Extremum Estimation when the Predictors are Estimated from Large Panels

Jushan Bai* Serena Ng †

June 25, 2008

Abstract

Much is written about the use of factors estimated by the method of principal components from large panels in linear regression models. In this paper, we provide an analysis for non-linear estimation and establish the conditions under which the estimated factors can be treated as though they were observable. The results can be used to estimate probabilities as in probit type analysis as well as classification of observations into types conditional on covariates. Comparison with traditional generated regressors is also made.

*Department of Economics, NYU, 269 Mercer St, New York, NY 10003 Email: Jushan.Bai@nyu.edu.

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

We also acknowledge financial support from the NSF (grants SES-0137084, SES-0136923, SES-0549978)

1 Introduction

The textbook treatment of generated regressors holds that although the coefficient estimates are consistent, the standard errors must adjust for the fact that the regressors are being estimated in a preliminary step. It turns out that whether this correction for standard error is necessary depends on how the preliminary step is performed. When the generated regressors are factors estimated from a panel of data with a large number of cross-section units (N) and a large number of time series observations T), we showed in Bai and Ng (2006a) that the estimated factors can be treated as though they are the true factors in linear regression models. The results provide the basis for the inferential theory of linear factor augmented regressors.

The linear model has broad uses, including diffusion forecasting based upon the single equation equation $y_{t+1} = \widetilde{W}_t'\delta + \varepsilon_t$, and vector autoregressions in $(y_t, \widetilde{W}_t')$, referred to as a FAVAR by Bernanke and Boivin (2003), where $\widetilde{W}_t$ consists of observable and estimated latent variables. However, many problems cannot be analyzed in a linear framework. To predict whether the economy is in a recession, and to quantify default risk, for example, we need to provide a binary classification based on the observed variables. When the number of observed covariates is small, a dynamic probit can be considered. As far as we are aware of, there does not exist a suitable framework when the number of informative covariates is large. Markov switching regressions of the type considered in Hamilton (1989) is also widely used in empirical applications. Many have extended the model to allow the transition probabilities to be state dependent, while using a small number of observed variables to proxy for the latent state. The restriction to a small number of observed variables is a consequence of a lack of a statistical framework to analyze non-linear models with latent variables.

One might also be interested in conditional moment restrictions that are non-linear in parameters. When the moment condition involves a latent variable, it would be tempting to 'construct' a regressor from a finite number of observed variables to serve as proxy for the latent variable. But such a proxy variable is the latent variable contaminated with an error that will not vanish. While estimation by instrumental variables will yield consistent estimates if the regression model is linear, the estimates can be inconsistent when the first step regression model is non-linear. Indeed, treatment of measurement error in non-linear models is a non-trivial problem even in an i.i.d. setting and remains an area of active research.

In this paper, we adopt a different approach. We posit that there is a large panel of

data from which the space spanned by the latent variable can be consistently estimated. We provide an analysis of extremum estimators when one or more of the regressors are factors estimated from a large dimensional panel by the method of principal components. We show that when the generated regressors are the estimated factors, sequential estimation is not a problem, provided $\frac{T^{5/8}}{N} \to 0$ as $N, T \to \infty$. The result is general and holds for M estimators including maximum likelihood estimation (MLE), non-linear least squares (NLS), and quantile regressors. Probit analysis is thus included as a special case. The result also applies to GMM estimators based upon a set of orthogonality conditions, as well as minimum-distance estimators (MDS) in which restricted estimates are formed from the unrestricted ones.

The driving force behind our result is that the first step estimation error vanishes at a faster rate than in conventional generated regressor problems, so that not only can we get consistent estimates of the parameters in the second step regression, there is not even a need to correct for standard errors under the assumptions of our analysis. In effect, the latent factors can be treated as though it is observed provided that N and T are both large. For linear models, we showed in Bai and Ng (2006a) that the factor estimates can be treated as though they are known if $\sqrt{T}/N \to 0$. The requirement that $T^{5/8}/N \to 0$ in the case of M estimation studied here implies that a larger N will be necessary for the sampling error in the factor estimates not to affect the second step estimation of the parameter of interest. Therefore, as in instrumental variable estimation, consistent estimation of models with latent variables that are non-linear in parameters may not be possible even when consistent estimation is possible in linear models. We note that $T^{5/8}$ may be replaced by $T^{\frac{1}{2}+\delta}$, where $\delta > 0$ can be arbitrarily small provided that the data possess moments of high enough orders.

2 Preliminaries

Suppose we observe $z_t = (y_t, x_t, F_t)$; $t = 1, \ldots T$. The data are generated according to a model with a set of finite dimensional unknown parameters θ. We assume that there is a function $Q_0(\theta)$ that is uniquely maximized at θ_0 for some $\theta_0 \in \Theta$, where Θ is the parameter space and is compact. Consider estimating θ as

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\, Q_T(\theta).$$

Following the literature, we refer to $\widehat{\theta}$ as an extremum estimator. We will separately consider two types of extremum estimators: M and GMM estimators. If

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} m_t(z_t, \theta),$$

the extremum estimator $\widehat{\theta}$ is an M estimator. Suppose z_t is iid with density $f(z_t|\theta)$. Let $m_t(\theta) = \log f(z_t|\theta)$. Then the maximum likelihood estimator (MLE) estimator of θ is an M estimator. Suppose the regression model is $y_t = h(x_t, F_t, \theta) + \varepsilon_t = h(W_t, \theta) + \varepsilon_t$, where $W_t = (x_t', F_t')'$. Letting

$$m_t(z_t, \theta) = -\Big[y_t - h(x_t, F_t, \theta)\Big]^2$$

yields the non-linear least squares estimator, which is an M estimator. If $h(W_t, \theta) = W_t'\theta$, the linear regression model obtains. LAD (least absolute deviation) and quantile regressions can also be cast in terms of M estimators.

Suppose $Eg(z_i, \theta) = 0$ if and only if $\theta = \theta_0$ and

$$Q_T(\theta) = -\Big[\frac{1}{T}\sum_{t=1}^{T} g(z_t, \theta)\Big]' W_T \Big[\frac{1}{T}\sum_{t=1}^{T} g(z_t, \theta)\Big],$$

the extremum estimator $\widehat{\theta}$ is a GMM estimator, where W_T is a weighting matrix. In consumption based asset pricing example of Hansen and Singleton (1982) x_t is an asset's return, y_t is consumption ratio $y_t = c_t/c_{t-1}$, and F_t a vector of instruments; β a rate of time preference, and γ a risk version parameter. Let $\theta = (\beta, \gamma)$ and $z_t = (y_t, x_t, F_t)$. Then

$$g(z_t) = F_t(\beta x_t y_t^{\gamma} - 1).$$

Included in the GMM class is the minimum distance estimator (MDE) where given an initial estimator $\widehat{\pi}$ for $\pi = \rho(\theta_0)$,

$$Q_T(\theta) = -[\widehat{\pi} - \rho(\theta)]' W_T (\widehat{\pi} - \rho(\theta)]$$

We assume the regularity conditions stated in Newey and Mcfadden (1994) hold, so that $Q_T(\theta)$ converges uniformly to $Q_0(\theta)$ on Θ. These assumptions immediately imply consistency of $\widehat{\theta}$ for θ_0. Since the objective function $Q_T(\theta)$ is smooth with respect to θ by assumption, the maximizer $\widehat{\theta}$ solves for the first order condition (FOC)

$$\nabla_\theta Q_T(\widehat{\theta}) = 0.$$

For future reference, we will represent this first order condition alternatively as

$$\frac{1}{T}\sum_{t=1}^{T} h(z_t, \widehat{\theta}) = 0.$$

2.1 The Generated Regressor Problem

This paper considers the problem that a component of z_t is not observable. This component is denoted by F_t. Instead, an estimated F_t is used in the estimation of θ. We study the effect of estimation of F_t on the inference of θ.

