(\widehat{\beta}_{FIV} - \beta^0)\xrightarrow{d}N(0,\Omega)$$

where $\Omega = \text{plim}(S_{\widetilde{F}x}'(\breve{S})^{-1}S_{\widetilde{F}x})^{-1}$. *Furthermore,* $J = T\bar{g}(\widehat{\beta}_{FIV})'\,\breve{S}^{-1}\,\bar{g}(\widehat{\beta}_{FIV})\xrightarrow{d}\chi^2_{r-K}$.

Theorem 1 establishes consistency and asymptotic normality of the GMM estimator when $\widetilde{F}_t$ are used as instruments, and when the observed instruments are not weak.[2] Just as if F_t was observed, $\widehat{\beta}_{FIV}$ reduces to $(\widetilde{F}'x)^{-1}\widetilde{F}'y$ and is the instrumental variable estimator in an exactly identified model with $K = r$. It is the two-stage least squares (2SLS) estimator, i.e., $\widehat{\beta}_{FIV} = (x'P_{\widetilde{F}}x)^{-1}x'P_{\widetilde{F}}y$, under conditional homoskedasticity. Furthermore, T times the value of the objective function is asymptotically χ^2 distributed with $r - K$ degrees of freedom. Essentially, if both N and T are large, estimation and inference can proceed as though F_t was observed. Other estimators such as obtained by minimizing mean-squared error considered in Carrasco (2006) and Hausman et al. (2006), as well as LIML and JIVE, can also be derived. Since $\widetilde{F}_t$ can be used as though it was F_t, we expect a factor based version of these estimators will remain valid, but analyzing their properties is beyond the scope of the present analysis.

The essence behind Theorem 1 is that $\widetilde{F}_t$ is estimating a rotation of F_t, denoted by HF_t, where H is an $r \times r$ invertible matrix. If F_t is a vector of valid instruments, then HF_t is also a vector of valid instruments and will give rise to an identical estimator. To show $\widetilde{F}_t$ will lead to the same estimator (asymptotically only), we need to establish

$$T^{-1/2}\sum_{t=1}^{T}(\widetilde{F}_t - HF_t)\varepsilon_t = o_p(1). \tag{4}$$

This result is given in Lemma 1 in the appendix. In fact, it can be shown that $\widetilde{F}_t - HF_t$ is equal to $D\frac{1}{N}\sum_{i=1}^{N}\lambda_i e_{it}$ plus a term that is negligible, where matrix D depends on N and T and is $O_p(1)$. Thus $T^{-1/2}\sum_{t=1}^{T}(\widetilde{F}_t - HF_t)\varepsilon_t \simeq DN^{-1/2}\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda_i e_{it}\varepsilon_t$. Thus if ε_t and e_{it} are independent, then the left hand side of (4) is $O_p(N^{-1/2}) = o_p(1)$.

[2] Irrelevant instruments are allowed in the sense that some factor loadings λ_i can be zero. All needed is $\frac{1}{N}\sum_{i=1}^{N}\lambda_i\lambda_i'\xrightarrow{p}\Sigma_\Lambda > 0$, as in Assumption A(b). The analysis should also go through when the instruments are not too weak in the sense of Hahn and Kuersteiner (2002). A weak factor model in which all factor loadings λ_i is of $O(N^{-\alpha})$ $(\alpha > 0)$ necessitates a different asymptotic framework and is considered in Kapetanios and Marcellino (2006).

Remark 1: Theorem 1 is derived under the assumption $E(\varepsilon_t e_{it}) = 0$ for all i and t so that all instruments are valid. The assumption is, however, not necessary under a data rich environment. Suppose that $E(\varepsilon_t e_{it}) \neq 0$ for all i so that none of the instruments are valid. When N is fixed, the instrument variable estimator based on z_t will not be consistent. But with a large N and under the assumption that $\sum_{i=1}^N |E(\varepsilon_t e_{it})| \leq M < \infty$ for all N with M not depending on N, Theorem 1 still holds provided that $\sqrt{T}/N \to 0$. Suppose $\gamma_i = E(e_{it}\varepsilon_t) \neq 0$, then

$$T^{-1/2}N^{-1}\sum_{t=1}^T\sum_{i=1}^N \lambda_i e_{it}\varepsilon_t = N^{-1/2}\frac{1}{\sqrt{TN}}\sum_{t=1}^T\sum_{i=1}^N \lambda_i[e_{it}\varepsilon_t - E(e_{it}\varepsilon_t)] + \sqrt{T}N^{-1}\sum_{i=1}^N \lambda_i\gamma_i$$

The first term on the right hand side is $N^{-1/2}O_p(1) = o_p(1)$. For the second term, since $E\|\lambda_i\| \leq M$ by assumption, the absolute value of the second term is bounded in expectation by $M\sqrt{T}N^{-1}\sum_{i=1}^N |\gamma_i|$. Thus if $\sum_{i=1}^N |\gamma_i|$ is bounded and $\sqrt{T}/N \to 0$, then the second term is also $o_p(1)$, implying that (4) still holds. In fact, $\sum_{i=1}^N |\gamma_i|$ is allowed to go to infinity, all that is needed is the product $(\sqrt{T}/N)\sum_{i=1}^N |\gamma_i| \to 0$. This would be impossible when N is fixed as long as there exists an i such that $\gamma_i \neq 0$. This result highlights the benefit of working in a data rich environment.

Remark 2: The assumption that $N \to \infty$ ensures consistent estimation of the factor space and is a key feature of the data rich environment. But even with N fixed, we can always mechanically construct $\widetilde{F}_t$ as the principal components of z_t. Under the assumption that all the instruments are valid, the resulting FIV estimator is still consistent because linear combinations of valid instruments remain valid instruments. However, when invalid instruments satisfying the condition of Remark 1 is permitted, consistent estimation will not be possible unless N is large.

In early work, Kloek and Mennes (1960) were concerned with situations when N is large relative to the given T (in their case, 30) so that the first stage estimation is inefficient. These authors motivated principal components as a practical dimension reduction device. Amemiya (1966) and more recently Carrasco (2006) provided different statistical justifications for the approach without reference to a factor structure. In contrast, we motivated principal components as a method that consistently estimates the space spanned by the ideal instruments with the goal of developing a theory for inference. Our asymptotic theory necessitates a factor structure on z_t and it is because of this structure that leads to the following:

Proposition 1 *Let z_{2t} be a subset of the observed instruments such that z_{2t} is $r \times 1$. Let $m_t = z_{2t}(y_t - x_t'\beta)$ with $\sqrt{T}\bar{m} \xrightarrow{d} N(0, Q)$. Let $\widehat{\beta}_{IV}$ be the minimizer of $\bar{m}'(\breve{Q})^{-1}\bar{m}$ with the property that $\sqrt{T}(\widehat{\beta}_{IV} - \beta^0) \xrightarrow{d} N(0, \Omega_{IV})$. If $var(e_{it}) \geq c > 0$ for all i in the z_{2t} set, then as $N, T \to \infty$,*

$$\Omega_{IV} - \Omega_{FIV} > 0$$

where Ω_{FIV} denotes the asymptotic variance of $\widehat{\beta}_{FIV}$, i.e., Ω in Theorem 1.

Proposition 1 says that when each observed instrument is measured with error, then in a data rich environment, $\widehat{\beta}_{FIV}$ is more efficient than $\widehat{\beta}_{IV}$, which uses an equal number of z_{2t} as instruments. The intuition is straightforward. The observed instruments are the ideal instruments contaminated with errors while $\widetilde{F}$ is consistent for the ideal instrument space. Pooling information across the observed variables washes out the noise to generate more efficient instruments for x_{2t}. Proposition 1 rules out cases when the observed variables are perfect instruments. This may seem restrictive, but is not unrealistic as researchers cannot be expected to isolate the perfect instruments, even if they exist. Of course, when the model assumptions do not hold, such as if the factor structure is weak (e.g., factor loadings $\lambda_i = O_p(N^{-1/2}) \to 0$ as N increases), Proposition 1 will not necessarily hold.

