INSTRUMENTAL VARIABLE ESTIMATION IN A DATA RICH ENVIRONMENT

Jushan Bai* Serena Ng †

November 13, 2006

Comments Welcome

Abstract

We consider estimation of parameters in a regression model in which the endogenous regressors are just a few of the many other endogenous variables driven by a small number of unobservable exogenous common shocks. We show the method of principal components can be used to estimate factors that can be used as instrumental variables. These are not only valid instruments, they are more efficient than the observed variables in our framework. Consistency and asymptotic normality of the single equation factor instrumental variable estimator (FIV) is established. We also show that consistent estimates can be obtained from large panel data regressions even when there is no valid instrument in a conventional sense. To reduce the bias that might arise from using too many instruments, we use boosting to select out the most relevant ones. Boosting necessitates a stopping rule. We derive the condition on the stopping parameter that arises from boosting estimated factors instead of observed variables.

Keywords: factor models, weak instruments, boosting.
JEL classification: C1, C2, C3, C4

*Department of Economics, NYU, 269 Mercer St, New York, NY 10003 Email: Jushan.Bai@nyu.edu.
†Department of Economics, University of Michigan, Ann Arbor, MI 48109 Email: Serena.Ng@umich.edu
We also acknowledge financial support from the NSF (grants SES-0137084, SES-0136923, SES-0549978)

1 Introduction

The primary purpose of structural econometric modeling is to explain how endogenous variables evolve according to fundamental processes such as taste shocks, policy, and productivity variables. To completely characterize the behavior and the evolution of a particular endogenous variable in a data consistent manner, the economist needs to estimate the structural parameters of the model. It is well known that because these parameters are often coefficients attached to endogenous variables, endogeneity bias invalidates least squares estimation. There is a long history and continuing interest in estimation by instrumental variables.[1] In this paper, we suggest a new way of constructing instrumental variables. We show that if we have a large panel of data and the common variations in the panel data coincide with those that underlie the endogenous regressors, 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 estimates are $\sqrt{T}$ consistent and 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 even when there is no valid instrument in the conventional sense.

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

An alternative view can also be developed by noting that the variables as defined in an economic

[1] See, for example, Andrews et al. (2006) and the references therein.

model may not coincide exactly with how the measured data are defined. For example, non-durable consumption is often used to estimate preference parameters, but non-durable consumption ignores service flows, which the model's notion of consumption includes. As is well known, measurement error in the regressors will invalidate least squares estimation, but estimation by instrumental variables will yield consistent estimates. 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.

In conventional analysis, variables that are weakly exogenous for the parameters of interest are valid instruments. For estimation of parameters of a small sub-system that is part of a large system, the number of potentially relevant instruments can be very large because there are many variables that are weakly exogenous for the parameters of interest. It is well recognized that use of all potentially relevant instruments in the first stage of two-stage least squares estimation will lead to a degrees of freedom problem. This motivates Kloek and Mennes (1960) to construct a small number of principal components from the predetermined variables as instruments. Our methodology is similar in some ways, but we put more structure on the predetermined variables. Our point of departure is that if the variables in the system are driven by common sources of variations, then the ideal instruments for the endogenous variables in the system are their common components. Thus, while we have many valid instruments, each is merely a noisy indicator of the ideal instrument that we do not observe. We use a factor approach to estimate the feasible instruments space from the space spanned by the observed instruments. In the terminology of Bernanke and Boivin (2003), what we propose is a way to construct instrumental variables in a 'data rich environment'. Favero and Marcellino (2001) used estimated factors as instruments to estimate forward looking Taylor rules with the motivation that the factors contain more information than a small number of series. Here, we provide a formal analysis and show that the estimated factors are more efficient instruments than the observed variables. As far as we are aware of, Kapetanios and Marcellino (2006) is the only other paper that considers using estimated factors as instruments. Their framework assumes that there are many observed weak instruments having a weak factor structure. In contrast, we assume that there are many observed instruments with an identifiable factor structure, but that some of the factors maybe unnecessary for estimating the parameters of interest. As such, we adopt standard instead of weak instrument asymptotics. A further point of departure is we consider single equation as well as large systems estimation.

Based on biased considerations, our methodology suggests but do not insist that we first estimate the feasible instrument space, and from it, select a set of most relevant instruments. These relevant instruments are then used to identify and estimate the structural parameters of interest. The variable selection methodology we consider is 'boosting'. Boosting is a statistical procedure that performs subset variable selection and shrinkage simultaneously to improve prediction. It has primarily been used in bio-statistics and machine learning analysis as a classification and model fitting device. Consideration of boosting as a method for selecting instrumental variables appears to be new. Boosting can also be used to select observed instruments and is thus of interest in its own right. We show here that when the instruments are factors estimated from a large panel of data, it can recover the space spanned by the relevant instruments, but that the boosting stopping rule has an upper bound because of the sampling error from estimation of the factors.

The rest of this paper is organized as follows. Section 2 presents the framework for estimation using the feasible instrument set. Section 3 presents boosting as a method for selecting the relevant set from the feasible set. Simulations and illustrations 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

Consider a system of structural equations where for $g = 1, \dots G$ and $t = 1, \dots T$, the endogenous variable y_{gt} is specified as a function of a $K_g \times 1$ vector of regressors x_{gt}:

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

We assume equation g is correctly specified for every g. The parameter vector of interest is $\beta_g = (\beta_{1g}', \beta_{2g}')'$ and corresponds to the coefficients on the regressors $x_{gt} = (x_{1gt}, x_{2gt})$, where the exogenous and predetermined regressors are collected into a $K_{1g} \times 1$ vector x_{1gt}. The $K_{2g} \times 1$ vector x_{2gt} is endogenous in the sense that $E(x_{2gt}\varepsilon_{gt}) \neq 0$ and the least squares estimator suffers from endogeneity bias. We assume that for each g,

$$x_{2gt} = \Psi_g F_{gt} + u_{gt}$$

where Ψ_g is a $K_{2g} \times r_g$ matrix, F_{gt} is a $r_g \times 1$ vector of fundamental variables, and r_g is a small number. Endogeneity arises when $E(F_{gt}\varepsilon_{gt}) = 0$ but $E(u_{gt}\varepsilon_t) \neq 0$. This induces a non-zero correlation between x_{2gt} and ε_{gt}. The reduced form in terms of (x_{1g}, F_g) is

$$y_{gt} = \beta_{1g}' x_{1gt} + \pi_g' F_{gt} + v_{gt}$$

where $\pi_i = \Psi_g \beta_{2g}$ and $v_{gt} = \beta_g' u_{gt} + \varepsilon_{gt}$. If F_{gt} was observed, β can be estimated, for example, by using F_{gt} to instrument x_{gt}. Here, we assume that F_{gt} is not observed. Let F_t be the vector that collects the linearly independent F_{gt} $(g = 1, 2, ..., G)$.

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

$$z_{it} = \lambda_i' F_t + e_{it}. \tag{1}$$

The $r \times 1$ vector F_t above is a set of common factors, λ_i is the corresponding vector of 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 F_t and with ε_{gt}. Neither e_{it} nor F_t is observed. Viewed from the factor model perspective, x_{gt} is just K_{2g} of the many other variables in the economic system that has a common component and an idiosyncratic component.

Although z, like x_2, is driven by F, we assume e_{it} is uncorrelated with ε_{gt}, and z_{it} is correlated with x_{2gt} through F_t. Thus, z is weakly exogenous for β, and z constitutes a large panel of observed valid instruments. While valid, z_{it} is a 'noisy' instrument for each x_{2gt} because the ideal instrument for x_{2gt} is $\Psi_g F_{gt}$. When the context is clear, we will simply refer to F_t as instruments instead of 'factor-based instruments'. We cannot use F_t only because it is not observed. Subsection 2.1 therefore begins by replacing F_t with its consistent estimates, $\widetilde{F}_t$. This forms a 'feasible' instrument set. The single equation and systems equation estimators (FIV and PFIV) are proposed. For the purpose of bias reduction, Section 3 considers an fIV estimator that uses $\widetilde{f}$, a subset of $\widetilde{F}$ as instruments. We then discuss how to go from the feasible to the relevant instrument set.

2.1 Estimating F_t

We assume that the (static) factors are estimated from the panel of data consisting of z_{it}, $i = 1, \ldots N, t = 1, \ldots T$ by the method of principal components. Let $Z_i = (z_{i1}, z_{2t}, ..., z_{NT})'$ be the $T \times 1$ matrix for the ith cross-section regressors, and let $Z = (Z_1, Z_2, ..., Z_N)$, which is $T \times N$. The estimated factors, denoted $\widetilde{F} = (\widetilde{F}_1, ..., \widetilde{F}_T)$, is a $T \times r$ matrix consisting of r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the matrix $ZZ'/(TN)$ in decreasing order. Then $\widetilde{\Lambda} = (\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_N)' = Z'\widetilde{F}/T$, and $\widetilde{e} = Z - \widetilde{F}\widetilde{\Lambda}'$. Also let $\widetilde{V}$ be the $r \times r$ diagonal matrix consisting of the r largest eigenvalues of $ZZ'/(TN)$. Hereafter, variables denoted a 'tilde' are (based on) principal component estimates, while 'hatted' variables are estimated from the regression model.

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. λ_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 $N \to \infty$.

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 $\frac{1}{NT}\sum_{i,j,t,s=1}|\sigma_{ij,ts}| \leq M$.

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

c.iv For each t, $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i e_{it} \xrightarrow{d} N(0,\Gamma_t)$, where $\Gamma_t = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}E(\lambda_i\lambda_j{}' e_{it}e_{jt})$.

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

e . For each $g = 1,\ldots G$ and for each $i = 1,\ldots N$, $E(\varepsilon_{gt}e_{it}) = 0$.

