Estimation of a Factor-Augmented Linear Panel Data Model with Applications Using Student Achievement Data[*]

Matthew Harding[†], Carlos Lamarche[‡] and Chris Muris[§]

June 9, 2019

Abstract

This paper investigates the estimation of latent factors in a factor-augmented panel model when the cross-section and time series dimensions of the panel are large. The paper proposes a new class of estimators that are linear combinations of instrumental variable estimators obtained from a large number of internally generated instruments. We demonstrate that recent theoretical and methodological developments at the intersection of machine learning and econometrics can be adapted to address the poor finite sample performance of existing methods for the factors when a large number of instruments are employed. We establish consistency of the estimator and we carry out simulation studies to investigate the small sample behavior of the new method. The simulation study shows that the proposed approach improves the performance of existing methods on the estimation of unknown factors, and it can be used to complement several consistent estimators of slope parameters in a factor-augmented panel data model. Lastly, we consider two empirical applications using administrative data of college and high school students.

Keywords: Factor Model; Panel Data; Instrumental Variables; Administrative data

[*] The authors would like to thank Atsushi Inoue and Ana Herrera for comments and useful conversations. We thank Chris Heaton for sharing his programs for estimation of the approximate factor model.

[†] Department of Economics and Department of Statistics, University of California - Irvine, 3207 Social Science Plaza B, Irvine, CA 92697; Phone (949) 824 1511; Email: harding1@uci.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, University of Bristol, The Priory Road Complex, 12 Priory Rd, Bristol BS8 1TU, United Kingdom. Email: chris.muris@bristol.ac.uk.

1. Introduction

In recent years, there has been a renewed interest in factor models. The new literature includes extensions of the classical factor model of Anderson and Rubin (1956), estimation and inference of large factor models for high-dimensional data sets (Bai and Wang 2016), and novel applications in economics, finance, and psychology. In economics, the identification and estimation of new factor models has received substantial attention in a number of areas from macro-finance to labor economics (Bernanke et al. 2005, Harding 2007, Kim and Opka 2014). Recent important work has studied the role of cognition, personality traits, and academic motivation on child development (Cunha and Heckman 2007, Cunha and Heckman 2008, Borghans et al. 2008, Heckman, Pinto and Savelyev 2013). Factor-augmented regressions as in Stock and Watson (1999, 2002) are known to improve forecast of macroeconomic time series such as inflation and industrial production, and in panel data econometrics, one popular interpretation treats the latent factors as a generalization of the traditional individual fixed effects model (Pesaran 2006, Bai 2009, Su, Jin, and Zhang 2014, and Harding and Lamarche 2014, among others).

While the consistent estimation of the parameters of the observed variables in the factor-augmented model is made possible by a number of approaches, the estimation of the latent factor structure has received less attention. Bai and Ng (2010) estimate the parameter vector of the observed variables using factors as instrumental variables, and these factors are obtained using principal component analysis. Papers by Hägglund (1982) and Heaton and Solo (2012), among others, proposed to estimate latent factors in panel data models using internally constructed instruments, an idea that can be traced back to the work of Madansky (1964). However, there are two known issues with the approach. In applications to high-dimensional data, including panel data with large cross-sectional dimension (N) and time-series dimension (T), the large number of available instrumental variables (IVs) creates finite sample bias similar to the one originated by using too many instruments (see Hahn and Hausman 2003, Hansen, Hausman and Newey 2008, Bekker 1994). A second issue arises from identification of the factor model and the selection of the "first" measure or partition when the observed measurements per subject do not have a predetermined or natural order. Consequently, practitioners often view the application of instrumental variables to the estimation of factor models as inferior to alternative methods.

This paper introduces a class of estimators that are linear combinations of consistent estimators obtained with a large number of internally generated instruments. The method accommodates recent econometric developments for high-dimensional models to address the poor finite sample performance of existing instrumental variable approaches. We propose two-step estimators in which the first step estimates a high-dimensional model using regularization techniques (e.g., Tibshirani 1996, Belloni, Chen, Chernozhukov and Hansen 2012) or a weighted average of consistent estimators (Chen, Jacho-Chavez and Linton 2016). In the second step, we obtain a weighted average of estimators that are indexed by the number of possible normalizations or partitions of the factor model. We show that the two-step estimator is consistent under similar assumptions to the ones used in the literature. We also investigate the small sample behavior of the new method. We show that the proposed approach improves the performance of existing methods on the estimation of unknown factors, and it can be used to complement several consistent estimators of slope parameters. The estimation of the parameters of the observed variables using instrumental variables is investigated in Bai and Ng (2010), Harding and Lamarche (2011), Ahn, Lee, and Schmidt (2013), and Robertson and Sarafidis (2015), among others.

We consider two empirical applications of the method to the estimation of models of educational attainment using administrative data of college and high school students. Although factor models for test scores are not new (e.g., Borghans et al. 2008), we are not aware of studies that use administrative data on high school and college students. In our first illustration of the approach, we estimate a factor model using administrative data from a higher education institution in Europe (De Giorgi, Pellizzari and Woolston 2012). The findings reveal that the average grade distribution of students in economics and finance is less dispersed in later years relative to the first year in the program. The evidence suggests that latent abilities of students who remained in the program improved over time. The second application employs data from Angrist et al. (2002) to evaluate the impact of an educational voucher program implemented in Latin America. We find that educational vouchers for private schooling improved the average student performance in several subjects, although it created a wider gap between weak and strong students.

This paper is organized as follows. The next section introduces the approach and the proposed estimator. The section also establishes the consistency of the estimator. Section 3 provides Monte Carlo experiments to investigate the small sample performance of the proposed estimator. Section 4 demonstrates how the estimator can be used in practice by exploring applications using administrative data. Section 5 concludes.

2. Model and Method

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

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{it}, \tag{2.1}$$

where $y_{it} \in \mathbb{R}$ is the response variable for subject i at time t, $\boldsymbol{x}_{it} \in \mathbb{R}^p$ is a vector of independent variables, $\boldsymbol{\beta} \in \mathcal{B} \subseteq R^p$ is an unknown parameter vector, $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{ir})' \in \mathbb{R}^r$ is a vector of loadings, and $\boldsymbol{f}_t = (f_{t1}, f_{t2}, \dots, f_{tr})' \in \mathbb{R}^r$ is a vector of factors. The number of factors is known and equal to r. The variable u_{it} is the error term for $1 \leq i \leq N$ and $1 \leq t \leq T$, satisfying $E(u_{it}) = 0$ for all t and $\text{Cov}(u_{is}, u_{jt}) = 0$ for all (s,t) such that $s \neq t$ or $i \neq j$.

In this paper, we are interested in the estimation of the vector of factor loadings, $\boldsymbol{\lambda}_i$, and the vector of factors, $\boldsymbol{f}_t$. Pesaran (2006) and Bai (2009), among others, provide consistent estimators for $\boldsymbol{\beta}$. We consider first the estimation of $\boldsymbol{f}_t$, because the proposed procedure uses a large number of cross-sectional units N which is common for applied researchers interested in using micro-economic data. Once the estimator of $\boldsymbol{f}_t$ is available, it is straightforward to construct an estimator for $\boldsymbol{\lambda}_i$. In Section 4 as an illustration of the approach, we first estimate college course-specific factor variables in a model of academic achievement, and then we estimate subject-specific factors in a model of test scores using a sample of N students.

2.1. Background

It is convenient to assume temporarily that $\boldsymbol{\beta}$ is known and define

$$R_{it} := y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\beta} = \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{it}. \tag{2.2}$$

In the empirical applications considered in Section 4, we consistently estimate $\boldsymbol{\beta}$ in a first stage, and then concentrate on estimation of the factor $\boldsymbol{f}_t$ for $1 \leq t \leq T$ and loading $\boldsymbol{\lambda}_i$ for $1 \leq i \leq N$.

