Combining Instrumental Variable Estimators for a Panel Model with Factors

Matthew Harding[*]
Department of Economics,
University of California at Irvine,
Email: harding1@uci.edu

Carlos Lamarche
Department of Economics,
University of Kentucky,
Email: clamarche@uky.edu

Chris Muris
Department of Economics,
McMaster University,
Email: muerisc@mcmaster.ca

February 10, 2024

Abstract

We address the estimation of factor-augmented panel data models using observed measurements to proxy for unobserved factors or loadings and explore the use of internal instruments to address the resulting endogeneity. The main challenge consists in that economic theory rarely provides insights into which measurements to choose as proxies when several are available. To overcome this problem, we propose a new class of estimators that are linear combinations of instrumental variable estimators and establish large sample results. We also show that an optimal weighting scheme exists, leading to efficiency gains relative to an instrumental variable estimator. Simulations show that the proposed approach performs better than existing methods. We illustrate the new method using data on test scores across US school districts.

JEL: C23; C26; C38; I21
Keywords: Factor Model; Panel Data; Instrumental Variables; Administrative data

[*] The authors would like to thank Irene Botosaru, Rajeev Darolia, Atsushi Inoue, Ana Herrera, Arturas Juodis, Jeff Racine, Youngki Shin, Jeff Wooldridge, and participants at the 2022 Midwest Econometrics Group Conference and the 28th International Panel Data Conference, for comments and useful conversations. We thank Chris Heaton for sharing his programs for estimation of the approximate factor model.

1 Introduction

Applications of factor models in economics, finance, and psychology continue to be extremely popular. In economics, the identification and estimation of factor models has received substantial attention in a number of areas, from macro-finance to labor economics and development (Bernanke et al. 2005, Kim & Oka 2014, Attanasio et al. 2020). Important work has studied the role of cognition, personality traits, and academic motivation on child development (Cunha & Heckman 2007, 2008, Borghans et al. 2008, Heckman et al. 2013). Factor-augmented regressions as in Stock & Watson (1999, 2002) are known to improve forecasts of macroeconomic time series such as inflation and industrial production. The literature also includes new models for high-dimensional data sets (Bai & Wang 2016), and methods for panels with large cross-sectional and time-series dimensions, following the influential work by Pesaran (2006) and Bai (2009). In panel data econometrics, one popular interpretation treats the latent factors as a generalization of traditional fixed effects models (Harding & Lamarche 2014, Chudik & Pesaran 2015, Moon & Weidner 2015, 2017, Ando & Bai 2016, 2017, Juodis & Sarafidis 2018, Harding et al. 2020, among others).

In this paper, we focus on a class of estimators that use internally generated instruments. Papers by Heaton & Solo (2012), and Juodis & Sarafidis (2020), among others, also propose to estimate similar models using internally constructed instruments, an idea that can be traced back to the work of Madansky (1964). These estimators use outcome variables to proxy the factors (or loadings) and then use other outcome variables as instruments. Slope parameters are identified, but the factors are not identified without further restrictions. In some cases, identification is achieved through the use of dedicated measurements, where *a priori* knowledge is used to associate certain measurements uniquely with specific factors (see Cunha et al. 2021, as an example of a test associated uniquely with a given skill).

A challenge for this approach is that the selection of the measurements to use as a proxy for the factors (or loadings) is, in many cases, arbitrary. Economic theory often does not provide a framework to select measurements or proxies when several are available. To address this challenge, we propose an estimator that combines information from each subset of measurements used to proxy the factors. We demonstrate that there exists a combination of estimators that improves the efficiency of an estimator that uses an arbitrary subset, and we show that the proposed estimator is consistent and asymptotically normal under standard conditions. Therefore, the proposed approach eliminates subjective choices made by practitioners, while improving the efficiency of existing instrumental variable estimators.

This paper develops a new class of estimators that are simple to implement and offer practitioners objective evidence when they lack concrete guidance on their choices. The estimation of slope parameters using instrumental variables is investigated in Bai & Ng (2010), Harding & Lamarche (2011), Ahn et al. (2013), Robertson & Sarafidis (2015), Juodis & Sarafidis (2020), Norkutė et al. (2021), and Juodis & Sarafidis (2022), among others. This literature uses instrumental variables (IVs) based on defactored covariates via principal components, proxies based on observables, or external instruments. Holtz-Eakin et al. (1988) consider a panel data model with one factor and identify reduced-form parameters using instrumental variables. On the other hand, the latent factor structure is estimated in Madansky (1964), Hägglund (1982), and Heaton & Solo (2012). This literature uses internal instruments based on response variables. In contrast to the existing literature using IVs in the estimation of pure factor models, we extend this approach to factor-augmented panel data models and develop the corresponding estimation theory.

Motivating example. Throughout the paper, we illustrate our results using an example based on the empirical application. We use district-level administrative data on test scores by subject in grades 3 to 8 in over 2,000 school districts. We estimate a model for educational attainment in mathematics in middle school, $y_{gj} = \boldsymbol{x}'_{gj}\boldsymbol{\gamma} + \lambda_g f_j + u_{gj}$, where y_{gj} is the average normalized test score in district g in grade j, $\boldsymbol{x}_{gj}$ is a vector of regressors, λ_g is a latent factor loading, f_j is a factor, and u_{gj} is the error term. In our dataset, there are 6 possible choices of measurements to proxy λ_g. Each proxy creates classical measurement error bias, which can be addressed by using other measurements of academic achievement as internal IVs.

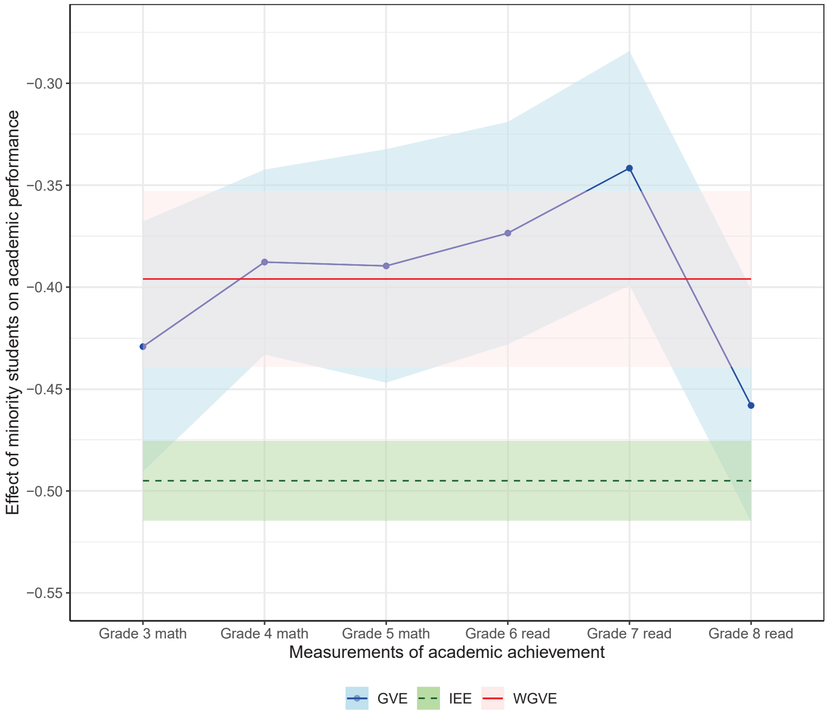

Figure 1: *The effect of percentage of minority students on middle school math. The available proxies for the district specific loading are math test scores (grade 3 to 5) and reading test scores (grade 6 to 8). GVE denotes group variable estimator, IEE denotes interactive effects, and WGVE denotes the weighted group variable estimator.*

The impact of these choices on the estimation of the first coefficient, γ_1, is documented in Figure 1. We report point-estimates and 95% confidence intervals obtained by employing the proposed methods, and we also report results obtained by the interactive effects (IEE) estimator (Bai 2009) for comparison. The group variable estimator (GVE), which is presented in Section 2, depends on choosing a measurement of achievement to proxy λ_g arbitrarily from the available set. The weighted group variable estimator (WGVE) proposed in Section 3 combines estimates based on all the proxies, producing results that do not vary by the choice of measurement. WGVE offers an improvement in the precision of the estimator in comparison to GVE.

Monte Carlo explorations of the small sample performance of the proposed estimators are discussed in Section 4. The fully developed empirical illustration of the proposed approach is presented in Section 5 highlighting insights into academic performance and the heterogeneous nature of school districts in the US. Section 6 concludes. Mathematical proofs are offered in the online appendix.

