Econometric Theory, **26**, 2010, 1577–1606.
doi:10.1017/S0266466609990727

INSTRUMENTAL VARIABLE ESTIMATION IN A DATA RICH ENVIRONMENT

JUSHAN BAI
Columbia University
and
Central University of Finance and Economics

SERENA NG
Columbia University

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 and they are more efficient than the observed variables in our framework. Whereas standard optimal generalized method of moments estimator using a large number of instruments is biased and can be inconsistent, the factor instrumental variable estimator (FIV) is shown to be consistent and asymptotically normal, even if the number of instruments exceeds the sample size. Furthermore, FIV remains consistent even if the observed variables are invalid instruments as long as the unobserved common components are valid instruments. We also consider estimating panel data models in which all regressors are endogenous but share exogenous common factors. We show that valid instruments can be constructed from the endogenous regressors. Although single equation FIV requires no bias correction, the faster convergence rate of the panel estimator is such that a bias correction is necessary to obtain a zero-centered normal distribution.

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. When the parameters of interest are 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

This paper was presented at Columbia, Duke, Harvard/MIT, Michigan, Queen's, Yale, UCSD, UCR, UPenn, Wisconsin, Institute of Statistics at Universite Catholique de Louvain, and SETA in Hong Kong. We thank seminar participants, Guido Kuersteiner (the co-editor), and two anonymous referees for many helpful comments and suggestions. We also acknowledge financial support from the NSF (grants SES-0551275 and SES-0549978). Address Correspondence to Jushan Bai, Department of Economics, Columbia University, MC 3308, 1022 IAB, 420 West 118 St., New York, NY 10027, USA; e-mail: Jushan.Bai@Columbia.edu.

© Cambridge University Press 2010 0266-4666/10 $15.00

example, Andrews, Moreira, and Stock (2006) and the references therein. This paper is also concerned with the quality of instruments but has a different focus. 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 if these variables and the endogenous regressors share some common exogenous factors, then 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 panel data models, we show that if N and T are both large, 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, Hallin, Lippi, and Reichlin (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.

It is well recognized that using too many potentially relevant instruments in the first stage of two-stage least squares estimation will induce bias. 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, although 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 noise indicators. We use a factor approach to estimate the feasible instruments space from the observed instruments. The resulting factor-based instrumental variable estimator is denoted FIV.

Our framework of many instruments is different from that of the existing literature, such as Bekker (1994), Donald and Newey (2001), and Chao and Swanson (2005), among many others. In those analyses, no structure is imposed on the many and weak instruments. Also, although the theoretical setup allows the number of instruments to increase with the sample size, the number of instruments is smaller than the number of observations. This is evident from the simulations reported in these studies. In contrast, our analysis allows the number of instruments to exceed the number of observations. This is possible because of the structure we impose on the panel of instruments. Not every application will satisfy the assumptions of our analysis, but when these assumptions are satisfied, our framework allows irrelevant and even invalid instruments. In the terminology of Bernanke and Boivin (2003), what we propose is a way to construct valid instrumental variables in a "data rich environment."

In macroeconomic analysis, the "data rich" environment is commonly encountered, as lots of variables are available over a long time span. These macroeconomic panels of data also tend to have a factor structure, as indicated by common comovements in a large number of variables. The factor framework has a long history in macroeconomic modeling. 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 and are thus better instruments. Here, we provide a formal analysis and show that the estimated factors are more efficient instruments than the observed variables. Our analysis is confined to cases in which the model is linear in the endogenous regressors, though we permit nonlinear instrumental variable estimation when the nonlinearity is induced by parameter restrictions. Nonlinear instrumental variable estimation is a more involved problem even when the instruments are observed, and this issue is not dealt with in our analysis.

As far as we are aware, Kapetanios and Marcellino (2006) is the only other paper that considers using estimated factors as instruments. The authors' framework assumes that there are many observed 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. As will be made clear later, the strong factor asymptotics does not preclude the possibility that some series only have a weak factor structure (in fact, some factor loadings can be zero). We also consider the case of "invalid instruments" and compare efficiency of the FIV with the traditional optimal generalized method of moments (GMM) estimator. We show that the latter is inconsistent unless the ratio of the number of instruments and the sample size goes to zero, and that FIV is as efficient as the bias-corrected optimal GMM. Whereas Kapetanios and Marcellino (2006) focused on single equation models, we also consider a simultaneous equations system and show that valid instruments can be constructed from endogenous regressors.

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 instruments. Simulations are given in Section 4. Section 5 concludes, and proofs of the results are contained in the Appendixes.

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 and x_{2t}. The $K_2 \times 1$ vector x_{2t} is endogenous in the sense that $\mathrm{E}(x_{2t}\varepsilon_t) \neq 0$ 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 an $r \times 1$ vector of fundamental variables satisfying $\mathrm{E}(F_t\varepsilon_t) = 0$, and $r \geq K_2$ is a small number. The assumption that $\mathrm{E}(F_t\varepsilon_t) = 0$ is required for F_t to be valid instruments. The assumption $r \geq K_2$ is analogous to the order condition that the number of instruments is at least as large as the number of parameters to be estimated. Endogeneity arises when $\mathrm{E}(u_t\varepsilon_t) \neq 0$. This induces a nonzero correlation between x_{2t} and ε_t. Equation (2) can be modified to allow variables other than F_t to be present. For example, if $x_{2t} = \Psi' F_t + \Gamma' W_t + u_t$ with W_t being observable and exogenous, the obvious extension is to use $(F_t', W_t')'$ as instruments. The thrust of the analysis remain valid.

If F_t were observed, $\beta = (\beta_1', \beta_2')'$ could be estimated, for example, by using F_t to instrument x_{2t}. Our point of departure is that the ideal instrument vector F_t is not observed. We assume that there is a "large" panel of data other than lags of the endogenous variables, $z_{1t}, \ldots, z_{Nt}$, that are weakly exogenous for β and are generated as follows:

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

The $r \times 1$ vector F_t 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}, and 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 have a common component and an idiosyncratic component.

2.1. Assumptions and Estimation of F_t

Although the variable z_{it}, like x_{2t}, is driven by F_t, e_{it} is uncorrelated with ε_t by assumption, 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. Although valid, z_{it} is a "noisy" instrument for each x_{2t} because the ideal instrument for x_{2t} is F_t. We cannot use F_t in estimation only because it is not observed. The idea is to use estimated F_t as instrument. When the context is clear, we will simply refer to F_t as instruments instead of as "factor-based instruments."

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 $z_t = (z_{1t}, z_{2t}, \ldots, z_{Nt})'$ be the $N \times 1$ vector of the instrumental variables and let $Z = (z_1, z_2, \ldots, z_T)$, which is $N \times T$. We define $F = (F_1, \ldots, F_T)'$ to be the $T \times r$ factor matrix and

$\Lambda = (\lambda_1, \ldots, \lambda_N)'$ to be the $N \times r$ factor loading matrix. The estimated factors, denoted $\tilde{F} = (\tilde{F}_1, \ldots, \tilde{F}_T)'$, are 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 $\tilde{\Lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)' = Z\tilde{F}/T$, which is $N \times r$, is an estimate for the factor loading matrix Λ. Let $\tilde{e} = Z - \tilde{\Lambda}\tilde{F}'$ be the residual matrix $(N \times T)$. Also let $\tilde{V}$ 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 components estimates associated with the factor model (3), and hatted variables are estimated from the regression model. The following assumption is concerned with the factor model (3).

Assumption A.

(a) $\mathrm{E}\|F_t\|^4 \leq M$ and $1/T \sum_{t=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F > 0$, an $r \times r$ nonrandom matrix.

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

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

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

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

(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 $\tilde{F}$ and $\tilde{\Lambda}$ as estimators for $F = (F_1, \ldots, F_T)'$ and $\Lambda = (\lambda_1, \ldots, \lambda_N)'$, respectively. Assumptions A(a) and A(b) imply the existence of r factors, as the largest r population eigenvalues of Σ_Z will increase with N, whereas the remaining eigenvalues are bounded; see Chamberlain and Rothschild (1983). We also assume that r is fixed, which is appropriate given that in the examples that motivate our analysis, the number of macroeconomic shocks (typically technology, hours worked, taste, and fiscal and monetary policy) is quite small. Although allowing r to increase with N and T is possible, the theoretical results on large-dimensional factor models available to date all assume that r is fixed.