The single equation set up extends naturally to a system of equations. Suppose there are G equations, where G is finite. For $g = 1, \ldots G$, and $t = 1, \ldots T$,

$$y_{gt} = x_{gt}'\beta_g + \varepsilon_{gt}$$

where x_{gt} is $K_g \times 1$. As an example of $G = 2$, (y_1, y_2) could be aggregate consumption and earnings, while the endogenous regressor is hours worked. Let $\widetilde{F}_{gt}$ be the $r_g \times 1$ vector of instruments for the g-th equation, $g = 1, \ldots G$, and let $r = \sum_g r_g$. Then g_t is a $r \times 1$ vector of stacked up moment conditions. Assuming that for each $g = 1, \ldots G$, the $r_g \times K_g$ moment matrix $E(\widetilde{F}_{gt}x_{gt}')$ is of full column rank, Theorem 1 still holds, but the $r \times r$ matrix S is now the asymptotic variance of the stacked up moment conditions. Note that this need not be a block diagonal matrix. Likewise, $S_{\widetilde{F}x}$ is a $K \times r$ matrix. If each equation has a regressor matrix of the same size and uses the same number of instruments, the $S_{\widetilde{F}x}$ matrix under systems estimation will be G times bigger, just as when F_t is observed. See, for example, Hayashi (2000).

2.3 A Control Function Interpretation

We have motivated the FIV as a method of constructing more efficient instruments, but the estimator can also be motivated in a different way. Under the assumed data generating process, ie $x_{2t} = \Psi' F_t + u_t$, the non-zero correlation between x_{2t} and ε_t arises because $\text{cov}(u_t, \varepsilon_t) \neq 0$. We can decompose ε_t into a component that is correlated with u_t, and a component that is not. Let

$$\varepsilon_t = u_t'\gamma + \varepsilon_{t|u}$$

where $\varepsilon_{t|u}$ is orthogonal to u_t and thus x_{2t}. We can rewrite the regression $y_t = x_{1t}\beta_1 + x_{2t}\beta_2 + \varepsilon_t$ as

$$y_t = x_t'\beta + u_t'\gamma + \varepsilon_{t|u}$$

If F_t was observed, we would estimate the reduced form for x_{2t} to yield fitted residuals $\widehat{u}_t$. Then least squares estimation of

$$y_t = x_t'\beta + \widehat{u}_t'\gamma + error$$

not only provides a test for endogeneity bias, it also provides estimates of β that are numerically identical to two stage least squares with F_t as instruments. This way of using the fitted residuals to control endogeneity bias is sometimes referred to as a 'control function' approach as in Hausman (1978).

In our setting, we cannot estimate the reduced form for x_{2t} because F_t is not observed. Indeed, if we only observe x_{2t}, and $x_{2t} = \Psi' F_t + u_t$, there is no hope of identifying the two components in x_{2t}. However, we have a panel of data Z with a factor structure, and $\widetilde{F}_t$ are consistent estimates of F_t up to a linear transformation. The control function approach remains feasible in our data rich environment and consists of three steps. In step one, we obtain $\widetilde{F}_t$. In step 2, for each $i = 1, \ldots K_2$, least squares estimation of

$$x_{2it} = \widetilde{F}_t'\Psi_i + u_{it}$$

will yield $\sqrt{T}$ consistent estimates of Ψ_i, from which we obtain $\widehat{u}_t$. Least squares estimation of

$$y_t = x_{1t}'\beta_1 + x_{2t}'\beta_2 + \widetilde{u}_t'\gamma + \varepsilon_t^u \tag{5}$$

will yield $\sqrt{T}$ consistent estimates of β. It is straightforward to show that the estimate is again numerically identical to 2SLS with $\widetilde{F}_t$ as instruments. In this regard, the FIV is a control function estimator. But the 2SLS is a special case of the FIV that is efficient only under conditional homoskedasticity. Thus, the FIV can be viewed as an efficient alternative to controlling endogeneity when conditional homoskedasticity does not hold or may not be appropriate. The control function approach also highlights the difference between the FIV and the IV. With the IV, u_t is estimated from regressing x_{2t} on z_{2t}, where z_{2t} are noisy indicators of F_t. With the FIV, u_t is estimated from regressing x_{2t} on a consistent estimate of F_t and is thus more efficient than the IV.

3 Panel Data and Large Simultaneous Equations System

Consider a large panel data regression model and assume for simplicity that there are no predetermined variables. For $i = 1, 2, ..., N, t = 1, 2, ..., T$ with N and T both large, let

$$y_{it} = x_{it}'\beta + \varepsilon_{it}$$

where x_{it} is $K \times 1$. This is a large simultaneous equation system since we allow

$$E(x_{it}\varepsilon_{it}) \neq 0$$

for all i and t. Therefore, the pooled OLS estimator

$$\widehat{\beta}_{POLS} = (\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x_{it}')^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}y_{it}$$

is inconsistent. Unlike the single equation system, we do not need the existence of valid instruments z_{it}. When N is large, x_{it} can play the role of z_{it} despite the fact that none of x_{it} is a valid instrument in the conventional sense. As in the single equation case, the regressors are influenced by the common factors,

$$x_{it} = \Lambda_i' F_t + u_{it} = C_{it} + u_{it}$$

where Λ_i is a matrix of $r \times K$, F_t is $r \times 1$ with $r \geq K$. We assume ε_{it} is correlated with u_{it} but not with F_t so that $E(F_t\varepsilon_{it}) = 0$. The loading Λ_i can be treated as a constant or random; when it is regarded as random, we assume ε_{it} is independent of it. Therefore we have

$$E(C_{it}\varepsilon_{it}) = 0.$$

In this panel data setting, the common component $C_{it} = \Lambda_i' F_t$ is the ideal instrument for x_{it}. As we will see later, it is a more effective instrument than F_t in terms of convergence rate and the mean squared errors of the estimator. Again, C_{it} is not available, but it can be estimated.

Let $X_i = (x_{i1}, x_{it}, ..., x_{iT})'$ be a $T \times K$ matrix of regressors for the ith cross-section unit, so that $X = (X_1, X_2, ..., X_N)$ is $T \times (NK)$. Let Λ be a $(NK) \times r$ matrix while F is $T \times r$. Let

$\widetilde{F}$ be the principal component estimate of F from the matrix XX', as explained in Section 2.1 with Z replaced by X. Let $\widetilde{C}_{it} = \widetilde{\Lambda}_i'\widetilde{F}_t$, which is $K \times 1$.

Consider the pooled two-stage least-squares estimator with $\widetilde{C}_{it}$ as instruments

$$\widehat{\beta}_{PFIV} = \Big(\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{C}_{it}x_{it}'\Big)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{C}_{it}y_{it}. \tag{6}$$

To study the properties of this estimator, we need the following assumptions:

Assumption A': Same as Assumption A (a-d) with three changes. Part (b) holds with λ_i replaced by Λ_i; part (c) holds with e_{it} replaced by each component of u_{it} (note that u_{it} is a vector). In addition, we assume u_{it} are independent over i.

