

# Instrumental Variable Estimation in a Data Rich Environment

Jushan Bai\*      Serena Ng †

April 28, 2008

## 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 and they are more efficient than the observed variables in our framework. While standard optimal GMM using a large number of instruments is biased and can be inconsistent, 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 with invalid instruments, provided that the unobserved common components are valid instruments. 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. While 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.

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, Columbia University, 420 W 118 St., IAB 1019, New York, NY 10027 Email: Serena.Ng@columbia.edu.

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 (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. 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 example, Andrews et al. (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 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.

It is well recognized that use of too many potentially relevant instruments in the first stage of two-stage least squares estimation will induce bias. This motivates Kloeck 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 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), Chao and Swanson (2005), among many others. In those analysis, no structure is imposed on the many and weak instruments. Also, while 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. 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 co-movements in a large number of variables. As will be discussed again below, 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.

As far as we are aware, Kapetanios and Marcellino (2006) is the only other paper that considers using estimated factors as instruments. An incomplete draft of their paper was brought to our attention when the first draft of our paper was circulated. The scope of their paper has since been broadened to extend our work to allow the endogenous regressors to depend on variables other than the factors. The primary difference, however, is that their 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 below, 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 compare efficiency of the FIV with the traditional optimal GMM. 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 at least as efficient as the biased corrected optimal GMM. A further point of departure is that we consider high dimensional simultaneous equations system in which there exist no valid instrument in the conventional sense. Invalid instruments are allowed in our analysis even in the single equation framework,

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  $\Psi'$  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. 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  $E(F_t\varepsilon_t) = 0$  but  $E(u_t\varepsilon_t) \neq 0$ . This induces a non-zero correlation between  $x_{2t}$  and  $\varepsilon_t$ . 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,  $z_{1t}, \dots, z_{Nt}$  that are weakly exogenous for  $\beta$  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,  $\lambda_i$  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  $\varepsilon_t$ . 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.

There are at least two ways to motivate using the common factors as instruments. The first is that co-movements observed in economic time series arise because of common shocks. Early work by Sargent and Sims (1977) recognized that the NBER's framework for business cycle analysis fits naturally into the framework of index (factor) models. They illustrated this (p.50) using the simple multiplier-accelerator model where consumption (C), investment (I), and government spending (G) all depends on GDP, so that any subset of  $(GDP, G, C, I)$  forms a one (real) index model. Adding an equation for the price level, and equations for the demand and supply of money adds another (nominal) index. This index model setup motivated much of the recent interest in large dimensional factor models, except that the recent literature considers not just a handful of series, but one or two hundred variables. A large number of studies have documented evidence for common static and dynamic factors in macroeconomic data, both for the U.S. and abroad. See Breitung and Eickmeier (2005) and Reichlin (2003) for a review of recent empirical work.

The factor model 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. As such, the data are repeated measures of latent variables observed with errors. For example, let  $y_t$  be an asset return. According to asset pricing theory,  $y_t$  depends on the market portfolio,  $R_{mt}^*$ , which is not observed. We observe a proxy  $R_{mt}$  (such as the SP 500 index), where  $R_{mt} = R_{mt}^* + u_t$ . A regression of  $y_t$  on  $R_{mt}$  will give inconsistent estimates due to a classical measurement error problem. But theory suggests that other assets  $z_{it}$  are also determined by  $R_{mt}^*$ . That is to say,  $z_{it} = \lambda_i' R_{mt}^* + e_{it}$ . Putting  $F_t = R_{mt}^*$  and  $x_{2t} = R_{mt}$  yields the econometric model under consideration. 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.

The analysis of Boivin and Giannoni (2006) shows how a factor model can emerge from a DSGE model in a data rich environment. Let  $Y_t$  be a vector of endogenous variables, which together with some predetermined variables and exogenous shocks form a (linearized) rational

expectations system. Let  $S_t$  be a vector of state variables that include the predetermined and exogenous shocks. The solution for  $Y_t$  to the rational expectations system is linear in  $S_t$ , i.e.  $Y_t = \Lambda_Y S_t$ . When  $Y_t$  is measured with errors denoted  $e_{Y,t}$ , the reduced form solution is  $Y_t = \Lambda_Y S_t + e_{Y,t}$ . Let  $y_{1t}$  be one of the elements of  $Y_t$  and let  $Y_{2t}$  be a sub-vector of  $Y_t$  (excluding  $y_{1t}$ ), and let  $W_t$  be a subset of predetermined variables. Consider estimating the linear model  $y_{1t} = \beta' Y_{2t} + \alpha' W_t + \varepsilon_t$ . By definition, the regressor  $Y_{2t}$  is endogenous (with or without measurement errors). Boivin and Giannoni argue that there is a large number of indicator variables  $X_t$  for  $S_t$  (e.g., asset prices, commodity prices, monetary aggregates, output, etc.) such that  $X_t = \Lambda_X S_t + e_{X,t}$ . In our framework,  $x_{2t}$  corresponds to  $Y_{2t}$ ,  $W_t$  corresponds to  $x_{1t}$ , while  $z_{it}$  corresponds to the elements of  $X_t$ , and  $F_t$  corresponds to  $S_t$ .

## 2.1 Assumptions and Estimation of $F_t$

Although the variables  $z_{it}$ , like  $x_{2t}$ , are driven by  $F_t$ ,  $e_{it}$  is uncorrelated with  $\varepsilon_t$  by assumption, and  $z_{it}$  is correlated with  $x_{2t}$  through  $F_t$ . Thus,  $z_{it}$  is weakly exogenous for  $\beta$ , 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  $F_t$  in estimation only because it is not observed. The idea is to use estimated  $F_t$  as instrument.