The assumption that $\Lambda'\Lambda/N > 0$ means that there are r identifiable factors or, in this context, that the instruments are strong. The assumption does not, however, preclude the possibility that some of the series have weak factor loadings. Consider, for example, the case of one factor and $\lambda_i \sim N(0, \sigma_\lambda^2)$. Although many of the factor loadings will be close to zero, $\frac{1}{N}\sum_{i=1}^{N} \lambda_i^2$ has a positive limit. In contrast, the weak instrument setup of Kapetanios and Marcellino (2006) assumes $\lambda_i = \lambda_i^0/N^a$ for some $a \geq 0$. Under this assumption, also considered in Onatski (2006), there

can be little separation between the rth and the $(r+1)$th eigenvalues of Σ_Z, the population covariance matrix of Z. The factors are then not identifiable from the population eigenvalues for large a. It is debatable whether the strong or the weak factor structure is a better characterization of the macroeconomic panel data we work with. It should not, however, come as a surprise that the weak factor assumption would lead to different results. We proceed with the strong factor assumption as stated in Assumption A(b), which is also the assumption used in the majority of the work in this literature. The strong factor assumption alone is not enough to ensure that F_t is a relevant instrument. For this, we also need $\Psi'\Psi > 0$, which we assume.

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 independent and identically distributed (i.i.d.), then Assumptions 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 Assumptions A(a)–(c) hold. However, we assume no dependence between the factor loadings and the factors or between the factors and the idiosyncratic errors, and so on, 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. Define $\varepsilon_t(\beta) = y_t - x_t'\beta$ and let $\varepsilon_t = \varepsilon_t(\beta^0)$.

Assumption B.

(a) $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{E}|\varepsilon_t|^{4+\delta} < \infty$ for some $\delta > 0$. The vector process $g_t(\beta^0) = F_t^+\varepsilon_t$ satisfies $\mathrm{E}[g_t(\beta^0)] = 0$ with $\mathrm{E}[g_t(\beta)] \neq 0$ when $\beta \neq \beta^0$. Let $\bar{g}^0 = 1/T\sum_{t=1}^T F_t^+\varepsilon_t$ 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 $\Psi'\Psi > 0$, $\mathrm{E}(F_t u_t) = 0$, $\mathrm{E}(u_t\varepsilon_t) \neq 0$, and $\mathrm{E}(F_t\varepsilon_t) = 0$.

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

Part (a) of Assumption B 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 $T^{-1}\sum_{t=1}^T\sum_{s=1}^T \mathrm{E}[F_t^+F_s^{+\prime}\varepsilon_t\varepsilon_s]$. However, to focus on the main idea, we assume $F_t^+\varepsilon_t$ 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 nonzero loading on x_{2t} and that $\mathrm{E}(F_{jt}\varepsilon_t) = 0$ for each $j = 1, \ldots, r$. As both conditions hold under the assumption of our analysis, F_t is the ideal but infeasible instrument for x_{2t}.

The requirement that $\Psi'\Psi > 0$ means that the factors attribute a nondegenerate fraction of the variations in the endogenous variable in question. Under Assumption B(b), F_t is exogenous and relevant and hence satisfies instrument validity. Part (c) assumes that the correlation between the instruments and the endogenous regressor comes 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 2 in Section 2.2.

In empirical work, it is common practice to use past values of the observed variables as instruments. In the present setup, this can be justified only if x_{2t} is serially correlated (for instrument relevance) and ε_t must be uncorrelated with the past observations (for instrument validity).[1] 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 the lags of x_{2t}.

Lags of F_t should provide no further information about x_{2t} once conditioned on F_t. This raises the question of whether lags of x_{2t} have information beyond F_t, and this depends on u_t. Given the factor structure, lags of x_{2t} can be better instruments only if u_t contributes to the dynamics in x_{2t} and ε_t is uncorrelated with the lags of u_t.

2.2. A Feasible Factor Instrumental Variable Estimator

Although the ideal instrument under our setup is F_t, it is unobservable. We suggest using $\tilde{F}_t$ 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 $\tilde{F}_t$. It is understood that when x_{1t} is present, the results still go through upon replacing $\tilde{F}_t$ in the estimator that follows by $\tilde{F}_t^+ = (x_{1t}', \tilde{F}_t')'$.

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

$$\begin{aligned}\beta_{FIV}^* &= \underset{\beta}{\operatorname{argmin}}\ \bar{g}(\beta)' W_T \bar{g}(\beta) \\ &= (S_{\tilde{F}x}' W_T S_{\tilde{F}x})^{-1} S_{\tilde{F}x}' W_T S_{\tilde{F}y},\end{aligned}$$

where $S_{\tilde{F}x} = 1/T \sum_{t=1}^T \tilde{F}_t x_t'$. Let $\varepsilon_t^* = y_t - x_t' \beta_{FIV}^*$ and let $S^* = 1/T \sum_{t=1}^T \tilde{F}_t \tilde{F}_t' (\varepsilon_t^*)^2$. Then the efficient GMM estimator, which is our main focus, is to let $W_T = S^{*-1}$, giving

$$\hat{\beta}_{FIV} = (S_{\tilde{F}x}' S^{*-1} S_{\tilde{F}x})^{-1} S_{\tilde{F}x}' S^{*-1} S_{\tilde{F}'y}.$$

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

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

where $\Omega_{FIV} = \operatorname{plim}(S_{\tilde{F}x}' (S^*)^{-1} S_{\tilde{F}x})^{-1} = \Omega_{Fx}' (S^0)^{-1} \Omega_{Fx}$, *with* $\Omega_{Fx} = \operatorname{plim} 1/T \sum_{t=1}^T F_t x_t'$ *and* S^0 *as defined in Assumption B.*

Theorem 1 establishes consistency and asymptotic normality of the GMM estimator when $\tilde{F}_t$ are used as instruments and when the observed instruments are

not weak.[2] Just as if F_t were observed, $\hat{\beta}_{FIV}$ reduces to $(\tilde{F}'x)^{-1}\tilde{F}'y$ and is the instrumental variable estimator in an exactly identified model with $K = r$. It is the two-stage least squares (2SLS) estimator, that is, $\hat{\beta}_{FIV} = (x'P_{\tilde{F}}x)^{-1}x'P_{\tilde{F}}y$, under conditional homoskedasticity. Furthermore, $J = T\,\bar{g}(\hat{\beta}_{FIV})'\,S^{*-1}\,\bar{g}(\hat{\beta}_{FIV}) \xrightarrow{d} \chi^2_{r-K}$ 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 were observed. The procedure proposed by Carrasco (2006) is similar to ours, but no factor structure is assumed.[3] Other estimators such as those in Hausman, Newey, and Woutersen (2006), in addition to limited information maximum likelihood and jackknife instrumental variables estimator, can also be derived. Because $\tilde{F}_t$ can be used as though it were F_t, we expect that a factor-based version of these estimators will remain valid, but analyzing their properties is beyond the scope of this paper.

The essence behind Theorem 1 is that $\tilde{F}_t$ 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 that $\tilde{F}_t$ will lead to the same estimator (asymptotically only), we need to establish

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

This result is given in Lemma A1 in Appendix A. In fact, it can be shown that $\tilde{F}_t - HF_t$ is equal to $D\frac{1}{N}\sum_{i=1}^{N}\lambda_i e_{it}$ plus a term that is negligible, where the matrix D depends on N and T and is $O_p(1)$. Thus $T^{-1/2}\sum_{t=1}^{T}(\tilde{F}_t - HF_t)\varepsilon_t \simeq DN^{-1/2}\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda_i e_{it}\varepsilon_t$. If ε_t and e_{it} are independent, the left-hand side of (4) is $O_p(N^{-1/2}) = o_p(1)$.