Assumption B':

a. $E(\varepsilon_{it}) = 0$, $E|\varepsilon_{it}|^{4+\delta} < M < \infty$ for all i, t, for some $\delta > 0$; ε_{it} are independent over i.

b. $x_{it} = \Lambda_i'F_t + u_{it}$; $E(u_{it}\varepsilon_{it}) \neq 0$; ε_{it} is independent of F_t and Λ_i.

c. $(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}\varepsilon_{it} \xrightarrow{d} N(0, S)$, where S is the long-run covariance of the sequence $\xi_t = N^{-1/2}\sum_{i=1}^{N} C_{it}\varepsilon_{it}$, defined as

$$S = \lim_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} E(C_{it}C_{is}'\varepsilon_{it}\varepsilon_{is}).$$

Theorem 2 *Suppose Assumptions A' and B' hold. As $N, T \to \infty$, we have*

(i) $\widehat{\beta}_{PFIV} - \beta^0 = O_p(T^{-1}) + O_p(N^{-1})$ *and thus* $\widehat{\beta}_{PFIV} \xrightarrow{p} \beta^0$.

(ii) If $T/N \to \tau > 0$, *then*

$$\sqrt{NT}(\widehat{\beta}_{PFIV} - \beta^0) \xrightarrow{d} N(\tau^{1/2}\Delta_1^0 + \tau^{-1/2}\Delta_2^0,\ \Omega)$$

where $\Omega = \text{plim}[S_{\widetilde{x}\widetilde{x}}]^{-1}S[S_{\widetilde{x}\widetilde{x}}]^{-1}$ *with* $S_{\widetilde{x}\widetilde{x}} = (NT)^{-1}\sum_{i=1}^{N}\widetilde{C}_{it}x_{it}'$, *and* Δ_1^0 *and* Δ_2^0 *are defined in the appendix.*

Theorem 2 establishes that the estimator $\widehat{\beta}_{PFIV}$ is consistent for β as $N, T \to \infty$. Remarkably, there can be no instrument in the conventional sense, yet, we can still consistently estimate the large simultaneous equations system.[3] In a very rich data environment, the information in the data collectively permits consistent instrumental variable estimation under much weaker conditions on the individual instruments. Because the bias is of order $\max[N^{-1}, T^{-1}]$, the effect of the bias on $\widehat{\beta}_{PFIV}$ can be expected to vanish quickly.

If C_{it} is known, asymptotic normality simply follows from Assumption B′(c) and there will be no bias. However, C_{it} is not observed, and biases arise from the estimation of C_{it}. More precisely, $\widetilde{C}_{it}$ contains u_{it} which is correlated with ε_{it}, and is the underlying reason for biases. When T and N are of comparable magnitudes, $\widehat{\beta}_{PFIV}$ is $\sqrt{NT}$ consistent and asymptotically normal, but the limiting distribution is not centered at zero, as shown in part (ii) of Theorem 2.

A biased-corrected estimator can be considered to recenter the asymptotic distribution to zero for small N and T. For this purpose, we assume that ε_{it} are serially uncorrelated.[4] Let

$$\widehat{\delta}_1 = \left(\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{K} \widetilde{\Lambda}_i' \widetilde{V}^{-1} \widetilde{\lambda}_{i,k} \widetilde{u}_{it,k} \widehat{\varepsilon}_{it} \right), \quad \text{and} \quad \widehat{\Delta}_1 = (S_{\widetilde{x}\widetilde{x}})^{-1} \widehat{\delta}_1$$

$$\widehat{\delta}_2 = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{u}_{it} \widetilde{F}_t' \widetilde{F}_t \widehat{\varepsilon}_{it} \right), \quad \text{and} \quad \widehat{\Delta}_2 = (S_{\widetilde{x}\widetilde{x}})^{-1} \widehat{\delta}_2,$$

---

[3] This estimator can be easily extended to include additional regressors that are uncorrelated with ε_{it}. For example, $y_{it} = x_{1it}'\beta_1 + x_{2it}'\beta_2 + \varepsilon_{it}$ with x_{1it} being exogenous. We estimate $\widetilde{F}$ and $\widetilde{\Lambda}$ from x_2 alone. Then the pooled 2SLS is simply

$$\widehat{\beta}_{PFIV} = \Big(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{Z}_{it} x_{it}' \Big)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{Z}_{it} y_{it}$$

where $\widetilde{Z}_{it} = (x_{1it}', \widetilde{C}_{it}')'$. It is noted that equation (6) can be written alternatively as

$$\widehat{\beta}_{PFIV} = \Big(\sum_{i=1}^{N} X_i' P_{\widetilde{F}} X_i \Big)^{-1} \sum_{i=1}^{N} X_i' P_{\widetilde{F}} Y_i$$

where $Y_i = (y_{i1}, y_{i2}, ..., y_{iT})'$ is $(T \times 1)$. This follows from the fact that $(\widetilde{C}_{i1}, \widetilde{C}_{i2}, ..., \widetilde{C}_{iT})' = P_{\widetilde{F}} X_i = \widetilde{F}\widetilde{\Lambda}_i$. However, this representation is not easily amendable in the presence of additional regressors x_{1it}.

[4] It is possible to construct biased-corrected estimators when ε_{it} is serially correlated. The bias correction involves estimating a long-run covariance matrix, denoted by Υ. The estimated long-run covariance $\widehat{\Upsilon}$ must have a convergence rate satisfying $\sqrt{N/T}(\widehat{\Upsilon} - \Upsilon) = o_p(1)$. Assuming $T^{1/4}(\widehat{\Upsilon} - \Upsilon) = o_p(1)$, this implies the requirement that $N/T^{3/2} \to 0$ instead of $N/T^2 \to 0$ under no serial correlation.

where $\widetilde{u}_{it} = x_{it} - \widetilde{C}_{it}$, $\widehat{\varepsilon}_{it} = y_{it} - x'_{it}\widehat{\beta}_{PFIV}$, and $S_{\widetilde{x}\widetilde{x}} = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{C}_{it}x'_{it}$. The estimated bias is[5]

$$\widehat{\Delta} = \frac{1}{N}\widehat{\Delta}_1 + \frac{1}{T}\widehat{\Delta}_2.$$

Corollary 1 *Suppose Assumptions A′ and B′ hold. If ε_{it} are serially uncorrelated, $T/N^2 \to 0$, and $N/T^2 \to 0$, then*

$$\sqrt{NT}(\widehat{\beta}_{PFIV} - \widehat{\Delta} - \beta^0) \xrightarrow{d} N(0, \Omega).$$

Both $\widehat{\beta}_{PFIV}$ and its bias-corrected variant are $\sqrt{NT}$ consistent. One can expect the estimators to be more precise than the single equation estimates because of the fast rate of convergence. However, while $\widehat{\beta}_{PFIV}$ is expected to be sufficiently precise in terms of the mean squared errors, the bias corrected estimator, $\widehat{\beta}^{+}_{PFIV} = \widehat{\beta}_{PFIV} - \widehat{\Delta}$ should provide more accurate inference in terms of the t statistic because it is properly re-centered around zero.

It is worth noting that the PFIV estimator is different from the traditional panel IV estimator that uses $\widetilde{F}$ as instruments. Such an estimator, PTFIV, would be constructed as

$$\widehat{\beta}_{PTFIV} = \left(S'_{\widetilde{F}x}\breve{S}^{-1}S_{\widetilde{F}x}\right)^{-1} S'_{\widetilde{F}x}\breve{S}^{-1}S_{\widetilde{F}y}$$

where $S_{\widetilde{F}x} = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{F}_t x'_{it}$, and $\breve{S} = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{F}_t \widetilde{F}'_t \breve{e}^2_{it}$, $\breve{e}_{it}$ is based on a preliminary estimate of β using a $r \times r$ positive definite weighting matrix. However, the probability limit of $S_{\widetilde{F}x}$ is $\Sigma_{Fx} = E(\lambda_i)'\Sigma_F$, which can be singular if $E(\lambda_i) = 0$, and in that case the estimator is only $\sqrt{T}$ consistent. The $\widehat{\beta}_{PTFIV}$ is $\sqrt{NT}$ consistent only if one assumes a full column rank for Σ_{Fx}. In contrast, the proposed estimator uses the moment $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T x_{it}c'_{it} = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T c_{it}c'_{it} + o_p(1) > 0$ and is always $\sqrt{NT}$ consistent, without the extra rank condition.