Equation (2.2) is 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 $\boldsymbol{R}_i = (R_{i1}, R_{i2}, \dots, R_{iT})'$ be a $T \times 1$ vector of observations for subject $1 \leq i \leq N$. 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 $T - r$. We then have the following representation of equation (2.2):

$$\boldsymbol{R}_{i,(1)} = \boldsymbol{F}_{(1)}\boldsymbol{\lambda}_i + \boldsymbol{u}_{i,(1)} \tag{2.3}$$

$$\boldsymbol{R}_{i,(2)} = \boldsymbol{F}_{(2)}\boldsymbol{\lambda}_i + \boldsymbol{u}_{i,(2)} \tag{2.4}$$

where $\boldsymbol{F}_{(1)}$ is an invertible $r \times r$ matrix, $\boldsymbol{F}_{(2)}$ is a $(T-r) \times r$ matrix, $\boldsymbol{u}_{i,(1)}$ is an r dimensional vector, and $\boldsymbol{u}_{i,(2)}$ is a $T-r$ dimensional vector. Solving for $\boldsymbol{\lambda}_i$ in equation (2.3) gives,

$$\boldsymbol{\lambda}_i = \boldsymbol{F}_{(1)}^{-1} \left(\boldsymbol{R}_{i,(1)} - \boldsymbol{u}_{i,(1)} \right),$$

and replacing it in equation (2.4), we have that:

$$\boldsymbol{R}_{i,(2)} = \boldsymbol{\theta}_{(2)} \boldsymbol{R}_{i,(1)} + \boldsymbol{V}_i, \tag{2.5}$$

where $\boldsymbol{\theta}_{(2)} = \boldsymbol{F}_{(2)} \boldsymbol{F}_{(1)}^{-1}$ is a $(T-r) \times r$ matrix of reduced form coefficients, and $\boldsymbol{V}_i = \boldsymbol{u}_{i,(2)} - \boldsymbol{F}_{(2)} \boldsymbol{F}_{(1)}^{-1} \boldsymbol{u}_{i,(1)}$ is the reduced form error term.

In model (2.2), $\boldsymbol{\lambda}_i$ and $\boldsymbol{f}_t$ are identified up to a non-singular transformation. In the literature of instrumental variables, it is common to normalize the upper $r \times r$ block of a matrix of loadings (Bai and Wang 2016), which in our case, it is equivalent to consider the following normalization $\boldsymbol{F}_{(1)} = \boldsymbol{I}_r$. Several other normalizations are also possible, including 'fixing the rotation' imposed by principal component analysis (PCA). In our model (2.5), we define a normalization for the first elements of $\boldsymbol{R}_{i,(1)}$. This means that without loss of generality, we can choose r measurements of R_{it} to be proxies of the factor loadings measured with error.

After adopting *a* normalization, the parameter $\boldsymbol{\theta}_{(2)}$ in equation (2.5) can be estimated by linear regression methods using a sample of N cross-sectional observations. It is known that these results would be biased since the components of $\boldsymbol{V}_i$ are correlated with $\boldsymbol{R}_{i,(1)}$ through $\boldsymbol{u}_{i,(1)}$. On the other hand, a number of instrumental variables approaches have been proposed to consistently estimate $\boldsymbol{\theta}_{(2)}$ (Madansky 1964, Hägglund 1982, Heaton and Solo 2012). Following closely Heaton and Solo (2012), consider one row of equation (2.5) with t taken from the second partition (i.e., $t > r$):

$$R_{it} = \boldsymbol{R}'_{i,(1)} \boldsymbol{\theta}_{(2t)} + V_{it}, \tag{2.6}$$

where $\boldsymbol{\theta}_{(2t)} = (\theta_{(2t)1}, \theta_{(2t)2}, \ldots, \theta_{(2t)r})' \in \mathbb{R}^r$, and denote by $\boldsymbol{Z}_{it} \in \mathbb{R}^r$ a vector of available instruments for $\boldsymbol{R}_{i,(1)}$. Then, after multiplying (2.6) by $\boldsymbol{Z}_{it}$, we obtain,

$$E(\boldsymbol{Z}_{it} R_{it} - \boldsymbol{Z}_{it} \boldsymbol{R}'_{i,(1)} \boldsymbol{\theta}_{(2t)}) = E(\boldsymbol{Z}_{it} V_{it}) = 0, \tag{2.7}$$

and the parameter $\boldsymbol{\theta}_{(2t)}$ is identified under Assumptions (A.1)-(A.3) below. These conditions are standard assumptions for identification when the error covariance matrix is diagonal.

(A.1) The error term u_{it} is cross-sectionally and serially independent, i.e. $E\left[u_{it}\right] = 0$ for all (i,t), and $E\left[u_{it} u_{js}\right] = 0$ for all $(i,t) \neq (j,s)$. The variance of the error term is bounded, that is then there exists constants $\kappa > 0$ and $K > 0$ such that $\kappa < \sigma_u^2 \leq K < \infty$. We denote $\mathrm{Var}(\boldsymbol{u}) = \sigma_u^2 \boldsymbol{I} = \boldsymbol{\Sigma}$, where $\boldsymbol{u}$ is a NT dimensional vector.

(A.2) The vector of factor loadings $\boldsymbol{\lambda}_i$ is uncorrelated with the error term u_{it} and $\boldsymbol{f}_t$, in the sense that $E\left[\lambda_{ij}u_{it}\right] = 0$ and $E\left[\lambda_{ij}f_{tj}\right] = 0$ for all t and for all $j = 1, \cdots, r$. Also, the $r \times r$ matrix $\boldsymbol{\Sigma}_\Lambda \equiv \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$ exists and it is bounded.

(A.3) The factors $\boldsymbol{f}_t$ and the error term u_{it} are independent for $1 \leq t \leq T$. Also, let $\boldsymbol{f} = (\boldsymbol{f}_1, \cdots, \boldsymbol{f}_T)'$ be the $T \times r$ matrix that collects all the factors. There exists an $r \times r$ submatrix of $\boldsymbol{f}$ that has full rank.

(A.4) Sufficient number of time periods are available: $T \geq 2r + 1$.

Under these conditions, the vector of instruments can be generated internally. Let $\mathcal{T}_{(1)} = \{t : R_{it}$ is an element of $\boldsymbol{R}_{i,(1)}\}$ and $\mathcal{T}_{(3t)} = \{j : j \notin \mathcal{T}_{(1)}$ and $j \neq t\}$. Note that the set $\mathcal{T}_{(3t)}$ defines $(T - r - 1)$ possible instruments and the j-th variable R_{ij} is a valid instrument if and only if $j \in \mathcal{T}_{(3t)}$.

For instance, consider that $t > r$ and the model in equation (2.7) is exactly identified by $\boldsymbol{Z}_{it} = (R_{it+1}, \ldots, R_{it+r})'$. It follows that, if we consider one instrument $R_{it+1,k}$ for $1 \leq k \leq r$,

$$E\left(R_{it+k}V_{it}\right) = E\left(\left(\lambda_{ik}f_{it+k} + u_{it+k}\right)\left(u_{it} - \sum_{j=1}^{r} f_{tj}u_{ij}\right)\right) = 0,$$

under Assumptions (A.1)-(A.3) and factors that are serially independent as in the models estimated in Section 4. The latter condition can be relaxed by assuming exponentially decaying coefficients or by considering an approximate factor model. Note too that an instrumental variable procedure will be consistent if the loadings are estimated using instrumental variables under the assumption that the loadings are independent for all i and $1 \leq j \leq N$. We call the instrumental variable estimator that uses the instruments in $\mathcal{T}_{(3t)}$ the grouped instrumental variable estimator for a factor model (GIVE).

The application of instrumental variables approaches to the estimation of factor models is generally regarded as inferior to alternative methods such as PCA. There are two well known issues. First, the cardinality of $\mathcal{T}_{(3t)}$ increases with $T \to \infty$. In large T panels, when many instruments are available, the procedure suffers from a finite sample bias problem similar to the one investigated in Hansen, Hausman and Newey (2008). Second, there is a large number of possible partitions of $\boldsymbol{R}_i$ in equations (2.3)-(2.4), and the number of these partitions (i.e., normalizations) tends to infinity as $T \to \infty$. In some panel data applications, the time dimension imposes a natural order that minimizes the issue of how to partition $\boldsymbol{R}_i$. In other cases, as in the applications with student achievement data presented below in Example 1 and Section 4, it is less clear. We illustrate the issue with the following example.