Remark 1. Theorem 1 assumes that the number of factors r is known. The asymptotic distribution still holds with a consistent estimator $\hat{r}$.[4] Let $\hat{\beta}_{FIV,\hat{r}}$ denote the FIV estimator using an estimated r. To see that $\hat{\beta}_{FIV,\hat{r}}$ has the same limiting distribution as $\hat{\beta}_{FIV,r}$, consider

$$P(\sqrt{T}(\hat{\beta}_{FIV,\hat{r}} - \beta) \le s) = P(\sqrt{T}(\hat{\beta}_{FIV,\hat{r}} - \beta) \le s|\hat{r} = r)P(\hat{r} = r) + P(\sqrt{T}(\hat{\beta}_{FIV,\hat{r}} - \beta) \le s|\hat{r} \ne r)P(\hat{r} \ne r).$$

Because $P(\hat{r} = r) \to 1$ and $P(\hat{r} \ne r) \to 0$, the second term on the right-hand side converges to zero, and the first term is equal to $P(\sqrt{T}(\hat{\beta}_{FIV,\hat{r}} - \beta) \le s|\hat{r} = r)$ $[1 + o(1)]$. Furthermore, conditional on $\hat{r} = r$, $\hat{\beta}_{FIV,\hat{r}} = \hat{\beta}_{FIV,r}$. Thus

$$|P(\sqrt{T}(\hat{\beta}_{FIV,\hat{r}} - \beta) \le s) - P(\sqrt{T}(\hat{\beta}_{FIV,r} - \beta) \le s)| \to 0.$$

Remark 2. Theorem 1 is derived under the assumption that $\mathrm{E}(\varepsilon_t e_{it}) = 0$ for all i and t so that all instruments are valid. This assumption is, however, not necessary

under a data rich environment. Suppose that $\mathrm{E}(\varepsilon_t e_{it}) \neq 0$ for all i so that z_{it} cannot be used as instruments. When N is fixed, using z_t will not consistently estimate β. But with a large N and under the assumption that $\sum_{i=1}^{N} |\mathrm{E}(\varepsilon_t e_{it})| \leq M < \infty$ for all N with M not depending on N, Theorem 1 still holds provided that $\sqrt{T}/N \to 0$. To see this, let $\gamma_i = \mathrm{E}(e_{it}\varepsilon_t) \neq 0$. Then

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

$$+\sqrt{T}N^{-1}\sum_{i=1}^{N}\lambda_i\gamma_i.$$

The first term on the right-hand side is $N^{-1/2}O_p(1) = o_p(1)$. Because $\mathrm{E}\|\lambda_i\| \leq M$ 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 $\sqrt{T}/N \to 0$, 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 and there exists an i such that $\gamma_i \neq 0$.

Remark 3. 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 $\tilde{F}_t$ 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, consistent estimation will not be possible unless N is large when variables satisfying the condition of Remark 2 are permitted. This underscores the benefit of working in a data rich environment.

Remark 4. The single equation setup 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, and the endogenous regressor is hours worked. Let $\tilde{F}_{gt}$ be the $r_g \times 1$ vector of instruments for the gth equation, $g = 1, \ldots, G$, and let $r = \sum_g r_g$. Then g_t is an $r \times 1$ vector of stacked up moments. Assuming that for each $g = 1, \ldots, G$, the $r_g \times K_g$ moment matrix $\mathrm{E}(\tilde{F}_{gt}x_{gt}')$ 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 moments. Note that this need not be a block diagonal matrix. Likewise, $S_{\tilde{F}x}$ 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 $S_{\tilde{F}x}$ 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, (DGP), that is, $x_{2t} = \Psi' F_t + u_t$, the nonzero 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 were observed, we would estimate the reduced form for x_{2t} to yield fitted residuals $\hat{u}_t$. Then least squares estimation of

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

not only provides a test for endogeneity bias, it also provides estimates of β that are numerically identical to 2SLS 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. See 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 $\tilde{F}_t$ 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 1, we obtain $\tilde{F}_t$. In step 2, for each $i = 1, \ldots, K_2$, least squares estimation of

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

will yield $\sqrt{T}$ consistent estimates of Ψ_i, from which we obtain $\tilde{u}_t$. In step 3, least squares estimation of

$$y_t = x_{1t}'\beta_1 + x_{2t}'\beta_2 + \tilde{u}_t'\gamma + \varepsilon_t^u \qquad \textbf{(5)}$$

will yield $\sqrt{T}$ consistent estimates of β. It is straightforward to show that the estimate is again numerically identical to 2SLS with $\tilde{F}_t$ 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 instrumental variable estimator. With the instrumental variable estimator, 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 instrumental variable estimator.

2.4. Optimality of the Feasible FIV

In early work, Kloek and Mennes (1960) were concerned with situations when N is large relative to the given T (in their case, $T = 30$) so that the first-stage estimation is inefficient. 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. It can be shown that when each observed instrument is measured with error, then under Assumptions A and B, $\hat{\beta}_{FIV}$ is more efficient than $\hat{\beta}_{IV}$, which uses an equal (or greater) number of z_{2t} as instruments.[5] The intuition is straightforward. The observed instruments are the ideal instruments contaminated with errors, and $\tilde{F}$ 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}.

One can also construct a GMM estimator that directly uses all N observed instruments $z_t = (z_{1t}, \ldots, z_{Nt})'$. This estimator was considered in Meng, Hu, and Bai (2007) in the context of estimating "betas" in asset returns when the market return is measured with errors. Because of the large number of instruments, the bias of the GMM estimator can be large. Instead of the unconstrained weighting matrix $(Z'Z)^-$ (generalized inverse of $Z'Z$), they proposed using an identity weighting matrix in the presence of many instruments, which yields a $\sqrt{T}$ consistent estimator.

We now provide an analysis of an optimal GMM estimator, defined as a GMM estimator whose weighting matrix is constructed to exploit the factor structure in the data.[6] More precisely, the estimator uses N moment conditions $\mathrm{E}(z_t\varepsilon_t) = 0$ and a weighting matrix constructed as

$$W = \mathrm{E}(z_t z_t' \varepsilon_t^2) = \sigma_\varepsilon^2[\Lambda \Sigma_F \Lambda' + D],$$

where D is assumed to be diagonal for ease of analysis. We can estimate W by $\hat{W} = \hat{\sigma}_\varepsilon^2[\hat{\Lambda}\hat{\Sigma}_F\hat{\Lambda}' + \hat{D}]$, from which the inverse of $\hat{W}$ can be easily computed. The optimal GMM estimator becomes

$$\hat{\beta}_{GMM} = (X'Z\hat{W}^{-1}Z'X)^{-1}(X'Z\hat{W}^{-1}Z'Y).$$

Let $S_{zx} = (1/T)Z'X$. The asymptotic variance is given by

$$\Omega_{GMM} = \operatorname{plim}\left(S_{zx}'\hat{W}^{-1}S_{zx}\right)^{-1}.$$

PROPOSITION 1. *Assume that z_t is stationary and ε_t^2 is uncorrelated with $z_t z_t'$. Then*

(i) $\hat{\beta}_{GMM} - \beta^0 = O_p(N/T)$. If $N/T \to 0$, then

(ii) $\sqrt{T}(\hat{\beta}_{GMM} - \beta^0 - (N/T)d) \xrightarrow{d} N(0, \Omega_{GMM})$, *and*
(iii) $\Omega_{GMM} = \Omega_{FIV}$, *where* $(N/T)d$ *is the bias term given in the proof in Appendix B.*

In general, the optimal GMM estimator is biased and asymptotically inefficient, confirming the finite-sample results found in Meng et al. (2007). The estimator has a bias in the order of N/T. The estimator is consistent only if $N/T \to 0$. It is interesting to note that the inconsistency is not due to the estimation of a large-dimensional weighting matrix W. It is inconsistent when N and T are comparable even if W is known. The bias-corrected optimal GMM has the same asymptotic covariance as that of the FIV if $N/T \to 0$, in which case the FIV is as efficient as the bias-corrected optimal GMM. However, to obtain consistency and asymptotic normality, the FIV requires neither bias correction nor N/T going to zero. It is not difficult to show that the GMM estimator with an identity weighting matrix, although consistent and asymptotically normal, is also less efficient than the FIV. It would be interesting but more demanding to compare FIV with the estimator proposed recently by Kuersteiner and Okui (2007) that is based on the average predicted value of the endogenous variables.

The GMM estimator in Proposition 1 uses all of the instruments. Given N instruments, one has the option to choose a smaller subset of instruments. Let $K \leq N$ be the number of instruments chosen. Donald and Newey (2001) show that if $K \to \infty$ and $K^2/T \to 0$, then one obtains an unbiased and asymptotically efficient estimator with an asymptotic variance that equals Ω_{GMM}. When choosing a smaller number of instruments from a larger set, one faces the issue of which subset to use among the 2^N possible instrument sets. The results of Donald and Newey essentially assume that the ordering of instrument variables is known so that the information criterion can be used. The FIV estimator uses all of the instruments and does not require the instruments to be ordered. Instead of using a selection criterion, dimension reduction is achieved through principal components.

3. PANEL DATA AND LARGE SIMULTANEOUS EQUATIONS SYSTEM

In this section we consider a large panel data regression model in which all regressors are endogenous. The presence of exogenous or predetermined regressors is discussed later.[7] For $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$ with N and T both large, let

$$y_{it} = x_{it}'\beta + \varepsilon_{it}, \tag{6}$$

where x_{it} is $K \times 1$ and $\mathrm{E}(x_{it}\varepsilon_{it}) \neq 0$ for all i and t. The same framework was also used in Wooldridge (2005). Equation (6) could be in first differenced form, as in Arellano and Bond (1991). This is a large simultaneous equation system because

we allow $\mathrm{E}(x_{it}\varepsilon_{it}) \neq 0$. The pooled ordinary least squares (POLS) estimator

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

is inconsistent. Unlike with 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, provided the regressors are driven by the common factors,

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

Here, Λ_i is a matrix of $r \times K$, and 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 $\mathrm{E}(F_t\varepsilon_{it}) = 0$. 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

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

As an example, let y_{it} be factor demand by firm i. If x_{it} are factor prices facing firm i, or revenue of firm i, they will be determined simultaneously with y_{it}. The economic model fits into our framework if factor prices are correlated across firms and each firm's revenue covaries with the business cycle. Spatial and cross-country studies in which the regressors have common variations can also be considered.