We estimate the factors from a panel of instruments  $z_{it}$ ,  $i = 1, \dots, N, t = 1, \dots, T$ , by the method of principal components. Let  $z_t = (z_{1t}, z_{2t}, \dots, z_{Nt})'$  be the  $N \times 1$  vector of the instrumental variables, and let  $Z = (z_1, z_2, \dots, z_T)$ , which is  $N \times T$ . We define  $F = (F_1, \dots, F_T)'$  to be the  $T \times r$  factor matrix, and  $\Lambda = (\lambda_1, \dots, \lambda_N)'$  to be the  $N \times r$  factor loading matrix. The estimated factors, denoted  $\tilde{F} = (\tilde{F}_1, \dots, \tilde{F}_T)'$ , 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  $\tilde{\Lambda} = (\tilde{\lambda}_1, \dots, \tilde{\lambda}_N)'$  is  $N \times r$ , is an estimate for the factor loading matrix  $\Lambda$ . 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 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.  $E\|F_t\|^4 \leq M$  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.  $\lambda_i$  is either deterministic such that  $\|\lambda_i\| \leq M$ , or it is stochastic such that  $E\|\lambda_i\|^4 \leq M$ . In either case,  $N^{-1}\Lambda'\Lambda \xrightarrow{p} \Sigma_\Lambda > 0$ , a  $r \times r$  non-random matrix, as  $N \rightarrow \infty$ .
- c.i  $E(e_{it}) = 0$ ,  $E|e_{it}|^8 \leq M$ .
- c.ii  $E(e_{it}e_{js}) = \sigma_{ij,ts}$ ,  $|\sigma_{ij,ts}| \leq \bar{\sigma}_{ij}$  for all  $(t, s)$  and  $|\sigma_{ij,ts}| \leq \tau_{ts}$  for all  $(i, j)$  such that  $\frac{1}{N} \sum_{i,j=1}^N \bar{\sigma}_{ij} \leq M$ ,  $\frac{1}{T} \sum_{t,s=1}^T \tau_{ts} \leq M$ , and  $\frac{1}{NT} \sum_{i,j,t,s=1}^N |\sigma_{ij,ts}| \leq M$ .
- c.iii For every  $(t, s)$ ,  $E|N^{-1/2} \sum_{i=1}^N [e_{is}e_{it} - E(e_{is}e_{it})]|^4 \leq M$ .
- 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  $F = (F_1, \dots, F_T)'$  and  $\Lambda = (\lambda_1, \dots, \lambda_N)'$ , respectively. These assumptions imply the existence of  $r$  factors, as  $r$  population eigenvalues of  $\Sigma_Z$  will increase with  $N$ , while 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, 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  $a \geq 1/2$ . Under this assumption, also considered in Onatski (2006), there is little separation between the  $r$  and the  $r+1$  eigenvalue of  $\Sigma_Z$ , the population covariance matrix of  $Z$ . The factors are then not identifiable from the population eigenvalues under this assumption. It is debatable whether the strong or the weak factor structure is a better characterization of 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 (b), which is also the assumption used in the majority of the work in this literature.

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,  $\lambda_i$  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  $\beta^0$  denote the true value of  $\beta$ . Define  $\varepsilon_t(\beta) = y_t - x'_t\beta$  and let  $\varepsilon_t = \varepsilon_t(\beta^0)$ .

### 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  $E(g_t(\beta)) \neq 0$  when  $\beta \neq \beta^0$ . 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  $E(u_t \varepsilon_t) \neq 0$ .
- 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  $\beta^0$ . In general,  $S^0$  is the limit of  $T^{-1} \sum_{t=1}^T \sum_{s=1}^T E[F_t^+ F_s^{+'} \varepsilon_t \varepsilon_s]$ . 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  $T^{-1} \sum_{t=1}^T F_t^+ F_t^{+'} \varepsilon_t^2$ . Heteroskedasticity of  $\varepsilon_t$  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  $\Psi'\Psi > 0$  means that the factors attribute a non-degenerate fraction of the variations in the endogenous variable in question. 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 2 below.

In empirical work, it is common practice to use past values of the observed variables as instruments. To justify this, 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  $\varepsilon_t$  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.<sup>1</sup>

As 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  $u_t$  contribute to the dynamics in  $x_{2t}$ . If  $u_t$  is serially uncorrelated, lags of  $x_{2t}$  may be weakly correlated with  $x_{2t}$ . In this case, the problem of weak observed instruments can be circumvented by using  $F_t$  as instruments directly.

## 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  $\beta$ . The reason why the moments  $g_t$  are not used to estimate  $\beta$  is that  $F_t$  is not observed. We suggest to use  $\tilde{F}_t$  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 below by  $\tilde{F}_t^+ = (x'_{1t}, \tilde{F}_t)'$ .