Suppose F_t is not observed, but $F_t = \ell(U_t, \gamma)$, where ℓ is a known function, U_t is observed, and γ is unknown parameter. Then $h(y_t, x_t, F_t, \theta) = h(y_t, x_t, \ell(U_t, \gamma), \theta) = h(z_t, \theta, \gamma)$ with $z_t = (y_t, x_t, U_t)$. By redefining z_t, we can account for the fact that F_t is not observed. Denote $\widehat{F}_t = \ell(U_t, \widehat{\gamma})$, the first order condition becomes

$$\frac{1}{T}\sum_{t=1}^{T} h(\widehat{z}_t, \theta) = 0$$

where $h(\widehat{z}_t, \theta) = h(y_t, x_t, \widehat{F}_t, \theta)$.

As an example, suppose F_t is the core rate of inflation which is not observed. One might specify inflation as a function of a small set of determinants, U_t. Then $\widehat{F}_t = U_t'\widehat{\gamma}$ is an estimate of core inflation. In the second step, $\widehat{F}_t$ enters a non-linear regression. In this conventional sequential estimation, one finds $\widehat{\theta}$ by solving

$$\frac{1}{T}\sum_{t=1}^{T} h(z_t, \theta, \widehat{\gamma}) = 0 \tag{1}$$

where $\widehat{\gamma}$ is a first step estimator by solving, say

$$\frac{1}{T}\sum_{t=1}^{T} g(z_t, \gamma) = 0$$

The estimator of $(\widehat{\theta}, \widehat{\gamma})$ can be viewed as a joint GMM estimator with stacked moments (see Newey and McFadden, 1994 for discussion). It can also be seen as a sequential estimator since $\widehat{\gamma}$ is obtained in a first step, and is substituted in for the γ in the second step. Newey (1983) formulates such a sequential estimation problem in a GMM framework and derived the asymptotic variance of the second step estimator . His results assumes that the first step estimation yields a $\sqrt{T}$ consistent estimates of γ. In general the first step estimation of γ will affect the limiting distribution of $\widehat{\theta}$.

Our point of departure is that F_t is completely unknown, rather than unknown up to a finite number of parameters. Let r be the dimension of F_t. We assume there is a panel data set (of N variables each with T observations) that is linked with F_t. We replace F_t, in the first step by the principal component estimates from the panel data. This departure has two implications for the second step estimation. First, our first step estimates $\widetilde{F}$, which is a $T \times r$ matrix, has the number of parameters increasing with T. Second, even though we estimate an increasing number of parameters in the first step, $\widetilde{F}_t$ is $\sqrt{N}$ consistent provided $N, T \to \infty$. Our main objective is to show that this allows us to treat $\widetilde{F}_t$ as though it is observed in the second step.

3 Estimation of the Factors

We assume that we have at our disposal for the first step a large panel of the data, x_{it}, that obeys a 'approximate' factor structure. That is, for $i = 1, \ldots N, t = 1, \ldots T$,

$$x_{it} = \lambda_i' F_t + e_{it}$$

where F_t is a $r \times 1$ vector of factors with λ_i as the corresponding factor loadings, and neither F_t nor λ_i is observed. Following the factor analysis literature, we call $\lambda_i' F_t$ is the common component and e_{it} is the idiosyncratic error. In matrix notation, the factor model is $X = F\Lambda' + e$, where X is a $T \times N$ matrix of stationary data, $F = (F_1, ..., F_T)'$ is $T \times r$, r is the number of common factors, $\Lambda = (\lambda_1, ..., \lambda_N)'$ is $N \times r$, and e is a $T \times N$ error matrix.

Our problem more precisely stated is to find $(\widetilde{F}, \widetilde{\theta})$ such that

$$\begin{aligned} (\widetilde{F}, \widetilde{\Lambda}) &= \min_{F,\Lambda} \sum_{t=1}^{T} \sum_{i=1}^{N} (x_{it} - \lambda_i' F_t)'(x_{it} - \lambda_i' F_t) \\ & \frac{1}{T} \sum_{t=1}^{T} h(\widetilde{z}_t, \widetilde{\theta}) = 0 \end{aligned}$$

where $\widetilde{z}_t = (y_t, x_t', \widetilde{F}_t')'$. The solution to the problem in the first step is the principal component estimator. Let $\widetilde{F} = (\widetilde{F}_1, ..., \widetilde{F}_T)$ be the matrix consisting of r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the matrix $XX'/(TN)$ in decreasing order. Then $\widetilde{\Lambda} = (\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_N)') = X'\widetilde{F}/T$, and $\widetilde{e} = X - \widetilde{F}\widetilde{\Lambda}'$. Also let $\widetilde{V}$ be the $r \times r$ diagonal matrix consisting of the r largest eigenvalues of $XX'/(TN)$, and $H = \widetilde{V}^{-1}(\widetilde{F}'F/T)(\Lambda'\Lambda/N)$.

For the factor estimation, we make the following assumptions.

Assumption A

A1: $E\|F_t\|^8 \leq M$ and $\frac{1}{T}\sum_{t=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F > 0$, an $r \times r$ non-random matrix.

A2: The loading λ_i is either deterministic such that $\|\lambda_i\| \leq M$, or it is stochastic such that $E\|\lambda_i\|^8 \leq M$. In either case, $N^{-1}\Lambda'\Lambda \xrightarrow{p} \Sigma_\Lambda > 0$, an $r \times r$ non-random matrix, as $N \to \infty$.

A3: e_{it} is weakly correlated, both over time and in the cross-section dimension.

A4: $\{\lambda_i\}$, $\{F_t\}$, and $\{e_{it}\}$ are three mutually independent groups. Dependence within each group is allowed.

A5: $E\|N^{-1/2}\sum_{i=1}^{N} \lambda_i e_{it}\|^8 \leq M$ for all t.

Assumption A concerns the factor model. A.1 and A.2 together imply r common factors. Assumption A.3 is formally given in Bai and Ng (2006a). It allows for heteroskedasticity and weak time series and cross section dependence in the idiosyncratic component, leading to the *approximate factor structure* of Chamberlain and Rothschild (1983). These assumptions are more general than a strict factor model. Assumption A.4 is standard in factor analysis.

Numerous authors have analyzed the properties of factor estimation. See, for example, Stock and Watson (2002). The results most relevant for the present analysis are summarized in the following lemma, proved in Bai and Ng (2002) and Bai (2003).