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. As C_{it} is not available, it needs to be estimated. Let $X_i = (x_{i1}, x_{it}, \ldots, x_{iT})'$ be a $T \times K$ matrix of regressors for the ith cross-section unit, so that $X = (X_1, X_2, \ldots, X_N)$ is $T \times (NK)$. Let Λ be a $(NK) \times r$ matrix whereas F is $T \times r$. Let $\tilde{F}$ be the principal component estimate of F from the matrix XX', as explained in Section 2.1, with Z replaced by X. Let $\tilde{C}_{it} = \tilde{\Lambda}_i' \tilde{F}_t$, which is $K \times 1$.

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

$$\hat{\beta}_{PFIV} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{C}_{it}x_{it}'\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{C}_{it}y_{it}. \tag{7}$$

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

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

(b) $x_{it} = \Lambda_i' F_t + u_{it}$; $\mathrm{E}(u_{it}\varepsilon_{it}) \neq 0$; 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

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

THEOREM 2. *Suppose Assumptions A′ and B′ hold. As $N, T \to \infty$, we have*

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

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

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

where $\Omega_{PFIV} = \mathrm{plim}[S_{\tilde{x}\tilde{x}}]^{-1} S [S_{\tilde{x}\tilde{x}}]^{-1}$ *with* $S_{\tilde{x}\tilde{x}} = (NT)^{-1} \sum_{i=1}^{N} \tilde{C}_{it} x_{it}'$ *and* Δ_1^0 *and* Δ_2^0 *are defined in Appendix C.*

Theorem 2 establishes that the estimator $\hat{\beta}_{PFIV}$ is consistent for β as $N, T \to \infty$. Even though there are no instruments in the conventional sense, we can still consistently estimate the large simultaneous equations system under the model assumptions.[8] Because the bias is of order $\max[N^{-1}, T^{-1}]$, the effect of the bias on $\hat{\beta}_{PFIV}$ 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, $\tilde{C}_{it}$ contains u_{it}, which is correlated with ε_{it}, the underlying reason for biases. When T and N are of comparable magnitudes, $\hat{\beta}_{PFIV}$ 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 bias-corrected estimator can be considered to recenter the asymptotic distribution to zero if we assume that ε_{it} are serially uncorrelated.[9] Let

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

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

where $\tilde{u}_{it} = x_{it} - \tilde{C}_{it}$, $\hat{\varepsilon}_{it} = y_{it} - x_{it}' \hat{\beta}_{PFIV}$, and $S_{\tilde{x}\tilde{x}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{C}_{it} x_{it}'$. The estimated bias is[10]

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

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*

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

Both $\hat{\beta}_{PFIV}$ 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, although $\hat{\beta}_{PFIV}$ is expected to be sufficiently precise in terms of the mean squared errors, the bias-corrected estimator, $\hat{\beta}^+_{PFIV} = \hat{\beta}_{PFIV} - \hat{\Delta}$ should provide more accurate inference in terms of the t statistic because it is properly recentered around zero.

Remark 5. The analysis is easily extended to models with fixed effects. All that is needed is to demean the data first and then proceed as usual. Consider

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}, \qquad x_{it} = \mu_i + \lambda'_i F_t + u_{it}.$$

Demeaning gives

$$\dot{y}_{it} = \dot{x}'_{it}\beta + \dot{\varepsilon}_{it}, \qquad \dot{x}_{it} = \lambda'_i \dot{F}_t + \dot{u}_{it},$$

where $\dot{y}_{it} = y_{it} - \bar{y}_i$ with $\bar{y}_i = T^{-1}\sum_{t=1}^T y_{it}$ and other dotted variables are defined in the same manner. The instrument is now $\dot{C}_{it} = \lambda'_i \dot{F}_t$. The limiting distribution also has the same form as before, except that variables are demeaned. Because T is large by assumption, the bias induced by demeaning is negligible. To analyze the limiting distribution, $\dot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$ can be replaced by ε_{it}. This follows from the result that $\sum_{t=1}^T \dot{F}_t \equiv 0$ so that

$$(NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T \dot{C}_{it}\dot{\varepsilon}_{it} \equiv (NT)^{-1/2}\sum_{i=1}^N\sum_{t=1}^T \dot{C}_{it}\varepsilon_{it}. \tag{8}$$

Similar to Assumption B'(c), under the assumption that the right-hand side in expression (8) has a normal limiting distribution,[11] say, $N(0, \dot{S})$, Theorem 2 still holds with limiting variance

$$\dot{\Omega}_{PFIV} = \text{plim}[\dot{S}_{\tilde{x}\tilde{x}}]^{-1}\dot{S}[\dot{S}_{\tilde{x}\tilde{x}}]^{-1},$$

where $\dot{S}_{\tilde{x}\tilde{x}} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widetilde{\dot{C}_{it}}\dot{x}'_{it}$. The detailed analysis will not be presented. Suffice it to mention that once the data are demeaned, exactly the same computation is performed including the bias correction.

Remark 6. The PFIV estimator is different from the traditional panel instrumental variable estimator that uses $\tilde{F}$ as instruments. Such an estimator, PTFIV, would be constructed as

$$\hat{\beta}_{PTFIV} = \left(S'_{\tilde{F}x}S^{*-1}S_{\tilde{F}x}\right)^{-1}S'_{\tilde{F}x}S^{*-1}S_{\tilde{F}y},$$

where $S_{\tilde{F}x} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{F}_t x_{it}'$ and $S^* = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{F}_t\tilde{F}_t'(e_{it}^*)^2$; e_{it}^* is based on a preliminary estimate of β using an $r \times r$ positive definite weighting matrix. However, the probability limit of $S_{\tilde{F}x}$ is $\Sigma_{Fx} = \mathrm{E}(\lambda_i)'\Sigma_F$, which can be singular if $\mathrm{E}(\lambda_i) = 0$, and in that case the estimator is only $\sqrt{T}$ consistent. The $\hat{\beta}_{PTFIV}$ 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 $\tilde{F}^+ = [x_1 \;\; \tilde{F}]$ as instruments, where $\tilde{F}$ is $T \times r$.[12] We also consider an estimator with $\tilde{f}^+ = [x_1 \;\; \tilde{f}]$ as instruments, where the dimension of $\tilde{f}$ is $T \times r\max$ with $r\max > r$. This estimator is denoted as fIV. The GMM estimator uses an identity weighting matrix in the first step to yield β^*. For the sake of comparison, we also report results of two other estimators. The first is a GMM estimator using a set of observed variables most closely correlated with x_2 and is of the same dimension as $\tilde{F}$. These instruments are determined by the R^2 from regressions of x_2 on both x_1 and one instrument. This estimator is labeled IV. The second is ordinary least squares (OLS), which does not account for endogeneity bias.

We consider three DGPs. In all cases,

$$z_{it} = \lambda_{iz}'F_t + \sqrt{r}\sigma_z e_{it},$$

$$F_{jt} = \rho_j F_{jt-1} + \eta_{jt} \qquad j = 1, \ldots r,$$

where $e_{it} \sim N(0,1)$, $\eta_{jt} \sim N(0,1)$, $\lambda_{iz} \sim N(0, I_r)$, $\rho_j \sim U(0.2, 0.8)$, 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

$$y_t = x_{1t}'\beta_1 + x_{2t}'\beta_2 + \sigma_y\varepsilon_t,$$

$$x_{i1t} = \alpha_x x_{i1,t-1} + v_{it}, \qquad i = 1, \ldots, K_1,$$

$$x_{i2t} = \lambda_{i2}'F_t + u_{it}, \qquad i = 1, \ldots, K_2,$$

where $\varepsilon_t = \frac{1}{\sqrt{2}}(\tilde{\varepsilon}_t^2 - 1)$ and $u_{it} = \frac{1}{\sqrt{2}}(\tilde{u}_{it}^2 - 1)$; $(\tilde{\varepsilon}_t, \tilde{u}_t')' \sim N(0_{K_2+1}, \Sigma)$ where $\mathrm{diag}(\Sigma) = 1$, $\Sigma(j,1) = \Sigma(1,j) \sim U(0.3, 0.6)$, and zero for other entries. This means that $\tilde{\varepsilon}_t$ is correlated with $\tilde{u}_{it}$ with covariance $\Sigma(1,i)$ but $\tilde{u}_{it}$ and $\tilde{u}_{jt}$ are uncorrelated $(i \neq j)$. We assume $\alpha_x \sim U(0.2, 0.8)$, $v_{it} \sim N(0,1)$ and uncorrelated with $\tilde{u}_{jt}$ and $\tilde{\varepsilon}_t$. By construction, the errors are heteroskedastic. The parameter σ_y^2 is set to be $K_1\bar{\sigma}_{x1}^2 + K_2\bar{\sigma}_{x2}^2$, where $\bar{\sigma}_{x_j}^2$ is the average variance of x_{jt}, $j = 1, 2$. This puts the noise-to-signal ratio in the primary equation to be of roughly one-half.