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

$$\begin{aligned} \check{\beta}_{FIV} &= \underset{\beta}{\operatorname{argmin}} \bar{g}(\beta)' W_T \bar{g}(\beta) \\ &= (S'_{\tilde{F}_x} W_T S_{\tilde{F}_x})^{-1} S'_{\tilde{F}_x} W_T S_{\tilde{F}_y} \end{aligned}$$

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

$$\hat{\beta}_{FIV} = (S'_{\tilde{F}_x} \check{S}^{-1} S_{\tilde{F}_x})^{-1} S'_{\tilde{F}_x} \check{S}^{-1} S_{\tilde{F}_y}.$$

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

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

---

<sup>1</sup>When  $x_{1t}$  is strongly exogenous such that  $E(x_{1t} \varepsilon_s) = 0$  for all  $t$  and  $s$ ,  $\varepsilon_t$  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  $\varepsilon_t$  is serially correlated, the lags of  $x_{2t}$  cannot be used as instruments since  $x_{2t-j}$  can be correlated with  $\varepsilon_{t-j}$ , which is correlated with  $\varepsilon_t$ .

where  $\Omega_{FIV} = \text{plim}(S'_{\tilde{F}_x}(\check{S})^{-1}S_{\tilde{F}_x})^{-1} = \Omega'_{F_x}(S^0)^{-1}\Omega_{F_x}$ , with  $\Omega_{F_x} = \text{plim} \frac{1}{T} \sum_{t=1}^T F_t x'_t$  and  $S^0$  being defined earlier.

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.<sup>2</sup> Just as if  $F_t$  was 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, i.e.,  $\hat{\beta}_{FIV} = (x'P_{\tilde{F}}x)^{-1}x'P_{\tilde{F}}y$ , under conditional homoskedasticity. Furthermore,  $J = T\bar{g}(\hat{\beta}_{FIV})' \check{S}^{-1} \bar{g}(\hat{\beta}_{FIV}) \xrightarrow{d} \chi^2_{r-K}$  is asymptotically  $\chi^2$  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. The procedure proposed by Carrasco (2006) is similar to ours, but no factor structure is assumed. 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. Other estimators such as those in Hausman et al. (2006), as well as LIML and JIVE, can also be derived. Since  $\tilde{F}_t$  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  $\tilde{F}_t$  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  $\tilde{F}_t$  will lead to the same estimator (asymptotically only), we need to establish

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

This result is given in Lemma 1 in the appendix. 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  $\varepsilon_t$  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  $\hat{r}$ . Let  $\hat{\beta}_{FIV,\hat{r}}$  denote the FIV estimator

---

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

using an estimated  $r$ . To see that  $\widehat{\beta}_{FIV,\widehat{r}}$  has the same limiting distribution as  $\widehat{\beta}_{FIV,r}$ , consider

$$\begin{aligned} P(\sqrt{T}(\widehat{\beta}_{FIV,\widehat{r}} - \beta) \leq s) &= P(\sqrt{T}(\widehat{\beta}_{FIV,\widehat{r}} - \beta) \leq s | \widehat{r} = r)P(\widehat{r} = r) \\ &\quad + P(\sqrt{T}(\widehat{\beta}_{FIV,\widehat{r}} - \beta) \leq s | \widehat{r} \neq r)P(\widehat{r} \neq r). \end{aligned}$$

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

$$|P(\sqrt{T}(\widehat{\beta}_{FIV,\widehat{r}} - \beta) \leq s) - P(\sqrt{T}(\widehat{\beta}_{FIV,r} - \beta) \leq s)| \rightarrow 0.$$

**Remark 2:** Theorem 1 is derived under the assumption that  $E(\varepsilon_t e_{it}) = 0$  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  $\sqrt{T}/N \rightarrow 0$ . To see this, let  $\gamma_i = E(e_{it}\varepsilon_t) \neq 0$ . 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  $N^{-1/2}O_p(1) = o_p(1)$ . 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  $\sqrt{T}/N \rightarrow 0$ , 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  $(\sqrt{T}/N) \sum_{i=1}^N |\gamma_i| \rightarrow 0$ . This would be impossible when  $N$  is fixed as long as there exists an  $i$  such that  $\gamma_i \neq 0$ .

**Remark 3:** The assumption that  $N \rightarrow \infty$  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 2 is permitted, consistent estimation will not be possible unless  $N$  is large. The benefit of working in a data rich environment cannot be overlooked.

**Remark 4:** The single equation set up extends naturally to a system of equations. Suppose there are  $G$  equations, where  $G$  is finite. For  $g = 1, \dots, G$ , and  $t = 1, \dots, T$ ,

$$y_{gt} = x'_{gt}\beta_g + \varepsilon_{gt}$$

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  $\tilde{F}_{gt}$  be the  $r_g \times 1$  vector of instruments for the  $g$ -th equation,  $g = 1, \dots, G$ , 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  $g = 1, \dots, G$ , the  $r_g \times K_g$  moment matrix  $E(\tilde{F}_{gt}x'_{gt})$  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 moment conditions. 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, ie  $x_{2t} = \Psi'F_t + u_t$ , the non-zero correlation between  $x_{2t}$  and  $\varepsilon_t$  arises because  $\text{cov}(u_t, \varepsilon_t) \neq 0$ . We can decompose  $\varepsilon_t$  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  $\hat{u}_t$ . Then least squares estimation of

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

not only provides a test for endogeneity bias, it also provides estimates of  $\beta$  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. 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 one, we obtain  $\tilde{F}_t$ . In step 2, for each  $i = 1, \dots, K_2$ , least squares estimation of

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

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

$$y_t = x'_{1t}\beta_1 + x'_{2t}\beta_2 + \tilde{u}'_t\gamma + \varepsilon_t^u \quad (5)$$

will yield  $\sqrt{T}$  consistent estimates of  $\beta$ . 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 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.