Lemma 1 *Let* $H = \widetilde{V}^{-1}(\widetilde{F}'F/T)(\Lambda'\Lambda/N)$ *and* $C_{NT}^2 = \min[N,T]$. *Under Assumption (A),*

(i) $\frac{1}{T}\sum_{t=1}^{T} \|\widetilde{F}_t - HF_t\|^2 = O_p(C_{NT}^{-2})$

(ii) If ξ_t *is uncorrelated with* e_{it} *for all* i *and* t *and* $E|\xi_t|^2 \leq M$ *for all* t, *then*

$$\frac{1}{T}\sum_{t=1}^{T}(\widetilde{F}_t - HF_t)\xi_t = O_p(C_{NT}^{-2})$$

Lemma 1 provides the basis of valid estimation and inference in linear regression models augmented with the estimated factors $\widetilde{F}_t$. It has been used in Bai and Ng (2006c), Bai and Ng (2006b), Bai and Ng (2006a) in differently motivated linear regression problems. However, for investigating the properties of non-linear estimators which is the objective of the present paper, we need a additional result.

Lemma 2 *Under the assumption that* $\max_{1\leq t\leq T}\|F_t\| = O_p(\alpha_T)$, *and* $T/N^2 \to 0$,

$$\max_{1\leq t\leq T}\|\widetilde{F}_t - HF_t\| = O_p(\alpha_T T^{-1}) + O_p(N^{-1/2}) + O_p(1)\max_{1\leq t\leq T}\frac{1}{N}\|\sum_{i=1}^{N}\lambda_i e_{it}\|$$

The proof of the above lemma is provided in the appendix. To establish consistency of extremum estimators, we need the sample objective function $Q_T(\theta)$ to converge uniformly to $Q_0(\theta)$ whose maximizer is θ_0. In our context, this requires uniform convergence of $\widetilde{F}_t$ to the space spanned by F_t. Lemma 2 shows that this convergence rate depends on α_T. If F_t are iid normal, then $\alpha_T = \log T$. If $E\|F_t\|^k \leq M$ for all t then, $\max_{1\leq t\leq T} \|F_t\| = O_p(T^{1/k})$ so that $\alpha_T = T^{1/k}$. Since we assumed $k = 8$, α_T can be taken as $T^{1/8}$. Together with assumption that $E\|N^{-1/2}\sum_{i=1}^N \lambda_i e_{it}\|^8 \leq M$ for all t, $\max_{1\leq t\leq T} \|N^{-1/2}\sum_{i=1}^N \lambda_i e_{it}\| \leq O_p(T^{1/8})$. Thus under our maintained assumptions, Lemma 2 can be restated as

$$\max_{1\leq t\leq T} \|\widetilde{F}_t - HF_t\| = O_p(\alpha_T T^{-1}) + O_p(T^{1/8})N^{-1/2} = o_p(1) \tag{2}$$

where the $o_p(1)$ requires $T^{1/4}/N \to 0$. It is not difficult to show that if the data are generated such that $F'F/T = I_r$ (or $E[F_tF_t'] = I_r$) and $N^{-1}\Lambda'\Lambda$ is a diagonal matrix with distinct elements arranged in decreasing order, then $H = I_r + O_p(C_{NT}^{-2})$. The preceding lemmas holds with H replaced by the identity matrix. In the following analysis, $H = I$ is assumed. In many cases, whether H is an identity matrix is unimportant. For example, in prediction problems, F_t and HF_t as predictors will give the same prediction. When used as instruments, F_t and HF_t contain the same information.

4 M estimators

The objective function of M estimators are of the form

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} m(z_t, \theta).$$

An application of it is the probit model. in which y_t is binary with $y_t = 0, 1$ and $P(y_t = 1|W_t) = \Phi(W_t'\theta)$, where $W_t = (x_t, F_t)$ is a vector of predictors, and $\Phi(v)$ is the cumulative distribution function (cdf) for a standard normal random variable. The density is given by

$$f(z_i|\theta) = \Phi(W_t'\theta)^{y_t}[1 - \Phi(W_t'\theta)]^{(1-y_t)}$$

Let

$$h(z_t, \theta) = \frac{\partial m(z_t, \theta)}{\partial \theta}$$

and

$$K(z_t, \theta) = \frac{\partial^2 m(z_t, \theta)}{\partial \theta \partial \theta'}.$$

Obviously, if $\widehat{\theta}$ maximizes Q_T, it must be that the first order condition holds. That is, $\bar{h}(\widehat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} h(z_t, \widehat{\theta}) = 0$ Under regularity conditions such as stated in Amemiya (1984), the estimator is consistent. If

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h(z_t, \theta_0) \xrightarrow{d} N(0, \Sigma)$$

where Σ is positive definite and $K_0 = E[K(z_t, \theta_0)]$ is non-singular, then

$$\sqrt{T}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, K_0^{-1}\Sigma K_0^{'-1}) \tag{3}$$

Replacing z_t by $\widetilde{z}_t = (y_t, x_t, \widetilde{F}_t)$ gives the feasible objective function

$$\widetilde{Q}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} m(\widetilde{z}_t, \theta).$$

Consider the estimator defined as

$$\widetilde{\theta} = \operatorname{argmax} \widetilde{Q}_T(\theta)$$

The linear factor augmented regressions considered in Bai and Ng (2006a) is a special case where $h(\widetilde{z}_t; \theta)$ is linear in θ. We now consider the general case when $h(\widetilde{z}_t, \theta)$ is non-linear in θ. The following assumptions will be made:

Assumption M1:

i) (a) Let θ_0 be an interior point of a compact set Θ. The function $Q_T(\theta)$ converges uniformly in probability to $Q(\theta)$ on Θ, and $Q(\theta)$ achieves its maximum at θ_0.

ii: $Q_T(\theta)$ is twice continuously differentiable at a neighborhood $\mathcal{N}$ of θ_0 , and

$$\sup_{\theta \in \mathcal{N}} |\frac{1}{T}\sum_{t=1}^{T} K(z_t, \theta) - K(\theta)| = o_p(1)$$

for some function $K(\theta)$. Let $K_0 = K(\theta_0)$ and K_0 is nonsingular.

iii: $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h(z_t, \theta_0) \xrightarrow{d} N(0, \Sigma)$, for some $\Sigma > 0$.

Assumption M1.(i) ensures that the extreme estimator is consistent for θ_0 and the remaining assumptions are for asymptotic normality of the estimator. Note that all these assumptions are imposed when the functions are evaluated at true F_t. For consistency of $\widetilde{\theta}$ with estimated F_t, we also need the following.

Assumption M2:

$$\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\|h(y_t,x_t,F_t^*,\theta)\|^2 = O_p(1) \tag{4}$$

uniformly in F_t^* that is in a neighborhood of F_t such that $\max_{1\le t\le T}\|F_t^* - F_t\| \le b_{NT}$ with $b_{NT}\to 0$.