2 Model and estimation

This paper considers the following model for $g = 1, 2, \cdots, G$ and $j = 1, 2, \cdots, J$:

$$y_{gj} = \boldsymbol{x}'_{gj}\boldsymbol{\gamma} + \boldsymbol{\lambda}'_g \boldsymbol{f}_j + u_{gj}, \tag{1}$$

where $y_{gj} \in \mathbb{R}$ is a response variable, $\boldsymbol{x}_{gj} \in \mathcal{X} \subseteq \mathbb{R}^{p_x \times 1}$ is a vector of independent variables, $\boldsymbol{\lambda}_g = (\lambda_{g,1}, \lambda_{g,2}, \ldots, \lambda_{g,r})' \in \Lambda \subseteq \mathbb{R}^{r \times 1}$ is a vector of factor loadings, $\boldsymbol{f}_j = (f_{j,1}, f_{j,2}, \ldots, f_{j,r})' \in F \subseteq \mathbb{R}^{r \times 1}$ is a vector of latent factors, and u_{gj} is an error term. The number of factors r does not need to be known, as one can determine the number of factors following a number of approaches (e.g., Bai & Ng 2002, Onatski 2010, Trapani 2018).

Standard notation for panel data factor models would have a dependent variable y_{it} with subscripts i for cross-section units and t for time. Instead, we use g (instead of i) to refer to clusters and j (instead of t) to refer to observations within a cluster. This emphasizes that the index j may not refer to time, but instead to any type of repeated observation within a cluster g. For instance, in the empirical application presented in Section 5, y_{gj} is the average test score in school district g corresponding to grade j. The number of observations J does not vary with g, although our analysis can be extended to accommodate this case.

The observations within g are grouped into three sets A, B, and C. Denote the cardinality of the set S by $|S| = m_S$. The set A contains the indexes for which we wish to estimate the factors. For instance, practitioners might be interested in the estimation of grade-level factors affecting academic achievement in middle school and not in elementary school, and vice versa. Thus, we are interested in the estimation of $\boldsymbol{\gamma}$ and $(\boldsymbol{f}_j, j \in A)$. Once estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{f}_j$ are available, it is straightforward to construct an estimator for $\boldsymbol{\lambda}_g$ (see, e.g., Heaton & Solo 2012, Bai & Ng 2013). The set B contains the indices of the observations we will use as proxies for the loadings. The set C contains the indices of the observations we will use as instrumental variables.

The model satisfies the following conditions:

A1. $\left\{\left(y_{gj}, \boldsymbol{x}'_{gj}\right), j = 1, 2, \cdots, J\right\}$ is independent across $g = 1, 2, \cdots, G$ conditional on $\boldsymbol{f}_j$ for $1 \leq j \leq J$.

A2. The variable y_{gj} is generated by model (1). The sets A, B, C are a partition of $\{1, 2, \cdots, J\}$, with $m_A \geq 1$ and $m_C \geq m_B \geq r$.

A3. $E\left(\boldsymbol{\lambda}_g u_{gj}\right) = \mathbf{0}$ for all g and for all $j \in A \cup B$; $E\left(u_{gh} u_{gj}\right) = 0$ for all g and for all $h \in (A \cup B)$ and $j \in C$; $E\left(\boldsymbol{x}_{gh} u_{gj}\right) = \mathbf{0}$ for all g and for all $1 \leq h, j \leq J$.

Assumption A1 is a sampling assumption that restricts dependence across g. Following A2, data are generated by model (1) and we can partition the observations as described above, with a sufficient number of instruments, m_C, and subsets of measurements, m_B. Lastly, Assumption A3 guarantees that the error term is not correlated with the loadings and independent variables, and that it is essentially serially uncorrelated. This assumption allows us to use observations from the set C as internal instruments. This assumption may be relaxed if external instruments are available.

2.1 Overview of methods and practical challenges

For ease of exposition, we set $m_B = r$, i.e., the number of observations in the subset B is equal to the number of factors. Everything that follows can be modified to allow for $m_B > r$, as shown in Section 2 of the online appendix. We start by collecting (1) over B, obtaining

$$\boldsymbol{y}_{gB} = \boldsymbol{x}_{gB}\boldsymbol{\gamma} + \boldsymbol{f}_B'\boldsymbol{\lambda}_g + \boldsymbol{u}_{gB}, \tag{2}$$

where a B subscript indicates that the elements were gathered to form $\boldsymbol{y}_{gB} = (y_{gj},\ j \in B) \in \mathbb{R}^{r\times 1}$, $\boldsymbol{f}_B = (\boldsymbol{f}_j,\ j \in B) \in \mathbb{R}^{r\times r}$, $\boldsymbol{x}_{gB} = (\boldsymbol{x}_{gj},\ j \in B) \in \mathbb{R}^{r\times p_x}$, and $\boldsymbol{u}_{gB} = (u_{gj},\ j \in B) \in \mathbb{R}^{r\times 1}$. It is standard in the literature to consider the following condition, which controls the behavior of $\boldsymbol{f}_B$ and guarantees that all parameters in equation (4) are well-defined.

A4. The $r \times r$ matrix of factors $\boldsymbol{f}_B$ is invertible.

Assuming invertibility of $\boldsymbol{f}_B$, we solve for $\boldsymbol{\lambda}_g$ in equation (2):

$$\boldsymbol{\lambda}_g = [\boldsymbol{f}_B']^{-1}(\boldsymbol{y}_{gB} - \boldsymbol{x}_{gB}\boldsymbol{\gamma} - \boldsymbol{u}_{gB}). \tag{3}$$

Substituting the representation of the loading (3) into equation (1), one obtains, for each $j \in A$:

$$y_{gj} = \boldsymbol{y}_{gB}'\boldsymbol{\theta}_{jB} + \boldsymbol{x}_{gj}'\boldsymbol{\gamma} - \boldsymbol{\theta}_{jB}'\boldsymbol{x}_{gB}\boldsymbol{\gamma} + \left(u_{gj} - \boldsymbol{\theta}_{jB}'\boldsymbol{u}_{gB}\right), \tag{4}$$

where $\boldsymbol{\theta}_{jB} = \boldsymbol{f}_B^{-1}\boldsymbol{f}_j$. By noting that $-\boldsymbol{\theta}_{jB}'\boldsymbol{x}_{gB}\boldsymbol{\gamma} = -\sum_{k=1}^{r}\theta_{j,k}\boldsymbol{x}_{g,k}'\boldsymbol{\gamma} = \boldsymbol{h}_{gB}'\boldsymbol{\delta}_{jB}$, where $\boldsymbol{h}_{gB} = \mathrm{vec}(\boldsymbol{x}_{gB}')$ and $\boldsymbol{\delta}_{jB} = -\boldsymbol{\theta}_{jB} \otimes \boldsymbol{\gamma}$, we obtain the following reduced form equation:

$$y_{gj} = \boldsymbol{y}_{gB}'\boldsymbol{\theta}_{jB} + \boldsymbol{x}_{gj}'\boldsymbol{\gamma} + \boldsymbol{h}_{gB}'\boldsymbol{\delta}_{jB} + v_{gj}, \tag{5}$$

where $v_{gj} = u_{gj} - \boldsymbol{\theta}'_{jB}\boldsymbol{u}_{gB}$.

Although $\boldsymbol{\theta}_{jB}$, $\boldsymbol{\delta}_{jB}$ and $\boldsymbol{\gamma}$ could be estimated by standard methods for linear models, the variable in the first term of equation (5), $\boldsymbol{y}_{jB}$, is endogenous because it is correlated with $\boldsymbol{u}_{gB}$, which appears as part of the error term, v_{gj}. We propose to estimate the parameters in equation (5) using internal instruments $\boldsymbol{y}_{gC}$. The use of internal instruments leads to identification of the slope parameter $\boldsymbol{\gamma}$, and the reduced form coefficients $\boldsymbol{\theta}_{jB}$ and $\boldsymbol{\delta}_{jB}$. The vector $\boldsymbol{\theta}_{jB}$ is a transformation of the factor $\boldsymbol{f}_j$, and similar to results established in the literature, factors are not separately identified without further restrictions. Moreover, it is easy to see that the within g correlation is different from zero. Even if we assume that the errors in (1) are independent within g for $j \in A$, then for $j, h \in A$ and $j \neq h$, $\text{Cov}(v_{gj}, v_{gh}) = \boldsymbol{\theta}'_{jB}\text{Var}(\boldsymbol{u}_{gB})\boldsymbol{\theta}_{hB}$. To handle this dependence, we employ the general framework developed in Hansen & Lee (2019).