## 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, 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  $r_2 \times 1$  observed instruments, where  $r_2 (\geq r)$  is arbitrary but fixed. Let  $m_t = z_{2t}(y_t - x'_t\beta)$  with  $\sqrt{T}\bar{m} \xrightarrow{d} N(0, Q)$ . Let  $\hat{\beta}_{IV}$  be the minimizer of*

$\bar{m}'(\check{Q})^{-1}\bar{m}$  with the property that  $\sqrt{T}(\hat{\beta}_{IV} - \beta^0) \xrightarrow{d} N(0, \Omega_{IV})$ . If  $\text{var}(e_{it}) \geq c > 0$  for all  $i$  in the  $z_{2t}$  set, then as  $N, T \rightarrow \infty$ ,

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

where  $\Omega_{FIV}$  is the asymptotic variance of  $\hat{\beta}_{FIV}$  defined in Theorem 1.

An IV estimator that uses a large number of instruments is highly biased and can be inconsistent. For example, if  $N \geq T$ , then 2SLS becomes OLS, which is inconsistent. We therefore compare two estimators whose bias is of the same order. Proposition 1 says that when each observed instrument is measured with error, then in a data rich environment,  $\hat{\beta}_{FIV}$  is more efficient than  $\hat{\beta}_{IV}$ , which uses an equal (or more) number of  $z_{2t}$  as instruments. The intuition is straightforward. The observed instruments are the ideal instruments contaminated with errors while  $\tilde{F}$  is consistent for the ideal instrument space. Pooling information across the observed variables washes out the noise to generate more efficient instruments for  $x_{2t}$ . Proposition 1 rules out cases when the observed variables are perfect instruments (i.e.,  $\text{var}(e_{it}) = 0$ ). This may seem restrictive, but is not unrealistic as researchers cannot be expected to isolate the perfect instruments, even if they exist. Alternatively, the strict inequality can be replaced by the result  $\Omega_{IV} \geq \Omega_{FIV}$  to allow instruments with  $\text{var}(e_{it}) = 0$ . If  $x_{2t} = \Psi'F_t + \Gamma'W_t + u_t$ , where  $W_t$  is an observable, exogenous, and fixed-dimensional vector, the natural instruments will be  $[\hat{F}, W]$  and the comparison instruments will be  $[z_2, W]$ . Then Proposition 1 still holds with strict inequality replaced by a non-strict one. Finally, when the model assumptions do not hold, such as if the factor structure is weak (e.g., factor loadings  $\lambda_i = O_p(N^{-1/2}) \rightarrow 0$  as  $N$  increases), Proposition 1 will not necessarily hold.

It is also of interest to compare the FIV with an IV estimator that directly uses all  $N$  observed instruments  $z_t = (z_{1t}, \dots, z_{Nt})'$ . This GMM estimator was considered in Meng et al. (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), they proposed to use an identity weighting matrix in the presence of many instruments, which yields a  $\sqrt{T}$ -consistent estimator. They also considered a weighting matrix that exploits the factor structure of asset returns. The resulting estimator did not have good finite sample properties, and the theoretical properties of the estimator were not explored.

Here, we 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. More

precisely, the estimator uses  $N$  moment conditions  $E(z_t \varepsilon_t) = 0$  and a weighting matrix constructed as

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

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

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

Let  $S_{zx} = \frac{1}{T} Z' X$ . The asymptotic variance is given by

$$\Omega_{GMM} = \text{plim} \left( S'_{zx} \widehat{W}^{-1} S_{zx} \right)^{-1}.$$

**Proposition 2** *Assumed that  $z_t$  is stationary and  $\varepsilon_t^2$  is uncorrelated with  $z_t z_t'$ . Then (i)  $\widehat{\beta}_{GMM} - \beta^0 = O_p(N/T)$ . If  $N/T \rightarrow 0$ , then (ii)  $\sqrt{T}(\widehat{\beta}_{GMM} - \beta^0 - \frac{N}{T}d) \xrightarrow{d} N(0, \Omega_{GMM})$ , and (iii)  $\Omega_{GMM} = \Omega_{FIV}$ , where  $(N/T)d$  is the bias term given in the proof.*

In 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  $N/T \rightarrow 0$ . 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  $N/T \rightarrow 0$ , 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, while 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 by Kuersteiner and Okui (2007). Their estimator is based on the average predicted value of the endogenous variables.

### 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 other than its own lags. For  $i = 1, 2, \dots, N, t = 1, 2, \dots, T$  with  $N$  and  $T$  both large, let

$$y_{it} = x'_{it} \beta + \varepsilon_{it} \tag{6}$$

where  $x_{it}$  is  $K \times 1$  with  $E(x_{it}\varepsilon_{it}) \neq 0$  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). As this is a large simultaneous equation system since we allow  $E(x_{it}\varepsilon_{it}) \neq 0$ . The pooled OLS estimator

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

is inconsistent. Unlike 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 influenced by the common factors,

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

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

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

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 co-vary 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. Again,  $C_{it}$  is not available, but it can be estimated. Let  $X_i = (x_{i1}, x_{it}, \dots, x_{iT})'$  be a  $T \times K$  matrix of regressors for the  $i$ th cross-section unit, so that  $X = (X_1, X_2, \dots, X_N)$  is  $T \times (NK)$ . Let  $\Lambda$  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

$$\widehat{\beta}_{PFIV} = \left( \sum_{i=1}^N \sum_{t=1}^T \widetilde{C}_{it}x'_{it} \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T \widetilde{C}_{it}y_{it}. \quad (7)$$

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  $\lambda_i$  replaced by  $\Lambda_i$ ; 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  $\delta > 0$ ;  $\varepsilon_{it}$  are independent over  $i$ .
- b.  $x_{it} = \Lambda_i' F_t + u_{it}$ ;  $E(u_{it}\varepsilon_{it}) \neq 0$ ;  $\varepsilon_{it}$  is independent of  $F_t$  and  $\Lambda_i$ .
- 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