Assumption A was used in Bai and Ng (2006) to show consistency of estimates of factor augmented regressions. Note that e_{it} is allowed to be cross-sectionally and serially correlated, but only weakly as stated under condition (A.c). Assumption (A.e) is specific to the present analysis. By assuming that the error of each primary equation is uncorrelated with the idiosyncratic errors of the panel, z is exogenous for β. The results most relevant for the present analysis are summarized in the following lemma:

Lemma 1 *Let $H = \widetilde{V}^{-1}(\widetilde{F}'F/T)(\Lambda'\Lambda/N)$. For $i = 1,\ldots N$, and let z_{it} be stationary with $E|z_{it}^4| < \infty$. Under Assumption (A) and as $N, T \to \infty$ (jointly):*

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

ii $T^{-1}(\widetilde{F} - HF)'\varepsilon_g = O_p(\min[N,T]^{-1})$, *for* $g = 1, 2, ..., G$;

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). As indicated in Lemma 1(i), we can only estimate the space spanned by the factors, and not F_t per se.

2.2 Single Equation

For the case of a single equation, we can drop the subscript g. The regression model is

$$\begin{aligned} y_t &= x_{1t}'\beta_1 + x_{2t}'\beta_2 + \varepsilon_t \\ &= x_t'\beta + \varepsilon_t \qquad (2) \\ x_{2t} &= \Psi' F_t + u_t \qquad (3) \end{aligned}$$

where $x_t = (x_{1t}', x_{2t}')'$ is $K \times 1$. The $K_1 \times 1$ regressors x_{1t}, which may include lags of y_t, are exogenous or predetermined while the $K_2 \times 1$ regressors x_{2t} are not. Thus $E(x_t\varepsilon_t) \neq 0$ because $E(x_{2t}\varepsilon_t) \neq 0$. We let $\beta^0 = (\beta_1^0, \beta_2^0)$ denote the true value of β.

In the following analysis, the instrument vector is really $\widetilde{F}_t^+ = (x_{it}', \widetilde{F}_t')'$. To fix ideas and for notational simplicity, we assume the absence of regressor x_1 ($K_1 = 0$) so that the instrument is $\widetilde{F}_t$. It is understood that when x_1 is present, the results still go through upon replacing $\widetilde{F}$ in the estimator below by $\widetilde{F}^+$.

Assumption B

a. $E(\varepsilon_t) = 0$, $E|\varepsilon_t|^{4+\delta} < \infty$ for some $\delta > 0$. The $r \times 1$ vector process $g_t(\beta^0) = F_t\varepsilon_t(\beta^0)$ satisfies $E[g_t(\beta^0)] \equiv E(g_t^0) = 0$ and $\sqrt{T}\bar{g}^0 \xrightarrow{d} N(0, S^0)$ where S^0 is the asymptotic variance of $\sqrt{T}\bar{g}^0$, and $\bar{g}^0 = \frac{1}{T}\sum_{t=1}^T F_t\varepsilon_t(\beta^0)$.

b. x_{1t} is a predetermined, covariance stationary processes with $E(x_{1t}\varepsilon_t) = 0$.

c. $x_{2t} = \Psi' F_t + u_t$ is covariance stationary and $\Psi'\Psi > 0$, $E(F_t u_t) = 0$, $E(u_t\varepsilon_t) \neq 0$.

Part (a) states that the model is correctly specified so that at the true value of β, denoted β^0, a set of orthogonality conditions hold. Heteroskedasticity of ε_t^0 is allowed and will be reflected in the asymptotic variance, S^0. Part (b) distinguishes a predetermined regressor from an endogenous one through their correlation with ε_t. Validity of F_t as an instrument requires that F_{jt} has a non-zero loading on x_{2t} for each $j = 1, \ldots r$. Thus, F_t forms the ideal but infeasible instrument set.

Assumptions A and B are sufficient to analyze how F_t can be exploited in estimation when $z_{jt}, j = 1, \ldots N$ are valid instruments by weakly exogeneity. In certain cases, lags of F_t can also serve as instruments, though in general, lags of F_t should provide no further information about x_2 once conditioned on F_t. When u_t is serially uncorrelated and all the dynamics in x_{2kt} are due to F_t, then lags of F_t are always better instruments than lags of x_{2kt}. Lags of x_2 can be better instruments only if F_t do not contribute at all to the dynamics in x_2.

Lags of the observed variables in the equation of interest are often used as instruments in empirical work. To compare our proposed estimator with estimators currently in use, we also need to make precise when lags of observed instruments can be used in conventional IV estimation.

Condition C: (a) $\Gamma_{x2k}(j) = E(x_{2kt}x_{2kt-j}) \neq 0$ for some $j > 1$. and (b) $E(\varepsilon_t|I_{t-1}) = 0$ where $I_{t-1} = \{x_{1t-j}, x_{2t-j}, y_{t-j}\}_{j=1}^{t-1}$.

In order to use past values of the observed variables as instruments, ε_t must be uncorrelated with the past observations, as required by part (b). Furthermore, x_{2t} must be serially correlated as required by part (a). Note that if lags of x_2 are valid instruments, they are always better instruments than lags of y because the latter are correlated with x_2 only through the correlation between x_2 and its past values. Not every x_{2kt} needs to satisfy Condition C. Those x_{2kt} that are serially uncorrelated will fail (a). When x_{1t} is strongly exogenous such that $x_{1t}\varepsilon_s = 0$ for all t, and s (this situation of course rules out x_{1t} being the lag of y), ε_t itself can be serially correlated of unknown form. When this occurs, the lags of x_{2t} cannot be used as instruments since x_{2t-j} can be correlated with ε_{t-j}, which is correlated with ε_t. If x_{1t} is simply the lags of y, then as argued earlier, it can be used as instrument though the lags of x_2 are better instruments. When x_{1t} does not include the lags of y, it can be valid instrument only if it is correlated with x_2 through the common component F_t.

The conventional treatment of endogeneity bias is to use lags of y, x_1 and x_2 as instruments for x_2 and invoke Condition C. Our point of departure is to note that g_t contains all the information about β. The reason why the moments g_t are not used to estimate β is that F_t is not observed. Define $\widetilde{g}_t(\beta) = \widetilde{F}_t\varepsilon_t(\beta)$. Consider estimating β using the r moment conditions $\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^T \widetilde{F}_t\varepsilon_t$. Let W_T be a $r \times r$ positive definite weighting matrix. Where appropriate, the dependence of $\bar{g}$ on β will be suppressed. The linear GMM estimator is defined as

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

where $S_{\widetilde{F}x} = \frac{1}{T}\sum_{t=1}^T \widetilde{F}_t x_t'$. Let $\breve{\varepsilon}_t = y_t - x_t'\breve{\beta}_{FIV}$ and let $\breve{S}$ be a consistent estimate of S based upon $\breve{g}_t = \widetilde{F}_t\breve{\varepsilon}_t$. Then the efficient GMM estimator, which we will focus, is to let $W_T = \breve{S}^{-1}$, giving

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

Theorem 1 *Let β^0 be $K \times 1$ vector of true values of β and let $\widetilde{F}_t$ be a $r \times 1$ vector of common factors estimated from the $T \times N$ panel of data, z. Let $g_t = \widetilde{F}_t\varepsilon_t$, $\bar{g} = \frac{1}{T}\sum_{t=1}^T g_t$, and let $\breve{S} = \frac{1}{T}\sum_{t=1}^T(\breve{g}_t\breve{g}_t')$, $\breve{g}_t = \widetilde{F}_t\breve{\varepsilon}_t$, and $\breve{\varepsilon}_t = y_t - x_t'\breve{\beta}_{FIV}$. Let $\widehat{\beta}_{FIV}$ minimizes $\bar{g}(\beta)'\breve{S}^{-1}\bar{g}(\beta)$. Under Assumptions A and B, $\widehat{\beta}_{FIV} = \beta^0 + o_p(1)$. Let $\widehat{S} = \frac{1}{T}\sum_{t=1}^T \widehat{g}_t\widehat{g}_t'$, $\widehat{g}_t = \widetilde{F}_t\widehat{\varepsilon}_t$ with $\widehat{\varepsilon}_t = y_t - x_t'\widehat{\beta}_{FIV}$. Then $\widehat{S} \xrightarrow{p} S$, where $\sqrt{T}\bar{g} \xrightarrow{d} N(0, S)$. If, in addition, $\frac{\sqrt{T}}{N} \to 0$ as $N, T \to \infty$,*

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

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

Theorem 1 establishes consistency and asymptotic normality of the GMM estimator when $\widetilde{F}_t$ are used as instruments, and is a direct consequence of Lemma 1. Just as if F was observed, $\widehat{\beta}_{FIV}$ reduces to $(\widetilde{F}'x)^{-1}\widetilde{F}'y$ and is the instrumental variable estimator in an exactly identified model with $K = r$. It is the two-stage least squares (2SLS) estimator, i.e., $\widehat{\beta}_{FIV} = (x'P_{\widetilde{F}}x)^{-1}x'P_{\widetilde{F}}y$, under conditional homoskedasticity. Furthermore, T times the value of the objective function is asymptotically χ^2 distributed with $r-K$ degrees of freedom. Essentially, if $\frac{\sqrt{T}}{N} \to 0$, estimation and inference can proceed as though F_t was observed.

Because we have a large panel of valid instruments, and mechanically speaking, $\widetilde{F}$ are the principal components of z, one might be tempted to interpret the FIV as a principal components approach to instrumental variables, such as considered in Kloek and Mennes (1960), and in Galbraith and Zinde-Walsh (2005). For one thing, the N and T we consider are much larger than in Kloek and Mennes (1960). These authors were concerned with situations when N is large to the given T (such as 30) so that the first stage estimation is inefficient. For another, these authors motivated principal components as a practical dimension reduction device. 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, and it is because of this structure that leads to the following.