Let $\boldsymbol{y}_{gA}$ and $\boldsymbol{v}_{gA}$ denote vectors of dimension $m_A \times 1$, and $\boldsymbol{x}_{gA}$ denote a matrix of dimension $m_A \times p_x$. It is convenient to stack the equations corresponding to set A, and write the system as

$$\boldsymbol{y}_{gA} = \left(\boldsymbol{I}_{m_A} \otimes \boldsymbol{y}'_{gB}\right)\boldsymbol{\theta}_{AB} + \boldsymbol{x}_{gA}\boldsymbol{\gamma} + \left(\boldsymbol{I}_{m_A} \otimes \boldsymbol{h}'_{gB}\right)\boldsymbol{\delta}_{AB} + \boldsymbol{v}_{gA}, \tag{6}$$

where $\boldsymbol{\theta}_{AB} = (\boldsymbol{\theta}_{jB},\ j \in A) \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{p_y \times 1}$, where $p_y = m_A r$, and $\boldsymbol{\delta}_{AB} = (\boldsymbol{\delta}_{jB}, j \in A) \in \boldsymbol{\Delta} \subseteq \mathbb{R}^{p_x p_y \times 1}$. For notational convenience, we define the dependent variable $\boldsymbol{y}_g := \boldsymbol{y}_{gA}$, the explanatory variables

$$\boldsymbol{X}_g := \begin{bmatrix} \boldsymbol{I}_{m_A} \otimes \boldsymbol{y}'_{gB} & \boldsymbol{x}_{gA} & \boldsymbol{I}_{m_A} \otimes \boldsymbol{h}'_{gB} \end{bmatrix},$$

and the instruments $\boldsymbol{Z}_g := \begin{bmatrix} \boldsymbol{I}_{m_A} \otimes \boldsymbol{y}'_{gC} & \boldsymbol{x}_{gA} & \boldsymbol{I}_{m_A} \otimes \boldsymbol{h}'_{gB} \end{bmatrix}$. The matrix $\boldsymbol{X}_g$ is of dimension $m_A \times k_x$, where $k_x = p_y + p_x(1 + p_y)$, and $\boldsymbol{Z}_g$ is of dimension $m_A \times k_z$, where $k_z = p_z + p_x(1 + p_y)$ and $p_z = m_A m_C$. By Assumption A2, $m_C \geq r$, and therefore, $k_z \geq k_x$.

Furthermore, let $\boldsymbol{\beta} := (\boldsymbol{\theta}'_{AB}, \boldsymbol{\gamma}', \boldsymbol{\delta}'_{AB})'$ and $\boldsymbol{e}_g := \boldsymbol{v}_{gA}$, so that (6) can be written as $\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{e}_g$.

The corresponding two-stage least squares (2SLS) estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{Z}_g \left(\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{Z}_g \left(\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{y}_g. \tag{7}$$

We denote the estimator in (7) as Group Variable Estimator (GVE) because it is based on grouping observations according to the three partitions in Assumption A2.

The GVE estimator has a number of attractive features. First, the implementation is trivial. The estimator belongs to the class of 2SLS estimators with internal instruments for the linear system (6). Second, we can easily accommodate additive fixed effects and endogenous regressors $\boldsymbol{x}_{gj}$ that are correlated with the error term u_{gj} by using external instruments.

Motivating example (cont.) We estimate a model for educational attainment in mathematics in middle school using data from G districts in grades 3 to 8. In this application, $J = 9$, where the first three math test scores corresponds to elementary school (3rd to 5th grade), the next three scores corresponds to middle school math, and the last 3 are measurements corresponding to reading scores in middle school. (The three elementary school reading tests are ignored as insufficiently relevant for middle school math.) Thus, $A = \{4, 5, 6\}$ corresponding to grades 6, 7 and 8 in mathematics. If we select $B = \{1\}$ (3rd grade math) as shown in Figure 1 and consequently $C = \{2, 3, 7, 8, 9\}$, the GVE estimate of γ_1 is -0.429, while the interactive effects (IEE) estimate is -0.495. These results suggest that a ten percent increase in the proportion of minority students decreases average math scores by 13 to 17 percent of a standard deviation. Despite its practical simplicity, the GVE estimator has an important drawback. It depends on the choice of B even if the set A is held fixed. For instance, in Figure 1, if a practitioner chooses $B = \{8\}$ (7th grade reading) instead of $B = \{1\}$ (3rd grade math), the estimate of the slope effect is -0.342 which represents a 20 percent increase. The solution we pursue in Section 3 is to average over multiple subsets.

2.2 Large sample results

We now establish conditions under which the estimator in (7) is consistent and asymptotically normal. The results are obtained for fixed J, and therefore A, B, C are fixed too. We will leverage the fact that our estimator can be viewed as a 2SLS estimator for clustered data. This allows us to use the asymptotic theory in Hansen & Lee (2019), in particular their results for 2SLS estimation in Theorems 8 and 9.

To state our results, define the total number of observations $n = G \times m_A$ and let

$$\boldsymbol{Q}_n = \frac{1}{n}\sum_{g=1}^{G} E\left[\boldsymbol{Z}_g'\boldsymbol{X}_g\right], \quad \boldsymbol{W}_n = \frac{1}{n}\sum_{g=1}^{G} E\left[\boldsymbol{Z}_g'\boldsymbol{Z}_g\right], \quad \boldsymbol{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G} E\left[\boldsymbol{Z}_g'\boldsymbol{e}_g\boldsymbol{e}_g'\boldsymbol{Z}_g\right],$$

$$\boldsymbol{V}_n = \left(\boldsymbol{Q}_n'\boldsymbol{W}_n^{-1}\boldsymbol{Q}_n\right)^{-1}\boldsymbol{Q}_n'\boldsymbol{W}_n^{-1}\boldsymbol{\Omega}_n\boldsymbol{W}_n^{-1}\boldsymbol{Q}_n\left(\boldsymbol{Q}_n'\boldsymbol{W}_n^{-1}\boldsymbol{Q}_n\right)^{-1}.$$

We consider the following assumptions:

A5. For some $s > 2$, $\sup_{g,j} E\left|y_{gj}\right|^{2s} < \infty$ and $\sup_{g,j,k} E\left|x_{gj,k}\right|^{2s} < \infty$.

A6. Let $\zeta_{\min}(\cdot)$ denote the smallest eigenvalue, and K_W and K_Ω be constants. Then, (i) $\boldsymbol{Q}_n$ has full rank and $\zeta_{\min}\left(\boldsymbol{W}_n\right) \geq K_W > 0$, and (ii) $\zeta_{\min}\left(\boldsymbol{\Omega}_n\right) \geq K_\Omega > 0$.

The result in Theorem 1 is obtained considering several standard assumptions. Assumption A5 is a boundedness condition on the regressors and outcome variable that allows for distributional heterogeneity, and is sufficient for Hansen & Lee (2019)'s central limit theorem. The first part of Assumption A6 asks for sufficient correlation of the instruments with the regressors, and the second part introduces a standard condition that guarantees a well-defined limiting distribution.

Then, we have the following result:

Theorem 1. *Under Assumptions A1-A6(i), as $G \to \infty$, the GVE defined in* (7) *is consistent, i.e.* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$. *Moreover, under Assumptions A1-A6, as $G \to \infty$,*

$$\boldsymbol{V}_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right).$$

The proof of Theorem 1 is presented in the online appendix, and consists of verifying the conditions for Theorems 8 and 9 in Hansen & Lee (2019). Their requirement that the observations from each g are asymptotically negligible (cf. their Assumptions 1, 2) is automatically satisfied, as our panel is balanced by construction, and because we use the same fixed set A for each g.

3 Combining GVE estimators

The issue of multiple available partitions deserves further treatment as there are many situations where economic theory does not offer concrete guidance. In those situations, practitioners face a possibly large number of subsets B to choose from. Figure 1 illustrates that it is not clear a priori how to select measurements to proxy the loadings following equation (3), leading to important practical questions. Considering the first partition might be arbitrary, as noted in a series of recent papers (Attanasio, Meghir & Nix 2020, Del Bono, Kinsler & Pavan 2020), see also the discussion by Freyberger (2021) in a related context.