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

**Theorem 2** Suppose Assumptions A' and B' hold. As  $N, T \rightarrow \infty$ , 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  $T/N \rightarrow \tau > 0$ , then

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

where  $\Omega_{PFIV} = \text{plim}[S_{\widetilde{x}\widetilde{x}}]^{-1} S [S_{\widetilde{x}\widetilde{x}}]^{-1}$  with  $S_{\widetilde{x}\widetilde{x}} = (NT)^{-1} \sum_{i=1}^N \widetilde{C}_{it}' x'_{it}$ , and  $\Delta_1^0$  and  $\Delta_2^0$  are defined in the appendix.

Theorem 2 establishes that the estimator  $\widehat{\beta}_{PFIV}$  is consistent for  $\beta$  as  $N, T \rightarrow \infty$ . Remarkably, there can be no instrument in the conventional sense, yet, we can still consistently estimate the large simultaneous equations system.<sup>3</sup> In a very rich data environment, the

<sup>3</sup> This estimator can be easily extended to include additional regressors that are uncorrelated with  $\varepsilon_{it}$ . 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

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

where  $\widetilde{Z}_{it} = (x'_{1it}, \widetilde{C}'_{it})'$ . It is noted that equation (7) can be written alternatively as

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

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

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  $\varepsilon_{it}$ , 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  $T$  if we assume that  $\varepsilon_{it}$  are serially uncorrelated.<sup>4</sup> Let

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

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

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<sup>5</sup>

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

**Corollary 1** *Suppose Assumptions A' and B' hold. If  $\varepsilon_{it}$  are serially uncorrelated,  $T/N^2 \rightarrow 0$ , and  $N/T^2 \rightarrow 0$ , then*

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

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

<sup>4</sup>It is possible to construct biased-corrected estimators when  $\varepsilon_{it}$  is serially correlated. The bias correction involves estimating a long-run covariance matrix, denoted by  $\Upsilon$ . 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  $N/T^{3/2} \rightarrow 0$  instead of  $N/T^2 \rightarrow 0$  under no serial correlation.

<sup>5</sup>In the presence of exogenous regressors  $x_{1it}$  as in footnote 3, the corresponding terms become

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

A small sample adjustment can also be made by using  $NT - (N + T)r$  instead of  $NT$  when computing  $\widehat{\delta}_1$  and  $\widehat{\delta}_2$ , where  $r(N + T)$  is the number of parameters used to estimate  $\widehat{u}_{it}$ .

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.

**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

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

De-meaning gives

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

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 demeaned model has the same form as before. The instrument is now  $\dot{C}_{it} = \lambda'_i \dot{F}_t$ . The limiting distribution also has the same form, except that variables are demeaned. In analyzing the limiting distribution,  $\dot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$  can be replaced by  $\varepsilon_{it}$ . This follows from the result that  $\sum_{t=1}^T \dot{F}_t \equiv 0$  so that

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

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

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

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

**Remark 6:** 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

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

where  $S_{\widetilde{F}x} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \widetilde{F}_t x'_{it}$ , and  $\check{S} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \widetilde{F}_t \widetilde{F}'_t \check{c}_{it}$ ,  $\check{c}_{it}$  is based on a preliminary estimate of  $\beta$  using a  $r \times r$  positive definite weighting matrix. However, the probability limit of  $S_{\widetilde{F}x}$  is  $\Sigma_{F_x} = E(\lambda_i)' \Sigma_F$ , 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  $\Sigma_{F_x}$ . 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  $T \times r$ .<sup>6</sup> 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 fIV. The GMM estimator uses an identity weighting matrix in the first step to yield  $\check{\beta}$ . 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  $\tilde{F}$ . 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,

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

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

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

with  $\varepsilon_t = \frac{1}{\sqrt{2}}(\tilde{\varepsilon}_t^2 - 1)$  and  $u_{it} = \frac{1}{\sqrt{2}}(\tilde{u}_{it}^2 - 1)$ . We assume  $\alpha_x \sim U(.2, .8)$ ,  $v_{it} \sim N(0, 1)$  and uncorrelated with  $\tilde{u}_{jt}$  and  $\tilde{\varepsilon}_t$ . Furthermore,  $(\tilde{\varepsilon}_t, \tilde{u}_t)' \sim N(0_{K_2+1}, \Sigma)$  where  $diag(\Sigma) = 1$ ,  $\Sigma(j, 1) = \Sigma(1, j) \sim U(.3, .6)$ , and zero elsewhere. 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 ( $i \neq j$ ). By construction, the errors are heteroskedastic. The parameter  $\sigma_y^2$  is set to  $K_1\bar{\sigma}_{x_1}^2 + K_2\bar{\sigma}_{x_2}^2$  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  $\beta_2$ . We considered various values of  $K_2$ ,  $\sigma_z$ , and  $r$ . The results are reported in Table 1 with  $K_2 = 1$ , and  $\sigma_z = 3$ . This is the least favorable situation

---

<sup>6</sup>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  $r$ ,  $r$  can be treated as known.

