Estimation of a Factor-Augmented Panel Data Model Through Instrumental Variables

Matthew Harding*, Carlos Lamarche† and Chris Muris‡

May 15, 2015

Abstract

This paper proposes an estimator for the latent factors in a factor-augmented panel model. The procedure relies on internally generated instruments. The simulation study shows that the proposed approach improves the performance of existing methods in cases with weak factors.

JEL: C23, C26.
Keywords: Factor Model; Panel Data; Instrumental variables

1. Introduction

This paper introduces a new instrumental variables (IV) based approach to the estimation of the latent structure in factor-augmented panel data models. The estimation of these models has recently received substantial attention due their popularity in a number of areas from macro-finance to labor economics (Bernanke, et. al. 2005; Kim and Opka, 2014). One popular interpretation treats the latent factors as a generalization of the traditional individual fixed effects model consisting of interactive fixed effects (Bai, 2009). The interest in this modeling approach is evidenced by the large number of papers on this topic ranging from new specification tests (Su, Jin, Zhang, 2014) to extensions to quantile regression (Harding and Lamarche, 2014).[1]

While the consistent estimation of the parameters of the observed variables in the factor-augmented model is made possible by a number of approaches, most notably those of Bai (2009) and Pesaran

*Sanford School of Public Policy, Duke University, 140 Science Drive, Durham, NC 27708; Phone: (919) 613-1306; Fax: (919) 613-0539; Email: matthew.harding@duke.edu
†Department of Economics, University of Kentucky, 335A Gatton College of Business and Economics, Lexington, KY 40506-0034; Phone: (859) 257 3371; Email: clamarche@uky.edu
‡Department of Economics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6; Phone: (778) 782-9395; Email: cmuris@sfu.ca
[1]See also Ahn, Lee, and Schmidt (2013), Robertson and Sarafidis (2015), and Harding and Lamarche (2011) for generalized method of moment and IV estimation of a related class of panel data models with endogeneity.

(2006), the estimation of the latent factor structure has received less attention. Note in particular that the approach of Pesaran (2006) constructs a proxy for the latent factors but does not explicitly estimate the factors. But in many applications (e.g. forecasting) it is also necessary to estimate the latent factors (Bai and Ng, 2006) .

This paper introduces a simple approach to the estimation of latent factors in these panel data models using internally constructed instruments. This approach can be traced back to the work of Madansky (1964). The connection to factor-augmented panel data models was first pointed out by Harding (2007). Below we introduce the method and provide Monte-Carlo evidence on its finite sample performance. It is important to note that our proposed approach can build on the approach of Pesaran (2006) to estimate the latent factors, but also improves the mean squared error of the estimator of the factors proposed by Bai (2009).

2. Model and Method

This paper considers the following factor-augmented panel data model for $i = 1, \ldots, N$ cross-sectional units and $t = 1, \ldots, T$ time periods:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{it}, \tag{1}$$

where y_{it} is the response variable for subject i at time t, $\boldsymbol{x}_{it}$ is a p-dimensional vector of independent variables, $\boldsymbol{\lambda}_i$ is a vector of loadings, $\boldsymbol{f}_t$ is a vector of factors, and u_{it} is the error term. In this paper, we are interested in the estimation of an $r \times 1$ vector of factor loadings, $\boldsymbol{\lambda}_i$, and an $r \times 1$ vector of factors, $\boldsymbol{f}_t$. We assume that the number of factors is known and equal to r and we focus on the estimation of the f_{jt}'s. Below, we first construct an estimator for $\boldsymbol{f}_t$. Once this estimator is available, it is straightforward to construct an estimator for $\boldsymbol{\lambda}_i$.

Consistent estimators for $\boldsymbol{\beta}$ were developed by Pesaran (2006) and Bai (2009), among others. It is therefore convenient to assume temporarily that $\boldsymbol{\beta}$ is known and define

$$R_{it} = y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\beta} = \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{it}. \tag{2}$$

Thus, we have a standard factor model where the variable R_{it} is observed and the other variables $\boldsymbol{\lambda}_i$, $\boldsymbol{f}_t$, and u_{it} are latent. Let $\boldsymbol{R}_i = (R_{i1}, R_{i2}, \ldots, R_{iT})'$ be a $T \times 1$ vector of observations for subject i. Moreover, let $\boldsymbol{R}_{i,(1)}$ be a partition of $\boldsymbol{R}_i$ of dimension r and $\boldsymbol{R}_{i,(2)}$ the remaining partition of dimension $T - r$. It follows that we have the following representation of the error term of the model