There are Q_J ways of choosing the subset B in equation (2) with $m_B = r$ elements, where

$$Q_J = \binom{J - m_A}{m_B} = \frac{(J - m_A)!}{m_B!\ (J - m_A - m_B)!} \leq 2^{J - m_A}. \tag{8}$$

Let $q \in \{1, 2, \ldots, Q_J\}$ index a choice of $B \subset \{1, 2, \ldots, J\} \setminus A$ from the collection of all possible sets $\{B_1, B_2, \ldots, B_{Q_J}\}$. For each q, we can construct an estimator

$$\widehat{\boldsymbol{\beta}}_q = \left(\sum_{g=1}^{G} \boldsymbol{X}'_{g,q} \boldsymbol{Z}_{g,q} \left(\sum_{g=1}^{G} \boldsymbol{Z}'_{g,q} \boldsymbol{Z}_{g,q} \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}'_{g,q} \boldsymbol{X}_{g,q} \right)^{-1} \sum_{g=1}^{G} \boldsymbol{X}'_{g,q} \boldsymbol{Z}_{g,q} \left(\sum_{g=1}^{G} \boldsymbol{Z}'_{g,q} \boldsymbol{Z}_{g,q} \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}'_{g,q} \boldsymbol{y}_g, \tag{9}$$

where $\boldsymbol{y}_g$ is as before, and the explanatory variables and instruments are constructed from levels of y_{gj} over the sets B_q and $C_q \subseteq \{1, 2, \cdots, J\} \setminus (A \cup B_q)$ corresponding to partition q:

$$\boldsymbol{X}_{g,q} = \begin{bmatrix} \boldsymbol{I}_{m_A} \otimes \boldsymbol{y}'_{gB_q} & \boldsymbol{x}_{gA} & \boldsymbol{I}_{m_A} \otimes \boldsymbol{h}'_{gB_q} \end{bmatrix}, \text{ and } \boldsymbol{Z}_{g,q} = \begin{bmatrix} \boldsymbol{I}_{m_A} \otimes \boldsymbol{y}'_{gC_q} & \boldsymbol{x}_{gA} & \boldsymbol{I}_{m_A} \otimes \boldsymbol{h}'_{gB_q} \end{bmatrix}.$$

Before introducing the main estimator in the next section, we obtain the joint distribution of $(\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \cdots, \widehat{\boldsymbol{\beta}}_{Q_J}')'$ in Theorem 2 below. To establish the result, let $\boldsymbol{e}_{g,q} = \boldsymbol{y}_g - \boldsymbol{X}_{g,q}\boldsymbol{\beta}_q$, and consider the following definitions:

$$\boldsymbol{Q}_{n,q} = \frac{1}{n}\sum_{g=1}^{G} E\left[\boldsymbol{Z}_{g,q}'\boldsymbol{X}_{g,q}\right], \quad \boldsymbol{W}_{n,q} = \frac{1}{n}\sum_{g=1}^{G} E\left[\boldsymbol{Z}_{g,q}'\boldsymbol{Z}_{g,q}\right], \quad \boldsymbol{\Omega}_{n,ql} = \frac{1}{n}\sum_{g=1}^{G} E\left[\boldsymbol{Z}_{g,q}'\boldsymbol{e}_{g,q}\boldsymbol{e}_{g,l}'\boldsymbol{Z}_{g,l}\right],$$

$$\boldsymbol{\Sigma}_{n,ql} = \left(\boldsymbol{Q}_{n,q}'\boldsymbol{W}_{n,q}^{-1}\boldsymbol{Q}_{n,q}\right)^{-1}\boldsymbol{Q}_{n,q}'\boldsymbol{W}_{n,q}^{-1}\boldsymbol{\Omega}_{n,ql}\boldsymbol{W}_{n,l}^{-1}\boldsymbol{Q}_{n,l}\left(\boldsymbol{Q}_{n,l}'\boldsymbol{W}_{n,l}^{-1}\boldsymbol{Q}_{n,l}\right)^{-1},$$

and let $\boldsymbol{\Sigma}_n$ be a $k_x Q_J \times k_x Q_J$ matrix with typical block $\boldsymbol{\Sigma}_{n,ql}$, that is $\boldsymbol{\Sigma}_n = [\boldsymbol{\Sigma}_{n,ql}]$. Similarly, let $\underline{\boldsymbol{\Omega}}_n = [\boldsymbol{\Omega}_{n,ql}]$ and $\Xi_n = (I_{Q_J} \otimes \boldsymbol{H})\Sigma_n(I_{Q_J} \otimes \boldsymbol{H})'$, where $\boldsymbol{H} = \left[\mathbf{0}_{p_x \times p_y}\ I_{p_x}\ \mathbf{0}_{p_x \times p_x p_y}\right]$ selects the elements corresponding to $\boldsymbol{\gamma}$ from $\boldsymbol{\beta}_q = (\boldsymbol{\theta}_{AB_q}', \boldsymbol{\gamma}', \boldsymbol{\delta}_{AB_q}')'$. Below, we suppress the dependence of the reduced form parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ on A, and we will write $\boldsymbol{\beta}_q = (\boldsymbol{\theta}_q', \boldsymbol{\gamma}', \boldsymbol{\delta}_q')'$.

Consider the following assumptions:

B1. Condition A3 holds for B_q and C_q for $1 \leq q \leq Q_J$, and Condition A4 in Theorem 1 also holds for $1 \leq q \leq Q_J$, so that, for each q, $\boldsymbol{f}_{B_q}$ is an invertible matrix of dimension $r \times r$.

B2. Let $\zeta_{\min}(\cdot)$ denote the smallest eigenvalue, and let K_W and K_Ω be constants. (i) Condition A6(i) in Theorem 1 holds for all $1 \leq q \leq Q_J$, so that $\boldsymbol{Q}_{n,q}$ has full rank, and $\zeta_{\min}\left(\boldsymbol{W}_{n,q}\right) \geq K_W > 0$. (ii) Additionally, $\zeta_{\min}\left(\underline{\boldsymbol{\Omega}}_n\right) \geq K_\Omega > 0$.

Assumptions B1 and B2 are generalizations of Assumptions A3, A4 and A6 in Theorem 1. The first part of B1 implies that the error term is independent of the loadings and serially uncorrelated, implying the validity of instruments in C_q across the Q_J ways of choosing B_q. The second part of condition B1 imposes restrictions to generate suitable non-singular transformations across all partitions. In practice, the second part of Condition B1 might not hold for all q. In this case, one can restrict the set of q for which the assumption is expected to hold. Lastly, condition B2 guarantees a well-defined asymptotic distribution across feasible subsets.

Let $\underline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \cdots, \boldsymbol{\beta}_{Q_J}'\right)'$ and $\widehat{\underline{\boldsymbol{\beta}}}$ denote the corresponding estimator. The following result establishes the joint distribution of a fixed number of GVEs:

Theorem 2. *Under Assumptions A1, A2, A5, B1 and B2, as* $G \to \infty$,

$$\boldsymbol{\Sigma}_n^{-1/2}\sqrt{n}\left(\underline{\widehat{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right).$$

The proof of this result builds on Theorem 1. Because $\underline{\widehat{\gamma}} = (\widehat{\gamma}_1', \widehat{\gamma}_2', \cdots, \widehat{\gamma}_{Q_J}')' = (I_{Q_J} \otimes \boldsymbol{H})\,\underline{\widehat{\boldsymbol{\beta}}}$, it follows immediately from Theorem 2 that $\underline{\widehat{\gamma}}$ has covariance $\boldsymbol{\Xi}_n$ and the joint distribution of $\underline{\widehat{\gamma}}$ is asymptotically Gaussian.

3.1 Estimation and parameter of interest

It is natural to consider weighted averages of the estimators over all possible partitions, and we will refer to any estimator from this class as a Weighted Group Variable Estimator (WGVE):

$$\widehat{\boldsymbol{\beta}}_W = \sum_{q=1}^{Q_J} W_{n,q}\widehat{\boldsymbol{\beta}}_q = \sum_{q=1}^{Q_J} \begin{bmatrix} W_{n,q}^{\theta} & 0 & 0 \\ 0 & W_{n,q}^{\gamma} & 0 \\ 0 & 0 & W_{n,q}^{\delta} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\theta}}_q \\ \widehat{\gamma}_q \\ \widehat{\boldsymbol{\delta}}_q \end{bmatrix}, \tag{10}$$

where $W_{n,q}$ are matrices of (possibly stochastic) weights that sum to the identity matrix, $W_{n,q}^{\theta}$ is a $p_y \times p_y$ matrix, $W_{n,q}^{\gamma}$ is a $p_x \times p_x$ matrix, and $W_{n,q}^{\delta}$ is a $p_y p_x \times p_y p_x$ matrix. The vector $\widehat{\boldsymbol{\beta}}_q$ is the GVE as in (9), and (10) defines a class of instrumental variable estimators by the weighting matrix $\{W_{n,q}, q = 1, 2, \ldots, Q_J\}$. Below, we allow for different weighting schemes. In the case of equal weights, $\widehat{\boldsymbol{\beta}}_M = Q_J^{-1}\sum_{q=1}^{Q_J}\widehat{\boldsymbol{\beta}}_q$, which is defined as the Mean Group Variable Estimator (MGVE).