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  $\varepsilon$  and thus indicates the degree of endogeneity. Under the assumed parametrization, this correlation is around .2. The true value of  $\beta_2$  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 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

$$y_t = \beta_1 + x_{2t}'\beta_2 + \varepsilon_t. \quad (8)$$

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 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  $\beta_2^* = \Lambda_x'^{-1}\Lambda_y$  is  $L \times 1$  and  $\varepsilon^* = e_y - e_x\beta_2^*$ . For given  $\Lambda_x$ , we then solve for  $\Lambda_y$  such that  $\beta_2^* = (1'_{K_2}, 0'_{L-K_2})$ . The  $x_{2t}$  in (8) corresponds to the first  $K_2$  columns of  $X_{2t}$ . This also implies that the true value of every element of  $\beta_2$  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  $\Lambda_x$  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  $\beta_2^0 = 1$ . The results are reported in Table 2. Unlike Example 1, the correlation between  $x_{2t}$  and  $\varepsilon_t$  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  $\beta$  by panel regressions. The DGP is

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

where  $F_t$  is again  $r \times 1$ ,  $\rho_i$  is the correlation between  $\varepsilon_{it}$  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 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  $S_{\bar{F}_x}$  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, are considered in a companion 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	na	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	na	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	na	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	na	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  $\hat{\beta}_2$ :  $\beta_2^0 = 1$

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  $\tilde{F}$  and  $\tilde{f}$  as instruments. These are of dimensions  $r$  and  $r_{max} = r + 2$ , 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  $\hat{\beta}_2$  for panel data,  $\beta_2^0 = 1$ .

T	N	r	$\rho_{x_2\varepsilon}$	PFIV	PFIV <sup>+</sup>	PfIV	PfIV <sup>+</sup>	PTFIV	POLS	$t_{\hat{\beta}_{PFIV}}$	$t_{\hat{\beta}_{PFIV^+}}$	$t_{\hat{\beta}_{PTFIV}}$
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  $\tilde{C}_{it} = \tilde{\lambda}'_i \tilde{F}_t$  and  $\tilde{c}_{it} = \tilde{\lambda}'_i \tilde{f}_t$  as instruments, respectively. The PFIV<sup>+</sup> and PfIV<sup>+</sup> are biased-corrected estimators.  $\tilde{F}_t$  is  $r \times 1$ , and  $\tilde{f}_t$  is  $rmax \times 1$  with  $rmax = r + 2$ . PTFIV is the ‘traditional’ panel IV estimator that uses  $\tilde{F}_t$  as instruments.

## Appendix I: 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  $N, T \rightarrow \infty$ ,*

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

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

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

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  $\tilde{g}_t(\beta^0) = \tilde{F}_t \varepsilon_t$  and  $\bar{g} = \frac{1}{T} \sum_{t=1}^T \tilde{g}_t(\beta^0)$ . Then

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

Now

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

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  $N, T \rightarrow \infty$ . By assumption,  $T^{-1/2} \sum_{t=1}^T F_t \varepsilon_t \xrightarrow{d} N(0, S^0)$ . Thus  $\sqrt{T} \bar{g} \xrightarrow{d} N(0, H_0 S^0 H_0')$ , where  $H_0 = \text{plim } H$ . But  $\text{plim } \check{S} = H_0 S^0 H_0'$ . This implies that  $\check{S}^{-1/2} \sqrt{T} \bar{g} \xrightarrow{d} N(0, I)$ . Furthermore,  $S_{\tilde{F}x} = \frac{1}{T} \tilde{F}' x = \frac{1}{T} H' F' x + o_p(1) \xrightarrow{p} H_0' \Omega_{Fx}$ , where  $\Omega_{Fx}$  is the probability limit of  $\frac{1}{T} F' x = \frac{1}{T} \sum_{t=1}^T F_t x'_t$ . Thus  $S'_{\tilde{F}x} \check{S}^{-1} S_{\tilde{F}x} \xrightarrow{p} \Omega'_{Fx} (S^0)^{-1} \Omega_{Fx}$ . Summarizing result, 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 that using the true  $F$  as instruments.

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

**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  $\sqrt{T}/N \rightarrow 0$ . The rest of the proof is identical to that of Theorem 1.

**Proof of Proposition 1:** Without loss of generality, we assume homoskedasticity for  $\varepsilon_t$ . 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  $F = (F_1, \dots, F_T)'$ , and  $x$  and  $u$  are  $T \times 1$  vectors. Let  $z_2$  be a  $T \times r_2$  matrix consisting of  $r_2$  valid instruments ( $r_2$  is fixed and  $r_2 \geq r$ ) 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  $r_2$  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,

$$T^{-1}x'P_2x = T^{-1}x'P_2F\Psi + o_p(1), \tag{A.1}$$

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),

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

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  $F'M_2F = 0$ ).

## Appendix II: Proof of Proposition 2

We first show that  $\Omega_{GMM} = \Omega_{FIV}$  if  $N/T \rightarrow 0$ . 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 hold with  $W$  being estimated. From  $Z = F\Lambda' + e$  with  $e = (e_1, e_2, \dots, e_T)$ , we can write

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

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, using

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

we have

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

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

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

since  $\Lambda'D^{-1}\Lambda = O(N)$ . Noting that  $X'F/T \xrightarrow{p} \Sigma_{xF}$ , we have

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

The above is equal to the inverse of the asymptotic matrix of the FIV estimator, see the proof of Theorem 1. For term  $b$ , again using the expression of  $W^{-1}$ ,

$$(X'e/T)W^{-1}(e'X/T) = \frac{1}{T^2} X'eD^{-1}e'X - \frac{1}{T} X'eD^{-1}\Lambda[\Sigma_F^{-1} + \Lambda'D^{-1}\Lambda]^{-1}\Lambda'D^{-1}e'X/T = b_1 + b_2$$