Example 1. Suppose y_{it} in equation (2.1) is the grade of student i in course t. As in Section 4.2, consider that the total number of courses is equal to $T = 3$, with data available for mathematics (t_1), reading (t_2), and writing (t_3) courses. We are interested in estimating how instructors' quality in these courses affects academic performance. Assume that $r = 1$, and suppose we are interested in the scalar, normalized factor $\theta_3 = f_3/f_1$. In this case, we can set $R_{it} = R_{i3}$ and we let $R_{i,(1)}$ be equal to R_{i1}. The instrument is then equal to R_{i2}. Under these simplifications, we have a straightforward expression for the FIV estimator,

$$\hat{\theta}_3 = \frac{\frac{1}{N}\sum_{i=1}^{N} R_{i2}R_{i3}}{\frac{1}{N}\sum_{i=1}^{N} R_{i2}R_{i1}}. \tag{2.8}$$

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

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

under Assumptions (A.1)-(A.4). A similar derivation can be employed for the numerator in equation (2.8). As a result,

$$\hat{\theta}_3 \xrightarrow{p} \frac{f_2 f_3 E\left(\lambda_i^2\right)}{f_2 f_1 E\left(\lambda_i^2\right)} = \frac{f_3}{f_1} =: \theta_3. \tag{2.9}$$

It is clear that $\hat{\theta}_3$ converges to f_3 if $f_1 = 1$, which corresponds to a normalization of the variables. The parameter θ_3 can also be estimated by normalizing f_2 and using R_{i1} as an instrument for R_{i2}.

In this example, it is not clear a priori whether to normalize based on mathematics, reading, or writing. The lack of a natural order over time leads to practical questions on how to select the normalization and partition. In practice, there are $\binom{T-1}{r}$ such ways of choosing such a normalization, and $Q = \binom{T}{r}$ possible partitions.

2.2. Estimation in large panels

As discussed before, there are two reasons for which an IV approach for the estimation of the latent factor structure of model (2.1) is complicated in panels with large N and T. First, there are $Q = \binom{T}{r}$ ways of implementing the estimator, with $Q \to \infty$ as $T \to \infty$. Second, the asymptotic theory is complicated by the fact that the number of instruments diverges for each one of the Q

estimators as $T \to \infty$. In this subsection, we outline the main contribution of this paper, which is to propose a method that simultaneously addresses these challenges.

Recall the process of partitioning $\boldsymbol{R}_i = (R_{i1}, R_{i2}, \ldots, R_{iT})' \in \mathbb{R}^T$ for $1 \leq i \leq N$ described in the previous subsection. Consider one such partition $q \in \{1, \ldots, Q\}$, and denote the resulting subvectors by $\left(\boldsymbol{R}_{i,(1q)}, \boldsymbol{R}_{i,(2q)}\right)$. Analogously, define the partitioned factors by $\left(\boldsymbol{F}_{(1q)}, \boldsymbol{F}_{(2q)}\right)$. For a given q, pick a t such that R_{it} is an element of $\boldsymbol{R}_{i,(2q)}$, and write

$$\boldsymbol{R}_t = \boldsymbol{R}_{(1q)}\boldsymbol{\theta}_{tq} + \boldsymbol{V}_t, \tag{2.10}$$

where $\boldsymbol{R}_t = (R_{1t}, R_{2t}, \ldots, R_{Nt})' \in \mathbb{R}^N$, $\boldsymbol{R}_{(1q)}$ is an $N \times r$ matrix, and $\boldsymbol{V}_t \in \mathbb{R}^N$. Denote by $\boldsymbol{R}_{(3tq)}$ an $N \times (T - r - 1)$ matrix of instruments for $\boldsymbol{R}_{(1q)}$ in the regression equation for time period t using partition q, and a projection matrix as:

$$\boldsymbol{P}_{(3tq)} = \boldsymbol{R}_{(3tq)}(\boldsymbol{R}'_{(3tq)}\boldsymbol{R}_{(3tq)})^{-1}\boldsymbol{R}'_{(3tq)}. \tag{2.11}$$

Also, define,

$$G_{Nq}(\boldsymbol{\theta}_{tq}) := \frac{1}{N}\boldsymbol{R}'_{(1q)}\boldsymbol{P}_{(3tq)}(\boldsymbol{R}_t - \boldsymbol{R}_{(1q)}\boldsymbol{\theta}_{tq}) \tag{2.12}$$

with $G_{tq}(\boldsymbol{\theta}_t) := E(G_{Nq}(\boldsymbol{\theta}_{tq}))$. The estimator $\hat{\boldsymbol{\theta}}_{tq}$ naturally satisfies:

$$G_{Nq}(\hat{\boldsymbol{\theta}}_{tq}) = o_p(N^{-1/2}). \tag{2.13}$$

We propose to estimate $\boldsymbol{\theta}_{tq}$ in two steps. In the first step, we obtain $\hat{\boldsymbol{\theta}}_{tq}$ for $q = 1, 2, \ldots, Q$. In the second step, we compute a consistent estimator of $\boldsymbol{\theta}_t$ using a linear combination of a possibly infinite dimensional set of consistent estimators obtained in the first step.

We begin by describing the second step:

Mean grouped variable estimator [*Step 2*]: Suppose $\hat{\boldsymbol{\theta}}_{tq}$ is available for $q = 1, 2, \ldots, Q$ and each $\hat{\boldsymbol{\theta}}_{tq}$ satisfies (2.13). Then, for a given t, we obtain:

$$\hat{\boldsymbol{\theta}}_t = \sum_{q=1}^{Q} W_{Tq}\hat{\boldsymbol{\theta}}_{tq}, \tag{2.14}$$

where W_{Tq} is the weight given to the q-th estimator. We call the estimator defined in equation (2.14) the mean grouped variable estimator (MGVE). The selection of W_{Tq} is discussed in Section 2.3.

Grouped variable estimator [*Step 1*]: In this step, we estimate $\boldsymbol{\theta}_{tq}$. In the case of small T, we estimate $\boldsymbol{\theta}_{tq}$ by the estimator $\hat{\boldsymbol{\theta}}_{tq}$ defined in (2.13) using N cross-sectional units. In large T

panels, we propose a penalized estimator but we provide other alternatives that practitioners can use depending on the specific application.

Method 1: Let $\hat{\boldsymbol{R}}_{l,(1q)} := \boldsymbol{R}_{(3tq)}\hat{\boldsymbol{\pi}}_{l,tq}$ for $l = 1, 2, \ldots, r$, where $\hat{\boldsymbol{\pi}}_{l,tq}$ is a Lasso estimator defined as,

$$\hat{\boldsymbol{\pi}}_{l,tq} = \arg\min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \sum_{i=1}^{N} (R_{i,l,(1q)} - \boldsymbol{R}'_{i,(3tq)}\boldsymbol{\pi}_{l,tq})^2 + \frac{\lambda_l}{N}\|\Upsilon_l \boldsymbol{\pi}_{l,tq}\|_1, \tag{2.15}$$

where the parameter set $\boldsymbol{\Pi} \subseteq \mathbb{R}^{(T-r-1)}$ and $\|\boldsymbol{a}\|_1$ is the standard ℓ_1-norm defined as $\|\boldsymbol{a}\|_1 = \sum_i |a_i|$ for a generic constant a_i. The penalty loadings Υ_l and λ_l can be selected as in Belloni, Chen, Chernozhukov and Hansen (2012).