Proposition 1 *Let z_2 be a subset of r of the N observed instruments $(z_{1t}, \ldots z_{Nt})$. Let $m_t = z_{2t}(y_t - x_t'\beta)$ with $\sqrt{T}\bar{m} \xrightarrow{d} N(0, Q)$. Let $\widehat{\beta}_{IV}$ minimizes $\bar{m}'(\breve{Q})^{-1}\bar{m}$ be the GMM estimator with the property that $\sqrt{T}(\widehat{\beta}_{IV} - \beta^0) \xrightarrow{d} N(0, Avar(\widehat{\beta}_{IV}))$. Then*

$$Avar(\widehat{\beta}_{IV}) - Avar(\widehat{\beta}_{FIV}) \geq 0.$$

Proposition 1 says that $\widehat{\beta}_{FIV}$ is more efficient than $\widehat{\beta}_{IV}$, which uses an equal number of z_2 as instruments. The intuition is straightforward. The observed instruments are the ideal instruments contaminated with errors while $\widetilde{F}$ is consistent for the ideal instrument space. More efficient instruments thus lead to more efficient estimates. The FIV estimator can be especially useful when we observe a large number of individually valid but noisy instruments in the sense of Hahn and Kuersteiner (2002). Pooling information across the observed variables washes out the noise to generate more efficient instruments for x_2.

In some instances, the structural coefficients β_1, β_2 are non-linear functions of the deep parameters, say, θ. Although Theorem 1 is presented as a result of linear estimation, $\widetilde{F}_t$ can still be used as instruments in GMM estimation with cross-parameter restrictions. This is because $\widetilde{F}$ is used to instrument x_2 and not functions of x_2. It is then straightforward to show that Theorem 1 holds with $S_{x\widetilde{F}}$ replaced by $G_{x\widetilde{F}}$, which is the derivative of $g_t = \widetilde{F}_t\varepsilon_t(\theta)$ with respect to θ. Details are omitted.

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

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

where x_{gt} is $K_g \times 1$. As an example of $G = 2$, (y_1, y_2) could be aggregate consumption and earnings, while the endogenous regressor is wages. Let $\widetilde{F}_{gt}$ be the $r_g \times 1$ vector of instruments for the g-th equation, $g = 1, \ldots G$, and let $r = \sum_g r_g$. Then g_t is a $r \times 1$ vector of stacked up moment conditions. Assuming that for each $g = 1, \ldots G$, the $r_g \times K_g$ moment matrix $E(\widetilde{F}_{gt} x'_{gt})$ is full column rank, Theorem 1 still holds, but the $r \times r$ matrix S is now the asymptotic variance of the stacked up moment conditions. Note that this need not be a block diagonal matrix. Likewise, $S_{\widetilde{F}x}$ is a $K \times r$ matrix. If each equation has a regressor matrix of the same size and uses the same number of instruments, the $S_{\widetilde{F}x}$ matrix under systems estimation will be G times bigger, just as when F 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 $x_2 = F\Psi' + u$, the non-zero correlation between x_2 and ε arises because $\text{cov}(u, \varepsilon) \neq 0$. We can decompose ε_t into a component that is correlated with u_t, and a component that does not. Let

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

where $\varepsilon_{t|u}$ is orthogonal to u_t and thus x_2. We can rewrite the regression $y = x_1\beta_1 + x_2\beta_2 + \varepsilon$ as

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

If F was observed, we can estimate the reduced form for x_2 to yield fitted residuals $\widehat{u}$. Then least squares estimation of

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

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

In our setting, we cannot estimate the reduced form for x_2 because F is not observed. Indeed, if we only observe x_2, and $x_2 = F\Psi' + u$, there is no hope of identifying the two components in x_2. However, we have a panel of data, z with a factor structure, and $\widetilde{F}_t$ are consistent estimates of F_t up to a linear transformation. The control function approach remains feasible in our data rich environment and consists of three steps. In step one, we obtain $\widetilde{F}$. In step 2, for each $i = 1, \ldots K_2$, least squares estimation of

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

will yield $\sqrt{T}$ consistent estimates of Ψ_i. By Bai (2003), $\widetilde{C}_{it} = \widetilde{F}_t'\widetilde{\Psi}_i$ is $\min[\sqrt{N}, \sqrt{T}]$ consistent for $C_{it} = F_t'\Psi_i$. It follows that $\widetilde{u}_{it} - u_{it} = O_p(\min[\sqrt{N}, \sqrt{T}]^{-1})$. Least squares estimation of

$$y_t = x_{1t}'\beta_1 + x_{2t}'\beta_2 + \widetilde{u}_t'\gamma + \varepsilon_t^u \tag{4}$$

will yield $\sqrt{T}$ consistent estimates of β. It is straightforward to show that the estimate is again numerically identical to 2SLS with $\widetilde{F}$ 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_2 on z_2, where z_2 are noisy indicators of F. With the FIV, u_t is estimated from regressing x_2 on a consistent estimate of F and is thus more efficient than the IV.

2.4 Panel Data and Large Simultaneous Equations System

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

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

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

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

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

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

is inconsistent. Unlike the single equation system, we do not need to assume 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. We continue to maintain the assumption that

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

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

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

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

Let $X_i = (x_{i1}, x_{2t}, ..., x_{NT})'$ be the $T \times K$ matrix for the ith cross-section regressors, so that $X = (X_1, X_2, ..., X_N)$ is $T \times (NK)$. Let Λ be a $(NK) \times r$ matrix while F is $T \times r$. Let $\widetilde{F}$ be the principal component estimate of F from the matrix XX', as explained in Section 2.1 with Z replaced by X. Let $\widetilde{C}_{it} = \widetilde{\Lambda}_i' \widetilde{F}_t$, which is $K \times 1$.

Consider the pooled two-stage least-squares estimator with $\widetilde{C}_{it}$ as instruments[2]

$$\widehat{\beta}_{PFIV} = \Big(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{C}_{it} x_{it}' \Big)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{C}_{it} y_{it}. \qquad (5)$$

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

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

Assumption B':

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

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

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

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

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

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

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

where $\widetilde{Z}_{it} = (x_{1it}', \widetilde{C}_{it}')'$. It is noted that equation (5) can be written alternatively as

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

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

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

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

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

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

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

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

A biased corrected estimator can be considered to recenter the asymptotic distribution to zero for small N and T. For this purpose, we assume that ε_{it} are serially uncorrelated.[3] 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[4]

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

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

[4]In the presence of exogenous regressors x_{1it} as in footnote 2, 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 δ_2, where $r(N+T)$ is the number of parameters used to estimate $\widehat{u}_{it}$.

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

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

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

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

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

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

In view of the reduced form $y_{it} = \beta'\Lambda_i F_t + \beta' u_{it} + \varepsilon_{it}$, one can actually estimate F by pooling X and Y. Theorem 2 can be extended to allow for other exogenous or predetermined regressors, such as gender, race, etc. For example, $y_{it} = w'_{it}\gamma + x'_{it}\beta + \varepsilon_{it}$ and w_{it} is independent of ε_{it} (or is predetermined). In this case, extracting F from Y as well as from X may not be desirable. If there is another large panel of data z that is informative about F, it too can be used together with X to estimate F.

3 Determining the Relevant Instruments

Theorems 1 and 2 are stated in terms of $\widetilde{F}_t$, which is a set of r estimated factor instruments. In a widely used macroeconomic panel of 132 series put together by Stock and Watson and used in Stock and Watson (2004) for example, researchers often found between 8 and 12 factors depending on the sample period. To understand why r can be much larger than the number factors suggested by economic analysis, we need to clarify that the method of principal components estimates the number of 'static' factors from Z, whereas economic analysis assumes that the number 'dynamic factors', say, q, is small. This distinction is most easily understood using an example: $z_{it} = \lambda_{i1}G_t + \lambda_{i2}G_{t-1} + e_{it}$.

Here, there is only one dynamic factor G_t because the population spectral density matrix of z is rank one. However, if we let $F_{1t} = G_t$, $F_{2t} = G_{t-1}$ and ignore the dynamic relation between F_{1t} and F_{2t}, we will have two static factors because the population covariance matrix of z has rank two. More generally, if we have q dynamic factors that are themselves moving-average of order s, we will end up with $r = q(s+1)$ static factors. Although knowledge of the dynamic factors are useful in some analysis, it is not necessary for the purpose of estimating β, and importantly, estimation and inference based on the estimated static factors are much better understood from a theoretical standpoint. The implication, however, is that the dimension of F can sometimes be quite large, though it is still much smaller than the dimension of Z, being N.

While $\widehat{\beta}_{FIV}$ is asymptotically unbiased and is more efficient than $\widehat{\beta}_{IV}$, it is biased in finite samples, just as when F is observed. A well known result is as follows:

Lemma 2 *Let $\widehat{\beta}_{GMM}$ be the linear GMM estimator obtained with r observed instruments. The bias of $\widehat{\beta}_{GMM}$ increases with r.*

Theorem 4.5 of Newey and Smith (2004) showed using higher order asymptotic expansions that the bias of the GMM estimator is linear in $r - K_2$, which is the number of overidentifying restrictions. Phillips (1983) showed that the rate at which $\widehat{\beta}_{2SLS}$ approaches normality depends on the true value of β and r. Hahn and Hausman (2002) showed that the expected bias of 2SLS is linear in r, the number of instruments. The result arises because $E(x'P_F\varepsilon/T) = E(u'P_F\varepsilon/T) = \sigma_{u\varepsilon}r/T$. When F is not observed, we can still expect the bias of $\widehat{\beta}_{FIV}$ to increase with r because $\widetilde{P}_{\widetilde{F}} - P_F = O_p(\min[\sqrt{N}, \sqrt{T}]^{-1})$. This motivates forming a set of L instruments with $K_2 \leq L \leq r$ by removing those elements of $\widetilde{F}_t$ that have weak predictive ability for x_2.