(1):

$$\boldsymbol{R}_{i,(1)} = \boldsymbol{f}'_{(1)}\boldsymbol{\lambda}_i + \boldsymbol{u}_{i,(1)} \tag{3}$$

$$\boldsymbol{R}_{i,(2)} = \boldsymbol{f}'_{(2)}\boldsymbol{\lambda}_i + \boldsymbol{u}_{i,(2)} \tag{4}$$

where $\boldsymbol{f}_{(1)}$ is an $r \times r$ matrix, $\boldsymbol{f}_{(2)}$ is a $(T-r) \times r$ matrix, $\boldsymbol{u}_{i,(1)}$ is a r dimensional vector, and $\boldsymbol{u}_{i,(2)}$ is a $T-r$ dimensional vector. We assume that the rank of $\boldsymbol{f}$ is r implying that the $r \times r$ matrix $\boldsymbol{f}_{(1)}$ is invertible. Solving for the nuisance parameter $\boldsymbol{\lambda}_i$ in equation (3) and replacing it in equation (4), we have that:

$$\boldsymbol{R}_{i,(2)} = \boldsymbol{\Pi}\boldsymbol{R}_{i,(1)} + \boldsymbol{V}_i, \tag{5}$$

where $\boldsymbol{\Pi} = \boldsymbol{f}'_{(2)}\boldsymbol{f}_{(1)}^{-1}$, $\boldsymbol{f}_{(1)}^{-1}$ denotes the inverse of $\boldsymbol{f}'_{(1)}$ and $\boldsymbol{V}_i = \boldsymbol{u}_{i,(2)} - \boldsymbol{f}'_{(2)}\boldsymbol{f}_{(1)}^{-1}\boldsymbol{u}_{i,(1)}$. The parameter $\boldsymbol{\Pi}$ can be estimated by linear regression methods on a sample of N observations but the results would be biased since $\boldsymbol{V}_i$ is correlated with $\boldsymbol{R}_{i,(1)}$ through $\boldsymbol{u}_{i,(1)}$. Consider one row of equation (5),

$$R_{it} = \boldsymbol{\Pi}_t\boldsymbol{R}_{i,(1)} + V_{it}, \tag{6}$$

where t is taken from the second partition, and denote by $\boldsymbol{Z}_{it}$ a vector of available instruments. Then we can estimate the r-dimensional row vector $\boldsymbol{\Pi}_t$ using two stage least squares with dependent variable R_{it}, endogenous regressors $\boldsymbol{R}_{i,(1)}$ and instrumental variables $\boldsymbol{Z}_{it}$.

As in Madansky (1964) and Harding (2007), it is possible to develop an estimation approach based on an instrumental variable strategy with instruments defined internally within the model. Assuming that the number of time periods T is fixed, let $\mathcal{T}_{(1)} = \{t : R_{it} \text{ is an element of } \boldsymbol{R}_{i,(1)}\}$ and $\mathcal{T}_{(3)} = \{s : s \notin \mathcal{T}_{(1)} \text{ and } s \neq t\}$. We observe a sample of $\{R_{it} : i = 1, ..., N;\ t = 1, ..., T\}$ where R_{it} is defined as in equation (2). We assume that $\text{Cov}(u_{is}, u_{jt}) = 0$ for all (s,t) such that $s \neq t$ or $i \neq j$. Moreover, we assume that u_{it} is independent of $\boldsymbol{\lambda}_i$ for each t and $\text{Cov}(\lambda_{ir}, \lambda_{jk}) = 0$ for all (i,j,r,k) with $i \neq j$. Under these conditions, the j-th element of the vector of instruments $\boldsymbol{Z}_i$ is $Z_{ij} = R_{ij}$ if $j \in \mathcal{T}_{(3)}$. We call this procedure factor instrumental variable estimator (FIV).