4 Simulations

In this section, we evaluate the effectiveness of the FIV using $\widetilde{F}^+ = [x_1 \;\; \widetilde{F}]$ as instruments, where $\widetilde{F}$ is $T \times r$.[6] We also consider an estimator with $\widetilde{f}^+ = [x_1 \;\; \widetilde{f}]$ as instruments, where

[5] In the presence of exogenous regressors x_{1it} as in footnote 3, the corresponding terms become

$$\widehat{\Delta}_1 = \Big(\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widetilde{Z}_{it}x'_{it}\Big)^{-1}\begin{bmatrix} 0 \\ \widehat{\delta}_1 \end{bmatrix}, \text{ and } \widehat{\Delta}_2 = \Big(\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widetilde{Z}_{it}x'_{it}\Big)^{-1}\begin{bmatrix} 0 \\ \widehat{\delta}_2 \end{bmatrix}.$$

A small sample adjustment can also be made by using $NT - (N+T)r$ instead of NT when computing $\widehat{\delta}_1$ and δ_2, where $r(N+T)$ is the number of parameters used to estimate $\widehat{u}_{it}$.

[6] In practice, the IC_2 criterion in Bai and Ng (2002) or the criterion of Hallin and Liska (2007) can be used to determine r. Since the estimated r is consistent for r, it can be shown that r can be treated as

the dimension of $\widetilde{f}$ is $T \times rmax$ with $rmax > r$. This estimator is denoted fIV. The GMM estimator uses an identity weighting matrix in the first step to yield $\breve{\beta}$. For the sake of comparison, we also report results of two other estimators. The first is a GMM using a set of observed variables most correlated with x_2 and is the same dimension as $\widetilde{F}$. These instruments are determined by the R^2 from regressions of x_2 on both x_1 and an instrument one at a time. This estimator is labeled IV. The second is OLS, which does not account for endogeneity bias.

We consider three data generating processes. In all cases,

$$\begin{aligned} z_{it} &= \lambda_{iz} F_t + \sqrt{r}\sigma_z e_{it} \\ F_{jt} &= \rho_j F_{jt-1} + \eta_{jt} \quad j = 1, \ldots r \end{aligned}$$

where $e_{it} \sim N(0,1)$, $\eta_{jt} \sim N(0,1)$, $\lambda_{iz} \sim N(0,1)$, $\rho_j \sim U(.2, .8)$, and $\sigma_z = 3$ for all i. The examples differ in how y_t, x_{1t}, and x_{2t} are generated.

Example 1 We modify the DGP of Moreira (2003). The equation of interest is

$$\begin{aligned} y_t &= x_{1t}'\beta_1 + x_{2t}'\beta_2 + \sigma_y \varepsilon_t \\ x_{i1t} &= \alpha_x x_{i1,t-1} + v_{it}, \quad i = 1, \ldots K_1 \\ x_{i2t} &= \lambda_{i2} F_t + u_{it}, \quad i = 1, \ldots K_2 \end{aligned}$$

with $\varepsilon_t = \frac{1}{\sqrt{2}}(\widetilde{\varepsilon}_t^2 - 1)$ and $u_{it} = \frac{1}{\sqrt{2}}(\widetilde{u}_{it}^2 - 1)$. We assume $\alpha_x \sim U(.2, .8)$, $v_{it} \sim N(0,1)$ and uncorrelated with $\widetilde{u}_{jt}$ and $\widetilde{\varepsilon}_t$. Furthermore, $(\widetilde{\varepsilon}_t, \widetilde{u}_t')' \sim N(0_{K2+1}, \Sigma)$ where $diag(\Sigma) = 1$, $\Sigma(j,1) = \Sigma(1,j) \sim U(.3, .6)$, and zero elsewhere. This means that $\widetilde{\varepsilon}_t$ is correlated with $\widetilde{u}_{it}$ with covariance $\Sigma(1,i)$ but $\widetilde{u}_{it}$ and $\widetilde{u}_{jt}$ are uncorrelated $(i \neq j)$. By construction, the errors are heteroskedastic. The parameter σ_y^2 is set to $K_1\bar{\sigma}_{x1}^2 + K_2\bar{\sigma}_{x2}^2$ where $\bar{\sigma}_{x_j}$ is the average variance of x_{jt}, $j = 1, 2$. This puts the noise-to-signal ratio in the primary equation of roughly one-half.

The parameter of interest is β_2. We considered various values of K_2, σ_z, and r. The results are reported in Table 1 with $K_2 = 1$, and $\sigma_z = 3$. This is the least favorable situation since the factors are less informative with a low common component to noise ratio. The column labeled $\rho_{x_2\varepsilon}$ is the correlation coefficient between x_2 and ε and thus indicates the degree of endogeneity. Under the assumed parametrization, this correlation is around .2. The true value of β_2 is 2, and the impact of endogeneity bias on OLS is immediately obvious. The

known.

estimators that use the factors as instruments are more precise. The factor based instruments dominate the IV either in bias or RMSE, if not both. The J test associated with the FIV is close to the nominal size of 5%, while the two-sided t statistic for testing $\beta_2 = 2$ has some size distortion when N, T are both small. The size distortions of both tests decrease with T.

Example 2 In this example, the regression model is

$$y_t = \beta_1 + x_{2t}'\beta_2 + \varepsilon_t. \tag{7}$$

The endogenous variables x_{2t} are spanned by L factors, while the panel of observed instruments is spanned by r factors and $r \geq L$. To generate data with this structure, let F be a $T \times r$ matrix of iid $N(0,1)$ variables and let $F(:, 1:L)$ be a $T \times L$ matrix consisting of the columns 1 to L of F. We simulate a $T \times 1$ vector y, a $T \times N$ matrix Z, and a $T \times L$ matrix X_2 as

$$\begin{aligned} y &= F(:, 1:L)\Lambda_y' + \sigma_y e_y \\ X_2 &= F(:, 1:L)\Lambda_x' + e_x \end{aligned}$$

where $e_{jt} \sim N(0, \sigma_j^2)$, $\sigma_j^2 \sim U(\sigma_l, \sigma_h)$. Now if $F(:, 1:L)$ is L dimensional, it can be represented in terms of *any* L variables spanned by the these factors. Thus, using $F(:, 1:L) = (X_2 - e_x)\Lambda_x'^{-1}$ yields

$$\begin{aligned} y &= X_2\Lambda_x^{-1}\Lambda_y + e_y - e_x\Lambda_x^{-1}\Lambda_y \\ &= X_2\beta_2^* + \varepsilon^* \end{aligned}$$

where $\beta^* = \Lambda_x'^{-1}\Lambda_y$ is $L \times 1$ and $\varepsilon^* = e_y - e_x\beta^*$. For given Λ_x, we then solve for Λ_y such that $\beta_2^* = (1_{K_2}', 0_{L-K_2}')$. The x_{2t} in (7) corresponds to the first K_2 columns of X_{2t}. This also implies that the true value of every element of β_2 is unity. The endogeneity bias is $\beta'\text{cov}(e_x)\beta$. For the loadings, we assume $\lambda_z \sim N(0_N, I_N)$. The elements of the $L \times L$ matrix Λ_x are drawn from the $N(1,1)$ distribution. Written in terms of r factors, $X_2 = F(:, 1:r)\Lambda_x^{(r)}$ where $\Lambda_x^{(r)}$ only has the first $L \times L$ positions being non-zero. Viewed this way, the first L factors are the relevant factors.