By assumption, $x_{2t} = \Psi' F_t + u_t$ and F_t is r dimensional. But not every F_{jt} needs to be of first order importance in predicting x_2. We call those $F_{jt} \subset F_t$ with Ψ_j taking on small values the 'relatively weak' instruments. Hahn and Kuersteiner (2002) defined relatively weak instruments as those with Ψ_j in the $T^{-\delta}$ neighborhood of zero with $\delta < \frac{1}{2}$ and showed that standard asymptotic results hold when such instruments are used to perform 2SLS. Let $f \subset F$ be the set of L instruments that remain after the relatively weak instruments are removed from F, and with the property that they predict the endogenous regressors x_2 as well as when the relatively weak instruments are present. That is to say,

$$\Phi(F) = E(x_2|F) \approx \Phi(f) = E(x_2|f).$$

We refer to f as the 'relevant instruments'. We do not allow for weak instruments with in the sense of Staiger and Stock (1997); such a case is considered in Kapetanios and Marcellino (2006). Now consider the estimator

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

As $\widehat{\beta}_{fIV}$ is a special case of $\widehat{\beta}_{FIV}$ when $r = L$, it is consistent and asymptotically normal. It can also be shown that $\widehat{\beta}_{FIV}$ has a smaller variance. But the bias of $\widehat{\beta}_{fIV}$ should be smaller than $\widehat{\beta}_{FIV}$, and $\widehat{\beta}_{fIV}$ can be desirable in finite samples.

3.1 Selecting $\widetilde{f}$ From $\widetilde{F}$

In the case of large panel regressions, the factors in x_{it} are of direct interest because we do not need additional instruments. Then selecting $\widetilde{g}$ is the same as selecting the number of factors, and we can use the criteria developed in Bai and Ng (2002). For the case of a single equation, we need a different approach. We will proceed to discuss how to select $\widetilde{f}$ from $\widetilde{F}$ to keep the analysis focused, though our proposed approach also applies, and is in fact just as useful when selecting z_2 from a large number of observed instruments Z.

To motivate the approach, consider first the simple case when F is observed, and there is no predetermined variable. From

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

with $E(u_t\varepsilon_t) \neq 0$, we see that the semi-optimal instrument for x_2 here is actually the one-dimensional common component, C_2, though F is the set of instrumental variables that permit the construction of C_2. But C_2 is just the conditional mean, $\Phi(F) = E(x_2|F)$. Thus in the terminology of 2SLS, the first stage estimation should be to uncover $\Phi(F)$. But by the definition of relevant instruments and under linearity,

$$C_{2t} \approx \psi' f_t.$$

The preferred instrument from the point of view of bias is therefore f, since $L \leq r$. In this sense, determining f is the same as determining those factors that load non-trivially on x_2 with the goal of uncovering $E(x_2|F)$.

When there are predetermined regressors and F is not observed, we want to fit x_2 by $\widetilde{F}_t$ controlling for the explanatory power of x_1 for x_2. To this end, let $X_2 = M_1 x_2$ and $\widetilde{G} = M_1\widetilde{F}$, where $M_1 = I - x_1(x_1'x_1)^{-1}x_1$. Thus, x_2 and $\widetilde{G}$ are the residuals from regressing x_2 and $\widetilde{F}$, respectively, on x_1. Selecting the best subset $\widetilde{f}$ out of $\widetilde{F}$ after controlling for x_1 is then the same as selecting $\widetilde{g}$ from $\widetilde{G}$ such that $\widetilde{g}$ has the most explanatory power for X_2. The problem is now one of variable selection.

Oone way is to proceed along the lines of Andrews and Lu (2001) or Donald and Newey (2001), who proposed using information type criteria to select the number of moment conditions in a GMM setting. However, if $\widetilde{G}$ has r columns, there would be 2^r models to evaluate, and as just mentioned, r, being the number of static factors, can be large. Now if there is a way to pre-order the components of $\widetilde{G}$, we would only need to evaluate r sets of $\widetilde{G}$. But the only available ordering we can exploit is that of $\widetilde{F}$, since by construction, these are r factors ordered such that $\widetilde{F}_{1t}$ explains the most variance in z_{it}, $\widetilde{F}_{rt}$ explains the most variance not explained by $\widetilde{F}_{1t}, \ldots \widetilde{F}_{r-1.t}$ and so on, with $\widetilde{F}_{jt}$ orthogonal to $\widetilde{F}_{kt}$ for $j \neq k$. However, of interest is not what explains the panel of instruments z_{it} per se, but what explains x_2. Factors that have strong explanatory power for Z need not be good instruments for x_2. Furthermore, $\widetilde{G}_t$ need not preserve the ranking of $\widetilde{F}_t$ once the effects of x_1 are partialled out. We now suggest boosting as a variable selection procedure that does not require pre-ordering the instruments.

3.2 Selecting $\widetilde{f}$ by Boosting

Boosting was initially introduced by Freund (1995) and Schapire (1990) to the machine learning literature as a classification device. It is recognized for its ability to find predictors that improves prediction without overfitting and for its low mis-classification error rate. Buhlmann and Hothorn (2006) provide an excellent introduction to boosting from a statistical perspective. For our purpose, it is best to think of boosting as a procedure that performs model selection and coefficient shrinkage simultaneously. This means that variables not selected are set to zero, as opposed to being shrunk to zero. In consequence, boosting results in very sparse models. Here, we want to use boosting to select the most relevant instruments from a large set of feasible instruments with no natural ordering to the instruments in the set.

The specific L_2 boost algorithm we use is based on component-wise least squares. Component-wise boosting was considered by Buhlmann and Yu (2003) when the number of predictors is large. Instead of evaluating sets of regressors one set at a time, the regressors are evaluated one at a time. Under component-wise boosting the r-th instrument is as likely to be chosen as the first; the ordering does not matter. The algorithm for fitting $\Phi(\widetilde{G})$, the conditional mean of a variable X_2 $(T \times 1)$ with a set of r predictors $\widetilde{G}$ is as follows.

1 let $\widehat{\phi}_0 = \bar{X}_2$;

2 for $m = 1, \ldots M$

  a for $t = 1, \ldots T$, let $u_t = X_{2t} - \widehat{\phi}_{m-1}$ be the 'current residuals'.

  b for each $i = 1, \ldots r$, regress the current residual vector u on $\widetilde{G}_{.,i}$ (the i-th regressor) to obtain $\widehat{b}_i$. Compute the $\widehat{e}_{.,i} = u - \widetilde{G}_{.,i}\widehat{b}_i$ as well as $SSR_i = \widehat{e}'_{.,i}\widehat{e}_{.,i}$;

  c let i^* be such that $SSR_{i^*} = \min_{i \in [1,\ldots r]} SSR_i$;

d let $\widehat{\phi}_m = \widetilde{G}_{.,i^*}\widehat{b}_{i^*}$

3 for $t = 1, \ldots T$, update $\widehat{\Phi}_{t,m} = \widehat{\Phi}_{t,m-1} + \nu\widehat{\phi}_{t,m}$, where $0 \leq \nu \leq 1$ is the step length.

Component-wise L_2 boost is nothing more than repeatedly fitting least squares to the current residuals and selecting at each step the predictor that minimizes the sum of square residuals. Note that component wise L_2 boosting selects one predictor at each iteration, but the same predictor can be selected more than once during the M iterations. This means that boosting makes many small adjustments, rather than accepting the predictor one and for all. This seems to play a role in the ability of boosting not to overfit. Step 3 is not directly relevant since we are interested in boosting as a selector, and not the fit it produces.

After m steps, boosting produces $\widehat{\Phi}_m(\widetilde{G}) = \widetilde{G}\widehat{\delta}_m$. If ι_{i^*} is a selection vector that is unity at position i^* and zero otherwise, then $\widehat{\delta}_m$ can be shown to follow the recursion.

$$\widehat{\delta}_m = \widehat{\delta}_{m-1} + \iota_{i^*} \odot \widehat{b},$$

where $\odot$ denotes element by element multiplication. Thus, $\widehat{\delta}_m$ and $\widehat{\delta}_{m-1}$ differ only in the i^*-th position. If X_2 is $T \times K_2$, the algorithm needs to be repeated K_2 times.

3.3 Determining the Stopping Rule

The boosting estimate of the conditional mean can be rewritten as $\widehat{\Phi}(\widetilde{G}) = B_m Y$, where

$$\begin{aligned} B_m &= B_{m-1} + \nu P_{\widetilde{G}}^{(m)}(I_T - B_{m-1}) = I_T - (I_T - \nu P_{\widetilde{G}}^{(1)})(I - \nu P_z^{(2)}) \cdots (I_T - \nu P_{\widetilde{G}}^{(M)}) && (6) \\ &= I_T - \prod_{j=1}^{m}(I_T - \nu P_{\widetilde{G}}^{(m)}) && (7) \end{aligned}$$

where $P_{\widetilde{G}}^m = \widetilde{G}_{.,m}(\widetilde{G}'_{.,m}\widetilde{G}_{.,m})^{-1}\widetilde{G}'_{.,m}$, the projection matrix based upon the regressor selected at the m-th step. A distinctive feature of boosting is that it will produce a sparse solution when the underlying model structure is sparse. In our context, a sparse structure occurs when there are many relatively weak instruments and Ψ has small values in possibly many positions.

If M (the number of boosting iterations) tends to infinity, we will eventually have a saturated model in which case all predictors are used. The sparseness of $\widehat{\delta}_M$ is possible only if boosting is stopped at the 'appropriate' time. In the literature, M is known as the stopping rule. Because we apply the boosting algorithm to potential predictors that are themselves estimated, we need to take care of one detail.