If $T - r - 1 >> N$, for each (t, q), we replace $\boldsymbol{R}'_{(1q)}\boldsymbol{P}_{(3tq)}$ in equation (2.12) by the $N \times r$ matrix $\hat{\boldsymbol{R}}_{(1q)}$ formed by columns $\hat{\boldsymbol{R}}_{l,(1q)}$ for $l = 1, 2, \ldots, r$. The solution gives,

$$\hat{\boldsymbol{\theta}}_{tq} = (\hat{\boldsymbol{R}}'_{(1q)}\boldsymbol{R}_{(1q)})^{-1}\hat{\boldsymbol{R}}'_{(1q)}\boldsymbol{R}_t. \tag{2.16}$$

Under regularity conditions, $\hat{\boldsymbol{\theta}}_{tq}$ is $\sqrt{N}$-consistent and asymptotically normal as demonstrated in Belloni, Chen, Chernozhukov and Hansen (2012).

Method 2: Let $j \in \mathcal{J}$ denote the number of exact identified models (2.7) and $\mathcal{J}$ is a set of possibly infinite cardinality as $T \to \infty$. We denote the cardinality of the set $\mathcal{J}_T$ by $C(T, r) = \binom{T-1}{r}$, which represents the total number of combinations of exact identified models. Let $C^*(T, r) < C(T, r)$ for all T, r as $T \to \infty$. One can interpret $C^*(T, r)$ as the upper bound of a set that divides strong and weak instruments, after ordering the subsets of instrumental variable estimators. Moreover, let $\boldsymbol{R}_{j,(3tq)}$ be a $N \times r$ matrix of instruments where $j = 1, 2, \ldots, C(T, r)$.

Define

$$\hat{\boldsymbol{\theta}}^{j}_{tq} = (\boldsymbol{R}'_{j,(3tq)}\boldsymbol{R}_{(1q)})^{-1}\boldsymbol{R}'_{j,(3tq)}\boldsymbol{R}_t, \tag{2.17}$$

and obtain,

$$\hat{\boldsymbol{\theta}}_{tq} = \sum_{j=1}^{C(T,r)} w_{Tj}\hat{\boldsymbol{\theta}}^{j}_{tq}, \tag{2.18}$$

where w_{Tj} is a weight matrix, possibly stochastic, given to the j-th estimator. To combine the estimators across all the $C(T, r)$ available estimators, including determining the weighting matrix in applications, we draw on the recent work by Chen, Jacho-Chavez and Linton (2016).

The estimators defined in (2.16) and (2.18) are labeled grouped variable estimators (GVE).

Procedure 1

Depending on the dimensions of the panel, MGVE estimates factors or loadings using two different first-stage methods:

1. Group variables in $\mathcal{T}_{(1)}$, $\mathcal{T}_{(2)}$, and $\mathcal{T}_{(3)}$ according to (2.3) and (2.4), and obtain residuals R_{it} based on a consistent estimator of $\boldsymbol{\beta}$ in equation (2.1).
2. 2.a If $T < N$, regress $R_{it} \in \mathcal{T}_{(2)}$ on $R_{it} \in \mathcal{T}_{(1)}$ using $R_{it} \in \mathcal{T}_{(3)}$ as IVs.
   2.b If $T > N$, regress $R_{it} \in \mathcal{T}_{(2)}$ on $R_{it} \in \mathcal{T}_{(1)}$ using $R_{it} \in \tilde{\mathcal{T}}_{(3)}$ which includes grouped instrumental variables selected by Lasso.
3. Repeat 1. and 2. for all possible Q groups
4. Obtain the MGVE estimator as in (2.14) using the estimates in 3.

Procedure 2

If $T > N$, replace 2.b by:

2.b.i Regress $R_{it} \in \mathcal{T}_{(2)}$ on $R_{it} \in \mathcal{T}_{(1)}$ using $R_{it} \in \check{\mathcal{T}}_{(3)}$ which includes strong instrumental variables.

2.b.ii Obtain the GVE estimator as a weighted average using the estimates in 2.b.i.

The difference between the estimator defined in (2.16) and the estimator defined in (2.18) is that the latter estimator does not impose sparsity or approximate sparsity as in Assumption (A.5) below. That is, $\boldsymbol{R}_{l,(1q)} = \boldsymbol{R}_{(3tq)}\boldsymbol{\pi}_{l,tq} + \boldsymbol{\epsilon}_{l,tq}$, allowing all elements in the vector $\boldsymbol{\pi}_{l,tq}$ to be non-zero for $1 \leq l \leq r$. On the other hand, the estimator (2.18) is computed by truncating the sum of estimators to length $C^*(T, r) < C(T, r)$ when $T \to \infty$ and considering each subset of IVs corresponding to exactly identified models.

The next section discusses the specific conditions needed for consistency, including assumptions on the truncation parameter.

2.3. **Large Sample Properties**

This section investigates the large sample properties of the proposed approaches. We establish consistency of the GVE estimators with an increasing number of instruments first, and then we show consistency of the MGVE that uses a linear combination of consistent GVEs.

In addition to regularity conditions (A.1)-(A.4), we consider the following conditions:

(A.5) The conditional expectations $E(R_{l,(1q)}|\boldsymbol{R}_{(3tq)})$ for $l = 1, 2, \dots r$ and $q = 1, 2, \dots, Q$ are well approximated by a linear model and they are function of s_{tq} instruments such that

$$\max_{1\leq l\leq r} \|\boldsymbol{\pi}_{l,tq}\|_0 = \max_{1\leq l\leq r} \sum_j 1\{\pi_{lj,tq} \neq 0\} \leq s_{tq} << T$$

(A.6) The sequence of deterministic weights $\{w_{Tj}\}$ satisfies $\sum_{j\in\mathcal{J}_T} w_{Tj} = \boldsymbol{I}_r$.

Assumption (A.5) is similar to Condition AS in Belloni et al. (2012). This is a key condition for the performance of the Lasso estimator, as it requires a sparse model in terms of the instruments. The small set of IVs guarantees that the estimation of $\boldsymbol{\theta}_t$ is not affected by the first stage. This assumption can be relaxed by considering an approximate sparse model as in Belloni et al. (2012). Assumption (A.6) is a stronger condition to the one employed in Chen, Jacho-Chavez and Linton (2016). It is trivially satisfied in our case since we employ equal weighting. Chen, Jacho-Chavez and Linton (2016) require that with probability one, $\sup_T \sum_{j\in\mathcal{J}_T} \|w_{Tj}\| < \infty$ but this is not needed in our application to the estimation of factor models. The assumptions on the weights are typically satisfied in applications, including when the sequence of weights are non-random.

Theorem 1. *Under Assumptions (A.1)-(A.5), the estimator $\hat{\boldsymbol{\theta}}_{tj}$ defined in equation* (2.16) *is consistent for $\boldsymbol{\theta}_t$. Moreover, under Assumptions (A.1)-(A.4) and (A.6), the estimator $\hat{\boldsymbol{\theta}}_{tj}$ defined in equation* (2.18) *is consistent for $\boldsymbol{\theta}_t$.*

The result in Theorem 1 establishes that the variants of the GVE estimator are consistent, and therefore, we can concentrate later on showing the consistency of the linear combination of consistent estimators that diverges to infinity. The results of Theorem 1 follow directly from Chen, Jacho-Chavez and Linton (2016) and Belloni, Chen, Chernozhukov and Hansen (2012). It is important to emphasize that some of the conditions for the application of the regularization methods can be relaxed, as described in Belloni, Chen, Chernozhukov and Hansen (2012).