We estimate $rmax = r + 2$ factors and report simulations for $K_2 = 1$ with $\beta_2^0 = 2$. The results are reported in Table 2. Unlike Example 1, the correlation between x_{2t} and ε_t is now negative. In this example, the IV is actually more biased than OLS. The factor IV estimators again perform well.

Example 3 Here, we consider estimation of β by panel regressions. The DGP is

$$\begin{aligned} y_{it} &= \beta_1 + \beta_2 x_{it} + \varepsilon_{it} \\ x_{it} &= \lambda_i' F_t + \sqrt{r}\, u_{it} \\ \rho_i &= corr(\varepsilon_{it}, u_{it}) \sim U(.3, .6). \end{aligned}$$

where F_t is again $r \times 1$, ρ_i is the correlation between ε_{it} and u_{it}. We set the true value of $\beta = (\beta_1, \beta_2)' = (0, 1)'$ but include an intercept in the regression. According to Theorem 2, we can use the factors estimated from x_{it} to instrument themselves. For the PFIV, we use r factors. We also consider an estimator, denoted PfIV, which uses $rmax = r + 2$ factors. Note that these estimates are not corrected for bias in order to to show that the bias is of second order importance. For the sake of comparison, we also consider PTFIV. Note that in this example, $E(\lambda_i) = 0$ and the PTFIV should be more volatile (larger variance) because $S_{\widetilde{F}x}$ can be near singular.

The results are reported in Table 3. As expected, the pooled POLS estimator is quite severely biased. The PTFIV has noticeably larger RMSE than the three factor based estimators, which are all centered around the true value. The PFIV has smaller bias than the PfIV with no increase in variance. Even with $\min[N, T]$ as small as 25, the PFIV is quite precise. Increasing N and/or T clearly improves precision even without bias correction. Because the PFIV has a small variance, the t test becomes very sensitive to small departures of the estimate from the true value. Thus, without bias correction, the t test based on the PFIV has important size distortions. The bias-corrected test is, however, much more accurate though there is still size distortions when r is large. The t-statistics based on OLS will have much higher distortions (not reported). The test based on PTFIV is much closer to the nominal size of 5% regardless of r, primarily because the variance of the estimator is much larger than the PFIV. In terms of MSE. The PFIV is clearly the estimator of choice.

Summing up, we have reported results for the FIV which uses the true number of factors underlying the endogenous variable x_2, and the fIV which uses more instruments than is necessary. While the results do not show significant difference, using too many factors can sometimes increase bias but may reduce mean-squared error. This suggests further research on choosing the number of factors. Instead of using the suggested information criteria, selecting relevant factors via boosting is an alternative. A further alternative is to directly choose instruments from the observed ones or use the regularization approach of Carrasco (2006) without assuming a factor structure. Whether we use estimated factors or Z as instruments, it is open issue how to select the most relevant ones from many valid instruments that have no natural ordering. This problem, along with empirical applications, will be reported in a separate paper, Bai and Ng (2007).

5 Conclusion

This paper provides a new way of using the estimated factors not previously considered in either the factor analysis or the instrumental variables literature. We take as starting point that in a data rich environment, there are many instruments that are weakly exogenous for the parameters of interest. Pooling the information across instruments enables us to construct factor based instruments that are not only valid, but are more strongly correlated with the endogenous variable than each individually observed instrument. The result is a factor based instrumental variable estimator (FIV) that is more efficient. For large simultaneous systems, we show that valid instruments can be constructed from invalid ones. Whereas the correlation between a particular instrument and the endogenous regressor may be weak, the estimated factors are less susceptible to this problem under our maintained assumption that variables in the system have a factor structure.

Table 1: Finite Sample Properties of $\widehat{\beta}_2$, $\beta_2^0 = 2$.

T	N	r	rmax	$\rho_{x_2\varepsilon}$	FIV	fIV	IV	OLS	J_F	t_F	J_f	t_f
					Mean/RMSE							
50	50	1	2	0.38	1.97	2.00	2.18	2.73	0.00	0.06	0.04	0.07
					0.41	0.39	0.45	0.85				
100	50	1	2	0.35	1.98	2.00	2.06	2.67	0.00	0.05	0.04	0.06
					0.25	0.25	0.28	0.73				
100	100	1	2	0.32	2.00	2.01	2.05	2.59	0.00	0.05	0.05	0.06
					0.23	0.22	0.26	0.64				
200	100	1	2	0.28	2.01	2.01	2.03	2.50	0.00	0.06	0.04	0.06
					0.14	0.14	0.15	0.53				
50	50	2	4	0.56	2.04	2.15	2.57	3.18	0.05	0.09	0.04	0.14
					0.59	0.51	0.78	1.28				
100	50	2	4	0.52	2.01	2.05	2.23	3.08	0.04	0.06	0.03	0.09
					0.32	0.29	0.41	1.14				
100	100	2	4	0.52	2.01	2.04	2.23	3.07	0.05	0.08	0.05	0.10
					0.31	0.29	0.40	1.13				
200	100	2	4	0.50	2.00	2.03	2.04	3.04	0.05	0.06	0.05	0.07
					0.21	0.20	0.23	1.06				

Note: FIV and fIV are GMM estimators with $\widetilde{F}$ and $\widetilde{f}$ as instruments. These are of dimensions r and $rmax$, respectively. IV is the GMM estimator with z_2 as instruments, where z_2 is of dimension r and has the largest correlation with x_2.

Table 2: Finite Sample Properties of $\widehat{\beta}_2$: $\beta_2^0 = 1$

T	N	r	L	$\rho_{x_2\varepsilon}$	FIV	fIV	IV	OLS	J_F	t_F	J_f	t_f
					Mean/RMSE							
50	50	2	2	-0.43	1.01	0.99	0.94	0.72	0.04	0.08	0.03	0.10
					0.19	0.19	0.20	0.32				
100	50	2	2	-0.43	1.01	1.00	1.00	0.72	0.04	0.08	0.05	0.10
					0.13	0.14	0.14	0.30				
100	100	2	2	-0.68	0.99	0.94	0.81	0.29	0.05	0.09	0.07	0.15
					0.20	0.20	0.25	0.71				
200	100	2	2	-0.56	1.00	0.99	0.94	0.53	0.05	0.07	0.05	0.07
					0.10	0.10	0.13	0.48				
50	50	4	3	-0.56	0.96	0.92	0.85	0.57	0.04	0.11	0.04	0.17
					0.22	0.23	0.24	0.45				
100	50	4	3	-0.59	0.97	0.96	0.90	0.53	0.06	0.09	0.05	0.11
					0.15	0.15	0.17	0.48				
100	100	4	3	-0.61	0.97	0.95	0.86	0.50	0.06	0.09	0.06	0.13
					0.16	0.16	0.21	0.51				
200	100	4	3	-0.67	0.99	0.97	0.88	0.40	0.05	0.07	0.04	0.10
					0.13	0.13	0.17	0.60				

Note: FIV and fIV are GMM estimators with $\widetilde{F}$ and $\widetilde{f}$ as instruments. These are of dimensions r and $rmax = r + 2$, respectively. IV is the GMM estimator with z_2 as instruments, where z_2 is of dimension r and has the largest correlation with x_2.

Table 3: Finite Sample Properties of $\widehat{\beta}_2$ for panel data, $\beta_2^0 = 1$.