The parameter of interest is β_2. We consider 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 because 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 0.2. The true value of β_2 is 2, and the impact of endogeneity bias on OLS is immediately obvious. The estimators that use the factors as instruments are more precise. The factor-based instruments dominate the IV either in bias or root mean squared error (RMSE), if not both. The J test associated with the FIV is close to the nominal size of 5%, whereas the two-sided t statistic for testing $\beta_2 = 2$ has some size distortion when N and T are both small. The size distortions of both tests decrease with T.

Example 2

In this example, the regression model is

$$y_t = \beta_1 + x_{2t}'\beta_2 + \varepsilon_t. \tag{9}$$

The endogenous variables x_{2t} are spanned by L factors, whereas the panel of observed instruments is spanned by r factors and $r \geq L$. To generate data with

TABLE 1. Finite-sample properties of $\hat{\beta}_2$, $\beta_2^0 = 2$

					Mean/RMSE							
T	N	r	$r$max	$\rho_{x_2\varepsilon}$	FIV	fIV	IV	OLS	J_F	t_F	J_f	t_f
50	50	1	2	0.38	1.97	2.00	2.18	2.73	N.A.	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	N.A.	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	N.A.	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	N.A.	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 $\tilde{F}$ and $\tilde{f}$ as instruments. These are of dimensions r and $r$max, respectively. Here 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. The N.A. entries correspond to exact identification (no overidentifying restrictions).

this structure, 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

$$y = F(:,1:L)\Lambda_y + \sigma_y e_y,$$

$$X_2 = F(:,1:L)\Lambda_x + e_x,$$

where $e_{j,xt} \sim N(0,\sigma_j^2)$, $\sigma_j^2 = L$ $(j = 1,\ldots,L)$, $\sigma_y^2 = L$, and $e_{yt} \sim N(0,1)$. The factors F_t are AR(1) processes with dynamic coefficients uniformly distributed between 0.2 and 0.8. The L-dimensional factors $F(:,1:L)$ can be expressed as $F(:,1:L) = (X_2 - e_x)\Lambda_x^{-1}$. Thus

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

where $\beta_2^* = \Lambda_x^{-1}\Lambda_y$ is $L \times 1$ and $\varepsilon = \sigma_y e_y - e_x\beta_2^*$. 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 (9) 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 elements of the $L \times L$ matrix Λ_x are drawn from the $N(1,1)$ distribution. Written in terms of r factors, $X_2 = F(:,1:r)\Lambda_x^{(r)} + e_x$, where $\Lambda_x^{(r)}$ only has the first $L \times L$ positions being nonzero. Viewed this way, the first L factors are the relevant factors.

We estimate $r\max = r + 2$ factors and report simulations for $K_2 = 1$ with $\beta_2^0 = 1$. The results are reported in Table 2. Unlike in 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 instrumental variable estimators again perform well.

Example 3

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

$$\begin{aligned} y_{it} &= \mu_i + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \varepsilon_{it}, \\ x_{it,2} &= \lambda_i' F_t + \sqrt{r}\, u_{it}, \end{aligned}$$

$$\begin{pmatrix} \varepsilon_{it} \\ u_{it} \end{pmatrix} i.i.d.\ N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix} \right),$$

where $x_{it,1} = 1$ for all i; $\rho_i \sim U(0.3, 0.6)$. We set the true value of $\beta = (\beta_1, \beta_2)' = (0,1)'$ and draw $\mu_i \sim U(0,1)$ and $\lambda_i \sim N(0, I_r)$; F_t are generated as in earlier examples. For each i, the data are demeaned to control for fixed effects. An intercept is included in the regression in the demeaned data. According to Theorem 2, we can use the factors estimated from x_{it} (also demeaned) to instrument themselves. For the PFIV, we use r factors. We also consider an estimator, denoted PfIV, that uses $r\max = r + 2$ factors. Note that these estimates are not corrected for bias to show that the bias is of second-order importance. For the sake of comparison, we

TABLE 2. Finite-sample properties of $\hat{\beta}_2$, $\beta_2^0 = 1$

					Mean/RMSE							
T	N	r	L	$\rho_{x_2\varepsilon}$	FIV	fIV	IV	OLS	J_F	t_F	J_f	t_f
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 $\tilde{F}$ and $\tilde{f}$ as instruments. These are of dimensions r and $r\max = r+2$, respectively. Here 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.

also consider PTFIV. Note that in this example, $\mathrm{E}(\lambda_i) = 0$ and the PTFIV should be more volatile (larger variance) because $S_{\tilde{F}x}$ can be near singular.

The results are reported in Table 3. As expected, the POLS estimator is quite severely biased. The PTFIV has noticeably larger RMSE than the four 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 are still size distortions when r is large. The t statistics based on OLS 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 mean squared error, 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. Although the results do not show significant

TABLE 3. Finite-sample properties of $\hat{\beta}_2$ for panel data, $\beta_2^0 = 1$

				Mean/RMSE								
T	N	r	$\rho_{x_2\varepsilon}$	PFIV	PFIV$^+$	PfIV	PfIV$^+$	PTFIV	POLS	$t_{\hat{\beta}_{PFIV}}$	$t_{\hat{\beta}_{PFIV^+}}$	$t_{\hat{\beta}_{PTFIV}}$
15	15	2	0.29	1.06	1.04	1.09	1.08	1.12	1.11	0.41	0.25	0.12
				0.07	0.06	0.10	0.09	0.22	0.12			
25	25	2	0.29	1.03	1.01	1.06	1.04	1.08	1.11	0.37	0.13	0.08
				0.04	0.03	0.07	0.05	0.18	0.11			
25	50	2	0.30	1.02	1.01	1.05	1.03	1.08	1.11	0.38	0.11	0.08
				0.03	0.02	0.05	0.04	0.19	0.11			
50	25	2	0.29	1.02	1.01	1.04	1.03	1.07	1.10	0.37	0.11	0.12
				0.03	0.02	0.05	0.03	0.13	0.10			
50	50	2	0.29	1.01	1.00	1.03	1.02	1.06	1.10	0.31	0.07	0.08
				0.02	0.01	0.04	0.02	0.13	0.10			
100	50	2	0.30	1.01	1.00	1.02	1.01	1.05	1.10	0.34	0.07	0.10
				0.01	0.01	0.03	0.01	0.10	0.10			
50	100	2	0.29	1.01	1.00	1.03	1.01	1.04	1.10	0.34	0.07	0.06
				0.01	0.01	0.03	0.02	0.13	0.10			
100	100	2	0.29	1.01	1.00	1.02	1.01	1.04	1.10	0.29	0.06	0.08
				0.01	0.01	0.02	0.01	0.11	0.10			
15	15	4	0.29	1.07	1.06	1.08	1.07	1.09	1.08	0.84	0.70	0.18
				0.07	0.06	0.08	0.08	0.14	0.09			
25	25	4	0.29	1.05	1.03	1.06	1.05	1.07	1.08	0.91	0.56	0.18
				0.05	0.04	0.06	0.05	0.12	0.08			
25	50	4	0.29	1.04	1.02	1.05	1.04	1.06	1.08	0.91	0.46	0.12
				0.04	0.02	0.05	0.04	0.11	0.08			
50	25	4	0.28	1.03	1.02	1.05	1.03	1.05	1.08	0.88	0.39	0.16
				0.04	0.02	0.05	0.03	0.08	0.08			
50	50	4	0.29	1.02	1.01	1.04	1.02	1.04	1.08	0.88	0.23	0.12
				0.02	0.01	0.04	0.02	0.08	0.08			
100	50	4	0.29	1.02	1.00	1.03	1.01	1.03	1.08	0.89	0.18	0.14
				0.02	0.01	0.03	0.02	0.06	0.08			
50	100	4	0.28	1.02	1.00	1.03	1.01	1.03	1.08	0.88	0.19	0.08
				0.02	0.01	0.03	0.02	0.08	0.08			
100	100	4	0.29	1.01	1.00	1.02	1.01	1.03	1.08	0.85	0.11	0.11
				0.01	0.00	0.02	0.01	0.06	0.08			

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

difference, using too many factors can sometimes increase bias but may reduce mean squared error. Whether we use estimated factors or Z as instruments, it is an open issue how to select the most relevant instruments from many valid ones that have no natural ordering. This problem, along with empirical applications, is considered in a companion paper, Bai and Ng (2009).