Let $q \in \mathcal{Q}^*_T = \{1, 2, \dots, Q^*(T, r)\}$, where $Q^*(T, r) < Q(T, r)$ is a truncation parameter for the number of possible normalizations. There are no explicit conditions on $Q^*(T, r)$ other than condition (B.4) below, but it is allowed to grow slowly as N and T go to jointly to infinity. To make explicit the dependence of the maximum number of grouped estimators and T, consider a variation of the estimator introduced in (2.14) for $\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t$:

$$\hat{\boldsymbol{\theta}}_t = \sum_{q\in\mathcal{Q}^*_T} W_{Tq}\hat{\boldsymbol{\theta}}_{tq}, \tag{2.19}$$

We assume the following conditions:

(B.1) There exists a $\epsilon_N(\delta) > 0$ for $\delta > 0$ with $\epsilon_N(\delta) \to 0$ when $N \to \infty$ such that

$$\min_{q \in \mathcal{Q}_T^*} \inf_{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| > \delta} \|G_j(\boldsymbol{\theta}_t)\| \geq \epsilon_N(\delta) > 0.$$

(B.2) For $\epsilon_N(\delta)$, $Q^*(T,r)$ and $Q(T,r)$, with $Q^*(T,r) \to \infty$, $Q(T,r) \to \infty$ and $Q^*(T,r) < Q(T,r)$ as $T \to \infty$, there is a positive sequence $\alpha_{1N} = o(1)$ with $\sup_N(\alpha_{1N}/\epsilon_N(\delta)) < \infty$ such that

$$\max_{q \in \mathcal{Q}_T^*} \left(\|G_{Nq}(\hat{\boldsymbol{\theta}}_{tq})\| - \inf_{\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t} \|G_{Nj}(\boldsymbol{\theta}_t)\| \right) = o_p(\alpha_{1N}).$$

(B.3) For $\epsilon_N(\delta)$, $Q^*(T,r)$ and $Q(T,r)$, with $Q^*(T,r) \to \infty$, $Q(T,r) \to \infty$ and $Q^*(T,r) < Q(T,r)$ as $T \to \infty$, there is a positive sequence $\alpha_{2N} = o(1)$ with $\sup_N(\alpha_{2N}/\epsilon_N(\delta)) < \infty$ such that

$$\max_{q \in \mathcal{Q}_T^*} \sup_{\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t} \|G_{Nq}(\boldsymbol{\theta}_t) - G_q(\boldsymbol{\theta}_t)\| = o_p(\alpha_{2N}).$$

(B.4) The sequence of weights $\{W_{tq}\}_{q \in \mathcal{Q}}$ for $t = 1, 2, \ldots, T$ satisfy

$$\sum_{q \in \mathcal{Q}_T} W_{tq} = \boldsymbol{I}_r, \quad \sup_{T \leq 1} \sum_{q \in \mathcal{Q}_T} \|W_{tq}\| < \infty,$$

with probability 1. Also, as $Q(T,r) \to \infty$, $Q^*(T,r) \to \infty$ with $Q^*(T,r) < Q(T,r)$ and,

$$\sup_{T \geq 1} \sum_{q = Q^*(T,r)+1}^{Q(T,r)} \|W_{tq}\| \to 0.$$

Assumption (B.1) is a modified version of a standard condition for uniform convergence of estimators that minimize a criterion function (Theorem 5.9, van der Vaart 1998). The difference is that the condition is imposed on every normalization. Assumptions (B.2)-(B.3) are similar to conditions (A.3)-(A.4) in Chen, Jacho-Chavez and Linton (2016). These assumptions impose uniformity over $Q^*(T,r)$ normalizations, for a rate $\epsilon_N(\delta)$ that is determined by the sequence $Q^*(T,r) \to \infty$. Under conditions (A.1)-(A.4), $\epsilon_N(\delta)$ must be of larger order than $N^{-1/2}$, which is satisfied if $Q^*(T,r)$ grows slowly (i.e., $\log(T)$). See Lemma 1 in Chen, Jacho-Chavez and Linton (2016).

We now establish the consistency of the estimator.

Theorem 2. *Under Assumptions A and B, the estimator $\hat{\boldsymbol{\theta}}_t$ defined in* (2.14) *is consistent for $\boldsymbol{\theta}_t$.*

The use of weights for the combination of estimators in a linear fashion is naturally not new, although it is worth discussing two aspects of the method in this setting. First, one can naively bound $Q(T,r) \leq 2^T$ but the truncation parameter $Q(T,\tau)$ does not grow slowly to satisfy the conditions of Theorem 2 and Lemma 1 in Chen, Jacho-Chavez and Linton (2016). When r is fixed, as in our case, one can approximate $Q(T,r) \sim T^r/r!$ but it is reasonable to expect r to grow as

$T \to \infty$, which is out of the scope of this paper. Therefore, $Q^*(T,\tau)$ is required to grow slower than $\log(Q(T,r))$. Second, the computational cost of the procedure might not be small as it increases exponentially with the sample size. The truncation parameter $Q^*(T,\tau)$ should be determined to minimize computational time as well as to maximize efficiency gains. There are several options for practitioners. One can potentially investigate the marginal impact of an additional normalization in terms of the standard error of the estimator. Also, one can retain the estimators with the largest t-statistic or, the estimators whose t-statistics exceeds some predetermined level.

3. Monte Carlo simulation

In this section, we conduct an investigation of the small sample performance of the proposed approach in comparison with existing methods considering a series of simulation experiments. We use standard designs from the literature (e.g., Pesaran 2006). 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 with an initial condition $f_{1,-50} = 0$. 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. The parameters are assumed to be: $\beta_1 = \beta_2 = a_1 = 1$, $b_1 = 2$, $c_1 = 0.5$, and $\beta_0 = a_2 = b_2 = c_2 = 0$. Thus, our designs have one endogenous variable, x_1, and one exogenous variable, x_2.

We consider the following variations assuming $\rho = 0$ (Design A) and $\rho = 0.9$ (Design B):

Design 1: The loading $\lambda_{i,1}$ is an i.i.d. random variables distributed as Uniform i.e. $\mathcal{U}[0,1]$.
Design 2: The loading $\lambda_{i,1}$ is an i.i.d. random variables distributed as Gaussian 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), $\lambda_{1,i} = 0.5 \times \theta_i / \sum_i \theta_i$ for $i = m+1,...,N$, where $\theta_i \sim \mathcal{U}[0,1]$ and $m = 0.9N$.

While the focus of this investigation is on the estimation of the latent factor structure in the model, our approach relies on the availability of a consistent estimator for the observed part of the model in equation (2.1), and the ability to consistently obtain R_{it} in equation (2.2). We employ

		Estimators								
N	T	PCA			GIVE			MGVE		
		OLS	BAI	MGE	OLS	BAI	MGE	OLS	BAI	MGE
Design A.1: Loadings distributed as Uniform										
50	5	2.099	2.129	2.104	2.072	2.037	2.025	1.088	1.240	1.330
100	5	2.148	2.132	2.176	1.978	1.959	1.934	1.072	1.266	1.292
50	10	1.843	2.028	2.002	2.084	2.015	1.977	1.032	1.177	1.172
100	10	1.886	2.000	2.056	2.097	2.004	1.988	1.064	1.154	1.135
Design A.2: Loadings distributed as Gaussian										
50	5	1.965	2.063	2.199	2.139	2.087	1.774	1.224	1.217	1.365
100	5	1.787	1.940	1.992	2.180	2.112	1.727	1.296	1.273	1.410
50	10	1.958	2.257	2.198	2.117	1.986	2.011	1.192	1.074	1.092
100	10	1.818	1.941	2.080	2.203	2.116	2.100	1.255	1.169	1.168
Design A.3: Loadings distributed as Uniform and Gaussian										
50	5	2.084	2.234	1.980	2.176	2.012	1.726	1.165	1.162	1.223
100	5	2.011	2.120	2.103	2.247	2.175	1.894	1.339	1.250	1.435
50	10	1.723	1.960	1.999	1.948	1.801	1.775	1.206	1.068	1.091
100	10	1.782	1.821	1.881	2.306	2.205	2.219	1.222	1.122	1.096

Table 3.1. *Small sample performance of feasible MGVE estimator compared to PCA approach in the case that $\rho = 0$.*

three estimators to obtain the residuals: 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). For completeness, we compare the performance of these estimators for $(\beta_0, \beta_1, \beta_2)$ in Table B.1 (Appendix B).

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 $R_{i,t}$ in equation (2.2). Second, we apply either PCA, GIVE, or the MGVE estimator proposed in this paper to estimate the latent factor structure. In Tables 3.1 and 3.2, 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.