The WGVE is an extension of the method discussed in the previous section. In the first step, we obtain $\widehat{\boldsymbol{\beta}}_q$ for $q = 1, 2, \ldots, Q_J$. In the second step, we compute (10). As expected, a linear combination of a finite number of consistent and asymptotically normal estimators is consistent and asymptotically normal, as shown in Theorem 3 below.

The use of weights for linear combinations of estimators is not new (e.g., Pesaran 2006, Chen et al. 2016, Harding et al. 2020, among others). If $r = 1$ and $m_B = 1$, then $Q_J = J - m_A$, and one could set $W_{n,q} = Q_J^{-1} I_{k_x} = (J - m_A)^{-1} I_{k_x}$, and define the estimator as $\widehat{\boldsymbol{\beta}}_M = (J - m_A)^{-1} \sum_{q=1}^{J-m_A} \widehat{\boldsymbol{\beta}}_q$, which is similar to the common correlated effect mean group estimator of Pesaran (2006) (see also Brown, Schmidt & Wooldridge 2021). Pesaran's estimator averages over coefficients obtained from individual regressions and we average over coefficients obtained from different partitions. Moreover, the estimator (10) is similar to the ones investigated by Chen, Jacho-Chávez & Linton (2016) and Chalfin & McCrary (2018). For instance, Example 1 in Chen, Jacho-Chávez & Linton (2016) considers a similar instrumental variable estimator for a simultaneous equation model, and the optimal choice of weights makes a weighted instrumental variable estimator asymptotically equivalent to the classical 2SLS estimator. Our method might be seen as a generalization of the estimator proposed by Chalfin & McCrary (2018), which considers the case of $m_B = m_C = 1$ and equal weights. In addition to an estimator similar to the one defined in (9), they suggest reversing the role of $\boldsymbol{y}_{gB_q}$ and $\boldsymbol{y}_{gC_q}$, to then pool the estimates with the goal of improving efficiency.

Motivating example (cont.) There are $Q_J = \binom{9-3}{1} = 6$ sets B and there is no a priori practical guidance on selecting B. The range of GVE estimates in Figure 1 goes from -0.458 to -0.342, and the WGVE (with weights introduced in Section 3.3) is equal to -0.396.

The estimator defined in (10) estimates

$$\boldsymbol{\beta}_n = (\bar{\boldsymbol{\theta}}_n', \boldsymbol{\gamma}', \bar{\boldsymbol{\delta}}_n')', \tag{11}$$

where $\bar{\boldsymbol{\theta}}_n' = \sum_{q=1}^{Q_J} W_{n,q}^{\theta} \boldsymbol{\theta}_q$ and $\bar{\boldsymbol{\delta}}_n' = \sum_{q=1}^{Q_J} W_{n,q}^{\delta} \boldsymbol{\delta}_q$. The parameter (11) differs from the one estimated in Section 2. However, as in the case of one subset B, the use of instrumental variables leads to identification of the slope parameter $\boldsymbol{\gamma}$ and reduced form coefficients associated with the factors. This can be seen clearly in the following example:

Example 1. Consider the simplest version of model (1), $y_{gj} = \gamma x_{gj} + \lambda_g f_j + u_{gj}$, where $J = 3$ and $r = 1$. Assuming $A = \{1\}$, we have $k_x = 3$ because $p_x = p_y = 1$. There are two partitions,

$B_1 = \{2\}$ and $B_2 = \{3\}$, and thus, $Q_J = 2$, and the parameter $\boldsymbol{\beta}_n = W_{n,1}\boldsymbol{\beta}_1 + W_{n,2}\boldsymbol{\beta}_2$. For the case of equal weights, we set $W^{\theta}_{n,q} = W^{\gamma}_{n,q} = W^{\delta}_{n,q} = Q_J^{-1}$ and $W_{n,q} = Q_J^{-1} I_3 = \frac{1}{2} I_3$. Therefore,

$$\boldsymbol{\beta}_n = \frac{1}{2} I_3 \begin{bmatrix} \frac{f_1}{f_2} \\ \gamma \\ -\gamma\frac{f_1}{f_2} \end{bmatrix} + \frac{1}{2} I_3 \begin{bmatrix} \frac{f_1}{f_3} \\ \gamma \\ -\gamma\frac{f_1}{f_3} \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma \\ 0 \end{bmatrix} + \frac{1}{2}\left(\frac{1}{f_2} + \frac{1}{f_3}\right) \begin{bmatrix} f_1 \\ 0 \\ -\gamma f_1 \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma \\ 0 \end{bmatrix} + \omega \begin{bmatrix} f_1 \\ 0 \\ -\gamma f_1 \end{bmatrix},$$

where $\omega := \frac{1}{2}\sum_{q=1}^{2} \frac{1}{f_{B_q}}$. Thus, f_1 is identified up to a non-singular transformation, which requires that factors in B_1 and B_2 are bounded away from zero (as implied by Assumption B2).

It is well-known that the identification of $(\boldsymbol{\lambda}_g)$ and $(\boldsymbol{f}_j)$ requires at least r^2 restrictions (Bai and Ng, 2013, p. 19). Various restrictions have been imposed in the literature. For the case of $r = 1$, most of them amount to $\sum_{j=1}^{J} f_j^2 = 1$ (Bai and Ng, 2013, PC1 and PC2) or $\lambda_1 = 1$ (PC3). In Example 1, a PC3-type restriction is to set either $f_2 = 1$ or $f_3 = 1$. One way to think about our approach of averaging in Example 1 is that it uses the normalization

$$\omega = \frac{1}{Q_J}\sum_{q=1}^{Q_J} \frac{1}{f_{B_q}} = 1.$$

It is possible to achieve point identification of the parameters of the model if we set the first $r \times r$ block of factors equal to the identity matrix. This is illustrated in the following example:

Example 2. A suitable choice of $W_{n,q}$ in (11) is $W^{\theta}_{n,1} = I_{p_y}$, $\sum_{q=1}^{Q_J} W^{\gamma}_{n,q} = I_{p_x}$, and $W^{\delta}_{n,1} = I_{p_x p_y}$. In this case, $\boldsymbol{\beta}_n = (\boldsymbol{\theta}_1', \boldsymbol{\gamma}', \boldsymbol{\delta}_1')'$. Furthermore, under a PC3 restriction on the factors equal to $\boldsymbol{f}_{B_1} = I_r$ (Bai & Ng 2013, Heaton & Solo 2012, Heckman & Scheinkman 1987, among others), the parameter $\boldsymbol{\theta}_1$ is,

$$\boldsymbol{\theta}_1 = \begin{bmatrix} \boldsymbol{f}_{B_1}^{-1}\boldsymbol{f}_1 \\ \vdots \\ \boldsymbol{f}_{B_1}^{-1}\boldsymbol{f}_{m_A} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_1 \\ \vdots \\ \boldsymbol{f}_{m_A} \end{bmatrix}.$$

Lastly, it is important to mention that there are alternative transformations of model parameters that may be of interest. For instance, if we fix $B_q = B_1$ and q indexes the choice of instruments in the set C, the parameter of interest is $\boldsymbol{\beta}_n = (\boldsymbol{\theta}_1', \boldsymbol{\gamma}', \boldsymbol{\delta}_1')'$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_{j1}, j \in A)$ and $\boldsymbol{\theta}_{j1} = \boldsymbol{f}_{B_1}^{-1}\boldsymbol{f}_j$. The parameter can be estimated considering a weighted instrumental variable estimator (WIVE), $\widetilde{\boldsymbol{\beta}}_W = \sum_{q=1}^{|C|} W_{n,q}\widetilde{\boldsymbol{\beta}}_q$, where $\widetilde{\boldsymbol{\beta}}_q$ is similar to $\widehat{\boldsymbol{\beta}}_q$ in (9) but it uses

$$\boldsymbol{X}_{g,1} = \begin{bmatrix}\boldsymbol{I}_{m_A} \otimes \boldsymbol{y}_{gB_1}' & \boldsymbol{x}_{gA} & \boldsymbol{I}_{m_A} \otimes \boldsymbol{h}_{gB_1}'\end{bmatrix}, \text{ and } \boldsymbol{Z}_{g,q} = \begin{bmatrix}\boldsymbol{I}_{m_A} \otimes \boldsymbol{y}_{gC_q}' & \boldsymbol{x}_{gA} & \boldsymbol{I}_{m_A} \otimes \boldsymbol{h}_{gB_1}'\end{bmatrix}.$$

Motivating example (cont.) As in the case of the WGVE estimator, the implementation of the WIVE is simple. In the application considered in Figure 1, the point estimate is -0.444, which is slightly smaller than the point estimate -0.396 obtained by the WGVE estimator. Additional results are available upon request.