For example, we can take b_{NT} to be the righthand side of (2). Assumption M2 is easily verified for probit model. Here for simplicity, assume there is no x_t, and the only regressors are F_t. Then

$$h(y_t,F_t^*,\theta) = \; y_t\frac{\phi(\theta'F_t^*)}{\Phi(\theta'F_t^*)}F_t^* - (1-y_t)\frac{\phi(\theta'F_t^*)}{1-\Phi(\theta'F_t^*)}F_t^* \tag{5}$$

From $\phi(v)/\Phi(v)\le C(1+|v|)$ for some bounded C, $|y_t|\le 1$ and $\|\theta\|\le M$ for some M since Θ is compact, the first term on the right hand is bounded by

$$\begin{aligned} C(1+\|\theta'F_t^*\|)\|F_t^*\| \quad &\le \quad C\|F_t^*\| + M\|F_t^*\|^2 \\ &\le \quad C\|F_t^* - F_t\| + C\|F_t\| + 2\|F_t^* - F_t\|^2 M + 2\|F_t\|^2 M \\ &\le C \quad b_{NT} + C\|F_t\| + 2Mb_{NT}^2 + 2M\|\|F_t\|^2. \end{aligned}$$

Using $1-\Phi(v) = \Phi(-v)$ and $\phi(v) = \phi(-v)$, the second term on the right hand side of (5) has the same upper bound. It follows that if $E\|F_t\|^2$ is finite, and Assumption M2 holds.

Lemma 3 *Under assumptions A, M1, and M2,*

$$\widetilde{\theta}\xrightarrow{p}\theta_0$$

Proof: We show that $\widetilde{Q}_T(\theta)$ is uniformly close to $Q_T(\theta)$. By the mean value expansion,

$$\widetilde{Q}_T(\theta) = Q_T(\theta) + \frac{1}{T}\sum_{t=1}^{T} h(y_t,x_t,F_t^\dagger,\theta)'(\widetilde{F}_t - F_t)$$

where $F_t^\dagger$ is between F_t and $\widetilde{F}_t$. By Assumption M2

$$\|\frac{1}{T}\sum_{t=1}^{T} h(y_t,x_t,F_t^\dagger,\theta)'(\widetilde{F}_t - F_t)\| \le \Big(\frac{1}{T}\sum_{t=1}^{T}\|h(y_t,x_t,F_t^\dagger,\theta)\|^2\Big)^{1/2}\Big(\frac{1}{T}\sum_{t=1}^{T}\|\widetilde{F}_t - F_t\|^2\Big)^{1/2} = O_p(C_{NT}^{-1})$$

uniformly in θ since the first expression on the right hand side is uniformly bounded in θ and the second expression does not depend on θ and is $O_p(C_{NT}^{-1})$. Consistency follows from

the argument as in Amemiya (1984). Note that the extremum estimator satisfies the first order condition:

$$\frac{1}{T}\sum_{t=1}^{T} h(\widetilde{z}_t, \widetilde{\theta}) = 0$$

□

To derive the limiting distribution, additional assumptions are needed.

Assumption M3:

(i) $\xi_t = (\partial/\partial F_t)h(z_t, \theta_0)$ is uncorrelated with e_{it} and $E\|\xi_t\|^2 \leq M$ for all t.

(ii) For $j = 1, 2, ..., p = dim(\theta)$,

$$\frac{1}{T}\sum_{t=1}^{T} \|\frac{\partial^2 h_j(y_t, x_t, F_t^*, \theta^*)}{\partial F_t \partial F_t'}\|^2 = O_p(1)$$

$$\frac{1}{T}\sum_{t=1}^{T} \|\frac{\partial^2 h_j(y_t, x_t, F_t^*, \theta^*)}{\partial F_t \partial \theta'}\|^2 = O_p(1)$$

$$\frac{1}{T}\sum_{t=1}^{T} \|\frac{\partial^2 h_j(y_t, x_t, F_t^*, \theta^*)}{\partial \theta \partial \theta'}\|^2 = O_p(1)$$

where h_j is the jth component of h, $O_p(1)$ is uniform over F_t^* and θ^* such that $\max_{1\leq t\leq T} \|F_t^* - F_t\| \leq b_{NT}$ and $\|\theta^* - \theta\| \leq b_{NT}$ with $b_{NT} \to 0$.

A sufficient condition for M3(i) is that the z_t are independent of e_{it}. Because any function of z_t is also independent of e_{it} it follows that ξ_t will be independent of e_{it}. Conditions in M3(ii) are also easily verified for the probit model. Consider the second order derivative of h_j with respect to F_t and focus the first term on the right hand side of (5), since the second term is similar. Denote this term by h_{j1} (assume $y_t = 1$). Evaluating the derivative at F_t^* and θ^*, we have

$$\frac{\partial h_{j1}}{\partial F_t} = \Big[-\theta^{*\prime} F_t^* \frac{\phi}{\Phi} - \Big(\frac{\phi}{\Phi}\Big)^2 \Big]\theta^* F_{jt}^* - \frac{\phi}{\Phi}\iota_j$$

where ι_j is $q \times 1$ having zero elements except jth component being 1.

$$\begin{aligned}\frac{\partial^2 h_{j1}}{\partial F_t \partial F_t'} &= -\Big(\frac{\phi}{\Phi}\Big)\theta^*\theta^{*\prime} F_{jt}^* \\ &\quad -\Big((\theta^{*\prime} F_t^*)\frac{\phi}{\Phi} + \frac{\phi^2}{\Phi^2}\Big)\Big((\theta^{*\prime} F_t^*)\theta^*\theta^{*\prime} F_{jt}^* + \theta^*\iota_j' + \iota_j\theta^{*\prime} - 2\frac{\phi}{\Phi}\theta^*\theta^{*\prime} F_j^*\Big)\end{aligned}$$

From $\phi/\Phi \leq C(1 + \|\theta^*\|\|F_t^*\|)$, we have

$$\|\frac{\partial^2 h_{j1}}{\partial F_t \partial F_t'}\| \leq \sum_{j=0}^{4} C_j \|F_t^*\|^j$$

for some bounded C_j. Here we simply assume $\|\theta^*\|$ to be bounded because the parameter space is compact by assumption (in fact, this is not necessary). Now

$$\|F_t^*\|^4 \leq 4\|F_t^* - F_t\|^4 + 4\|F_t\|^4 \leq 4b_{NT}^4 + 4\|F_t\|^4.$$

Because the second order derivative of h_j involves the the fourth moment of $\|F_t\|$, the squared value of these derives involve the eighth moment of $\|F_t\|$. If follows that if $E\|F_t\|^8 \leq M$, then Assumption M3 holds.

Theorem 1 *Under assumptions A, M1-M3, and if* $T^{5/8}/N \to 0$, *then*

$$\sqrt{T}(\widetilde{\theta} - \theta_0) \xrightarrow{d} N(0, K_0^{-1}\Sigma K_0^{-1}).$$

The limiting distribution is the same as if F_t is observed. The proof is given in the appendix. The main task is to establish that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} h(\widetilde{z}_t, \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} h(z_t, \theta_0) + o_p(1).$$

5 GMM Estimation

Let W_T be a positive definite a weighting matrix that is consistent for some invertible W and $\widetilde{\theta} \xrightarrow{p} \theta_0$. Let

$$g_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} g(z_t, \theta).$$

Define the GMM estimator as

$$\widehat{\theta} = \underset{\theta}{\text{argmax}} \; -g_T(\theta)' W_T g_T(\theta)$$

Conditions for asymptotic normality of GMM estimator are given in Hansen (1982) and Newey and McFadden (1994).