					Estimators					
N	T		PCA			GIVE			MGVE	
		OLS	BAI	MGE	OLS	BAI	MGE	OLS	BAI	MGE
Design B.1: Loadings distributed as Uniform										
20	20	1.988	1.918	2.154	2.065	2.034	1.965	1.044	1.180	1.224
50	20	1.788	1.918	2.038	1.943	2.045	2.035	1.030	1.234	1.332
20	50	1.882	2.085	2.112	2.043	1.992	2.002	1.055	1.140	1.140
50	50	1.992	2.036	2.057	2.075	2.091	2.074	1.026	1.194	1.200
Design B.2: Loadings distributed as Gaussian										
20	20	1.833	2.020	1.943	2.099	1.929	1.888	1.217	1.219	1.211
50	20	1.796	2.079	2.080	2.093	1.958	1.985	1.356	1.363	1.370
20	50	1.896	2.059	2.059	2.209	2.085	2.056	1.214	1.155	1.153
50	50	1.869	2.000	2.060	2.093	1.874	1.888	1.251	1.182	1.181
Design B.3: Loadings distributed as Uniform and Gaussian										
20	20	1.711	1.928	1.923	2.110	1.984	1.996	1.237	1.246	1.232
50	20	1.791	1.881	1.803	2.170	2.100	2.085	1.374	1.386	1.386
20	50	1.877	1.981	1.924	2.113	2.031	2.027	1.185	1.121	1.127
50	50	1.920	2.079	2.079	2.222	2.039	2.037	1.246	1.152	1.148

Table 3.2. *Small sample performance of feasible MGVE estimator compared to PCA approach in the case that $\rho = 0.9$.*

Tables 3.1 and 3.2 show that the performance of the approach proposed in this paper leads to significant improvements in terms of MSE relative to the existing PCA and instrumental variable approaches. The results for the case of no serial correlation with small T, presented in Table 3.1, and the results for positive serial correlation with large T, presented in Table 3.2, offer similar conclusions.

4. Applications using administrative student data

This section presents two illustrative examples of the application of the method to the estimation of models of educational attainment. We apply the methodology to estimate factors that measure course/teacher quality, in order to understand the distribution of latent abilities of students. In the first section, we specifically investigate how the distribution of latent abilities changes over subsequent years of college education, while in the second section, we investigate the change of the distribution of student ability after the implementation of a voucher program designed to improve educational outcomes. The next sections adopt a one-factor model to satisfy condition (A.4).

4.1. **Application to college data**

Consider the following factor model of the effect of class size and socioeconomic class composition on educational attainment:

$$y_{ict} = \boldsymbol{d}_{ct}'\boldsymbol{\alpha} + \boldsymbol{x}_{ict}'\boldsymbol{\beta} + \lambda_i f_{ct} + u_{ict}, \tag{4.1}$$

where y is the average grade of student i in a class c at year t, and $\boldsymbol{d}$ is a vector of variables that includes class size, and measures of actual dispersion of gender and income in each class. The vector $\boldsymbol{x}$ includes indicators for gender, whether or not the student is a high income student, and a cognitive test score corresponding to a test the student took as part of the admission process.

In this model, λ_i is associated with student motivation and ability, and the factor f_{ct} is interpreted as measuring teaching quality of the course c taken in year t. Moreover, the term $\lambda_i f_{ct}$ represents the interaction between students' i motivation λ_i and the quality of the teacher in a class f_{ct}. High teaching quality can have a modest effect on the educational attainment of relatively unmotivated students, although it can dramatically affect performance among strong, motivated students.

To estimate the parameters $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$ and λ_i and f_{ct} in equation (4.1), we use data from administrative records of the economics and finance programs at Bocconi University. The university is an established higher education institution located in Milan, Italy. The data set includes information on course grades, background demographic and socioeconomic characteristics such us gender, family income, and pre-enrollment test scores. Additionally, the data set includes information on enrollment year, academic program, number of exams by academic year, official enrollment, official proportion of female students in each class, and official proportion of high income students in each class. We restrict our attention to students who matriculated in the 2000 academic year and took the same non-elective classes in the first three years of the program. See De Giorgi, Pellizzari and Woolston (2012) and Harding and Lamarche (2014) for additional details on the data.

As in De Giorgi, Pellizzari and Woolston (2012), we estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in equation (4.1) using instrumental variables generated by a random assignment of students into classes. Students were assigned to each class by the administration at Bocconi University, and therefore, the random assignment determine the actual class size, percentages of female students in a class and high income students in a class, which are considered endogenous variables in the literature. The use of the randomized assignment leads to consistent estimation of the coefficients in equation (4.1),

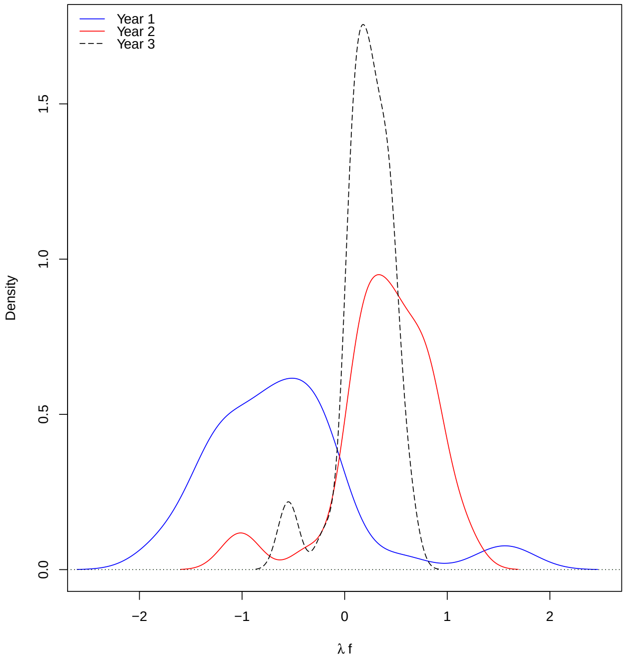

FIGURE 4.1. *Changes in the distribution of student's grade by year in the program.*

satisfying one of the conditions of our approach. In the second stage, we write $R_{ict} = \lambda_i f_{ct} + u_{ict}$, and we estimate the factors and loadings following the procedure described in Section 2.

		PCA	GVE	MGVE
Course 1	Year 1	-0.956	-1.379	-1.394
Course 2	Year 2	-1.040	0.418	0.904
Course 3	Year 3	-1.002	0.961	0.489

TABLE 4.1. *Factor loadings using the Bocconi's data. Principal Component Analysis is denoted by PCA, the grouped variable estimator by GVE, and the mean grouped variable estimator by MGVE.*

Table 4.1 shows the factors f_{ct} estimated using PCA, GVE, and MGVE. While PCA suggests that teacher/course quality f_{ct} does not seem to vary greatly over courses, the MGVE suggests that the quality of courses improves in the second and third year of the program. In Figure 4.2, we estimate the distribution of $\hat{\lambda}_i \hat{f}_{ct}$ by years in the program. It is interesting to see that the distribution is shifted to the right over time as expected, and by the third year, the conditional distribution of the average grade becomes less dispersed. This finding suggest that the students who remained in the program became less heterogeneous and their latent abilities improved over time. This information is naturally important for college administrators and for potential employers interested in recruiting college students.

4.2. **Investigating the effect of an educational policy**

In this section, we illustrate the use of our estimation approach to evaluate an educational policy using data previously employed in Angrist et al. (2002).

During past decades, several educational voucher programs were adopted in the U.S. and Latin America. In 1991, Colombia implemented an educational voucher program that provided resources to students interested in attending private schools. The vouchers were assigned using lotteries, and they were renewable as long as the students maintained satisfactory academic progress. The vast empirical literature has been concerned with the evaluation of the effect of the program on observable outcomes (see, Angrist et al. (2002), Angrist et al. (2006), and Lamarche (2011), among others), but the effect of the program on latent variables such as cognitive ability of students is unknown.

We estimate the following factor-augmented linear panel data model:

$$y_{is} = \alpha d_i + \boldsymbol{x}_{is}' \boldsymbol{\beta} + \lambda_i f_s + u_{is}, \tag{4.2}$$

where y_{is} is student's i test score in subject s and d_i indicates treatment status (i.e., whether student i won the lottery). The parameter α is the mean treatment effect of the program. The vector of independent variables is denoted by $\boldsymbol{x}_{it}$ and the error term by u_{is}. The loading λ_i measures students intrinsic ability or effort that determines performance in the three subjects, and the variable f_s is a subject specific effect that impacts student achievement.