In the next section, we establish conditions under which the WGVE is consistent and asymptotically normal. Later in Section 3.3, we investigate the selection of optimal weights for the estimator of the parameter $\boldsymbol{\gamma}$.

3.2 Theoretical properties

Since J is fixed, so are the number of subsets Q_J, and the number of proxy and instrumental variables m_B and m_C. Then, the number of instrumental variables and the number of estimators in (10) does not diverge. This is the case most relevant in the application using administrative data presented in Section 5 and in the recent econometric literature (see Juodis & Sarafidis 2018, 2020, Norkutė et al. 2021, for examples).

To establish the large sample results of the WGVE, the weight matrix must satisfy the following condition B3. The condition allows the use of different weighting schemes to improve the performance of the GVE and it is similar to the ones employed in the literature such as Pesaran (2006) and Chen, Jacho-Chávez & Linton (2016). Assumption B3 is similar to A1 and B4 in Chen, Jacho-Chávez & Linton (2016), and the second part of the assumption is needed in the case of random weights which are well approximated by a non-random sequence.

B3. The weights $\{W_{n,q}\}_{q=1}^{Q_J}$ satisfy (i) $\sum_{q=1}^{Q_J} W_{n,q} = I_{k_x}$ and $\sup_{n\geq 1}\sum_{q=1}^{Q_J}\|W_{n,q}\| < \infty$ w.p.1. Moreover, (ii) there exist deterministic weight matrices $\{W_{n,q}^0\}_{q=1}^{Q_J}$ satisfying $\sum_{q=1}^{Q_J}\|W_{n,q} - W_{n,q}^0\| = o_p(1)$ and $\sup_{n\geq 1}\sum_{q=1}^{Q_J}\|W_{n,q}^0\| < \infty$.

Define $\mathcal{V}_n = \sum_{q=1}^{Q_J}\sum_{l=1}^{Q_J} W_{n,q}\boldsymbol{\Sigma}_{n,ql}W_{n,l}'$. The next result builds on Theorem 1:

Theorem 3. *Under conditions A1, A2, A5, and B1-B3, as $G \to \infty$, the WGVE defined in* (10) *is consistent, i.e. $\widehat{\boldsymbol{\beta}}_W - \boldsymbol{\beta}_n \xrightarrow{p} 0$, and*

$$\mathcal{V}_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_W - \boldsymbol{\beta}_n\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right).$$

Thus, a WGVE is consistent and asymptotically normal. One advantage of WGVE over any given GVE is that WGVE does not require the researcher to select a group of measurements to proxy loadings, as in equation (3). An additional advantage of WGVE over GVE is that weights can be chosen to gain efficiency. We now turn to these efficiency gains.

3.3 Efficiency gains from combining GVE estimators

To discuss efficiency gains, we focus on the estimation of $\boldsymbol{\gamma}$, because this parameter is invariant to q. Given a choice of $p_x \times p_x$ weight matrices $W_{n,q}^{\gamma}$ for $q \in \{1, \cdots, Q_J\}$, such that $\sum_{q=1}^{Q_J} W_{n,q}^{\gamma} = I_{p_x}$, define the WGVE for $\boldsymbol{\gamma}$ as

$$\widehat{\boldsymbol{\gamma}}_W = \sum_{q=1}^{Q_J} W_{n,q}^{\gamma}\widehat{\boldsymbol{\gamma}}_q. \tag{12}$$

It follows from our previous result that $(\mathcal{V}_n^{\gamma})^{-1/2}\sqrt{n}\,(\widehat{\boldsymbol{\gamma}}_W - \boldsymbol{\gamma})$ is asymptotically Gaussian, where

$$\mathcal{V}_n^{\gamma} = W_n^{\gamma}\boldsymbol{\Xi}_n(W_n^{\gamma})', \tag{13}$$

the weight matrix $W_n^{\gamma} = \left(W_{n,1}^{\gamma}, W_{n,2}^{\gamma}, \ldots, W_{n,Q_J}^{\gamma}\right)$, and $\boldsymbol{\Xi}_n$ is the covariance matrix of the joint distribution of $\underline{\widehat{\boldsymbol{\gamma}}} = (\widehat{\boldsymbol{\gamma}}_1', \widehat{\boldsymbol{\gamma}}_2', \cdots, \widehat{\boldsymbol{\gamma}}_{Q_J}')'$ obtained from Theorem 2.

Within the class of estimators (12), we will call as *optimal weights* the weight matrices that minimize (13) with respect to the weights W_n^{γ} subject to $\sum_{q=1}^{Q_J} W_{n,q}^{\gamma} = I_{p_x}$. We are particularly interested in comparing the variance of the resulting estimator to that of any given GVE.

To state the following result, it is convenient to write $\underline{\boldsymbol{\gamma}} = R\boldsymbol{\gamma}$, where $R = (\iota_{Q_J} \otimes I_{p_x})$. Optimal weights can be obtained from the minimum distance estimation problem with objective function:

$$n\left(\widehat{\underline{\boldsymbol{\gamma}}} - \underline{\boldsymbol{\gamma}}\right)' \boldsymbol{\Xi}_n^{-1} \left(\widehat{\underline{\boldsymbol{\gamma}}} - \underline{\boldsymbol{\gamma}}\right) = n\left(\widehat{\underline{\boldsymbol{\gamma}}} - R\boldsymbol{\gamma}\right)' \boldsymbol{\Xi}_n^{-1} \left(\widehat{\underline{\boldsymbol{\gamma}}} - R\boldsymbol{\gamma}\right).$$

From standard results on optimal minimum distance estimation, we know that variance-minimizing weights are given by

$$W_{n,q}^{\gamma *} = \left[R'\boldsymbol{\Xi}_n^{-1}R\right]^{-1}\left[R'\boldsymbol{\Xi}_n^{-1}\right]_q, \tag{14}$$

where $[R'\boldsymbol{\Xi}_n^{-1}]_q$ is the q-th block of the matrix $R'\boldsymbol{\Xi}_n^{-1}$. It follows that the estimator $\widehat{\boldsymbol{\gamma}}_W^* = \sum_{q=1}^{Q_J} W_{n,q}^{\gamma *}\widehat{\boldsymbol{\gamma}}_q$ has covariance matrix,

$$\mathcal{V}_n^{\gamma *} = \sum_{q=1}^{Q_J}\sum_{l=1}^{Q_J} W_{n,q}^{\gamma *}\boldsymbol{\Xi}_{n,ql}(W_{n,l}^{\gamma *})' = \left[R'\boldsymbol{\Xi}_n^{-1}R\right]^{-1}. \tag{15}$$

The following result demonstrates that the WGVE of $\boldsymbol{\gamma}$ can improve on GVE by averaging over optimally chosen weights.

Theorem 4. *Under the Assumptions of Theorem 3, and if $\boldsymbol{\Xi}_n$ is nonsingular, then the WGVE defined in equation* (12) *with weights $W_{n,q}^{\gamma *}$ is at least as efficient as any GVE estimator $\widehat{\boldsymbol{\gamma}}_q$.*

The optimal weight and variance matrices can be estimated following closely Hansen & Lee (2019) and Section 6 in Chen, Jacho-Chávez & Linton (2016). To elaborate on the implications of Theorem 4, we now present an example that illustrates the potential gains of combining different GVEs.