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

if  $N/T \rightarrow 0$ . Furthermore,

$$b_2 = \frac{1}{T} X'eD^{-1}\Lambda = (N/T)^{1/2} \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \sigma_{i,e}^{-2} x_t \lambda'_i e_{it} = O_p((N/T)^{1/2}).$$

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  $A = \Lambda'D^{-1}\Lambda = O(N)$ , and omitting  $\sigma_\varepsilon^2$ ,

$$\begin{aligned}\Lambda'W^{-1} &= \Lambda'D^{-1} - A(\Sigma_F^{-1} + A)^{-1}\Lambda'D^{-1} = [I - A(\Sigma_F^{-1} + A)^{-1}]\Lambda'D^{-1} \\ &= A[A^{-1} - (\Sigma_F^{-1} + A)^{-1}]\Lambda'D^{-1} = (\Sigma_F + A^{-1})^{-1}A^{-1}\Lambda'D^{-1} \\ &= [\Sigma_F + O(N^{-1})]^{-1}(\Lambda'D^{-1}\Lambda/N)^{-1}\frac{1}{N}\Lambda'D^{-1} = O(1)\frac{1}{N}\Lambda'D^{-1}.\end{aligned}$$

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  $c$ .

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

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

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

$$T^{-2}X'ZW^{-1}Z'\varepsilon = O_p(N/T). \quad (\text{A.3})$$

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

$$\begin{aligned}T^{-2}X'ZW^{-1}Z'\varepsilon &= (X'F/T)\Lambda'W^{-1}\Lambda(F'\varepsilon/T) + T^{-2}X'eW^{-1}e'\varepsilon \\ &\quad + T^{-2}X'F\Lambda'W^{-1}e'\varepsilon + T^{-2}X'eW^{-1}\Lambda F'\varepsilon \\ &= I_1 + I_2 + I_3 + I_4.\end{aligned}$$

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

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

Consider  $I_2$ , assuming  $\sigma_\varepsilon^2 = 1$  for simplicity. Analogous to the analysis for the term  $b$  when we showed  $\Omega_{FIV} = \Omega_{GMM}$ ,

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

But

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

where  $\xi_i$  and  $\eta_i$  are implicitly defined. Note that  $\xi_i$  and  $\eta_i$  are dependent because  $x_t$  and  $\varepsilon_t$  are dependent. Let  $\gamma = \frac{1}{N} \sum_{i=1}^N E(\xi_i \eta_i)$ , and thus  $E(I_2) = \frac{N}{T} \gamma$ . 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  $\Lambda' W^{-1} = O(1) \frac{1}{N} \Lambda' D^{-1}$ ,  $I_3$  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

$$\widehat{\beta}_{GMM} - \beta^0 = \left( \frac{1}{T^2} X' Z W^{-1} Z' X \right)^{-1} \left[ I_1 + \frac{N}{T} \left( \frac{1}{N} \sum_{i=1}^N \xi_i \eta_i \right) + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right) \right]$$

where we recall that  $\sqrt{T} I_1 \xrightarrow{d} N(0, \Omega_{GMM}^{-1})$ . The second term in the squared bracket is  $O_p(N/T)$ , showing inconsistency of optimal GMM under large  $N$ .

**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  $\gamma = E(\xi_i \eta_i)$ , so that

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

where  $\widehat{\Omega}$  stands for  $(T^{-2}X'ZW^{-1}Z'X)^{-1}$  and  $d = \widehat{\Omega}\gamma$ . If  $N/T \rightarrow 0$ , the last three terms in the bracket multiplied by  $T^{1/2}$  are all  $o_p(1)$ , and as shown earlier,  $\widehat{\Omega} \xrightarrow{p} \Omega_{GMM}$ . It follows that if  $N/T \rightarrow 0$ ,

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

### Appendix III: Properties of PFIV

#### 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  $\widehat{\beta}_{PFIV} = \beta + S_{\widehat{x}\widehat{x}}^{-1} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \widehat{C}_{it} \varepsilon_{it}$ , it is sufficient to consider the limit of  $(NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \widehat{C}_{it} \varepsilon_{it}$ . Since  $(NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T C_{it} \varepsilon_{it} \xrightarrow{d} N(0, S)$ , 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

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

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

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

Denote  $G = V_{NT}^{-1} \left( \frac{1}{T} \widetilde{F}' F \right)$ , 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  $\varepsilon_{it}$  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

$$\begin{aligned} &(NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \Lambda'_i (\widetilde{F}_t - HF_t) \varepsilon_{it} \\ &= (T/N)^{1/2} \frac{1}{T} \sum_{t=1}^T \left( \frac{1}{\sqrt{N}} \sum_{i=1}^N \Lambda_i \varepsilon_{it} \right) G \left( \frac{1}{\sqrt{N}} \sum_{j=1}^N \sum_{k=1}^K \lambda_{j,k} u_{jt,k} \right) + o_p(1) \quad (\text{A.4}) \\ &= (T/N)^{1/2} O_p(1) \end{aligned}$$

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

$$\begin{aligned} (NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T (\tilde{\Lambda}_i - H^{-1}\Lambda_i)' H F_t \varepsilon_{it} &= (NT)^{-1} \frac{1}{T} \sum_{i=1}^N \sum_{s=1}^T u_{is} F'_s H' H \sum_{t=1}^T F_t \varepsilon_{it} + o_p(1) \\ &= (N/T)^{1/2} \frac{1}{N} \sum_{i=1}^N \left( \frac{1}{\sqrt{T}} \sum_{s=1}^T u_{is} F'_s \right) H' H \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T F_t \varepsilon_{it} \right) + o_p(1) \\ &= (N/T)^{1/2} O_p(1) \end{aligned} \quad (\text{A.5})$$