Proposition 2 *Let $\widehat{\Phi}(G)$ be the boosting estimate of the conditional mean when the factors are observed and let $\widehat{\Phi}(\widetilde{G})$ be obtained when the latent factors are replaced by the principal component estimates. Then $\widehat{\Phi}(\widetilde{G}) - \widehat{\Phi}(G) = o_p(1)$ if $\frac{M}{\min[\sqrt{N},\sqrt{T}]} \to 0$ as $N, T \to \infty$.*

The projection matrix formed using $\widetilde{G}$ has an error rate of $\min[\sqrt{N}, \sqrt{T}]^{-1}$. Repeated use of it thus accumulates sampling error. Proposition 1 puts discipline on how long the boosting algorithm can continue. Zhang and Yu (2005) and Buhlmann (2006) showed that when G is observed, boosting will consistently estimate the true conditional mean, even if the number of units in G increases with T. Thus if $\widehat{\Phi}(G) \xrightarrow{p} \Phi(G)$, we also have $\widehat{\Phi}(\widetilde{G}) \xrightarrow{p} \Phi(G)$.

In practice, cross-validation is often used to determine M when a researcher has access to training samples. But this is not often the case in time series economic applications. Let

$$df_m = trace(B_m).$$

Consider the information criterion

$$M = \underset{m=1,\dots \bar{M}}{\operatorname{argmin}} \; IC(m)$$

$$IC(m) = \log(\widehat{\sigma}_m^2) + \frac{A_T \cdot df_m}{T}$$

where in light of Proposition 2, we set the upper bound on M at

$$\bar{M} = c \cdot \min[N^{1/3}, T^{1/3}] \qquad c > 0.$$

The BIC obtains when $A_T = \log(T)$, and when $A_T = 2$, the criterion is as proposed in Buhlmann (2006). The primary departure from the standard AIC/BIC is that the complexity of the model is measured by the degrees of freedom, rather than by the number of predictors. In our experience, the degrees of freedom in a model with k predictors tends to be higher than k.

Boosting determines that $\widetilde{G}_{t,j}$ is an instrument if $\widehat{\delta}_{M,j}$ is non-zero, and $\widehat{\delta}_M$ is expected to be sparse if the number of truly relevant instruments is small. The sparseness of $\widehat{\delta}_M$ is a feature also shared by the LASSO estimator of Tibshirani (1996), defined as

$$\widehat{\delta}_L = \underset{\delta}{\operatorname{argmin}} \left\| X_2 - \widetilde{G}\delta_2) \right\|^2 + \lambda \sum_{j=1}^{\bar{r}} |\delta_j|.$$

That is, LASSO estimates δ subject to a L_1 penalty. Instrumental variable selection using a ridge penalty has been suggested by Okui (2004) to solve the 'many instrument variable' problem.[5] The ridge estimator differs from LASSO only in that the former replaces the L_1 by an L_2 penalty. Because of the nature of L_2 penalty, the coefficients can only be shrunk towards zero but will not usually be set to zero exactly. As shown in Tibshirani (1996), the L_1 penalty performs both subset variable selection and coefficient shrinkage simultaneously. Efron et al. (2004) showed that certain

[5]In a recent paper, Carrasco (2006) also takes a regularization approach to reduce the number of instruments.

forward-stage wise regressions can produce a solution path very similar to LASSO, and boosting is one such forward-stagewise regression. We consider boosting because the exact LASSO solution is numerically more difficult to solve, and that boosting has been found to have better properties than LASSO in the statistics literature.

Let L be the number of non-zero elements of $\widehat{\delta}_M$, and let $l_1, \ldots, l_L$ denote their position in $\widehat{\delta}_M$. Then $\widetilde{f}$ is defined as

$$\widetilde{f}_t = \{\widetilde{F}_{tl_1}, \widetilde{F}_{tl_2}, \ldots \widetilde{F}_{tl_L}.\}$$

This is the eventual set of instruments that is used in GMM estimation. As K_2 instruments are necessary for β_2 to be identified, it follows that if $L < K_2$, then β cannot be identified. In some analysis such as DSGE models, β are the structural coefficients which are functions of deep parameters of the model, say, θ with $\beta = h(\theta)$. To estimate θ, one can first estimate β, and then find estimates of θ that minimize the distance between $\widehat{\beta}$ and $h(\theta)$. It is then immediate that if the structural coefficients, β are not identified, there is no hope of identifying the deep parameters, θ. The L that emerges from boosting can be used to determine whether the necessary (but not sufficient) condition for identification of θ is satisfied.

We have motivated boosting as a method for selecting instruments when the instruments are estimated factors. But by assumption, Z is a set of variables that are weakly exogenous for the parameter of interest and are thus valid instruments. It follows that boosting can also be used to select observed instruments. When there is a large number of valid observed instruments (which exceeds one hundred in the empirical considered), boosting provides an effective and systematic method of instrument selection especially when the instruments have no natural ordering.

4 Finite Sample Properties

In this section, we study the finite sample properties of the fIV. It is implemented as follows:

1. estimate r factors by the method of principal components from the panel of data, Z. Form the potential instrument set, $\widetilde{F}$;
2. partial out the effect of x_1 from x_2 and $\widetilde{F}$ to yield X_2 and $\widetilde{G}$;
3. use boosting to determine $\widehat{\Phi}(\widetilde{G}) = \widetilde{G}\widehat{\delta}_M$, where M is determined by an information criterion with $\bar{M} = 10\min[N^{1/3}, T^{1/3}]$.
4. let $l_1, \ldots l_L$ be those positions in $\widehat{\delta}_M$ that are non-zero. Let $\widetilde{f}_t = (\widetilde{F}_{tl_1}, \ldots \widetilde{F}_{tl_L})$;
5. perform GMM estimation with $\widetilde{f}^+ = [x_1 \;\; \widetilde{f}]$ as instruments and an identity weighting matrix to yield $\breve{\beta}$. Let $\breve{S} = \frac{1}{T}\sum_{t=1}^{T} \breve{\varepsilon}_t^2 \widetilde{f}_t^+ \widetilde{f}_t^{+\prime}$, $\breve{\varepsilon}_t = y_t - x_t'\breve{\beta}$;
6. re-do GMM one more time using $W = \breve{S}^{-1}$ to yield $\widehat{\beta}_{fIV}$.

The FIV obtains when $\widetilde{f}^+ = \widetilde{F}^+$. For the sake of comparison, we report three other estimators. The first is labeled CFn, the control function estimator. It uses $\widetilde{f}_t^+$ as instruments, which amounts to using $W = \widehat{\sigma}_\varepsilon^{-2}(\widetilde{f}'^+ \widetilde{f}^+)$ as the weighting matrix. The second estimator is GMM with the first L variables in the panel of Z as observed instruments. This is labeled IV. The final estimator is OLS, which does not account for endogeneity bias. We compare the estimators using the mean estimate and the root-mean-squared error. The t statistic for testing β_2 is only reported for the fIV.

4.1 Simulations

We consider three data generating processes. In all cases,

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

where $e_{izt} \sim N(0,1)$, $e_{jft} \sim N(0,1)$, $\lambda_{iz} \sim N(0,1)$, $\rho_j \sim U(.2,.8)$. The examples differ in how y, x_1, and x_2 are generated.

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

$$\begin{aligned} y_t &= x_1\beta_1 + x_{2t}'\beta_2 + \sigma_y \varepsilon_t \\ x_{i1t} &= \alpha_x x_{it-1} + e_{ix1t}, \quad i = 1, \ldots K_1 \\ x_{i2t} &= \lambda_{i2} F_t + e_{ix2t} \quad i = 1, \ldots K_2 \end{aligned}$$

with $\varepsilon_t = \frac{1}{\sqrt{2}}(\widetilde{\varepsilon}_t^2 - 1)$ and $e_{ix2t} = \frac{1}{\sqrt{2}}(\widetilde{e}_{ix2t}^2 - 1)$. We assume $\rho_x \sim U(.2,.8)$, $e_{ix1t} \sim N(0,1)$ and uncorrelated with $\widetilde{e}_{2jt}$ and $\widetilde{\varepsilon}_t$. Furthermore, $(\widetilde{\varepsilon}, \widetilde{e}_{2t}') \sim N(0_{K2+1}, \Sigma)$ where $diag(\Sigma) = 1$, $\Sigma(j,1) = \Sigma(1,j) \sim U(.3,.6)$, and zero elsewhere. This means that $\widetilde{\varepsilon}_t$ is correlated with $\widetilde{e}_{ix2t}$ with covariance $\Sigma(1,i)$ but $\widetilde{e}_{ix2t}$ and $\widetilde{e}_{jx2t}$ is uncorrelated. By construction, the errors are heteroskedastic. The parameter σ_y^2 is set to $K_1\bar{\sigma}_{x1}^2 + K_2\bar{\sigma}_{x2}^2$ where $\bar{\sigma}_{x_j}$ is the average variance of x_{jt}, $j = 1, 2$. This puts the noise-to-signal ratio in the primary equation of roughly one-half.

The parameter of interest is β_{12}, the coefficient on the first endogenous variable. We considered various values of K_2, σ_z, and r. To evaluate the difference between the fIV and FIV, we vary $rmax$, the number of factors used in FIV, while boosting determines the number of factors used in fIV. The results are reported in Table 1 with $K_2 = 1$, and $\sigma_z = 3$. This is the least favorable situation since the factors are less informative with a low common component to noise ratio. The column labeled $\rho_{x_2\varepsilon}$ is the correlation coefficient between x_{12} and ε and thus indicates the degree of endogeneity. Under the assumed parametrization, this correlation is around .2. The true value of

β_{12} is 2, and the impact of endogeneity bias on OLS is immediately obvious. The three estimators that use the factors as instruments are less biased.

The factor based instruments dominate the IV either in bias or RMSE, if not both. The fIV generally has better properties than the CFn because the CFn imposes homoskedasticity when the data have heteroskedasstic errors. The J test associated with the fIV is close to the nominal size of 5%, while the two-sided t statistic for testing $\beta_{12} = 2$ has some size distortion when N, T are both small. The size distortions of both tests decrease with T.