Example 3. Consider the model in Example 1. The covariance matrix of $\widehat{\boldsymbol{\beta}}_W = \sum_{q=1}^{2} W_{n,q}\widehat{\boldsymbol{\beta}}_q$ is $\mathcal{V}_n = \sum_{q=1}^{2}\sum_{l=1}^{2} W_{n,q}\boldsymbol{\Sigma}_{n,ql}W_{n,l}'$. We set $W_{n,1}^{\theta} = W_{n,1}^{\delta} = 1$ and we assume $\gamma = 0$ and $\sigma_\lambda^2 = 1$ to simplify the expressions. Because $\boldsymbol{\Xi}_n$ in equation (14) is equal to

$$\boldsymbol{\Xi}_n = \frac{\sigma_u^2}{\sigma_x^2}\begin{bmatrix} \frac{f_1^2+f_2^2}{f_2^2} & 1 \\ 1 & \frac{f_1^2+f_3^2}{f_3^2} \end{bmatrix},$$

we obtain $W_{n,1}^{\gamma *} = f_2^2/(f_2^2+f_3^2) =: \tau_2$ and $W_{n,2}^{\gamma *} = f_3^2/(f_2^2+f_3^2) =: \tau_3$. Then, letting $\theta_1 = f_1/f_2$, $\theta_2 = f_1/f_3$, and $\kappa = (f_3^2+\sigma_u^2)/(f_2 f_3)^2$, the optimal covariance matrix of $\widehat{\boldsymbol{\beta}}_W^*$ is

$$\mathcal{V}_n^* = \frac{\sigma_u^2}{\sigma_x^2}\begin{bmatrix} (1+\theta_1^2)\kappa\sigma_x^2 & 0 & 0 \\ 0 & (1+\tau_2\theta_1^2) & 0 \\ 0 & 0 & (1+\theta_1^2) \end{bmatrix}.$$

The variance of $\boldsymbol{\gamma}_W^*$ can be compared with the variance of $\widehat{\gamma}_1$. Because $0 < \tau_2 < 1$ by Assumption B1, we obtain

$$[\boldsymbol{V}_{n,1}]_{2,2} = \text{var}(\widehat{\gamma}_1) = (1+\theta_1^2)\frac{\sigma_u^2}{\sigma_x^2} > \left(1+\tau_2\theta_1^2\right)\frac{\sigma_u^2}{\sigma_x^2} = \text{var}(\widehat{\gamma}_W^*) = [\mathcal{V}_n^*]_{2,2}.$$

Similarly, if we consider the GVE $\hat{\gamma}_2$ obtained from the second partition:

$$[\boldsymbol{V}_{n,2}]_{2,2} = \text{var}(\widehat{\gamma}_2) = (1+\theta_2^2)\frac{\sigma_u^2}{\sigma_x^2} > \left(1+\tau_3\theta_2^2\right)\frac{\sigma_u^2}{\sigma_x^2} = \text{var}(\widehat{\gamma}_W^*) = [\mathcal{V}_n^*]_{2,2}.$$

Therefore, the estimator of $\widehat{\gamma}_W^*$ is optimal in the sense of achieving the smallest variance in the class of GVE estimators, $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$.

For the case $\sigma_u^2 = \sigma_x^2 = 1$ and $f_1 = f_3 = 1$, Figure 2 displays the variance of $\widehat{\gamma}$ as a function of f_2 for the two GVEs, the MGVE, and the WGVE. Recall that MGVE uses equal weights and the WGVE uses optimal weights. The MGVE outperforms each GVE over a large range of values of f_2 but not everywhere. The WGVE with optimal weights dominates the other three estimators. The sole exception is $f_2 = 1$, where the variance of MGVE is equal to the variance of WGVE. This

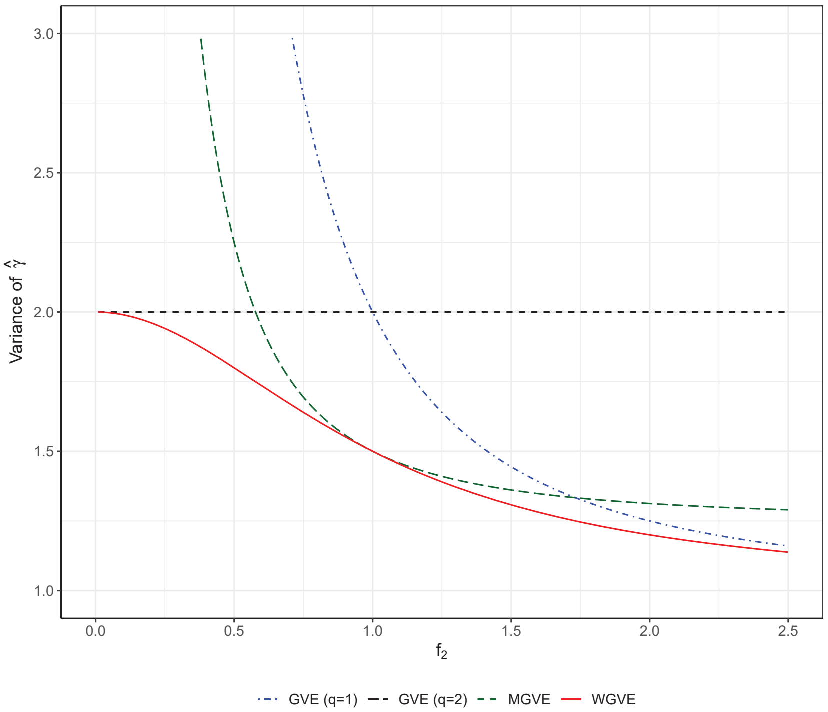

Figure 2: *Variance of GVE, MGVE, and WGVE estimators in Example 3. GVE denotes group variable estimator, MGVE denotes mean group variable estimator and WGVE is the weighted group variable estimator with optimal weights.*

is because, in this particular case,

$$\text{var}(\widehat{\gamma}_M) = \left(1 + \frac{f_2^2 + 1}{4f_2^2}\right) \geq \left(1 + \frac{1}{f_2^2 + 1}\right) = \text{var}(\widehat{\gamma}_W^*).$$

4 Simulation experiments

In this section, we investigate the finite sample performance of the proposed approaches in comparison to existing methods. We generate observations based on the following model used in Pesaran (2006):

$$y_{gj} = \gamma_0 + \gamma_1 x_{gj,1} + \gamma_2 x_{gj,2} + \lambda_{g,1} f_{j,1} + u_{gj}, \tag{16}$$

where $x_{gj,s} = a_j\lambda_{g,1} + b_j f_{j,1} + c_j\lambda_{g,1}f_{j,1} + \varepsilon_{gj,s}$, and $f_{j,1} = \rho f_{j-1,1} + \eta_j$ for $g = 1, 2, \ldots, G$ and $j = -S+1, \ldots, 0, 1, \ldots, J$. We assume that $J = 10$ and the researcher is interested in $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)'$ and $(f_{j,1}, j \in A)$ where $A = \{5, 6, \ldots, 10\}$. The error term in equation (16) is distributed as either Gaussian or t-student with 3 degrees of freedom (t_3). The loading $\lambda_{g,1}$ is drawn as an independent observation from a uniform distribution ranging from 0.5 to 3.5, and η_j is an i.i.d. random variable distributed as uniform $\mathcal{U}[0, 1]$. The error term $(\varepsilon_{gj,1}, \varepsilon_{gj,2})' \sim \mathcal{N}(0, \boldsymbol{I})$. Moreover, we set the parameters of the model to generate an endogenous variable, $x_{jg,1}$, and an exogenous variable, $x_{gj,2}$. The parameters are $\gamma_1 = \gamma_2 = a_1 = 1$, $b_1 = 2$, $c_1 = 0.5$, $\gamma_0 = a_2 = b_2 = c_2 = 0$, and $\rho = 0.8$. Lastly, we set $S = 50$ to minimize the effects of the initial values on the outcome, $f_{-49,1} = 1$.

We focus our investigation on the estimation of the slope coefficient γ_1 in equation (16). Section 3 in the online appendix reports results for the estimation of reduced form parameters. Table 1 shows the bias and root mean squared error (RMSE) of different estimators under sample sizes $G = \{200, 500, 1000\}$. The upper panel shows results for $u_{gj} \sim \mathcal{N}(0, 1)$, and the lower panel shows results for $u_{gj} \sim t_3$. We compare our estimators to existing approaches such as the estimator for an interactive effects model (IEE) proposed by Bai (2009) which uses PCA, the 2SLS estimator that uses internal instrumental variables, and a 2SLS estimator that uses the LASSO (LAS) estimator proposed by Belloni et al. (2012) in the first stage. The LAS estimator employs internal instruments and it is implemented using the R package `hdm` (Chernozhukov et al. 2016).