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 a starting point that in a data rich environment, there can be 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 the endogenous regressors. 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. It is important to emphasize again that having a large quantity of data will not solve all instrumental variables problems. Our assumptions require a factor structure with factors being valid instruments. Practitioners still need to ascertain that these assumptions are satisfied.

NOTES

1. When ε_t is serially correlated of unknown form, the lags of x_{2t} cannot be used as instruments because x_{2t-j} is correlated with ε_{t-j}, which is correlated with ε_t.

2. Irrelevant instruments are allowed in the sense that some factor loadings λ_i can be zero. All that is 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 are of $O(N^{-\alpha})$ ($\alpha > 0$) necessitates a different asymptotic framework and is considered in Kapetanios and Marcellino (2006).

3. In the Carrasco (2006) analysis, the instrument variables are transformed and ordered via the principal components method, and the number of principal components is selected by minimizing the mean squared errors.

4. Like all analysis that requires premodel selection, the critique of Leeb and Potscher (2008) applies. In particular, the finite-sample distributions of postmodel selection estimators typically depend on unknown model parameters in a complicated fashion. The convergence of the finite-sample distributions to their large-sample limits is typically not uniform with respect to the underlying parameters. Therefore, the asymptotic distribution can be a poor approximation for the finite-sample distributions for certain DGPs.

5. The proof is given in an earlier version of the paper.

6. Meng et al. (2007) also explored a weighting matrix that exploits the factor structure of asset returns. The resulting GMM estimator did not have good finite-sample properties, and the theoretical properties of the estimator were not explored.

7. See notes 8 and 10.

8. 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 $\tilde{F}$ and $\tilde{\Lambda}$ from x_2 alone. Then the pooled 2SLS is simply $\hat{\beta}_{PFIV} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{Z}_{it}x_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{Z}_{it}y_{it}$ where $\tilde{Z}_{it} = (x_{1it}', \tilde{C}_{it}')'$. Equation (7) can be written alternatively as $\hat{\beta}_{PFIV} = \left(\sum_{i=1}^{N}X_i'P_{\tilde{F}}X_i\right)^{-1}\sum_{i=1}^{N}X_i'P_{\tilde{F}}Y_i$ where $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ is $(T \times 1)$. This follows from the fact that $(\tilde{C}_{i1}, \tilde{C}_{i2}, \ldots,$

$\tilde{C}_{iT})' = P_{\tilde{F}} X_i = \tilde{F}\tilde{\Lambda}_i$. However, this representation is not easily amendable in the presence of additional regressors x_{1it}.

9. It is possible to construct bias-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 $\hat{\Upsilon}$ must have a convergence rate satisfying $\sqrt{N/T}(\hat{\Upsilon} - \Upsilon) = o_p(1)$. Assuming $T^{1/4}$ $(\hat{\Upsilon} - \Upsilon) = o_p(1)$, this implies the requirement that $N/T^{3/2} \to 0$ instead of $N/T^2 \to 0$ under no serial correlation.

10. In the presence of exogenous regressors x_{1it} as in note 8, the corresponding terms become $\hat{\Delta}_1 = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{Z}_{it}x_{it}'\right)^{-1}\begin{bmatrix}0\\ \hat{\delta}_1\end{bmatrix}$ and $\hat{\Delta}_2 = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{Z}_{it}x_{it}'\right)^{-1}\begin{bmatrix}0\\ \hat{\delta}_2\end{bmatrix}$. A small-sample adjustment can also be made by using $NT - (N+T)r$ instead of NT when computing $\hat{\delta}_1$ and $\hat{\delta}_2$, where $r(N+T)$ is the number of parameters used to estimate $\tilde{u}_{it}$.

11. Under Assumptions A′ and B', expression (8) also has a normal limiting distribution if either $N/T \to 0$, or $\mathrm{E}(\Lambda_i) = 0$ for all i, or $\mathrm{E}(F_s\varepsilon_{it}) = 0$ for all t and s. We thank the coeditor for pointing this out.

12. 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. Because the estimated r is consistent for r, r can be treated as known.

REFERENCES

Amemiya, T. (1966) On the use of principal components of independent variables in two-stage least squares estimation. *International Economic Review* 7, 283–303.

Amemiya, T. (1985) *Advanced Econometrics*. Harvard University Press.

Andrews, D., M. Moreira, & J. Stock (2006) Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* 74, 715–754.

Arellano, M. & S. Bond (1991) Some specification tests for panel data models: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58, 277–298.

Bai, J. (2003) Inferential theory for factor models of large dimensions. *Econometrica* 71, 135–172.

Bai, J. & S. Ng (2002) Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.

Bai, J. & S. Ng (2009) Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics* 1(1), article 4 (online).

Bekker, P.A. (1994) Alternative approximations to the distributions of instrumental variables estimators. *Econometrica* 63, 657–681.

Bernanke, B. & J. Boivin (2003) Monetary policy in a data rich environment. *Journal of Monetary Economics* 50, 525–546.

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

Chamberlain, G. & M. Rothschild (1983) Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1281–2304.

Chao, J. & N. Swanson (2005) Consistent estimation with a large number of instruments. *Econometrica* 73, 1673–1692.

Donald, S. & W. Newey (2001) Choosing the number of instruments. *Econometrica* 69, 1161–1192.

Favero, C. & M. Marcellino (2001) Large Datasets, Small Models, and Monetary Europe. IGIER, Working paper 208.

Forni, M., M. Hallin, M. Lippi, & L. Reichlin (2005) The generalized dynamic factor model, one sided estimation and forecasting. *Journal of the American Statistical Association* 100, 830–840.

Hahn, J. & G. Kuersteiner (2002) Discontinuities of weak instrument limiting distributions. *Economics Letters* 75, 325–331.

Hallin, M. & R. Liska (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., W. Newey, & T. Woutersen (2006) IV Estimation with Heteroskedasticity and Many Instruments. Manuscript, MIT.
Hayashi, F. (2000) *Econometrics*. Princeton University Press.
Kapetanios, G. & M. Marcellino (2006) Factor-GMM Estimation with Large Sets of Possibly Weak Instruments. Manuscript, Queen Mary University.
Kloek, T. & L. Mennes (1960) Simultaneous equations estimation based on principal components of predetermined variables. *Econometrica* 28, 46–61.
Kuersteiner, G. & R. Okui (2007) Estimator Averaging for Two stage Least Squares. Manuscript, University of California, Davis.
Leeb, H. & B. Potscher (2008) Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24, 338–376.
Meng, G., G. Hu, & J. Bai (2007) A Simple method for Estimating Betas when Factors are Measured with Error. Mimeo, Boston College.
Moreira, M. (2003) A conditional likelihood ratio test for structural models. *Econometrica* 71, 1027–1048.
Onatski, A. (2006) Asymptotic Distribution of the Principal Components Estimator of Large Factor Models when Factors Are Relatively Weak. Mimeo, Columbia University.
Stock, J.H. & M.W. Watson (2002) Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–1179.
Wooldridge, J. (2005) Instumental variables estimation with panel data. *Econometric Theory* 21, 865–869.

APPENDIX A: Properties of the FIV

To prove the main result we need the following lemma.

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

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

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

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

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

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

Proof. 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 that of part (ii), and the bound is tightened by using the uncorrelation assumption. The details are omitted. ■

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

$$\hat{\beta}_{FIV} - \beta^0 = (S'_{\tilde{F}x}S^{*-1}S_{\tilde{F}x})^{-1}S'_{\tilde{F}x}S^{*-1}\bar{g}.$$

Now

$$\sqrt{T}\bar{g} = T^{-1/2}\sum_{t=1}^{T}\tilde{F}_t\varepsilon_t$$

$$= T^{-1/2}\sum_{t=1}^{T}(\tilde{F}_t - HF_t)\varepsilon_t + HT^{-1/2}\sum_{t=1}^{T}F_t\varepsilon_t$$

$$= HT^{-1/2}\sum_{t=1}^{T}F_t\varepsilon_t + o_p(1).$$

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

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

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

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

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

APPENDIX B: Proof of Proposition 1

We first show that $\Omega_{GMM} = \Omega_{FIV}$ if $N/T \to 0$. For simplicity, we assume W is known. The idea is that even with a known weighting matrix, the optimal GMM is no more efficient than FIV. It can be shown that the same result holds with estimated W. The matrix Ω_{GMM}^{-1} is the limit of $(X'Z/T)W^{-1}(Z'X/T)$. From $Z = F\Lambda' + e$ with $e = (e_1, e_2, \ldots, e_T)$, we can write

$$(X'Z/T)W^{-1}(Z'X/T) = (X'F/T)\Lambda'W^{-1}\Lambda(F'X/T) + (X'e/T)W^{-1}(e'X/T)$$
$$+(X'F/T)\Lambda'W^{-1}(e'X/T) + (X'e/T)W^{-1}\Lambda(F'X/T)$$
$$= a + b + c + d.$$

We will show that the first term has a limit that is the inverse of the asymptotic variance of the FIV and the last three terms are each $o_p(1)$.