Example 2 In this example, all data are generated via the factor model. The regression model is

$$y_t = x'_{2t}\beta_2 + \varepsilon_t. \tag{8}$$

The endogenous variables x_2 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 T observations for y, a $T \times N$ matrix z, and a $T \times L$ matrix X_2 as

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

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

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

where $\beta^* = \Lambda_x'^{-1}\Lambda_y$ is $L \times 1$ and $\varepsilon^* = e_y - e_x\beta^*$. For given Λ_x, we then solve for Λ_y such that $\beta_2^* = (1'_{K_2}, 0'_{L-K_2})$. The x_2 in (8) corresponds to the first K_2 columns of X_2. This also implies that the true value of every element of β_2 is unity. The endogeneity bias is $\beta'\text{cov}(e_x)\beta$. For the loadings, we assume $\lambda_z \sim N(0_N, I_N)$. The elements of the $L \times L$ matrix Λ_x are drawn from the $N(1,1)$ distribution. Written in terms of r factors, $X_2 = F(:, 1:r)\Lambda_x^{(r)}$ where $\Lambda_x^{(r)}$ only has the first $L \times L$ positions being non-zero. Viewed this way, the first L factors are the relevant factors.

We estimate $rmax = r + 2$ factors and let boosting determine how many estimated factors to use as instruments. We perform simulations for $K_2 = 2$ with $\beta_2^0 = (1\ 2)'$. The parameter of interest is β_{12}, the coefficient on the first endogenous variable. The results are reported in Table 2. Unlike Example 1, the correlation between x_2 and ε is now negative. In this example, the IV is actually more biased than OLS. However, the factor IV estimators all perform well, with the FIV being the more biased than fIV and CFn. The CFn performs better than the fIV here because the data are homoskedadstic and imposing the restriction improves efficiency. The t and J tests have rejection rates close to the nominal size.

One of the purposes of considering the two examples above is to assess the usefulness of selecting $\widetilde{f}$ from $\widetilde{F}$, where $\widetilde{F}_t$ is $rmax$ dimensional. As can be seen from $\widehat{L}$, the average dimension of $\widetilde{f}_t$ is much smaller than $rmax$ and is much closer to r. This shows that boosting chooses the appropriate number of relevant instruments on average. For a given N and T, the bias of the FIV increases with $rmax$, as predicted by the theory. By using fewer instruments, the fIV has a smaller bias but a slight larger than RMSE than the FIV. In general, the fIV is more stable with respect to the size of the feasible instrument set.

Example 3 Here, we consider estimation of β by panel regressions. The data are generated as

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

where F_t is again $r \times 1$. We set the true value of β to (0,1) but include an intercept in the regression. According to Theorem 2, we can use the factors estimated from x_{it} to instrument itself. Because of the pooled nature of the estimator, boosting was not used. We simply estimate L factors, where L is determined by the IC_2 criterion developed in Bai and Ng (2002). For the PFIV, we use $r+2$ factors. The PCFn imposes homoskedasticity while the PfIV and PFIV do not. Note that these estimates are not corrected for bias in order to to show that the bias is of second order importance. We do not consider the PIV since there is no valid observed instrument in this example. For the sake of comparison, we also consider PTfIV. Note that in this example, $E(\lambda_i) = 0$ and the PTfIV should be more unstable because $S_{\widetilde{F}x}$ can be near singular.

The results are reported in Table 3. As expected, the pooled POLS estimator is quite severely biased. The PTfIV has noticeably larger RMSE than the three factor based estimators, which are all centered around the true value. The PfIV has smaller bias than the PFIV with no increase in variance. Even with $\min[N, T]$ as small as 25, the PfIV is quite precise. Increasing N and/or T clearly improves precision even without bias correction. Because the PfIV has a small variance, the t test becomes very sensitive to small departures of the estimate from the true value. Thus, without bias correction, the t test based on the PfIV has important size distortions. The bias-corrected test is, however, much more accurate though there is still size distortions when r is large. The 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.

4.2 Application

The Phillips curve has long played an important role in our understanding of inflation dynamics. The New Keynesian Phillips Curve (NKPC) is

$$\pi_t = \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + \lambda x_t + \varepsilon_t$$

where π_t is inflation, x is a measures the state of the economy and is often taken to be real marginal cost or the output gap. The equation allows for backward and forward looking expectations and is thus a hybrid Phillips curve. The NKPC implies that the orthogonality conditions

$$E[z_{t-1}(\pi_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1} - \lambda x_t)] = 0$$

should hold for any set of variables z_{t-1} available in period $t-1$. Gertler and Gali (1999) first estimated the equation with four lags of inflation, labor share, commodity price and wage inflation, long-short interest rate spared, and the output gap as measured by detrended log GDP as instruments. Many researchers have raised issues about these instruments.[6] Some suggest that they are 'weak', while others suggest that the estimates are potentially biased because of the large number of moment conditions. Theorem 1 suggests that the many instrument problem can be alleviated by taking a factor instrumental variable approach. In effect, we expand the space spanned by the instruments from variables within the model to variables outside of the model but within the economic system. Because $\widetilde{f}_t$ is low dimension, we also alleviate the many instrument problem.

We use data collected in Ludvigson and Ng (2005), which consist of quarterly observations on a panel of 209 quarterly macroeconomic series for the sample 1960:1-2002:4, where some are monthly series aggregated to quarterly levels. We remove two variables from the panel, the GDP deflator used to construct π_t, and unit labor cost. Real unit labor cost (RULC) is the first difference of log nominal labor cost divided by the GDP deflator. The remaining 207 series are then used to estimate 8 factors.[7] We consider 3 different models to evaluate different sets of instruments. These are listed in Table 4.

In the base case, the endogenous variable is future inflation. Boosting selects π_{t-2} and x_{t-1} as observed instruments for Models A and B, but π_{t-2}, $\widetilde{F}_{t-1,2}$, and $\widetilde{F}_{t-1,6}$ for Model C. As instruments

[6] See, for example, Kichian et al. (2004) and Ma (2002).

[7] The PCP criterion developed in Bai and Ng (2002) also chooses 8 factors.

used in the fIV, boosting selects factors 1,2,3,6, 8 for Model A and 1,2,3,6 for Model B. When $\widetilde{F}_{t-1}$ and $\widetilde{F}_{t-2}$ are both in the feasible instrument set, boosting selects factors 1,2,3,6 from $\widetilde{F}_{t-1}$ and factors 1,2 from $\widetilde{F}_{t-2}$. Notably, boosting drops 10 of the 16 factors considered in Model C. Although $\widetilde{F}$ are linear combinations of the true underlying factors, we can give some interpretation to the factors by considering the marginal R^2 of each factor in each series. This is obtained by regressing each series on the factors one at a time. This exercise reveals that the highest marginal R^2 associated with $\widetilde{F}_1$ is UTIL1 (capacity utilization), $\widetilde{F}_2$ is PUXHS (CPI excluding shelter), $\widetilde{F}_3$ is DLAGG (a composite index of seven leading indicators), $\widetilde{F}_6$ is GDEXIM (terms of trade), and $\widetilde{F}_8$ is GMXQF (exports minus imports). While the observed instruments such as output gap also contain information about capacity utility and inflation, boosting suggests that open economy variables have predictive power for inflation that could have been exploited.

The results are reported in Table 4. We find little effect of marginal cost on inflation dynamics though the evidence for forward looking behavior is strong. The point estimate of γ_f is slightly below .8. This is larger than the IV estimate of .7 which is slightly higher than those reported in the literature. See, for example, Gali et al. (2005). The J tests for the three models are 6.045, 4.911 and 8.424. We cannot reject the model at the 5% level. Because $\widetilde{F}_{1t}$ is highly correlated with real variables, we also used it in place of real marginal cost as x_t. The results are similar. We also considered an alternative specification that treats both x_t and π_{t+1} as endogenous variables without changing the set of feasible instruments. The results continue to find no statistically significant effect of real marginal cost on inflation and a $\widehat{\gamma}_f$ of around .7.

Instrumental variables estimation plays a central role in economic analysis, and there has been much research to resolve estimation and inference problems that arise from the number and properties of instruments. The NKPC is an example in which such problems arise. Rather than developing bias corrections and robust tests to improve finite sample inference, we suggest that constructing more efficient instruments can be a solution to these problems. As our analysis showed, there are indeed instruments in the feasible set that are not relevant, and the factor based instruments can produce point estimates that are different from the observed instruments. Conditional on the factor structure being true, the fIV produces more precise estimates because it uses more more efficient instruments.

5 Conclusion

In this paper, we take as starting point that in a data rich environment, there are many valid instruments that are weakly exogenous for the parameters of interest. Pooling the information across instruments enable 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 and a fIV estimator that further reduces bias. For a large simultaneous systems, we show that valid instruments can be constructed from invalid ones. We also suggest boosting as a useful procedure for selecting the relevant instruments from the feasible set. The resulting fIV estimator has good finite sample properties. The factor instruments can be used in placed of observed instrumental variables in hypothesis testing, so they can potentially improve inference. We leave this for future research to focus on our main point that $\widetilde{F}$ forms a better set of instruments than the observed instruments when the data admit a factor structure, and that $\widetilde{f}$ can further reduce bias in finite samples.

Table 1: Finite Sample Properties of $\widehat{\beta}$, $\beta_{12}^0 = 2, \sigma_z^2 = 3$.