Combining (A.4) and (A.5), 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.4) and (A.5). 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.4) is  $A_t G B_t$ , which is a vector. Thus

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

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

$$(A.4) = (T/N)^{1/2} \left( \frac{1}{T} \sum_{t=1}^T (B_t \otimes A_t) \right) \text{vec}(G)$$

Because of the cross-sectional independence assumption on  $\varepsilon_{it}$  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

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

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

$$(A.4) = (T/N)^{1/2} \delta_1 + o_p(1)$$

Let  $\delta_1^0$  denote the limit of  $\delta_1$ . If  $T/N \rightarrow \tau$ , it follows that

$$(A.4) \rightarrow \tau^{1/2} \delta_1^0$$

Next consider (A.5). 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 (A.5) can be rewritten as (ignoring the  $o_p(1)$  term):

$$(A.5) = (N/T)^{1/2} \left( \frac{1}{N} \sum_{i=1}^N (\Phi'_i \otimes \Theta_i) \right) \text{vec}(H'H)$$

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

$$(A.5) = (N/T)^{1/2} \Delta_2 + o_p(1)$$

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

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

Let

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

If  $N/T \rightarrow \tau$ , we have  $(A.5) \rightarrow \tau^{-1/2} \delta_2^0$ . 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 theorem shows that

$$\sqrt{NT}(\widehat{\beta}_{PFIV} - \beta) = S_{\widehat{x}\widehat{x}}^{-1} \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T C_{it} \varepsilon_{it} + \sqrt{T/N} S_{\widehat{x}\widehat{x}}^{-1} \delta_1 + \sqrt{N/T} S_{\widehat{x}\widehat{x}}^{-1} \delta_2 + o_p(1) \quad (\text{A.6})$$

It can be shown that  $\widehat{\Delta}_1 - S_{\widehat{x}\widehat{x}}^{-1} \delta_1 = O_p(\delta_{NT}^{-1})$  and  $\widehat{\Delta}_2 - S_{\widehat{x}\widehat{x}}^{-1} \delta_2 = O_p(\delta_{NT}^{-1})$ . These imply that  $(T/N)^{1/2}(\widehat{\Delta}_1 - S_{\widehat{x}\widehat{x}}^{-1} \delta_1) = o_p(1)$  if  $T/N^2 \rightarrow 0$ , and  $((N/T)^{1/2}(\widehat{\Delta}_2 - S_{\widehat{x}\widehat{x}}^{-1} \delta_2) = o_p(1)$  if  $N/T^2 \rightarrow 0$ . Thus, we can replace  $S_{\widehat{x}\widehat{x}}^{-1} \delta_1$  by  $\widehat{\Delta}_1$  and replace  $S_{\widehat{x}\widehat{x}}^{-1} \delta_2$  by  $\widehat{\Delta}_2$  in (A.6). Equivalently,

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

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.
- Arellano, M. and Bond, S. 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* **71:1**, 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.
- Bekker, P. A. 1994, Alternative Approximations to the Distributions of Instrumental Variables Estimators, *Econometrica* **63**, 657–681.
- Bernanke, B. and Boivin, J. 2003, Monetary Policy in a Data Rich Environment, *Journal of Monetary Economics* **50:3**, 525–546.
- Boivin, J. and Giannoni, M. 2006, DSGE Models in a Data Rich Environment, NBER WP 12772.
- Breitung, J. and Eickmeier, S. 2005, Dynamic Factor Models, Deutsche Bundesbank Discussion Paper 38/2005.
- Carrasco, M. 2006, A Regularization Approach to the Many Instruments Problem, mimeo, Université de Montreal.
- Chamberlain, G. and Rothschild, M. 1983, Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets, *Econometrica* **51**, 1281–2304.
- Chao, J. and Swanson, N. 2005, Consistent Estimation With a Large Number of Instruments, *Econometrica* **73**, 1673–1692.
- Donald, S. and Newey, W. 2001, Choosing the Number of Instruments, *Econometrica* **69:5**, 1161–1192.
- 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 Distributions, *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 University 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.
- Kuersteiner, G. and Okui, R. 2007, Estimator Averaging for Two Stage Least Squares, mimeo.
- Meng, G., Hu, G. and Bai, J. 2007, A Simple Method for Estimating Betas When Factors are Measured with Error, mimeo.
- Moreira, M. 2003, A Conditional Likelihood Ratio Test for Structural Models, *Econometrica* **71:4**, 1027–1048.
- Onatski, A. 2006, Asymptotic Distribution of the Principal Components Estimator of Large Factor Models when Factors are Relatively Weak, mimeo, Columbia University.
- Reichlin, L. 2003, Factor Models in Large Cross Sections of Time Series, in S. T. M. Dewatripoint, L. P. Hansen (ed.), *Advances in Economics and Econometrics: Theory and Applications, Vol. 111, 8th World Congress of the Econometric Society*, Cambridge University Press.
- Sargent, T. and Sims, C. 1977, Business Cycle Modelling without Pretending to have too much a Priori Economic Theory, in C. Sims (ed.), *New Methods in Business Cycle Research*, Federal Reserve Bank of Minneapolis, Minneapolis.
- 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.
- Wooldridge, J. 2005, Instrumental Variables Estimation with Panel Data, *Econometric Theory* **21**, 865–869.