Example: Consider the simpler case where $T = 3$ and $r = 1$ and suppose we are interested in the scalar, normalized factor $\Pi_3 = f_3/f_1$. In this case, we can set $R_{i,t} = R_{i,3}$ and we let $R_{i,(1)}$ be equal to $R_{i,1}$. The instrument Z_{it} is then equal to $R_{i,2}$. Under these simplifications, we have a straightforward expression for the FIV estimator,

$$\hat{\Pi}_3 = \frac{\frac{1}{n}\sum_{i=1}^{n} R_{i,2}R_{i,3}}{\frac{1}{n}\sum_{i=1}^{n} R_{i,2}R_{i,1}}. \tag{7}$$

Using this representation, it is easy to show that this procedure is consistent. First, note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} R_{i2}R_{i1} &= \frac{1}{n}\sum_{i=1}^{n}\left(f_2\lambda_i + u_{i2}\right)\left(f_1\lambda_i + u_{i1}\right) \\ &= f_2 f_1 \frac{1}{n}\sum_{i=1}^{n}\lambda_i^2 + f_2\frac{1}{n}\sum_{i=1}^{n}\lambda_i u_{i1} + f_1\frac{1}{n}\sum_{i=1}^{n}\lambda_i u_{i2} + \frac{1}{n}\sum_{i=1}^{n} u_{i1}u_{i2} \\ &\xrightarrow{p} f_2 f_1 E\left(\lambda_i^2\right)\end{aligned}$$

where the second, third and fourth terms are zero under the no serial correlation assumption and from the independence of the factor loadings from the error term. A similar derivation can be employed for the other terms in equation (7). As a result,

$$\hat{\Pi}_3 \xrightarrow{p} \frac{f_2 f_3 E\left(\lambda_i^2\right)}{f_2 f_1 E\left(\lambda_i^2\right)} = \frac{f_3}{f_1} \equiv \Pi_3. \tag{8}$$

Note that the FIV estimator depends on an arbitrary normalization associated with the choice of $\boldsymbol{R}_{i,(1)}$. Hence, there are $Q = \binom{T}{r}$ ways of choosing such a normalization. We now propose a consistent estimator whose performance is independent of the normalization, meanwhile improving efficiency by using information from different normalizations. This is done by averaging the results across all Q possible normalizations. We refer to this estimator as the average factor instrumental variable estimator (MFIV).

First, let q index a normalization, and define an estimator $\hat{\boldsymbol{\Pi}}_t(q)$. We collect the results across t in the $T \times r$ matrix $\hat{\boldsymbol{\Pi}}(q)$. Note that, by definition, $\hat{\boldsymbol{\Pi}}_{(1)} = \boldsymbol{I}_r$. Second, rotate the estimator $\hat{\boldsymbol{\Pi}}(q)$ for $\boldsymbol{\Pi}(q)$ into an estimator $\hat{\boldsymbol{f}}(q)$ for $\boldsymbol{f}$ using a normalization that is independent of q. As an example, we could use the normalization $\boldsymbol{f}\boldsymbol{f}' = \boldsymbol{I}_T$. Finally, define the MFIV as

$$\hat{\boldsymbol{f}} = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\boldsymbol{f}}(q). \tag{9}$$

3. Monte Carlo

We generate the dependent variable based on the following model:

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \lambda_{1,i} f_{1,t} + u_{it}, \\ x_{j,it} &= a_j \lambda_{1,i} + b_j f_{1,t} + c_j \lambda_{1,i} f_{1,t} + v_{j,it}, \\ f_{1,t} &= \rho f_{1,t-1} + \eta_t, \end{aligned}$$

for $j = \{1, 2\}$, $i = 1, \ldots, N$ and $t = -49, \ldots 0, \ldots T$ in the last equation. Following Pesaran (2006), we let $t_{-50} = 0$. The error terms are $(u_{it}, \boldsymbol{v}'_{it}, \eta_{it})' \sim \mathcal{N}(0, \boldsymbol{I})$ and the factor loadings $\lambda_{i,1}$ are generated using different distributions in order to investigate the impact of weak factors on the performance of the proposed estimator. We consider the following variations:

Design 1: The loadings $\lambda_{i,1}$ are i.i.d. random variables distributed as Uniform i.e. $\mathcal{U}[0, 1]$.
Design 2: The loadings $\lambda_{i,1}$'s are i.i.d. random variables distributed as Gaussian random variables i.e. $\mathcal{N}(0, 1)$.
Design 3: We generate the loadings $\lambda_{1,i} \sim \mathcal{N}(0, 1)$, for $i = 1, ..., m$ and, following Chudik, Pesaran, and Tosetti (2011),

$$\lambda_{1,i} = \frac{\theta_i}{2\sum_i \theta_i}, \quad \text{for } i = m + 1, ..., N,$$

where $\theta_i \sim \mathcal{U}[0, 1]$ and $m = 0.9N$.

Finally, the parameters are assumed to be: $\beta_1 = \beta_2 = a_1 = 1$, $b_1 = 2$, $c_1 = 0.5$, $\beta_0 = a_2 = b_2 = c_2 = 0$, and $\rho_f = 0.90$. Our design has one endogenous variable, x_1, and one exogenous variable, x_2.