Assumptions GMM1: (i) θ_0 is the interior point of Θ, where Θ is compact; (ii) $g_T(\theta)$ is bounded over Θ and $g_T(\theta)$ is continuously differential in a neighborhood $\mathcal{N}$ of θ_0; (iii) $\sqrt{T}g_T(\theta_0) \xrightarrow{d} N(0,\Omega)$; (iv) There is a $G_0(\theta)$ that is continuous differential at θ_0 and $\sup_{\theta\in\mathcal{N}} \|\nabla_\theta g_T(\theta) - G(\theta)\| \xrightarrow{p} 0$; (v) For $G_0 = G(\theta_0)$, $G_0'WG_0$ is non-singular; (vi), $W_T \xrightarrow{p} W > 0$.

Under GMM1,

$$\sqrt{T}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, (G_0'WG_0)^{-1}G_0'W\Omega WG_0(G_0'WG_0)^{-1})$$

If W_T is chosen such that $W_T \xrightarrow{p} \Omega^{-1}$, the optimal GMM is obtained and asymptotic covariance matrix reduces to $(G_0'\Omega^{-1}G_0)^{-1}$.

When z_t is not observed, we use $\widetilde{z}_t$ in place of z_t. Let

$$\widetilde{g}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(\widetilde{z}_t, \theta).$$

Define the feasible GMM by

$$\widetilde{\theta} = \underset{\theta}{\text{argmax}}\ -\widetilde{g}_T(\theta)'W_T\widetilde{g}_T(\theta)$$

We need to impose additional assumptions to establish the large sample properties of the GMM estimator.

Assumptions GMM2:

$$\sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \|\frac{\partial}{\partial F_t} g(y_t, x_t, F_t^*, \theta)\|^2 = O_p(1) \tag{6}$$

where $O_p(1)$ is uniform over F_t^* such that $\max_{1\le t\le T} \|F_t^* - F_t\| \le b_{NT}$ with $b_{NT} \to 0$.

Assumption GMM2 is similar to M2 imposed on the M-estimator. The assumptions is necessary for the sample objective function defined on $\widetilde{z}_t$, denoted by $\widetilde{Q}_T(\theta)$, to be uniformly close to $Q_T(\theta)$, the sample objective function if z_t was observed. For the Hansen-Singleton example, conditions for consistency and asymptotic normality for observable F_t are given in Newey and McFadden (1994). To verify GMM2, $\partial g/\partial F_t = \beta x_t y_t^\gamma - 1$, and it is bounded by $|\beta|\cdot|x_t|\cdot|y_t|^\gamma + 1$. Because γ is positive and Θ is compact, there exist γ_1 and γ_2 such that $\gamma \in [\gamma_1, \gamma_2]$. It follows that

$$\|\partial g/\partial F_t\| \le C|x_t|[|y_t|^{\gamma_1} + |y_t|^{\gamma_2}] + 1$$

If $E|x_t|^2$ and $E|y_t|^{2\gamma_2}$ are bounded, then Assumption GMM2 holds.

Lemma 4 *Under Assumptions A, GMM1 and GMM2,*

$$\widetilde{\theta} \xrightarrow{p} \theta_0$$

Proof: We show $\widetilde{Q}_T(\theta) = Q_T(\theta) + o_p(1)$, where $o_p(1)$ is uniform in θ. By mean-value expansion

$$\widetilde{g}_T(\theta) = g_T(\theta) + \frac{1}{T}\sum_{t=1}^{T} \frac{\partial}{\partial F_t} g(y_t, x_t, F_t^{\dagger}, \theta)(\widetilde{F}_t - F_t)$$

The second term on the right hand side is $o_p(1)$ uniformly in θ by the Cauchy-Schwarz inequality, Assumption GMM2, and Lemma 1(i). Thus $\widetilde{g}_T(\theta) = g_T(\theta) + o_p(1)$, and

$$\widetilde{Q}_T(\theta) = -[g_T(\theta) + o_p(1)]' W_T [g_T(\theta) + o_p(1)]$$

$$= Q_T(\theta) + [g_T(\theta) + o_p(1)] W_T o_p(1) + o_p(1).$$

Since $g_T(\theta)$ is bounded on Θ, $[g_T(\theta) + o_p(1)]' W_t o_p(1) = o_p(1)$. It follows that $\widetilde{Q}_T(\theta) = Q_T(\theta) + o_p(1)$. Consistency then follows from the usual argument. $\square$.

The following assumption is used for the limiting distribution of $\widetilde{\theta}$.

Assumptions GMM3:

(i) $\eta_t = (\partial/\partial F_t) g(z_t, \theta_0)$ is uncorrelated with e_{it}; $E\|\eta_t\|^2 \leq M$ for all t.

(ii) For $j = 1, 2, ..., p = dim(g_T)$,

$$\frac{1}{T}\sum_{t=1}^{T} \| \frac{\partial^2 g_j(y_t, x_t, F_t^*, \theta^*))}{\partial F_t \partial F_t'} \|^2 = O_p(1)$$

$$\frac{1}{T}\sum_{t=1}^{T} \| \frac{\partial^2 g_j(y_t, x_t, F_t^*, \theta^*))}{\partial F_t \partial \theta'} \|^2 = O_p(1)$$

where $O_p(1)$ is uniform over F_t^* and θ^* such that $\max_{1 \leq t \leq T} \|F_t^* - F_t\| \leq b_{NT}$ and $\|\theta^* - \theta\| \leq b_{NT}$ with $b_{NT} \to 0$.

These conditions are easy to verify for the Hansen-Singleton example: $g(z_t, \theta) = F_t(\beta x_t y_t^{\gamma} - 1)$. For (i), a sufficient condition is that z_t is independent of e_{it}. For (ii), since g is linear in F_t, the second order derivative is zero, so the first equation in GMM2(ii) is trivially true. For the second equation,

$$\frac{\partial^2 g_j}{\partial F_t \partial \beta} = x_t y_t^{\gamma^*}$$

and it is bounded by $|x_t|(|y_t|^{\gamma_0/2} + |y_t|^{2\gamma_0})$ because γ^* is in the neighborhood of γ^0 such that $\gamma_0/2 \leq \gamma^* \leq 2\gamma^0$. Furthermore,

$$\frac{\partial^2 g_j}{\partial F_t \partial \gamma} = \beta^* x_t \log(y_t) y_t^{\gamma^*}$$

it is bounded by $2|\beta_0| \cdot |x_t| \cdot |\log(y_t)|(|y_t|^{\gamma_0/2} + |y_t|^{2\gamma_0})$. Moreover, for $\kappa > 0$, there exists a small $\delta > 0$ and a large M such that $|\log(y_t) y_t^{\kappa}| \leq M(|y_t|^{\kappa-\delta} + |y_t|^{\kappa+\delta})$. It follows that if high enough moments of $|y_t|$ and $|x_t|$ exist, the second equation of GMM2(ii) also holds.