For the first term, from

$$W^{-1} = \sigma_\varepsilon^{-2}\left\{D^{-1} - D^{-1}\Lambda[\Sigma_F^{-1} + \Lambda'D^{-1}\Lambda]^{-1}\Lambda'D^{-1}\right\},$$

we have

$$\sigma_\varepsilon^2 \Lambda' W^{-1} \Lambda = A - A[\Sigma_F^{-1} + A]^{-1} A = A\Big(A^{-1} - [\Sigma_F^{-1} + A]^{-1}\Big)A,$$

where $A = \Lambda' D^{-1} \Lambda$. Using $A^{-1} - (A+B)^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1} A^{-1}$ (see Amemiya, 1985, p. 461) and with $B = \Sigma_F^{-1}$,

$$\Lambda' W^{-1} \Lambda = \sigma_\varepsilon^{-2}[\Sigma_F + (\Lambda' D^{-1} \Lambda)^{-1}]^{-1} = \sigma_\varepsilon^{-2} \Sigma_F^{-1} + O(N^{-1}),$$

because $(\Lambda' D^{-1} \Lambda)^{-1} = O(N^{-1})$, which is dominated by Σ_F. Noting that $X'F/T \xrightarrow{p} \Omega_{xF}$, we have

$$(X'F/T)\Lambda' W^{-1} \Lambda (F'X/T)\sigma_\varepsilon^{-2} \xrightarrow{p} \sigma_\varepsilon^{-2} \Omega_{xF} \Sigma_F^{-1} \Omega_{Fx}.$$

The preceding expression is equal to the inverse of the asymptotic matrix of the FIV estimator; see the proof of Theorem 1. That is,

$$\Omega_{FIV}^{-1} = \sigma_\varepsilon^{-2} \Omega_{xF} \Sigma_F^{-1} \Omega_{Fx}$$

since $S^0 = \sigma_\varepsilon^2 \Sigma_F$ under homoskedasticity of ε_t.

For term b, again using the expression of W^{-1},

$$(X'e/T)W^{-1}(e'X/T) = \frac{1}{T^2} X'eD^{-1}e'X$$

$$-\frac{1}{T} X'eD^{-1}\Lambda[\Sigma_F^{-1} + \Lambda' D^{-1}\Lambda]^{-1}\Lambda' D^{-1} e'X/T = b_1 + b_2.$$

Consider b_1.

$$b_1 = \frac{1}{T^2} X'eD^{-1}e'X = \frac{N}{T}\frac{1}{N}\sum_{i=1}^{N}\left(T^{-1/2}\sum_{t=1}^{T}\frac{1}{\sigma_{i,e}} x_t e_{it}\right)\left(T^{-1/2}\sum_{t=1}^{T}\frac{1}{\sigma_{i,e}} x_t' e_{it}\right)$$

$$= O_p(N/T) = o_p(1)$$

if $N/T \to 0$. Next, consider b_2. Note that

$$\frac{1}{T} X'eD^{-1}\Lambda = (N/T)^{1/2}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{i,e}^{-2} x_t \lambda_i' e_{it} = O_p((N/T)^{1/2}).$$

Moreover, $\Sigma_F^{-1} + \Lambda' D^{-1}\Lambda \geq \Lambda' D^{-1}\Lambda = O(N)$. Thus b_2 is bounded in norm by $O_p(N/T)O(N^{-1}) = O_p(1/T)$.

Consider c. Again, let $A = \Lambda' D^{-1}\Lambda = O(N)$. Omitting σ_ε^2,

$$\Lambda' W^{-1} = \Lambda' D^{-1} - A(\Sigma_F^{-1} + A)^{-1}\Lambda' D^{-1} = [I - A(\Sigma_F^{-1} + A)^{-1}]\Lambda' D^{-1}$$

$$= A[A^{-1} - (\Sigma_F^{-1} + A)^{-1}]\Lambda' D^{-1} = (\Sigma_F + A^{-1})^{-1} A^{-1}\Lambda' D^{-1}$$

$$= [\Sigma_F + O(N^{-1})]^{-1}(\Lambda' D^{-1}\Lambda/N)^{-1}\frac{1}{N}\Lambda' D^{-1} = O(1)\frac{1}{N}\Lambda' D^{-1}.$$

Thus c can be written as

$$c = (X'F/T)O_p(1)\frac{1}{NT}\Lambda' D^{-1}e'X = O_p\left(\frac{1}{\sqrt{NT}}\right)\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{i,e}^{-2}\lambda_i x_t' e_{it}$$

$$= O_p\left(\frac{1}{\sqrt{NT}}\right).$$

Finally, d has the same order of magnitude as c. In summary, we have shown that, when $N/T \to 0$,

$$\Omega_{GMM}^{-1} = \text{plim}(X'Z/T)W^{-1}(Z'X/T) = \Omega_{FIV}^{-1}.$$

Consistency. We next show that $\hat{\beta}_{GMM}$ is inconsistent if $N/T \to c > 0$, even if the optimal weighting matrix is known. Notice that

$$\hat{\beta}_{GMM} - \beta^0 = (X'ZW^{-1}Z'X)^{-1}(X'ZW^{-1}Z'\varepsilon).$$

It was shown earlier that $\Omega_{GMM}^{-1} = \text{plim}T^{-2}X'ZW^{-1}Z'X = \Omega_{FIV}^{-1}$ if $N/T \to c = 0$. If $c > 0$ but bounded, its limit becomes $\Omega_{FIV}^{-1} + \Upsilon$, where Υ is the limit of $T^{-2}X'eD^{-1}e'X$. We now argue that

$$T^{-2}X'ZW^{-1}Z'\varepsilon = O_p(N/T). \tag{B.1}$$

Again, from $Z = F\Lambda' + e$, the left-hand side of the preceding expression can be expressed as the sum of four terms:

$$T^{-2}X'ZW^{-1}Z'\varepsilon = (X'F/T)\Lambda' W^{-1}\Lambda(F'\varepsilon/T) + T^{-2}X'eW^{-1}e'\varepsilon$$

$$+ T^{-2}X'F\Lambda' W^{-1}e'\varepsilon + T^{-2}X'eW^{-1}\Lambda F'\varepsilon$$

$$= I_1 + I_2 + I_3 + I_4.$$

From $X'F/T = O_p(1)$, $\Lambda' W^{-1}\Lambda = O_p(1)$, and $F'\varepsilon/T = O_p(T^{-1/2})$, the term I_1 is $O_p(T^{-1/2})$. In fact, $X'F/T \to \Omega_{xF}$, $\Lambda' W^{-1}\Lambda \to \sigma_\varepsilon^{-2}\Sigma_F^{-1}$, and $F'\varepsilon/\sqrt{T} \to N(0, \sigma_\varepsilon^2\Sigma_F)$. Thus, we have

$$\sqrt{T}\, I_1 \xrightarrow{d} N(0, \Omega_{xF}\sigma_\varepsilon^{-2}\Sigma_F^{-1}\Omega_{Fx}) =^d N(0, \Omega_{FIV}^{-1}).$$

Consider I_2 and assume $\sigma_\varepsilon^2 = 1$ for simplicity. Analogous to the analysis for the term b for the case of $N/T \to 0$,

$$I_2 = \frac{1}{T^2}X'eD^{-1}e'\varepsilon - O_p(1/T).$$

But

$$\frac{1}{T^2}X'eD^{-1}e'\varepsilon = \left(\frac{N}{T}\right)\frac{1}{N}\sum_{i=1}^{N}\left[\left(T^{-1/2}\sum_{t=1}^{T}\frac{1}{\sigma_{i,e}}x_t e_{it}\right)\left(T^{-1/2}\sum_{t=1}^{T}\frac{1}{\sigma_{i,e}}\varepsilon_t e_{it}\right)\right]$$

$$= \frac{N}{T}\left(\frac{1}{N}\sum_{i=1}^{N}\xi_i\eta_i\right),$$

where ξ_i and η_i are implicitly defined. Note that ξ_i and η_i are dependent because x_t and ε_t are dependent. Let $\gamma = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}(\xi_i \eta_i)$, and thus $\mathrm{E}(I_2) = (N/T)\gamma$. This implies that the bias term is proportional to the number of instruments, a well-known result, at least for the case of fixed N. Thus, $I_2 = O_p(N/T)$.