While the focus of our approach is on the estimation of the latent factor structure in the model, it relies on the availability of a consistent estimator for the observed part of the model in equation 1 and the availability of consistent estimates of the residuals $R_{i,t}$ in equation 2. Thus, we first compare the performance of three estimators: the standard ordinary least square estimator (OLS), the estimator for an interactive effects model (BAI) proposed by Bai (2009), and the mean group estimator (MGE) developed by Pesaran (2006).

Table 1 presents the bias and root mean square error (RMSE) for the intercept, β_0, and slope parameters, β_1 and β_2, and provides evidence of the biases present in the application of existing methods for the coefficients associated with the endogenous variable and exogenous variable. As expected, OLS is significantly biased and while BAI and MGE perform well. In small samples, the relative performance of the estimators depends on how the factors, f_t are generated. As N and T grows however, these differences disappear and, as expected, these approaches offer similar RMSE values.

We now evaluate the performance of our proposed approach against a standard approach involving Principal Component Analysis (PCA). As previously discussed, in the context of a factor-augmented panel data model, we can conceive of the feasible estimation of the factor structure in two steps. First, a consistent estimate of the coefficients on the observed variables is necessary to generate

	N	T	Estimators: Intercept β_0			Slope β_1			Slope β_2		
			OLS	BAI	MGE	OLS	BAI	MGE	OLS	BAI	MGE
Design 1: Loadings distributed as Uniform											
Bias	50	20	-0.109	-0.001	0.002	0.162	0.037	0.004	0.001	0.001	0.000
RMSE	50	20	0.164	0.425	0.027	0.167	0.075	0.039	0.033	0.033	0.039
Bias	50	50	-0.102	0.000	0.000	0.176	0.006	0.000	0.001	0.001	0.001
RMSE	50	50	0.124	0.057	0.008	0.178	0.025	0.022	0.021	0.020	0.022
Bias	100	20	-0.109	0.002	0.000	0.162	0.008	-0.001	0.000	0.000	0.001
RMSE	100	20	0.161	0.111	0.021	0.167	0.032	0.029	0.024	0.024	0.029
Bias	100	50	-0.100	0.000	0.000	0.177	0.003	0.000	0.000	0.000	0.000
RMSE	100	50	0.121	0.041	0.007	0.178	0.015	0.015	0.015	0.014	0.016
Design 2: Loadings distributed as Gaussian											
Bias	50	20	-0.207	-0.003	0.000	0.080	0.003	0.004	-0.001	-0.001	-0.001
RMSE	50	20	0.305	0.060	0.045	0.188	0.035	0.041	0.045	0.033	0.039
Bias	50	50	-0.104	-0.001	0.000	0.078	0.000	0.002	0.000	0.001	0.000
RMSE	50	50	0.177	0.026	0.022	0.140	0.020	0.021	0.027	0.019	0.021
Bias	100	20	-0.212	-0.003	0.000	0.087	0.000	0.000	0.000	0.000	0.000
RMSE	100	20	0.313	0.039	0.033	0.192	0.024	0.028	0.030	0.024	0.029
Bias	100	50	-0.102	0.000	0.001	0.086	0.000	0.001	-0.001	-0.001	-0.001
RMSE	100	50	0.182	0.017	0.016	0.138	0.014	0.015	0.019	0.014	0.015
Design 3: Loadings distributed as Uniform and Gaussian											
Bias	50	20	-0.202	-0.003	0.001	0.085	0.001	0.002	-0.001	-0.003	-0.003
RMSE	50	20	0.296	0.055	0.043	0.184	0.036	0.041	0.042	0.033	0.040
Bias	50	50	-0.099	-0.001	0.001	0.081	0.001	0.002	0.001	0.001	0.000
RMSE	50	50	0.170	0.025	0.021	0.134	0.021	0.022	0.026	0.020	0.021
Bias	100	20	-0.200	-0.002	-0.002	0.084	0.000	0.002	-0.001	0.000	0.000
RMSE	100	20	0.297	0.038	0.031	0.182	0.023	0.027	0.031	0.023	0.028
Bias	100	50	-0.109	0.001	0.000	0.083	-0.001	0.000	0.000	0.000	0.000
RMSE	100	50	0.184	0.018	0.015	0.138	0.015	0.015	0.019	0.014	0.016

TABLE 1. *Small sample performance of panel data estimators. Results are based on 1000 replications.*

$R_{i,t}$ in equation 2. Second, we apply either PCA or the MFIV estimator proposed in this paper to estimate the latent factor structure. In Table 2, we present the RMSE from estimating the vector $(f_1, ..., f_T)'$. The root mean square error is computed as, $T^{-1}\sum_{t=1}^{T}(f_t - \hat{f}_t)^2$, where $\hat{f}_t$ denotes an estimator of the factor f_t. The table compares the performance of the two approaches PCA and MFIV.