Theorem 2 *Under Assumptions A, GMM1-GMM3, if* $T^{5/8}/N \to 0$,

$$\sqrt{T}(\widetilde{\theta} - \theta_0) \xrightarrow{d} N(0, (G_0' W G_0)^{-1} G_0' W \Omega W G_0 (G_0' W G_0)^{-1})$$

Theorem 2 establishes that $\widetilde{\theta}$ has the same asymptotic distribution as if F_t was observed. Thus, the usual results obtained for the GMM estimator applies. In particular, if z_t are iid, then $\Omega = E[g(z_t, \theta_0) g(z_t, \theta_0)']$. Define

$$W_T = \Big[\frac{1}{T} \sum_{t=1}^{T} g(\widetilde{z}_t, \breve{\theta}) g(\widetilde{z}_t, \breve{\theta})\Big]^{-1}$$

where $\breve{\theta}$ is an initial consistent estimator of θ_0 (with identity weighting matrix, say). It is easy to show $W_T \xrightarrow{p} \Omega^{-1}$. Then an efficient estimator can be obtained, and

$$\sqrt{T}(\widetilde{\theta} - \theta_0) \xrightarrow{d} N(0, (G_0' \Omega^{-1} G_0)^{-1})$$

When $g(z_t, \theta_0)$ is serially correlated, $\Omega = \lim var(\sqrt{T} g_T(\theta_0))$. A HAC type estimator for Ω is required.

6 Concluding Remarks

Theorems 1 and 2 imply that inference with generated regressors estimated from large dimensional panel will differ from conventional estimation with generated regressors. Suppose instead of $\widetilde{F}_t$, we use $\widehat{F}_t$, the generated regressor estimated from a small set of covariates. Then

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} h(\widehat{z}_t, \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} h(z_t, \theta_0) + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla_{F_t} h(z_t, \theta_0)(\widehat{F}_t - F_t) + o_p(1)$$

Suppose $\widehat{F}_t = U_t'\widehat{\gamma}$, then $\widehat{F}_t - F_t = U_t(\widehat{\gamma} - \gamma)$. Unless U_t has zero mean and is uncorrelated with $\nabla_{F_t} h(z_t, \theta_0)$, the second term on the right hand side is equal to

$$\left(\frac{1}{T}\sum_{t=1}^{T} \nabla_{F_t} h(z_t, \theta_0) U_t\right)\sqrt{T}(\widehat{\gamma} - \gamma) = O_p(1),$$

which is non-negligible. Hence, the asymptotic variance of $T^{-1/2}\sum_{t=1}^{T} h(z_t, \theta_0)$ is not the same as the asymptotic variance of $T^{-1/2}\sum_{t=1}^{T} h(\widehat{z}_t, \theta_0)$. Thus, as in linear models and the problem studied by Pagan (1984), sampling variability of the first step estimation affects the overall variance of $\widehat{\theta}$ and must be taken into account when the first step estimator is $\sqrt{T}$ consistent. In contrast, if $\widetilde{z}_t = (y_t, x_t, \widetilde{F}_t)$, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h(\widetilde{z}_t, \theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} h(z_t, \theta_0) + o_p(1)$, and F_t can be treated as known.

It is also important to remark that there is an implicitly assumption underlying the conventional way of estimating F_t, namely, that although we do not observe F_t, we observe a finite number of variables that completely determine it. Notably, if one or more of these determinants are missing from the first step regression, the $\widehat{F}_t$ that results is not a consistent estimate of F_t, and $\widehat{F}_t - F_t = O_p(1)$. This will in turn yield inconsistent estimates of θ in the second step. It is also not uncommon for practitioners to pick a handful of variables to proxy for the latent state variables. Inconsistent second step estimates will also result when the selected predictors are imperfect indicators of the latent variables, even when the omitted predictors are orthogonal to the included ones.

In regression analysis, it is often the case that researchers have many times more predictors on hand than they actually use. Only in few exceptional cases can one be really sure that the discarded variables are completely uninformative. The fact that $\widetilde{F}_t$ can be used in linear and non-linear regressions as though they are asymptotically the same as F_t enables researchers to use information in x_{it} in a parsimonious way.

Appendix: Proofs

Proof of Lemma 2: We use the following expression, see Bai and Ng (2002)

$$\widetilde{F}_t - HF_t = \frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t) + \frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st} + \frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\eta_{st} + \frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\xi_{st}$$

$$= I + II + III + IV$$

where

$$\zeta_{st} = \frac{1}{N}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})], \quad \eta_{st} = F_s'\Lambda' e_t/N, \quad \xi_{st} = F_t'\Lambda' e_s/N$$

Consider I, which can be rewritten as $T^{-1}\sum_{s=1}^{T}(\widetilde{F}_s - F_s)\gamma_N(s,t) + T^{-1}\sum_{s=1}^{T}F_s\gamma_N(s,t)$. The first term is bounded by $T^{-1/2}(T^{-1}\sum_{s=1}^{T}\|\widetilde{F}_s - F_s\|^2)^{1/2}(\sum_{s=1}^{T}|\gamma_N(s,t))^{1/2} = T^{-1/2}O_p(C_{NT}^{-1/2})$ because of $T^{-}\sum_{s=1}^{T}\|\widetilde{F}_s - F_s\|^2 = O_p(C_{NT}^{-2})$ and $\sum_{s=1}^{T}\gamma(s,t)$ being bounded uniformly in t by assumption. The second term is bounded by $T^{-1}\max_{1\le t\le T}\|F_s\|\sum_{s=1}^{T}|\gamma_N(s,t)| = T^{-1}O_p(\alpha_T)$.

Consider II, which can be rewritten as $T^{-1}\sum_{s=1}^{T}(\widetilde{F}_s - F_s)\zeta_{st} + T^{-1}\sum_{s=1}^{T}F_s\zeta_{st}$. The first term is bounded by $(T^{-1}\sum_{s=1}^{T}\|\widetilde{F}_s - F_s\|^2)^{1/2}(T^{-1}\sum_{s=1}^{T}\zeta_{st}^2)^{1/2} = O_p(C_{NT}^{-1})(T^{-1}\sum_{s=1}^{T}\zeta_{st}^2)^{1/2}$. And

$$T^{-1}\sum_{s=1}^{T}\zeta_{st}^2 = \frac{1}{N}\Big\{\frac{1}{T}\sum_{s=1}^{T}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]\Big)^2\Big\} = \frac{1}{N}\phi_t$$

where ϕ_t is defined as the term in the braces. By assumption, $E\|\phi_t\|^2 \le M$ for all t, thus maximum of ϕ_t is bounded by $O_p(T^{1/2})$. This implies that the maximum of $T^{-1}\sum_{s=1}^{T}\zeta_{st}^2$ is bounded by $T^{1/2}N^{-1}O_p(1)$. Taking the squared root, the maximum of the first term is by $O_p(C_{NT}^{-1})T^{1/4}N^{-1/2}$. The second term can be written as $(NT)^{-1/2}\rho_t$, where $\rho_t = (NT)^{-1/2}\sum_{s=1}^{T}\sum_{i=1}^{N}F_s[e_{is}e_{it} - E(e_{is}e_{it})]$. By assumption, $E\|\rho_t\|^2 \le M$ for all t, thus the maximum of ρ_t is $O_p(T^{1/2})$. This implies the maximum of the second term is $N^{-1/2}O_p(1)$..