T	N	r	$\rho_{x_2\varepsilon}$	PFIV	PFIV$^+$	PfIV	PfIV$^+$	PTFIV	POLS	$t_{\widehat{\beta}_{PFIV}}$	$t_{\widehat{\beta}_{PFIV^+}}$	$t_{\widehat{\beta}_{PTFIV}}$
				Mean/RMSE								
15	15	2	0.29	1.05	1.03	1.08	1.06	1.12	1.10	0.40	0.20	0.10
				0.07	0.05	0.09	0.07	0.22	0.11			
25	25	2	0.30	1.03	1.01	1.06	1.04	1.08	1.10	0.43	0.11	0.07
				0.04	0.02	0.06	0.04	0.18	0.10			
25	50	2	0.30	1.03	1.01	1.05	1.03	1.07	1.10	0.50	0.09	0.08
				0.03	0.02	0.05	0.03	0.17	0.10			
50	25	2	0.27	1.02	1.01	1.04	1.02	1.07	1.09	0.39	0.08	0.10
				0.03	0.02	0.04	0.03	0.13	0.10			
50	50	2	0.29	1.02	1.00	1.03	1.02	1.06	1.10	0.37	0.06	0.08
				0.02	0.01	0.03	0.02	0.13	0.10			
100	50	2	0.28	1.01	1.00	1.02	1.01	1.05	1.09	0.36	0.06	0.09
				0.01	0.01	0.02	0.01	0.10	0.09			
50	100	2	0.29	1.01	1.00	1.03	1.01	1.04	1.10	0.48	0.06	0.06
				0.01	0.01	0.03	0.01	0.13	0.10			
100	100	2	0.29	1.01	1.00	1.02	1.01	1.04	1.10	0.38	0.06	0.07
				0.01	0.00	0.02	0.01	0.11	0.10			
15	15	4	0.28	1.06	1.04	1.07	1.06	1.08	1.07	0.79	0.57	0.15
				0.06	0.05	0.07	0.06	0.14	0.08			
25	25	4	0.29	1.04	1.02	1.05	1.04	1.06	1.07	0.88	0.43	0.14
				0.04	0.03	0.05	0.04	0.11	0.07			
25	50	4	0.30	1.03	1.01	1.05	1.03	1.06	1.08	0.93	0.37	0.12
				0.04	0.02	0.05	0.03	0.10	0.08			
50	25	4	0.28	1.03	1.01	1.04	1.03	1.05	1.07	0.86	0.33	0.15
				0.03	0.02	0.04	0.03	0.08	0.07			
50	50	4	0.28	1.02	1.01	1.03	1.02	1.04	1.07	0.87	0.18	0.11
				0.02	0.01	0.03	0.02	0.08	0.07			
100	50	4	0.29	1.02	1.00	1.02	1.01	1.03	1.07	0.90	0.14	0.14
				0.02	0.01	0.03	0.01	0.06	0.07			
50	100	4	0.29	1.02	1.00	1.03	1.01	1.03	1.07	0.91	0.16	0.09
				0.02	0.01	0.03	0.01	0.08	0.07			
100	100	4	0.29	1.01	1.00	1.02	1.01	1.02	1.07	0.88	0.09	0.10
				0.01	0.00	0.02	0.01	0.05	0.07			

Note: PFIV and PfIV are panel instrumental variable estimators with $\widetilde{C}_{it} = \widetilde{\lambda}_i'\widetilde{F}_t$ and $\widetilde{c}_{it} = \widetilde{\lambda}_i'\widetilde{f}_t$ as instruments, respectively. The PFIV$^+$ and PfIV$^+$ are biased-corrected estimators. $\widetilde{F}_t$ is $r \times 1$, and $\widetilde{f}_t$ is $rmax \times 1$ with $rmax = r + 2$. PTFIV is the 'traditional' panel IV estimator that uses $\widetilde{F}_t$ as instruments.

Appendix

To prove the main result we need the following lemma:

Lemma A1 *Let $H = \widetilde{V}^{-1}(\widetilde{F}'F/T)(\Lambda'\Lambda/N)$. Under Assumption (A) and as $N, T \to \infty$,*

i $\frac{1}{T}\sum_{t=1}^{T} \|\widetilde{F}_t - HF_t\|^2 = O_p(\min[N,T]^{-1})$;

ii If there exists an $M < \infty$ such that $\sum_{i=1}^{N} |E(\varepsilon_t e_{it})| \le M$ for all N and t, then

$$T^{-1}\sum_{t=1}^{T}(\widetilde{F}_t - HF_t)\varepsilon_t = O_p(\min[N,T]^{-1})$$

iii If ε_t is uncorrelated with e_{it} for all i and t, then

$$T^{-1}\sum_{t=1}^{T}(\widetilde{F}_t - HF_t)\varepsilon_t = O_p(\frac{1}{\sqrt{NT}}) + O_p(T^{-1})$$

The proof of part (i) is in Bai and Ng (2002); the proof of part (ii) is the same as that of Lemma B.1 of Bai (2003). The proof of part (iii) is also the same as part (ii), and the bound is tightened by using the uncorrelation assumption. The details are omitted.

Proof of Theorem 1: Let $\widetilde{g}_t(\beta^0) = \widetilde{F}_t\varepsilon_t$ and $\bar{g} = \frac{1}{T}\sum_{t=1}^{T}\widetilde{g}_t(\beta^0)$. Then

$$\widehat{\beta}_{FIV} - \beta^0 = (S'_{\widetilde{F}x}\breve{S}^{-1}S_{\widetilde{F}x})^{-1}S'_{\widetilde{F}x}\breve{S}^{-1}\bar{g}.$$

Now

$$\begin{aligned}\sqrt{T}\bar{g} &= T^{-1/2}\sum_{t=1}^{T}\widetilde{F}_t\varepsilon_t \\ &= T^{-1/2}\sum_{t=1}^{T}(\widetilde{F}_t - HF_t)\varepsilon_t + HT^{-1/2}\sum_{t=1}^{T}F_t\varepsilon_t \\ &= HT^{-1/2}\sum_{t=1}^{T}F_t\varepsilon_t + o_p(1)\end{aligned}$$

By Lemma A1(iii), $T^{-1/2}\sum_{t=1}^{T}(\widetilde{F}_t - HF_t)\varepsilon_t = O_p(N^{-1/2}) + O_p(T^{-1/2}) = o_p(1)$, as $N, T \to \infty$. By assumption, $T^{-1/2}\sum_{t=1}^{T}F_t\varepsilon_t \xrightarrow{d} N(0, S^0)$. Thus $\sqrt{T}\bar{g} \xrightarrow{d} N(0, H_0S^0H_0')$, where $H_0 = \text{plim}\, H$. But $\text{plim}\, \breve{S} = H_0S^0H_0'$. This implies that $\breve{S}^{-1/2}\sqrt{T}\bar{g} \xrightarrow{d} N(0, I)$. Furthermore, $S_{\widetilde{F}x} = \frac{1}{T}\widetilde{F}'x = \frac{1}{T}H'F'x + o_p(1) \xrightarrow{p} H_0'\Omega_{Fx}$, where Ω_{Fx} is the probability limit of $\frac{1}{T}F'x = \frac{1}{T}\sum_{t=1}^{T}F_tx_t'$. Thus $S'_{\widetilde{F}x}\breve{S}^{-1}S_{\widetilde{F}x} \xrightarrow{p} \Omega'_{Fx}(S^0)^{-1}\Omega_{Fx}$. Summarizing result, we have

$$\sqrt{T}(\widehat{\beta}_{FIV} - \beta) \xrightarrow{d} N(0, (\Omega'_{Fx}(S^0)^{-1}\Omega_{Fx})^{-1})$$

Thus the limiting distribution coincides with that using the true F as instruments.