T	N	r	rmax	fIV	FIV	CFn	IV	OLS	$\rho_{x_2\varepsilon}$	$\widehat{L}$	J	$t_{\widehat{\beta}_{12}}$
				Mean/RMSE								
50	50	5	8	2.04	2.04	2.05	2.28	2.30	0.21	4.02	0.04	0.10
				0.27	0.24	0.29	1.62	0.39				
100	50	5	8	2.03	2.03	2.04	2.27	2.33	0.23	4.26	0.02	0.07
				0.19	0.18	0.20	3.46	0.38				
100	100	5	8	2.01	2.02	2.02	2.10	2.27	0.20	4.77	0.08	0.08
				0.16	0.15	0.18	1.81	0.32				
200	100	5	8	2.02	2.02	2.02	2.54	2.37	0.26	4.58	0.09	0.06
				0.14	0.13	0.15	5.52	0.40				
50	50	5	12	2.20	2.23	2.28	2.54	2.78	0.48	4.39	0.05	0.22
				0.39	0.37	0.45	2.34	0.86				
100	50	5	12	2.13	2.17	2.16	2.58	2.92	0.55	3.90	0.07	0.14
				0.29	0.28	0.32	2.86	0.97				
100	100	5	12	2.21	2.26	2.25	2.27	3.05	0.61	3.80	0.13	0.20
				0.38	0.36	0.42	1.23	1.09				
200	100	5	12	2.03	2.05	2.04	2.03	2.75	0.46	4.93	0.15	0.07
				0.15	0.14	0.16	0.86	0.77				

Table 2: Finite Sample Properties of $\widehat{\beta}$, $\beta_2^0 = 1$, $\sigma_z^2 = 3$.

T	N	r	L	fIV	FIV	CFn	IV	OLS	$\rho_{x_2\varepsilon}$	$\widehat{L}$	J	$t_{\widehat{\beta}_{12}}$
				Mean/RMSE								
50	50	4	4	1.04	1.06	1.04	1.16	0.92	-0.40	4.20	0.04	0.07
				1.33	1.16	1.30	2.07	0.50				
100	50	4	4	0.98	1.02	0.97	1.02	1.10	-0.38	4.10	0.04	0.04
				1.08	0.80	1.07	1.65	0.34				
100	100	4	4	1.02	1.03	1.02	0.99	0.93	-0.34	4.20	0.05	0.06
				0.71	0.71	0.70	1.38	0.39				
200	100	4	4	1.07	1.10	1.07	1.29	1.04	-0.46	4.25	0.05	0.06
				0.73	0.64	0.73	1.50	0.19				
50	50	6	4	1.01	1.06	1.04	1.25	0.95	-0.41	5.07	0.02	0.07
				1.78	1.31	1.79	2.78	0.46				
100	50	6	4	1.01	1.02	1.02	1.05	0.80	-0.37	5.52	0.04	0.07
				0.79	0.77	0.77	1.17	0.41				
100	100	6	4	1.05	1.06	1.05	1.07	0.85	-0.49	5.29	0.06	0.07
				0.40	0.40	0.39	0.97	0.26				
200	100	6	4	1.01	1.02	1.01	1.15	0.82	-0.50	5.99	0.06	0.06
				0.32	0.31	0.31	1.15	0.23				

Note: fIV are FIV are GMM estimators with $\widetilde{f}$ and $\widetilde{F}$ as instruments, where $\widetilde{f}_t$ and $\widetilde{F}_t$ are of dimension L and r, respectively. CFn is the 2SLS with $\widetilde{f}$ as instruments. IV is the GMM estimator with $z_2 \subset Z$ as instruments, where z_{2t} is the same dimension as $\widetilde{f}_t$. $\widetilde{f}_t$ is selected from $\widetilde{F}_t$ by boosting. The t statistic is based on $\widehat{\beta}_{fIV}$.

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

T	N	r	rmax	PfIV	PfIV$^+$	PFIV	PCFn	PTFIV	POLS	$\rho_{x_2\varepsilon}$	$t_{\widehat{\beta}_{PfIV}}$	$t_{\widehat{\beta}_{PfIV^+}}$	$t_{\widehat{\beta}_{PTfIV}}$
				Mean/RMSE									
15	15	2	4	1.05	1.02	1.08	1.05	1.12	1.10	0.29	0.34	0.17	0.10
				0.06	0.05	0.09	0.06	0.22	0.11				
25	25	2	4	1.03	1.01	1.06	1.03	1.08	1.10	0.30	0.38	0.10	0.07
				0.04	0.02	0.06	0.04	0.18	0.10				
25	50	2	4	1.03	1.01	1.05	1.03	1.07	1.10	0.30	0.48	0.08	0.08
				0.03	0.02	0.05	0.03	0.17	0.10				
50	25	2	4	1.02	1.01	1.04	1.02	1.07	1.09	0.27	0.39	0.09	0.10
				0.03	0.02	0.04	0.03	0.13	0.10				
50	50	2	4	1.02	1.00	1.03	1.02	1.06	1.10	0.29	0.36	0.06	0.08
				0.02	0.01	0.03	0.02	0.13	0.10				
100	50	2	4	1.01	1.00	1.02	1.01	1.05	1.09	0.28	0.36	0.06	0.09
				0.01	0.01	0.02	0.01	0.10	0.09				
50	100	2	4	1.01	1.00	1.03	1.01	1.04	1.10	0.29	0.48	0.06	0.06
				0.01	0.01	0.03	0.01	0.13	0.10				
100	100	2	4	1.01	1.00	1.02	1.01	1.04	1.10	0.29	0.38	0.06	0.07
				0.01	0.00	0.02	0.01	0.11	0.10				
15	15	4	6	1.04	1.03	1.07	1.04	1.08	1.07	0.28	0.44	0.30	0.15
				0.05	0.05	0.07	0.05	0.14	0.08				
25	25	4	6	1.03	1.01	1.05	1.03	1.06	1.07	0.29	0.44	0.16	0.14
				0.03	0.02	0.05	0.03	0.11	0.07				
25	50	4	6	1.03	1.01	1.05	1.03	1.06	1.08	0.30	0.74	0.20	0.12
				0.03	0.02	0.05	0.03	0.10	0.08				
50	25	4	6	1.02	1.01	1.04	1.02	1.05	1.07	0.28	0.70	0.22	0.15
				0.03	0.02	0.04	0.03	0.08	0.07				
50	50	4	6	1.02	1.01	1.03	1.02	1.04	1.07	0.28	0.79	0.15	0.11
				0.02	0.01	0.03	0.02	0.08	0.07				
100	50	4	6	1.02	1.00	1.02	1.02	1.03	1.07	0.29	0.90	0.14	0.14
				0.02	0.01	0.03	0.02	0.06	0.07				
50	100	4	6	1.02	1.00	1.03	1.02	1.03	1.07	0.29	0.90	0.16	0.09
				0.02	0.01	0.03	0.02	0.08	0.07				
100	100	4	6	1.01	1.00	1.02	1.01	1.02	1.07	0.29	0.88	0.09	0.10
				0.01	0.00	0.02	0.01	0.05	0.07				

Note: PfIV and PFIV are panel instrumental variable estimators with $\widetilde{c}_{it} = \widetilde{\lambda}_i' \widetilde{f}_t$ and $\widetilde{C}_{it} = \widetilde{\lambda}_i' \widetilde{F}_t$ as instruments. PfIV$^+$ is the biased-corrected estimator. $\widetilde{F}_t$ is $r \times 1$, and $\widetilde{f}_t$ is $L \times 1$, where L is determined by the PCP criterion. PCFn is the same as pfIV but imposes conditional homoskedasticity. PTFIV is the 'traditional' panel IV estimator that uses $\widetilde{f}$ as instruments.

Table 4: Phillips curve estimates

		fIV	t	FIV	t	CFn	t_{fIV}	IV	t_{fV}	OLS	t_{OLS}
A	γ_b	0.23	2.17	0.26	2.53	0.25	2.89	0.29	1.73	0.48	8.94
	λ	-0.05	-1.85	-0.04	-1.66	-0.04	-1.60	-0.03	-1.08	-0.01	-0.47
	γ_f	0.79	7.16	0.77	7.26	0.77	7.88	0.72	3.91	0.49	9.04
B	γ_b	0.27	2.07	0.24	1.97	0.22	1.98	0.28	2.55	0.48	8.94
	λ	-0.04	-1.39	-0.04	-1.52	-0.04	-1.66	-0.03	-1.20	-0.01	-0.47
	γ_f	0.76	5.35	0.78	5.73	0.80	6.37	0.74	6.51	0.49	9.04
C	γ_b	0.35	4.64	0.31	4.18	0.34	4.54	0.38	4.69	0.48	8.94
	λ	-0.02	-1.01	-0.02	-0.75	-0.03	-1.18	-0.03	-1.11	-0.01	-0.47
	γ_f	0.66	8.57	0.68	9.22	0.66	8.06	0.64	7.69	0.49	9.04

Instruments:

Model	fIV, FIV	IV
A	$\widetilde{F}_{t-1}$	$\pi_{t-2},\, x_{t-1}, x_{t-2}$
B	$\widetilde{F}_{t-2}$	IV(A)+ $\widetilde{F}_{t-1,1}, \widetilde{F}_{t-1,2}$
C	$\widetilde{F}_{t-1}, \widetilde{F}_{t-2}$	IV(A)+ $\widetilde{F}_{t-1}$

Appendix

Proof of theorem 1: Let $\widetilde{g}_t(\beta^0) = \widetilde{F}_t\varepsilon_t$, $\varepsilon_t^0 = y_t - x_t'\beta_0$, $\bar{g} = \frac{1}{T}\sum_{t=1}^T \widetilde{g}_t$. Then

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

Now

$$\begin{aligned}\sqrt{T}\bar{g} &= T^{-1/2}\sum_{t=1}^T \widetilde{F}_t\varepsilon_t \\ &= T^{-1/2}\sum_{t=1}^T(\widetilde{F}_t - HF_t^0)\varepsilon_t + HT^{-1/2}\sum_{t=1}^T F_t^0\varepsilon_t \\ &= HT^{-1/2}\sum_{t=1}^T F_t^0\varepsilon_t + o_p(1)\end{aligned}$$