We use Angrist et al. (2002)'s data that includes 284 students who took tests on mathematics, reading and writing. These tests were taken three years after the vouchers were distributed. To facilitate the comparison among subjects, the test scores are in standard deviation units. In addition to an indicator variable for whether the student won a voucher, we use the following independent variables: site dummies, strata indicators for whether the student lives in a neighborhood ranked on a scale of 1-6 from poorest to richest, an indicator for whether the interview was done by a house visit since telephones were used in the majority of the interviews, gender, age, and parents' schooling. We also include an indicator for survey form, because Angrist et al. (2002) data also incorporate responses obtained from a pilot survey designed to test questions and interviewing strategies.

	Control			Treatment		
	PCA	GVE	MGVE	PCA	GVE	MGVE
Mathematics	-0.735	-1.008	-1.019	-1.056	-0.577	-0.639
Reading	-1.158	1.363	1.359	-1.024	1.407	1.412
Writing	-1.057	-0.355	-0.340	-0.914	-0.830	-0.773

Table 4.2. *Factor loadings using the Colombian voucher data (PACES). Principal Component Analysis is denoted by PCA, the grouped variable estimator by GVE, and the mean grouped variable estimator by MGVE.*

Table 4.2 presents the factors for Mathematics, Reading and Writing. We estimate separately f_s for students in the control and treatment group, to document whether these factors are different by treatment status. The table also presents results using the estimation approaches considered in the previous sections. The results for Mathematics and Writing are qualitatively similar using PCA or MGVE, although MGVE reveals significant gains in the treatment group in Mathematics compared to other subjects.

Lastly, to summarize the effect of the program on the factor structure of a model of academic achievement, Figure 4.2 shows the distribution of student's latent ability by treatment status. The figure reveals that the educational policy implemented in Colombia improved latent cognitive outcomes, while increasing the gap between strong and weak students.

5. Conclusions and Discussion

Motivated by the application of a factor-augmented panel data model to administrative data of college and high school students, this paper investigates two of the most important practical challenges in the use of instrumental variable methods. One important limitation of the Madansky (1964) approach is that the number of instruments tends to infinity as the number of time series increases to infinity too. Thus, in "large" T panels, the number of instrumental variables can

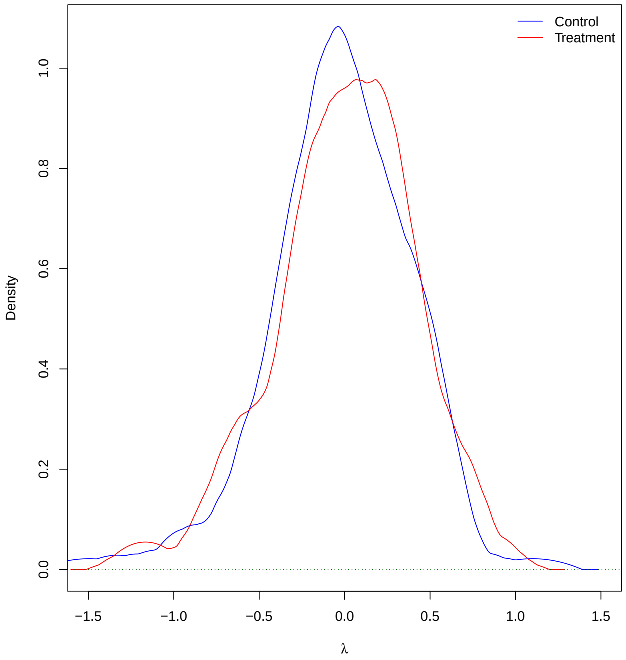

FIGURE 4.2. *The effect of winning a voucher for private schooling on student's ability.*

be larger than the number of cross-sectional observations N leading to poor finite sample properties. To overcome this problem, we propose a class of two-step estimators that build on the recent Lasso literature in the intersection of machine learning and econometrics. In the first stage, we estimate the model using newly developed regularization methods or a linear combination of estimators whose cardinality increases with the sample size. This provides a practical approach to the problem of using many instruments when $T >> N$ or $N >> T$ to estimate factors or loadings, respectively. In the second stage, we construct a linear combinations of the instrumental variable estimators considered in the first stage, while selecting the weights for each of them that satisfies conditions for consistency (i.e., the number of possible partitions is required to grow slower than $\log(Q(T, r)))$. The two-step procedure is consistent and it improves the efficiency of the first stage estimator by using information from different normalizations.

While the proposed approach is computationally intensive and 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 linear panel data models. The instruments are constructed internally from existing data, as illustrated in the empirical section. The application of the method to student administrative data reveals that the estimation of teacher-specific and student-specific latent variables in models of educational attainment is feasible and the analysis reveals interesting findings that are important for college and high school administrators. Further research may involve relaxing the identification assumptions to more general cases and to the extension of approximate factor models.

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.

Anderson, T. W., and H. Rubin (1956): "Statistical Inference in Factor Analysis," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, pp. 111–150. University of California Press.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92(5), 1535–1558.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *The American Economic Review*, 96(3), 847–862.

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

Bai, J., and S. Ng (2010): "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26(6), 15771606.

Bai, J., and P. Wang (2016): "Econometric Analysis of Large Factor Models," *Annual Review of Economics*, 8(1), 53–80.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657–681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80(6), 2369–2429.

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.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. t. Weel (2008): "The Economics and Psychology of Personality Traits," *Journal of Human Resources*, 43(4), 972–1059.

Chen, X., D. T. Jacho-Chvez, and O. Linton (2016): "Averaging of An Increasing Number of Moment Condition Estimators," *Econometric Theory*, 32(1), 3070.

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.

Cunha, F., and J. Heckman (2007): "The Technology of Skill Formation," *American Economic Review*, 97(2), 31–47.

Cunha, F., and J. J. Heckman (2008): "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources*, 43(4), 738–782.

De Giorgi, G., M. Pellizzari, and W. G. Woolston (2012): "Class Size and Class Heterogeneity," *Journal of the European Economic Association*, 10(4), 795–830.

Hägglund, G. (1982): "Factor analysis by instrumental variables methods," *Psychometrika*, 47(2), 209–222.

Hahn, J., and J. Hausman (2003): "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," *American Economic Review*, 93(2), 118–125.

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26(4), 398–422.

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.

Heaton, C., and V. Solo (2012): "Estimation of high-dimensional linear factor models with grouped variables," *Journal of Multivariate Analysis*, 105(1), 348 – 367.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103(6), 2052–86.

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.

Lamarche, C. (2011): "Measuring the incentives to learn in Colombia using new quantile regression approaches," *Journal of Development Economics*, 96(2), 278 – 288.

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.

Stock, J. H., and M. W. Watson (1999): "Forecasting inflation," *Journal of Monetary Economics*, 44(2), 293 – 335.

——— (2002): "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97(460), 1167–1179.

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

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

van der Vaart, A. (1998): *Asymptotic Statistics.* Cambridge University Press, New York.