Finally, because $\widetilde{F}_t$ is a vector of $r \times 1$ instruments, and β is $K \times 1$, the over-identification J test of Hansen (1982) has a limit of χ^2_{r-K}.

Proof of the Claim in Remark 1: Following the proof of Theorem 1, instead of invoking Lemma A1(iii), we use Lemma A1(ii) to obtain $T^{-1/2} \sum_{t=1}^{T} (\widetilde{F}_t - HF_t)\varepsilon_t = O_p(\sqrt{T}/\min[N, T])$, which is $o_p(1)$ provided that $\sqrt{T}/N \to 0$. The rest of the proof is identical to that of Theorem 1.

Proof of Proposition 1: Without loss of generality, we assume homoskedasticity for ε_t. In addition, we assume there is no x_1 so that $x = x_2$. Writing in vector format, equation (2) can be rewritten as $x = F\Psi + u$, where $F = (F_1, ..., F_T)'$, and x and u are $T \times 1$ vectors. Let z_2 be a $T \times r$ matrix consisting of r valid instruments from the N available instruments. Let P_2 be the projection matrix associated with z_2, i.e., $P_2 = z_2(z_2'z_2)^{-1}z_2$. Let $M_2 = I - P_2$. The asymptotic variance of the GMM estimator with a r observed variables as instruments is the probability limit of

$$\widehat{\Omega}_{IV} = \sigma_\varepsilon^2 (T^{-1}x'P_2x)^{-1}$$

The asymptotic variance of the FIV is the probability limit of

$$\widehat{\Omega}_{FIV} = \sigma_\varepsilon^2 (T^{-1}x'P_Fx)^{-1}.$$

Now $x = F\Psi + u$, $P_2x = P_2F\Psi + P_2u$. Thus,

$$T^{-1}x'P_2x_2 = T^{-1}x'P_2F\Psi + o_p(1), \tag{A.1}$$

where we have used $T^{-1}z_2'u = o_p(1)$, which follows from $E(z_{it}u_t) = 0$. Furthermore, $P_Fx = P_FF\Psi + P_Fu = F\Psi + P_Fu$ and from $I = M_2 + P_2$, we have

$$T^{-1}x'P_Fx = T^{-1}x'F\Psi + o_p(1) = T^{-1}x'(M_2 + P_2)F\Psi + o_p(1), \tag{A.2}$$

where $\frac{1}{T}x'P_Fu = o_p(1)$ because $E(F_tu_t) = 0$ and $T^{-1}F'u = o_p(1)$. Subtract (A.2) from (A.1),

$$\widehat{\Omega}_{IV}^{-1} - \widehat{\Omega}_{FIV}^{-1} = \sigma_\varepsilon^{-2}T^{-1}(x'P_2x) - \sigma_\varepsilon^{-2}T^{-1}(x'P_Fx)$$
$$= -\sigma_\varepsilon^{-2}T^{-1}(x'M_2F\Psi) + o_p(1) = -\sigma_\varepsilon^{-2}T^{-1}(x - u + u)'M_2F\Psi + o_p(1)$$
$$= -\sigma_\varepsilon^{-2}T^{-1}\Psi'F'M_2F\Psi + o_p(1) < 0,$$

where the last equality follows from $x - u = F\Psi$ and $T^{-1}u'M_2F = o_p(1)$. The limit of $T^{-1}F'M_2F$ is positive because z_2 can be written as $z_2 = F\Lambda_2 + e_2$ with $T^{-1}e_2'e_2 > 0$ under the assumption of the proposition (note that if $e_2 = 0$, then $F'M_2F = 0$).

Proof of Theorem 2, part(i): We shall show $\widehat{\beta}_{PFIV}-\beta=O_p(T^{-1})+O_p(N^{-1})$, equivalently, $\sqrt{NT}(\widehat{\beta}_{PFIV}-\beta)=O_p(\sqrt{N/T})+O_p(\sqrt{T/N})$. From $\widehat{\beta}_{PFIV}=\beta+S_{\tilde{x}\tilde{x}}^{-1}\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\widehat{C}_{it}\varepsilon_{it}$, it is sufficient to consider the limit of $(NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T\widehat{C}_{it}\varepsilon_{it}$. Because $(NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T C_{it}\varepsilon_{it}\xrightarrow{d}N(0,S)$, we need to show, for part (i)

$$(NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T(\widehat{C}_{it}-C_{it})\varepsilon_{it}=O_p(\sqrt{N/T})+O_p(\sqrt{T/N}).$$

Notice

$$\widehat{C}_{it}-C_{it}=\widetilde{\Lambda}_i'\widehat{F}_t-\Lambda_i'F_t=(\widetilde{\Lambda}_i-H^{-1}\Lambda_i)'\widehat{F}_t+\Lambda_i'(\widetilde{F}_t-HF_t)$$
$$=(\widetilde{\Lambda}_i-H^{-1}\Lambda_i)'(\widetilde{F}_t-HF_t)+(\widetilde{\Lambda}_i-H^{-1}\Lambda_i)'HF_t+\Lambda_i'(\widetilde{F}_t-HF_t)$$

The first term is dominated by the last two terms and can be ignored. Let $\Lambda_i=(\lambda_{i,1},...,\lambda_{i,k})$ $(r\times k)$ and $u_{it}=(u_{it,1},...,u_{it,K})'$ $(K\times 1)$. From Bai (2003), equations (A.5) and (A.6)

$$\widetilde{F}_t-HF_t=V_{NT}^{-1}(\frac{1}{T}\widetilde{F}'F)\frac{1}{NK}\sum_{j=1}^N\sum_{k=1}^K\lambda_{j,k}u_{jt,k}+O_p(\delta_{NT}^{-2})$$

Denote $G=V_{NT}^{-1}(\frac{1}{T}\widetilde{F}'F)$, which is $O_p(1)$, we have

$$(NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T\Lambda_i'(\widetilde{F}_t-HF_t)\varepsilon_{it}=(NT)^{-1/2}\sum_{t=1}^T\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N\sum_{k=1}^K\Lambda_i\varepsilon_{it}G\lambda_{j,k}u_{jt,k}+o_p(1)$$

Note that ε_{it} is scalar, thus commutable with all vectors and matrices. Here $\Lambda_i\varepsilon_{it}$ is understood as $\Lambda_i\otimes\varepsilon_{it}$, which is $K\times r$. We can rewrite the above as

$$\begin{aligned}&(NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T\Lambda_i'(\widetilde{F}_t-HF_t)\varepsilon_{it}\\&=(T/N)^{1/2}\frac{1}{T}\sum_{t=1}^T\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^N\Lambda_i\varepsilon_{it}\Big)G\Big(\frac{1}{\sqrt{N}}\sum_{j=1}^N\sum_{k=1}^K\lambda_{j,k}u_{jt,k}\Big)+o_p(1)\qquad\text{(A.3)}\\&=(T/N)^{1/2}O_p(1)\end{aligned}$$

Next, by (B.2) of Bai (2003),

$$\widetilde{\Lambda}_i-H^{-1}\Lambda_i=H\frac{1}{T}\sum_{s=1}^T F_su_{is}'+O_p(\delta_{NT}^{-2})$$

Thus

$$(NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T(\widetilde{\Lambda}_i-H^{-1}\Lambda_i)'HF_t\varepsilon_{it}=(NT)^{-1}\frac{1}{T}\sum_{i=1}^N\sum_{s=1}^T u_{is}F_s'H'H\sum_{t=1}^T F_t\varepsilon_{it}+o_p(1)$$

$$= (N/T)^{1/2}\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}u_{is}F_s'\Big)H'H\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t\varepsilon_{it}\Big) + o_p(1) \qquad \text{(A.4)}$$
$$= (N/T)^{1/2}O_p(1)$$

Combining (A.3) and (A.4), we prove part (i) of the theorem.