By Lemma 1(ii), $T^{-1/2}\sum_{t=1}^T(\widetilde{F}_t - HF_t^0)\varepsilon_t = O_p(\sqrt{T}/\min[N,T]) = o_p(1)$, provided that $\sqrt{T}/N \to 0$. By assumption, $T^{-1/2}\sum_{t=1}^T F_t^0\varepsilon_t \xrightarrow{d} N(0, S^0)$. Thus $\sqrt{T}\bar{g}\xrightarrow{d}N(0, H_0S^0H_0')$, where $H_0 = \text{plim}\, H$. But $\text{plim}\,\breve{S} = H_0S^0H_0'$. This implies that $\breve{S}^{-1/2}\sqrt{T}\bar{g}\xrightarrow{d}N(0,I)$. This further implies that

$$\sqrt{T}(\widehat{\beta}_{FIV} - \beta)\xrightarrow{d}N(0, \text{plim}(S_{\widetilde{f}x}'\breve{S}^{-1}S_{\widetilde{f}x})^{-1})$$

Proof of Proposition 1 Without loss of generality, assume there is no x_1 so that $x = x_2$. The asymptotic variance of the GMM estimator with a r observed variables as instruments is the probability limit of

$$\widehat{Avar}(\widehat{\beta}_{IV}) = \widehat{\sigma}_\varepsilon^2(T^{-1}x'P_zx)^{-1}$$

The asymptotic variance of the FIV when F is observed is the probability limit of

$$\widehat{Avar}^0(\widehat{\beta}_{FIV}) = \sigma_\varepsilon^2(T^{-1}x'P_Fx)^{-1}.$$

Now $x = F\Psi + u$, $P_zx = P_zF\Psi + P_zu$ and $T^{-1}P_zu = o_p(1)$. Thus,

$$T^{-1}x'P_zx_2 = T^{-1}x'P_zF\Psi + o_p(1).$$

Furthermore, $P_Fx = P_FF\Psi + u = F\Psi + o_p(1)$ and therefore

$$T^{-1}x'P_Fx = T^{-1}x'F\Psi + o_p(1) = T^{-1}x'(M_z + P_z)F\Psi + o_p(1).$$

Consider now

$$\widehat{Avar}(\widehat{\beta}_{IV})^{-1} - \widehat{Avar}^0(\widehat{\beta}_{FIV})^{-1} = T^{-1}(-x'M_zF\Psi) = -T^{-1}x'M_z(x-u)) = -T^{-1}x'M_zx + o_p(1) < 0.$$

The result holds when F is replaced by $\widetilde{F}$ since by Lemma 1, $\widehat{Avar}^0(\widehat{\beta}_{FIV}) - \widehat{Avar}(\widehat{\beta}_{FIV}) = O_p((\min[N,T])^{-1/2})$.

Proof of Proposition 1 We are interested in

$$\widetilde{B}_M - B_M = \prod_{m=1}^{M} \widetilde{a}_m - \prod_{m=1}^{M} a_m = (\widetilde{a}_1 - a_1)A_1 + B_1(\widetilde{a}_2 - a_2)A_2 + \ldots B_{M-1}(\widetilde{a}_M - a_M)A_M$$

where $\widetilde{a}_m = I_T - P_{\widetilde{G}}^{(m)}$, $a_m = I_T - P_G^{(m)}$, and $A_m = \prod_{j=m+1}^{M} \widetilde{a}_j$. But a_j and $\widetilde{a}_j$ are projection matrices whose largest eigenvalue is one, and thus $\|a_j\| \leq 1$ and $\|\widetilde{a}_j\| \leq 1$. It follows that $\|A_m\| \leq 1$ and $\|B_m\| \leq 1$ for all m. Furthermore, $(\widetilde{a}_m - a_m) = P_G^{(m)} - P_{\widetilde{G}}^{(m)}$ and $\|P_G^{(m)} - P_{\widetilde{G}}^{(m)}\| = O_p(\delta_{NT}^{-1})$ by Lemma A1 below. It follows that

$$\|B_{j-1}(\widetilde{a}_j - a_j)A_j\| \leq \|B_{j-1}\| \, \|\widetilde{a}_j - a_j\| \, \|A_j\| \leq \|\widetilde{a}_j - a_j\| = O_p(\delta_{NT}^{-1}),$$

and $\|\widetilde{B}_M - B_M\| = O_p(M/\delta_{NT})$.

Lemma A1 *Let $\widetilde{P} = \widetilde{F}(\widetilde{F}'\widetilde{F})^{-1}\widetilde{F}'$ where $\widetilde{F}$ is a $T \times r$ matrix of factors estimated from a $T \times N$ panel of data by the method of principal components. Also let $P = F(F'F)^{-1}F'$ where F is the factor matrix. From Lemma 2 of Bai and Ng (2002), with $\delta_{NT} = \min[\sqrt{N}, \sqrt{T}]$,*

$$\left\| \widetilde{P} - P \right\| = O_p(\delta_{NT}^{-1})$$

The above implies in our context that $\|P_{\widetilde{G}}^{(m)} - P_G^{(m)}\| = O_p(\delta_{NT}^{-1})$ for each m.

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 $\widehat{\beta}_{PFIV} = \beta + S_{\widetilde{x}\widetilde{x}}^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{C}_{it}\varepsilon_{it}$, 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

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

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

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

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

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

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

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

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

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

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

Thus

$$\begin{aligned}
(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widetilde{\Lambda}_i - H^{-1}\Lambda_i)'HF_t\varepsilon_{it} &= (NT)^{-1}\frac{1}{T}\sum_{i=1}^{N}\sum_{s=1}^{T}u_{is}F_s'H'H\sum_{t=1}^{T}F_t\varepsilon_{it} + o_p(1) \\
&= (N/T)^{1/2}\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}u_{is}F_s'\Big)H'H\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t\varepsilon_{it}\Big) + o_p(1) \qquad \text{(A.2)} \\
&= (N/T)^{1/2}O_p(1)
\end{aligned}$$

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

Proof of Theorem 2 part (ii). The biases equal to $S_{\widetilde{x}\widetilde{x}}^{-1}$ multiplied by the expected values of (A.1) and (A.2). 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.1) is A_tGB_t, which is a vector. Thus

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

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

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

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

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

Let

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

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

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

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

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

Next consider (A.2). Let

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

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

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

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

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

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

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

Let

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

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

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

then the asymptotic bias is

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

proving part (ii).

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

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

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

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

References

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

Andrews, D. W. K. and Lu, B. 2001, Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models, *Journal of Econometrics* **12**, 123–164.

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. 2006, Confidence Intervals for Diffusion Index Forecasts and Inference with Factor-Augmented Regressions, *Econometrica* **74:4**, 1133–1150.

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

Buhlmann, P. 2006, Boostinf for High-Dimensional Linear Models, *Annals of Statistics* **54:2**, 559–583.

Buhlmann, P. and Hothorn, T. 2006, Boosting: A Statistical Perspective, mimeo.

Buhlmann, P. and Yu, B. 2003, Boosting with the L_2 Loss: Regression and Classification, *Journal of the American Statistical Assocation* **98**, 324–339.

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

Donald, S. and Newey, W. 2001, Choosing the Number of Instruments, *Econometrica* **69:5**, 1161–1192.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004, Least Angle Regression, *Annals of Statistics* **32:2**, 407–499.

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

Freund, Y. 1995, Boosting a Weak Learning Algorithm by Majority, *Information and Computation* **121**, 256–285.

Galbraith, J. and Zinde-Walsh, V. 2005, Reduced Dimension Controls in Instrumental Variables Regression, mimeo, Mcgill University.

Gali, J., Gertler, M. and Lopez-Salido, D. 2005, Robustness of the Estimates of the Hybrid New Keynesian Phillips Curve, *Journal of Monetary Economics* **52**, 1107–1118.

Gertler, M. and Gali, J. 1999, Inflation Dynamics: A Structural Econometric Analysis, *Journal of Monetary Economics* **44**, 195–222.

Hahn, J. and Hausman, J. 2002, Notes on Bias in Estimators for Simultaneous Equations Models, *Economics Letters* **75**, 237–241.

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

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

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

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

Kichian, M., Dufour, J. M. and Khalaf, L. 2004, Are New Keynesian Phillips Curves Identified, Bank of Canada WP 2004-11.

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

Ludvigson, S. and Ng, S. 2005, The Empirical Risk Return Relation: A Factor Analysis Approach, *Journal of Financial Economics*.

Ma, A. 2002, GMM Estimation of the New Phillips Curve, *Economics Letters* **76**, 411–417.

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

Newey, W. and Smith, R. 2004, Higher Order Properties of GMM and Generalized Empirical Likeliood Estimators, *Econometrica* **71:1**, 219–255.

Okui, R. 2004, Shrinkage Methods for Instrumental Variable Estimation, mimeo.

Phillips, P. C. B. 1983, Exact Small Sample Theory in the Simultaneous Equations Models, *in* M. D. Intriligator and Z. Grilches (eds), *Handbook of Econometrics, Volume 1*, North Holland.

Schapire, R. E. 1990, The Strenght of Weak Learnability, *Machine Learning* **5**, 197–227.

Staiger, D. and Stock, J. H. 1997, Instrumental Variables Regression with Weak Instruments, *Econometrica* **65:3**, 557–586.

Stock, J. H. and Watson, M. W. 2004, Forecasting with Many Predictors, *Handbook of Forecasting.*

Tibshirani, R. 1996, Regression Shrinkage and Selection via the Lasso, *Journal of Royal Statistical Society Series B* **58:1**, 267–288.

Zhang, T. and Yu, B. 2005, Boosting with Early Stopping: Convergence and Consistency, *Annals of Statistics* **33:4**, 1538–1579.