		Design 1				Design 2				Design 3			
N	T	PCA		MFIV		PCA		MFIV		PCA		MFIV	
		BAI	MGE	BAI	MGE	BAI	MGE	BAI	MGE	BAI	MGE	BAI	MGE
50	20	1.92	1.99	1.09	1.08	1.90	1.93	1.33	1.33	1.99	1.95	1.29	1.29
50	50	1.99	1.99	1.05	1.05	1.91	1.93	1.19	1.20	2.15	2.16	1.16	1.16
100	20	1.90	1.88	1.11	1.11	2.05	2.06	1.37	1.37	2.01	1.98	1.36	1.36
100	50	1.97	1.96	1.07	1.07	1.98	1.98	1.18	1.18	1.92	1.92	1.20	1.20

TABLE 2. *Small sample performance of feasible MFIV estimator compared to PCA approach. Both estimators can utilize either the BAI or the MGE estimator as a first step to consistently estimate residuals.*

For each case we compare the RMSE of the estimator for the latent factors using either the BAI or the MGE estimators in the first step. Both first step approaches lead to a feasible strategy for the estimation of the factors. Note that while the BAI approach can also produce an estimate of the factors, the MGE estimator does not. Thus, we can think of our approach as useful for both refining the BAI estimator and also as an approach to estimating the latent structure in the MGE framework. Note, that here we omit to report the small sample performance of FIV. Although it exhibited good finite sample properties, overall it showed higher RMSE than the MFIV estimator in the simulations considered in this study.

Table 2 shows that the performance of the IV approach proposed in this paper leads to significant improvements in terms of MSE relative to the existing PCA approach. This is important and shows that not only can our approach be used to estimate the latent factors following the implementation of Pesaran's (2006) approach, but it can lead to significant improvements to Bai's (2009) estimator for the factors f_t.

4. **Conclusions**

This paper introduces a simple IV approach to the estimation of the latent structure in factor-augmented panel data models. While the identification relies on correctly specifying the dependence between the latent factors and the error term, it nevertheless leads to a simple approach to estimating the latent factors in the model considered by Pesaran (2006), in addition to providing a refinement of the approach in Bai (2009). We explore the finite sample performance of the feasible estimator in case of weak factor designs and conclude that the estimator performs very well. It is noteworthy that our approach does not require additional data and the instruments are constructed

internally from existing data. Further research may involve relaxing the identification assumptions to more general cases.

References

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013): "Panel data models with multiple time-varying individual effects," *Journal of Econometrics*, 174(1), 1–14.

Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77(4), 1229–1279.

Bai, J., and S. Ng (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74(4), 1133–1150.

Bernanke, B. S., J. Boivin, T. Doan, and P. S. Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics*, (120), 387–422.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011): "Weak and strong cross-section dependence and estimation of large panels," *The Econometrics Journal*, 14(1), C45–C90.

Harding, M., and C. Lamarche (2011): "Least squares estimation of a panel data model with multifactor error structure and endogenous covariates," *Economics Letters*, 111(3), 197 – 199.

——— (2014): "Estimating and testing a quantile regression model with interactive effects," *Journal of Econometrics*, 178, 101–113.

Harding, M. C. (2007): "Essays in econometrics and random matrix theory," Ph.D. thesis, Massachusetts Institute of Technology.

Kim, D., and T. Oka (2014): "Divorce Law Reforms And Divorce Rates In The Usa: An Interactive Fixed-Effects Approach," *Journal of Applied Econometrics*, 29(2), 231–245.

Madansky, A. (1964): "Instrumental variables in factor analysis," *Psychometrika*, 29(2), 105–113.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74(4), 967–1012.

Robertson, D., and V. Sarafidis (2015): "IV estimation of panels with factor residuals," *Journal of Econometrics*, 185(2), 526 – 541.

Su, L., S. Jin, and Y. Zhang (2014): "Specification test for panel data models with interactive fixed effects," *Journal of Econometrics*.