Consider I_3. The analysis is the same as that of c given earlier. From $\Lambda' W^{-1} = O(1) \times (1/N)\, \Lambda' D^{-1}$, I_3 can be written as

$$I_3 = (X'F/T)O_p(1)\frac{1}{NT}\Lambda' D^{-1} e'\varepsilon = O_p\left(\frac{1}{\sqrt{NT}}\right)\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{i,e}^{-2}\lambda_i e_{it}\varepsilon_t$$

$$= O_p\left(\frac{1}{\sqrt{NT}}\right).$$

Next consider I_4. The transpose of $\Lambda' W^{-1}$ gives

$$W^{-1}\Lambda = \frac{1}{N} D^{-1}\Lambda O(1).$$

Thus term I_4 can be written as

$$\frac{1}{TN} X' e D^{-1}\Lambda O_p(1) F'\varepsilon/T = O_p\left(\frac{1}{\sqrt{NT}}\right)\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{i,e}^{-2} x_t \lambda_i e_{it}\right)\left(\frac{1}{T}\sum_{t=1}^{T} F_t \varepsilon_t\right)$$

$$= O_p\left(\frac{1}{T\sqrt{N}}\right),$$

which is dominated by I_3. In summary

$$\hat{\beta}_{GMM} - \beta^0 = \left(\frac{1}{T^2} X' Z W^{-1} Z' X\right)^{-1}$$

$$\times \left[I_1 + \frac{N}{T}\left(\frac{1}{N}\sum_{i=1}^{N}\xi_i\eta_i\right) + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right)\right],$$

where we recall that $\sqrt{T} I_1 \xrightarrow{d} N(0, \Omega_{FIV}^{-1})$. The second term in square brackets is $O_p(N/T)$, showing inconsistency of optimal GMM when $N/T \to c > 0$.

Limiting Distribution of the Bias-Corrected Optimal GMM. Assume $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\xi_i\eta_i - \gamma) = O_p(1)$, where $\gamma = \mathrm{E}(\xi_i\eta_i)$, so that

$$\hat{\beta}_{GMM} - \beta^0 - \frac{N}{T} d = \hat{\Omega}\left[I_1 + O_p(\sqrt{N}/T) + O_p(T^{-1}) + O_p\left(\frac{1}{\sqrt{NT}}\right)\right],$$

where $\hat{\Omega}$ stands for $(T^{-2} X' Z W^{-1} Z' X)^{-1}$ and $d = \hat{\Omega}\gamma$. If $N/T \to 0$, the last three terms in brackets multiplied by $T^{1/2}$ are all $o_p(1)$. Now $\hat{\Omega} \xrightarrow{p} \Omega_{GMM}$ by definition and $\sqrt{T} I_1 \xrightarrow{d} N(0, \Omega_{GMM}^{-1})$ because $\Omega_{FIV} = \Omega_{GMM}$ for $N/T \to 0$, as shown earlier. It follows that if $N/T \to 0$,

$$\sqrt{T}(\hat{\beta}_{GMM} - \beta^0 - \frac{N}{T} d) \xrightarrow{d} N(0, \Omega_{GMM}). \qquad \blacksquare$$

APPENDIX C: Properties of PFIV

Proof of Theorem 2. (i) We shall show $\hat{\beta}_{PFIV} - \beta = O_p(T^{-1}) + O_p(N^{-1})$, or equivalently, $\sqrt{NT}(\hat{\beta}_{PFIV} - \beta) = O_p(\sqrt{N/T}) + O_p(\sqrt{T/N})$. From $\hat{\beta}_{PFIV} = \beta + S_{\tilde{x}\tilde{x}}^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{C}_{it}\varepsilon_{it}$, it is sufficient to consider the limit of $(NT)^{-1/2} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{C}_{it}\varepsilon_{it}$. 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

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

Notice

$$\begin{aligned}\hat{C}_{it} - C_{it} &= \tilde{\Lambda}_i'\tilde{F}_t - \Lambda_i' F_t = (\tilde{\Lambda}_i - H'^{-1}\Lambda_i)'\tilde{F}_t + \Lambda_i' H^{-1}(\tilde{F}_t - HF_t) \\ &= (\tilde{\Lambda}_i - H'^{-1}\Lambda_i)'(\tilde{F}_t - HF_t) + (\tilde{\Lambda}_i - H'^{-1}\Lambda_i)' HF_t + \Lambda_i' H^{-1}(\tilde{F}_t - HF_t).\end{aligned}$$

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

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

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

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

Note that ε_{it} is scalar and 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 preceding expression

$$\begin{aligned}&(NT)^{-1/2} \sum_{i=1}^{N} \sum_{t=1}^{T} \Lambda_i' H^{-1}(\tilde{F}_t - HF_t)\varepsilon_{it} \\ &\quad = (T/N)^{1/2} \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Lambda_i'\varepsilon_{it} \right) H^{-1} G \left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \sum_{k=1}^{K} \lambda_{j,k} u_{jt,k} \right) + o_p(1) \\ &\quad = (T/N)^{1/2} O_p(1). \qquad \textbf{(C.1)}\end{aligned}$$

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

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

Thus

$$(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{\Lambda}_i - H^{'-1}\Lambda_i)'HF_t\varepsilon_{it}$$

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

$$= (N/T)^{1/2}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}u_{is}F_s'\right)H'H\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t\varepsilon_{it}\right) + o_p(1)$$

$$= (N/T)^{1/2}O_p(1). \qquad \textbf{(C.2)}$$

Combining (C.1) and (C.2), we prove part (i) of the theorem.

(ii) The bias is equal to $S_{\tilde{x}\tilde{x}}^{-1}$ multiplied by the sum of the expected values of (C.1) and (C.2). We analyze these expected values subsequently. 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 (C.1) is $A_tH^{-1}GB_t$, which is a vector. Thus

$$A_tGB_t = \text{vec}(A_tH^{-1}GB_t) = (B_t'\otimes A_t)\text{vec}(H^{-1}G).$$

It follows that (again ignoring the $o_p(1)$ term)

$$\text{(C.1)} = (T/N)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}(B_t'\otimes A_t\right)\text{vec}(H^{-1}G).$$

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

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

Let

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

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

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

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

$$\text{(C.1)} \to \tau^{1/2}\delta_1^0.$$

Next consider (C.2). Let

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

then (C.2) can be rewritten as (ignoring the $o_p(1)$ term)

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

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 $\frac{1}{N}\sum_{i=1}^{N}[(\Phi_i'\otimes\Theta_i) - \text{E}(\Phi_i'\otimes\Theta_i)] = O_p(N^{-1/2})$, we have

$$\text{(C.2)} = (N/T)^{1/2}\delta_2 + o_p(1),$$

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

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

Let

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

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

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

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 proof of Theorem 2 shows that

$$\sqrt{NT}(\hat{\beta}_{PFIV}-\beta) = S_{\tilde{x}\tilde{x}}^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}C_{it}\varepsilon_{it} + \sqrt{T/N}S_{\tilde{x}\tilde{x}}^{-1}\delta_1 + \sqrt{N/T}S_{\tilde{x}\tilde{x}}^{-1}\delta_2 + o_p(1). \tag{C.3}$$

It can be shown that $\hat{\Delta}_1 - S_{\tilde{x}\tilde{x}}^{-1}\delta_1 = O_p(\delta_{NT}^{-1})$ and $\hat{\Delta}_2 - S_{\tilde{x}\tilde{x}}^{-1}\delta_2 = O_p(\delta_{NT}^{-1})$. These imply that $(T/N)^{1/2}(\hat{\Delta}_1 - S_{\tilde{x}\tilde{x}}^{-1}\delta_1) = o_p(1)$ if $T/N^2\to 0$ and $(N/T)^{1/2}(\hat{\Delta}_2 - S_{\tilde{x}\tilde{x}}^{-1}\delta_2) = o_p(1)$ if $N/T^2\to 0$. Thus, we can replace $S_{\tilde{x}\tilde{x}}^{-1}\delta_1$ by $\hat{\Delta}_1$ and replace $S_{\tilde{x}\tilde{x}}^{-1}\delta_2$ by $\hat{\Delta}_2$ in (C.3). Equivalently,

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

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