Consider III, which can be rewritten as $T^{-1}\sum_{s=1}^{T}(\widetilde{F}_s - F_s)\eta_{st} + T^{-1}\sum_{s=1}^{T}F_s\eta_{st}$. The first term is $\frac{1}{T}[\sum_{s=1}^{T}(\widetilde{F}_s - F_s)F_s']\Lambda' e_t/N = O_p(C_{NT}^{-2})\Lambda' e_t/N$ because $T^{-1}\sum_{s=1}^{T}(\widetilde{F}_s - F_s)F_s' = O_p(C_{NT}^{-2})$, see Lemma B2 of Bai (2003). Furthermore, the maximum of $\Lambda' e_t/N$ over t is bounded by $(T/N)^{-1/2}O_p(1)$. Thus the first term is $O_p(C_{NT}^{-2})(T/N)^{-1/2}$. The second term is $(\frac{1}{T}\sum_{s=1}^{T}F_sF_s')\frac{1}{N}\sum_{i=1}^{N}\lambda_i e_{it} = O_p(1)\frac{1}{N}\sum_{i=1}^{N}\lambda_i e_{it}$. Under the assumption that

$$E|N^{-1/2}\sum_{i=1}^{N}\|\lambda_i e_{it}\|^8 \le M$$

for all M, the maximum of $\frac{1}{N}\sum_{i=1}^{N}\lambda_i e_{it}$ is $O_p(T^{1/8})/N^{1/2}$.

Consider IV, which can be rewritten as $T^{-1}\sum_{s=1}^{T}(\widetilde{F}_s - F_s)\xi_{st} + T^{-1}\sum_{s=1}^{T}F_s\xi_{st}$. The first term is bounded by $(\frac{1}{T}\sum_{s=1}^{T}\|\widetilde{F}_s - F_s\|^2)^{1/2}(\frac{1}{T}\sum_{s=1}^{T}\xi_{st}^2)^{1/2}$. Now

$$\frac{1}{T}\sum_{s=1}^{T}\xi_{st}^2 \leq N^{-1}\frac{1}{T}\sum_{s=1}^{T}\|N^{-1/2}\sum_{i=1}^{N}\lambda_i e_{is}\|^2 \max_{1\leq t\leq T}\|F_t\|^2 = N^{-1}\alpha_T^2 O_p(1).$$

Thus the first term is equal to $O_p(C_{NT}^{-1})N^{-1/2}\alpha_T$. The second term is equal to

$$(NT)^{-1/2}\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{i=1}^{N}\sum F_s\lambda_i' e_{is}F_t,$$

which is bounded by $(NT)^{-1/2}O_p(1)\max_{1\leq t\leq T}\|F_t\| = (NT)^{-1/2}\alpha_T O_p(1)$. Thus the second term is dominated by the first.

Under the assumption that $T/N^2 \to 0$ and $\alpha_T \leq T^{1/2}$, with the exception of $\frac{1}{N}\sum_{i=1}^{N}\lambda_i e_{it}$ appearing in III, all terms are dominated by $O_p(\alpha_T/T) + O_p(N^{-1/2})$. This proves the lemma.

Proof of Theorem 1: Without loss of generality, assume $h(z_t,\theta)$ is a scalar; otherwise, consider each component of h. By Taylor expansion

$$\begin{aligned}0 = \frac{1}{T}\sum_{t=1}^{T}h(\widetilde{z}_t,\widetilde{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}h(z_t,\theta_0) + \frac{1}{T}\sum_{t=1}^{T}\xi_t'(\widetilde{F}_t - F_t) + \frac{1}{T}\sum_{t=1}^{T}H(z_t,\theta_0)(\widetilde{\theta}-\theta_0)\\ &+\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix}\widetilde{F}_t - F_t\\ \widetilde{\theta}-\theta_0\end{bmatrix}'\begin{bmatrix}\frac{\partial^2 h(z_t^\dagger,\theta^\dagger)}{\partial F_t\partial F_t'} & \frac{\partial^2 h(z_t^\dagger,\theta^\dagger)}{\partial F_t\partial\theta'}\\ \frac{\partial^2 h(z_t^\dagger,\theta^\dagger)}{\partial\theta\partial F_t'} & \frac{\partial^2 h(z_t^\dagger,\theta^\dagger)}{\partial\theta\partial\theta'}\end{bmatrix}\begin{bmatrix}\widetilde{F}_t - F_t\\ \widetilde{\theta}-\theta_0\end{bmatrix}\end{aligned} \tag{7}$$

where $F_t^\dagger$ is in between F_t and $\widetilde{F}_t$, and $\theta^\dagger$ is in between θ and $\widetilde{\theta}$, and $\xi_t = \frac{\partial h(z_t,\theta_0)}{\partial F_t}$. By assumption, ξ_t is uncorrelated with e_{it} for all i and t, by Lemma 1(ii)

$$\frac{1}{T}\sum_{t=1}^{T}\xi_t'(\widetilde{F}_t - F_t) = O_p(C_{NT}^{-2})$$

Consider the expression involving the matrix. The first block is

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}(\widetilde{F}_t - F_t)'\frac{\partial^2 h(z_t^\dagger,\theta^\dagger)}{\partial F_t\partial F_t'}(\widetilde{F}_t - F_t)\\ &\leq \max_{1\leq t\leq T}(\|\widetilde{F}_t - F_t\|)\frac{1}{T}\sum_{t=1}^{T}\|(\widetilde{F}_t - F_t)\|\|\frac{\partial^2 h(z_t^\dagger,\theta^\dagger)}{\partial F_t\partial F_t'})\|\\ &\leq \max_{1\leq t\leq T}(\|\widetilde{F}_t - F_t\|)\Big(\frac{1}{T}\sum_{t=1}^{T}\|\widetilde{F}_t - F_t\|^2\Big)^{1/2}\Big(\frac{1}{T}\sum_{t=1}^{T}\|\frac{\partial^2 h(z_t^\dagger,\theta^\dagger)}{\partial F_t\partial F_t'})\|\Big)^{1/2}\\ &= O_p(T^{1/8})N^{-1/2}O_p(C_{NT}^{-1})\end{aligned}$$

The last equality is due to (2). The above is $o_p(T^{-1/2})$ if $T^{5/8}/N \to 0$.