Proof of Theorem 2 part (ii): The biases equal to $S_{\tilde{x}\tilde{x}}^{-1}$ multiplied by the expected values of (A.3) and (A.4). We analyze these expected values below. Introduce

$$A_t = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Lambda_i\varepsilon_{it}, \quad \text{and} \quad B_t = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\sum_{k=1}^{K}\lambda_{j,k}u_{jt,k}$$

The summand in (A.3) is A_tGB_t, which is a vector. Thus

$$A_tGB_t = \text{vec}(A_tGB_t) = (B_t'\otimes A_t)\text{vec}(G)$$

it follows that (again ignoring the $o_p(1)$ term):

$$(A.3) = (T/N)^{1/2}\Big(\frac{1}{T}\sum_{t=1}^{T}(B_t\otimes A_t\Big)\text{vec}(G)$$

Because of the cross-sectional independence assumption on ε_{it} and on u_{it}, we have

$$E(B_t'\otimes A_t) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_{j,k}'\otimes\Lambda_i)E(u_{it,k}\varepsilon_{it})$$

Let

$$\delta_1 = \Big(\frac{1}{T}\sum_{t=1}^{T}E(B_t'\otimes A_t)\Big)\text{vec}(G) = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K}\Lambda_iG\lambda_{i,k}E(u_{it,k}\varepsilon_{it})$$

From $\frac{1}{T}\sum_{t=1}^{T}[(B_t'\otimes A_t) - E(B_t'\otimes A_t)] = O_p(T^{-1/2})$, it follows immediately that

$$(A.3) = (T/N)^{1/2}\delta_1 + o_p(1)$$

Let δ_1^0 denote the limit of δ_1. If $T/N \to \tau$, it follows that

$$(A.3) \to \tau^{1/2}\delta_1^0$$

Next consider (A.4). Let

$$\Theta_i = T^{-1/2}\sum_{s=1}^{T}u_{is}F_s' \;\; \text{and} \;\; \Phi_i = T^{-1/2}\sum_{t=1}^{T}F_t\varepsilon_{it}$$

then (A.4) can be rewritten as (ignoring the $o_p(1)$ term):

$$(A.4) = (N/T)^{1/2}\Big(\frac{1}{N}\sum_{i=1}^{N}(\Phi_i' \otimes \Theta_i)\Big)\text{vec}(H'H)$$

The expected value of $\Phi_i' \otimes \Theta_i$ contains the elements of the long-run variance of the vector sequence $\eta_t = (\text{vec}(u_{it}F_t)', F_t'\varepsilon_{it})'$. From $\frac{1}{N}\sum_{i=1}^{N}[(\Phi_i' \otimes \Theta_i) - E(\Phi_i' \otimes \Theta_i)] = O_p(N^{-1/2})$, we have

$$(A.4) = (N/T)^{1/2}\Delta_2 + o_p(1)$$

where $\delta_2 = \Big(\frac{1}{N}\sum_{i=1}^{N} E(\Phi_i' \otimes \Theta_i)\Big)\text{vec}(H'H)$. It can be shown that

$$H'H = (F'F/T)^{-1} + O_p(\delta_{NT}^{-2}) = \Sigma_F^{-1} + o_p(1)$$

Let

$$\delta_2^0 = \lim\Big(\frac{1}{N}\sum_{i=1}^{N} E(\Phi_i' \otimes \Theta_i)\Big)\Sigma_F^{-1}$$

If $N/T \to \tau$, we have $(A.4) \to \tau^{-1/2}\delta_2^0$. Denote

$$\Delta_1^0 = [\text{plim}\, S_{\widetilde{x}\widetilde{x}}]^{-1}\delta_1^0, \quad \text{and} \quad \Delta_2^0 = [\text{plim}\, S_{\widetilde{x}\widetilde{x}}]^{-1}\delta_2^0$$

then the asymptotic bias is

$$\tau^{1/2}\Delta_1^0 + \tau^{-1/2}\Delta_2^0,$$

proving part (ii).

Proof of Corollary 1: The analysis in part (ii) of the theorem shows that

$$\sqrt{NT}(\widehat{\beta}_{PFIV} - \beta) = S_{\widetilde{x}\widetilde{x}}^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}\varepsilon_{it} + \sqrt{T/N}S_{\widetilde{x}\widetilde{x}}^{-1}\delta_1 + \sqrt{N/T}S_{\widetilde{x}\widetilde{x}}^{-1}\delta_2 + o_p(1) \quad (A.5)$$

It can be shown that $\widehat{\Delta}_1 - S_{\widetilde{x}\widetilde{x}}^{-1}\delta_1 = O_p(\delta_{NT}^{-1})$ and $\widehat{\Delta}_2 - S_{\widetilde{x}\widetilde{x}}^{-1}\delta_2 = O_p(\delta_{NT}^{-1})$. These imply that $(T/N)^{1/2}(\widehat{\Delta}_1 - S_{\widetilde{x}\widetilde{x}}^{-1}\delta_1) = o_p(1)$ if $T/N^2 \to 0$, and $((N/T)^{1/2}(\widehat{\Delta}_2 - S_{\widetilde{x}\widetilde{x}}^{-1}\delta_2) = o_p(1)$ if $N/T^2 \to 0$. Thus, we can replace $S_{\widetilde{x}\widetilde{x}}^{-1}\delta_1$ by $\widehat{\Delta}_1$ and replace $S_{\widetilde{x}\widetilde{x}}^{-1}\delta_2$ by $\widehat{\Delta}_2$ in (A.5). Equivalently,

$$\sqrt{NT}(\widehat{\beta}_{PFIV} - \frac{1}{N}\widehat{\Delta}_1 - \frac{1}{T}\widehat{\Delta}_2 - \beta) = S_{\widetilde{x}\widetilde{x}}^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}\varepsilon_{it} + o_p(1).$$

Asymptotic normality of the biased corrected estimator follows from the asymptotic normality for $(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}\varepsilon_{it}$. This proves Corollary 1.

References

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

Andrews, D., Moreira, M. and Stock, J. 2006, Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression, *Econometrica* **74:3**, 715–754.

Bai, J. 2003, Inferential Theory for Factor Models of Large Dimensions, *Econometrica* **71:1**, 135–172.

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

Bai, J. and Ng, S. 2007, Selecting Instrumental Variables in a Data Rich Environment by Boosting, mimeo, University of Michigan.

Bernanke, B. and Boivin, J. 2003, Monetary Policy in a Data Rich Environment, *Journal of Monetary Economics* **50:3**, 525–546.

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

Favero, C. and Marcellino, M. 2001, Large Datasets, Small Models, and Monetary Europe, IGIER, Working Paper 208.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. 2005, The Generalized Dynamic Factor Model, One Sided Estimation and Forecasting, *Journal of the American Statistical Association* **100**, 830–840.

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

Hallin, M. and Liska, R. 2007, Determining the Number of Factors in the General Dynamic Factor Model, *Journal of the American Statistical Association* **102**, 603–617.

Hansen, L. P. 1982, Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* **50**, 1029–1054.

Hausman, J. 1978, Specification Tests in Econometrics, *Econometrica* **46**, 1251–1272.

Hausman, J., Newey, W. and Woutersen, T. 2006, IV Estimation with Heteroskedasticity and Many Instruments, mimeo, MIT.

Hayashi, F. 2000, *Econometrics*, Princeton Unversity Press, Princeton, N.J.

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

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

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

Stock, J. H. and Watson, M. W. 2002, Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association* **97**, 1167–1179.