Appendix A. Proofs

Proof of Theorem 1. The proof for the consistency of the estimator defined in (2.16) follows directly from Belloni, Chen, Chernozhukov and Hansen (2012) under condition (A.5) which is similar to sparsity condition AS in their paper. In terms of the behavior of $E(\boldsymbol{R}_{tq,(3)}\boldsymbol{R}'_{tq,(3)})$ in cases where $T-r-1>N$, it is straightforward to verify that conditions RE and SE are satisfied by the factor structure of the model under Assumptions (A.1)-(A.4). Similarly, the proof for the consistency of the estimator defined in (2.18) follows immediately from verifying that the conditions of Theorem 1 in Chen, Jacho-Chavez and Linton (2016) are met in our case. Condition (A.6) is similar to Assumption (A.1) in Chen, Jacho-Chavez and Linton (2016). Also, given a pair (t,q), $\boldsymbol{\theta}_{tq}\in\boldsymbol{\Theta}$, a compact set, and $G(\boldsymbol{\theta}_{tq})$ in equation (2.12) is continuous such that $G(\boldsymbol{\theta}_{0,tq})=0$. Therefore, conditions (A.2)-(A.4) in Chen, Jacho-Chavez and Linton (2016) are trivially satisfied. □

Proof of Theorem 2. The proof follows closely Theorem 1 in Chen, Jacho-Chavez and Linton (2016), thus we focus on the main differences. First, we write

$$\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\|\leq\sum_{q\in\mathcal{Q}_T^*}\|W_{Tq}\|\times\max_{q\in\mathcal{Q}_T^*}\|\hat{\boldsymbol{\theta}}_{tq}-\boldsymbol{\theta}_t\|+\sum_{q=Q_T^*(T,r)+1}^{Q(T,r)}\|W_{Tq}\|\Delta, \tag{A.1}$$

where Δ is the radius of the compact set $\boldsymbol{\Theta}$. By Assumption (B.4), the last term does $o_p(1)$ as $T\to\infty$, so we concentrate in the first term. To show that the estimator $\hat{\boldsymbol{\theta}}_t$, which based on a linear combination of consistent instrumental variable estimators, is consistent, we need to show that $\max_{q\in\mathcal{Q}_T^*}\|\hat{\boldsymbol{\theta}}_{tq}-\boldsymbol{\theta}_t\|=o_p(1)$ as N and T go jointly to ∞, because the weights are bounded by Assumption (B.4).

By Assumption (B.1), if $\max_{q\in\mathcal{Q}_T^*}\|\hat{\boldsymbol{\theta}}_{tq}-\boldsymbol{\theta}_t\|>\delta$, we have that $\|G_q(\hat{\boldsymbol{\theta}}_{tq})\|\geq\epsilon_N(\delta)$ for some q. Therefore,

$$\Pr\left(\max_{q\in\mathcal{Q}_T^*}\|\hat{\boldsymbol{\theta}}_{tq}-\boldsymbol{\theta}_t\|>\delta\right)\leq\Pr\left(\max_{q\in\mathcal{Q}_T^*}\|G_q(\hat{\boldsymbol{\theta}}_{tq})\|\geq\epsilon_N(\delta)\right). \tag{A.2}$$

For $\epsilon_N(\delta) > 0$,

$$\begin{aligned}\max_{q\in\mathcal{Q}_T^*}\|G_q(\hat{\boldsymbol{\theta}}_{tq})\| &\leq \max_{q\in\mathcal{Q}_T^*}\|G_q(\hat{\boldsymbol{\theta}}_{tq}) - G_{Nq}(\hat{\boldsymbol{\theta}}_{tq})\| + \max_{q\in\mathcal{Q}_T^*}\|G_{Nq}(\hat{\boldsymbol{\theta}}_{tq})\| \\ &\leq \max_{q\in\mathcal{Q}_T^*}\sup_{\theta_t\in\Theta_t}\|G_q(\boldsymbol{\theta}_t) - G_{Nq}(\boldsymbol{\theta}_t)\| + \max_{q\in\mathcal{Q}_T^*}\|G_{Nq}(\hat{\boldsymbol{\theta}}_{tq})\| \\ &= o_p(\alpha_{2N}) + \max_{q\in\mathcal{Q}_T^*}\|G_{Nq}(\hat{\boldsymbol{\theta}}_{tq})\|,\end{aligned}$$

by Assumption (B.3). By the triangle inequality, we can bound the second term as follows:

$$\begin{aligned}\max_{q\in\mathcal{Q}_T^*}\|G_{Nq}(\hat{\boldsymbol{\theta}}_{tq})\| &\leq \max_{q\in\mathcal{Q}_T^*}\left(\|G_{Nq}(\hat{\boldsymbol{\theta}}_{tq}) - \inf_{\theta_t\in\Theta}\|G_{Nq}(\boldsymbol{\theta}_t)\|\right) + \max_{q\in\mathcal{Q}_T^*}\inf_{\theta_t\in\Theta}\|G_{Nq}(\boldsymbol{\theta}_t)\| \\ &\leq \max_{q\in\mathcal{Q}_T^*}\left(\|G_{Nq}(\hat{\boldsymbol{\theta}}_{tq}) - \inf_{\theta_t\in\Theta}\|G_{Nq}(\boldsymbol{\theta}_t)\|\right) + \max_{q\in\mathcal{Q}_T^*}\|G_{Nq}(\boldsymbol{\theta}_{t0})\|.\end{aligned}$$

Note that the firm term is $o_p(\alpha_{1N})$ by Assumption (B.2) and the second term is zero by definition. Therefore,

$$\max_{q\in\mathcal{Q}_T^*}\|G_q(\hat{\boldsymbol{\theta}}_{tq})\| \leq o_p(\alpha_{1N}) + o_p(\alpha_{2N}) = o_p(\epsilon_N(\delta)),$$

with $\epsilon_N(\delta) \to 0$ as N and T go to infinity. □

Appendix B. Additional simulation results

In this section, we briefly compare the performance of three estimators considering Design B: 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 B.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.

			Estimators								
	N	T	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	20	20	-0.091	-0.004	0.005	0.161	0.089	0.001	-0.005	-0.007	-0.008
RMSE	20	20	0.155	0.355	0.043	0.167	0.127	0.060	0.053	0.056	0.061
Bias	50	20	-0.109	0.000	-0.003	0.158	0.036	0.001	0.000	-0.001	0.001
RMSE	50	20	0.178	0.384	0.029	0.165	0.073	0.037	0.031	0.031	0.038
Bias	20	50	-0.099	-0.020	-0.002	0.177	0.046	0.006	0.002	0.002	0.003
RMSE	20	50	0.123	0.090	0.013	0.179	0.085	0.035	0.031	0.033	0.037
Bias	50	50	-0.095	0.001	0.000	0.175	0.004	0.000	-0.002	-0.001	-0.001
RMSE	50	50	0.122	0.059	0.009	0.177	0.024	0.022	0.021	0.020	0.022
Design 2: Loadings distributed as Gaussian											
Bias	20	20	-0.219	-0.015	-0.007	0.087	0.017	0.013	-0.004	-0.004	-0.003
RMSE	20	20	0.325	0.122	0.068	0.196	0.062	0.063	0.068	0.054	0.060
Bias	50	20	-0.194	-0.002	-0.005	0.067	0.000	0.002	0.003	0.003	0.005
RMSE	50	20	0.294	0.048	0.043	0.184	0.033	0.036	0.046	0.036	0.041
Bias	20	50	-0.102	-0.004	0.002	0.074	0.001	0.007	0.005	0.002	0.001
RMSE	20	50	0.177	0.040	0.036	0.143	0.041	0.038	0.042	0.032	0.034
Bias	50	50	-0.129	-0.002	0.000	0.087	-0.002	0.000	0.001	0.000	-0.001
RMSE	50	50	0.210	0.025	0.024	0.154	0.022	0.022	0.028	0.021	0.022
Design 3: Loadings distributed as Uniform and Gaussian											
Bias	20	20	-0.178	-0.009	0.001	0.073	0.007	0.009	-0.004	-0.006	-0.006
RMSE	20	20	0.272	0.095	0.061	0.167	0.058	0.058	0.074	0.058	0.068
Bias	50	20	-0.173	-0.005	-0.004	0.064	0.000	0.002	0.001	0.002	0.001
RMSE	50	20	0.259	0.054	0.043	0.164	0.036	0.041	0.042	0.033	0.040
Bias	20	50	-0.087	-0.002	-0.001	0.069	0.008	0.012	-0.001	0.001	0.001
RMSE	20	50	0.146	0.043	0.032	0.118	0.038	0.038	0.040	0.032	0.035
Bias	50	50	-0.106	-0.002	0.000	0.086	-0.001	0.000	-0.002	0.000	0.001
RMSE	50	50	0.170	0.026	0.022	0.137	0.023	0.024	0.027	0.021	0.021

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