The last block and cross product terms are each $o_p(1)(\widetilde{\theta} - \theta_0)$. Thus we can rewrite equation (7) as

$$0 = \frac{1}{T}\sum_{t=1}^{T} h(z_t, \theta_0) + \left[\overline{K}_T(\theta_0) + o_p(1)\right](\widetilde{\theta} - \theta_0) + O_p(C_{NT}^{-2}) + o_p(T^{-1/2})$$

where $\overline{K}_T(\theta_0) = \frac{1}{T}\sum_{t=1}^{T} K(z_t, \theta_0) \xrightarrow{p} K_0 = EK(z_t, \theta_0)$. Multiplying $\sqrt{T}$, we obtain

$$\sqrt{T}(\widetilde{\theta} - \theta_0) = [\overline{K}_T(\theta_0) + o_p(1)]^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} h(z_t, \theta_0) + o_p(1)$$

where $\sqrt{T}O_p(C_{NT}^{-2}) = o_p(1)$ under $T^{5/8}/N \to 0$. It follows that

$$\sqrt{T}(\widetilde{\theta} - \theta_0) \xrightarrow{d} N(0, K_0^{-1}\Sigma K_0^{-1}).$$

Remark: For a linear model it is sufficient to have $\sqrt{T}/N \to 0$ instead of $T^{5/8}/N \to 0$. To see this, consider $m(z_t, \theta) = -(y_t - F_t'\theta)^2$ and $h(z_t, \theta) = 2(y_t - F_t'\theta)F_t$. Thus $h_j(z_t, \theta) = 2(y_t - F_t'\theta)F_{jt}$, and

$$\frac{\partial^2 h_j(z_t, \theta)}{\partial F_t \partial F_t'} = -2\theta\iota_j' - 2\iota_j\theta'$$

which is a constant matrix (does not depend on t). It follows that

$$\|\frac{1}{T}\sum_{t=1}^{T}(\widetilde{F}_t - F_t)' \frac{\partial^2 h_j(z_t^\dagger, \theta^\dagger)}{\partial F_t \partial F_t'}(\widetilde{F}_t - F_t)\| \leq M\frac{1}{T}\sum_{t=1}^{T}\|\widetilde{F}_t - F_t\|^2 = O_p(C_{NT}^{-2})$$

But $\sqrt{T}O_p(C_{NT}^{-2}) \xrightarrow{p} 0$ provided that $\sqrt{T}/N \to 0$.

Proof of Theorem 2. Under Assumption GMM1, the estimator $\widetilde{\theta}$ solves for the first order condition

$$\widetilde{G}_T(\widetilde{\theta})' W_T \widetilde{g}_T(\widetilde{\theta}) = 0$$

where $\widetilde{G}_T(\theta) = \nabla_\theta \widetilde{g}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \nabla_\theta g(\widetilde{z}_t, \theta)$. Expanding $\widetilde{g}_T(\widetilde{\theta})$ at θ_0

$$\widetilde{g}_T(\widetilde{\theta}) = \widetilde{g}_T(\theta_0) + \widetilde{G}_T(\bar{\theta})(\widetilde{\theta} - \theta_0).$$

where $\bar{\theta}$ is between θ_0 and $\widetilde{\theta}$. Solving for $\widetilde{\theta} - \theta_0$ gives

$$\sqrt{T}(\widetilde{\theta} - \theta_0) = [\widetilde{G}_T(\widetilde{\theta})' W_T \widetilde{G}_T(\bar{\theta})]^{-1} \widetilde{G}_T(\widetilde{\theta})' W_T \sqrt{T}\widetilde{g}_T(\theta_0)$$

It is sufficient to show

$$\widetilde{G}_T(\widetilde{\theta}) = G_T(\widetilde{\theta}) + o_p(1) \tag{8}$$

$$\widetilde{G}_T(\bar{\theta}) = G_T(\bar{\theta}) + o_p(1)$$

and

$$\sqrt{T}\widetilde{g}_T(\theta_0) = \sqrt{T}g_T(\theta_0) + o_p(1) \tag{9}$$

Result (8) implies $\widetilde{G}_T(\widetilde{\theta}) \xrightarrow{p} G(\theta_0) = G$ because $G_T(\widetilde{\theta}) \xrightarrow{p} G(\theta_0)$ as already argued in Newey and McFadden. Similarly, $\widetilde{G}_T(\bar{\theta}) \xrightarrow{p} G$. Result (9) implies asymptotic normality for $\sqrt{T}\widetilde{g}_T(\theta_0)$ since $\sqrt{T}g_T(\theta_0)$ is asymptotic normal by assumption. To prove (8), consider the jth column of $\widetilde{G}_T$, denoted by $\widetilde{G}_{jT}$

$$\widetilde{G}_{jT}(\widetilde{\theta}) = G_{jT}(\widetilde{\theta}) + \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 g_j(y_t, x_t, F_t^{\dagger}, \widetilde{\theta})}{\partial\theta\partial F_t'}(\widetilde{F}_t - F_t)$$

where $F_t^{\dagger}$ is in between F_t and $\widetilde{F}_t$. By the Cauchy-Schwarz inequality, Assumption GMM2, and $\frac{1}{T}\sum_{t=1}^{T} \|\widetilde{F}_t - F_t\|^2 = O_p(C_{NT}^{-2})$, the second term above is $O_p(C_{NT}^{-1}) = o_p(1)$. This proves (8). The equation below (8) can be proved in the same way.

Next consider (9). Its jth component, upon expansion, is equal to

$$\widetilde{g}_{jT}(\theta_0) = g_{jT}(\theta_0) + \frac{1}{T}\sum_{t=1}^{T} \eta_{jt}'(\widetilde{F}_t - F_t) + \frac{1}{T}\sum_{t=1}^{T} (\widetilde{F}_t - F_t)' \frac{\partial^2 g_j(y_t, x_t, F_t^{*}, \theta_0)}{\partial F_t \partial F_t'}(\widetilde{F}_t - F_t)$$

where $\eta_{jt} = (\partial/\partial F_t)g_j(z_t, \theta_0)$. The rest argument is the same as in the proof of Theorem 1. This implies that $\widetilde{g}_{jT}(\theta_0) = g_{jT}(\theta_0) + o_p(T^{-1/2})$ if $T^{5/8}/N \to 0$. This is equivalent to (9).

References

Amemiya, T. 1984, *Advanced Econometrics*, Harvard University Press, Cambridge, MA.

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

Bai, J. and Ng, S. 2006b, Evaluating Latent and Observed Factors in Macroeconomics and Finance, *Journal of Econometrics* **113:1-2**, 507–537.

Bai, J. and Ng, S. 2006c, Instrumental Variables in a Data Rich Environment, mimeo, University of Michigan.

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

Chamberlain, G. and Rothschild, M. 1983, Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets, *Econometrica* **51**, 1281–2304.

Hamilton, J. D. 1989, A New Approach to the Economic Analysis of Non-Stationary Time Series and the Business Cycle, *Econometrica* **57**, 357–384.

Hansen, L. and Singleton, K. 1982, Generalized Instrumental Variable Estimation of Non-linear Rational Expectations Models, *Econometrica* **50**, 1269–1286.

Newey, W. 1983, A Method of Momentse Interpretation of Sequential Estimators, *Economics Letters* **14**, 201–206.

Newey, W. and McFadden, D. 1994, Large Sample Estimators and Hypothesis Testing, *Handbook of Econometrics*, Vol. 4,Chapter 36, North Holland.

Pagan, A. 1984, Econometric Issues in the Analysis of Regressions with Generated Regressors, *International Economic Review* **25**, 221–247.

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