Table 1 also shows the bias and RMSE of the new estimators. GVE denotes the group variable estimator as in (7), MGVE denotes the mean weighted group variable estimator, and WGVE denotes the optimally weighted group variable estimator defined in (10). The GVE estimator is obtained considering $B = \{1\}$ and the MGVE uses $Q_J^{-1}I_{k_x}$ as weights. Because $A = \{5, 6, \ldots, 10\}$ in all the simulations, the MGVE and WGVE estimators are obtained based on $Q_J = \binom{J-m_A}{m_B} = \binom{10-6}{1} = 4$ estimators. Finally, the optimal weights for WGVE are estimated following a simple two step procedure. First, we estimate $\boldsymbol{\beta}_q$. Then, using residuals $\widehat{\boldsymbol{e}}_{g,q} = \boldsymbol{y}_g - \boldsymbol{X}_{g,q}\widehat{\boldsymbol{\beta}}_q$ for each g and

G		IEE	2SLS	LAS	GVE	MGVE	WGVE
Model with Gaussian Errors							
200	Bias	0.004	0.009	0.007	0.009	0.010	0.021
	RMSE	0.029	0.058	0.058	0.058	0.040	0.044
	Standard Error	0.025	0.046	0.047	0.056	0.039	0.038
	Coverage Probability	0.880	0.820	0.829	0.886	0.891	0.826
500	Bias	0.002	0.002	0.001	0.002	0.001	0.006
	RMSE	0.018	0.035	0.035	0.035	0.024	0.025
	Standard Error	0.016	0.029	0.030	0.035	0.025	0.025
	Coverage Probability	0.885	0.839	0.854	0.920	0.908	0.888
1000	Bias	0.001	0.000	0.000	0.000	0.000	0.003
	RMSE	0.014	0.026	0.026	0.026	0.019	0.019
	Standard Error	0.011	0.021	0.021	0.025	0.018	0.018
	Coverage Probability	0.853	0.815	0.823	0.894	0.892	0.884
Model with t_3 Errors							
200	Bias	0.115	0.036	0.027	0.036	0.036	0.067
	RMSE	0.175	0.122	0.190	0.122	0.088	0.104
	Standard Error	0.041	0.082	0.099	0.105	0.074	0.065
	Coverage Probability	0.467	0.757	0.781	0.841	0.823	0.692
500	Bias	0.086	0.014	0.010	0.014	0.015	0.030
	RMSE	0.143	0.076	0.075	0.076	0.051	0.058
	Standard Error	0.026	0.053	0.054	0.068	0.048	0.045
	Coverage Probability	0.524	0.770	0.794	0.876	0.868	0.804
1000	Bias	0.079	0.008	0.006	0.008	0.006	0.015
	RMSE	0.134	0.052	0.052	0.052	0.036	0.038
	Standard Error	0.019	0.038	0.039	0.049	0.035	0.034
	Coverage Probability	0.558	0.789	0.812	0.899	0.891	0.855

Table 1: *Finite sample performance of several estimators for γ_1. IEE refers to the interactive effects estimator, 2SLS denotes the instrumental variable estimator, LAS denotes a 2SLS estimator that uses LASSO in the first stage, GVE is the group variable estimator, MGVE the mean group variable estimator, and WGVE is the optimally weighted group variable estimator.*

q, we obtain $\widehat{\boldsymbol{\Xi}}_n$. It is straightforward to estimate (14) as $\widehat{W}_{n,q}^{\gamma *} = \left[R'\widehat{\boldsymbol{\Xi}}_n^{-1}R\right]^{-1}\left[R'\widehat{\boldsymbol{\Xi}}_n^{-1}\right]_q$, where $R = (\iota_{Q_J} \otimes I_{p_x}) = (\iota_4 \otimes I_3)$ in all simulations.

As expected, the results in Table 1 demonstrate that the IEE estimator offers excellent performance in models with Gaussian errors, but the estimator can be biased when $u_{gj} \sim t_3$. As expected, the performance of 2SLS and LAS estimators is similar, because the number of instruments is not large. The performance of the proposed GVEs is excellent and they offer the smallest RMSE in the class of instrumental variables estimators for a linear panel data model. For the model with t_3 errors, the GVEs offer better performance in terms of RMSE than IEE, 2SLS, and LAS.

Table 1 also reports standard errors and coverage probabilities for a nominal 90% confidence interval for a slope parameter γ_1. The coverage is constructed based on the standard error of the estimator. GVE offers larger standard errors, in general, than the 2SLS estimator, because it accounts for the within-cluster correlation in the reduced form equation. Moreover, also as expected, WGVE offer significant gains in precision relative to GVE. Overall, the new proposed approaches offer excellent performance when $G > 200$, and the coverage probability is close to the target 90 percent under different distributional assumptions, in contrast to the IEE.

5 Educational Opportunity in the US

Educational attainment depends on numerous observable economic factors such as school resources and parental income, but also on unobservable or difficult to measure factors such as teacher quality and student motivation. We are interested in situations where student ability and teacher quality interact, and wish to understand the impact of the latent educational quality and of the distribution of student ability. Therefore, applications in this area present themselves as natural test beds for our proposed methods. In this section, we illustrate the use of the proposed estimator and investigate how the distribution of school district heterogeneity changes across states. We find significant geographic variability of educational opportunity across the US.

5.1 Data

We investigate educational performance at the school district level in the US using the Stanford Education Data Archive (SEDA), which is commonly used to evaluate educational policies and practices (Reardon 2019, Fahle et al. 2021). SEDA provides nationally comparable scores in mathematics and reading for over 11,000 school districts in the majority of US states, from standardized tests administered from 3rd grade through 8th grade. SEDA also includes a range of variables for each grade and year separately, including the percentage of non-white students by grade, the percentage of students receiving free or reduced-price lunch, the percentage of adult population (over 25 years of age) with a college degree or higher, and the number of students in grade. The non-white students include the percentage of African-American, Asian-American, Native-American and Hispanic students. For the purpose of this empirical illustration, we will use data from the year 2018. Our final dataset includes average academic achievement from 2,033 school districts measured by standardized test scores in mathematics and reading.

5.2 Model

We model the district level test scores for middle school students using the following equation, which also accounts for the impact of latent school-district and grade-level heterogeneity on educational attainment using a one-factor specification:

$$y_{gj} = \boldsymbol{x}'_{gj}\boldsymbol{\gamma} + \lambda_g f_j + u_{gj}. \tag{17}$$

In our model, y_{gj} is the average normalized test score in district g in grade j and $\boldsymbol{x}_{gj}$ is a vector of control variables. Here, λ_g is a district loading that might be associated with district educational attainment, and the grade-level factor f_j can be interpreted as measuring educational attainment by grade j.

By including factors and loadings, we can account for unobserved grade and district character-

	Mathematics				Reading			
	OLS	IEE	GVE	WGVE	OLS	IEE	GVE	WGVE
Percent minority in class	-0.541	-0.495	-0.429	-0.396	-0.548	-0.504	-0.467	-0.435
	(0.010)	(0.010)	(0.031)	(0.022)	(0.009)	(0.010)	(0.024)	(0.021)
Percent free or reduced lunch in class	-0.281	-0.441	-0.557	-0.530	-0.237	-0.374	-0.454	-0.462
	(0.013)	(0.014)	(0.051)	(0.029)	(0.012)	(0.013)	(0.036)	(0.028)
Percent high education in district	1.162	0.814	0.865	0.916	1.129	0.834	0.957	0.988
	(0.028)	(0.029)	(0.073)	(0.050)	(0.025)	(0.027)	(0.056)	(0.047)
Logarithm of average class enrollment	0.010	0.017	0.002	0.008	0.015	0.021	0.005	0.009
	(0.002)	(0.002)	(0.004)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)
Grade 6 effect		-0.057	-0.229	-0.027		-0.067	-0.119	-0.119
			(0.217)	(0.026)			(0.163)	(0.053)
Grade 7 effect		-0.049	-0.884	-0.011		-0.028	-0.609	-0.040
			(0.275)	(0.049)			(0.190)	(0.052)
Grade 8 effect		0.155	0.510	0.157		0.137	0.455	0.147
			(0.260)	(0.045)			(0.184)	(0.048)

Table 2: *Estimated slopes and reduced form coefficients for the factor by grade. IEE denotes Bai's (2009) estimator, GVE denotes the group variable estimator, and WGVE is the optimally weighted group variable estimator. Standard errors are in parenthesis.*

istics. This is important, as the set of observed covariates is limited in this data set. We can think of district educational attainment as capturing the different district-specific but grade-invariant factors such as additional demographic or other drivers of test scores. In contrast, grade-level educational attainment is assumed to be constant across district but differing across grades and may capture unobserved factors such as the inherent variation in difficulty of the material in different grades. Note that the term $\lambda_g f_j$ represents the interaction between average educational performance in district g and average educational attainment in grade j. The interaction between these latent terms allows us to account for the fact that grade specific challenges faced by students are being addressed very differently in high and low performing districts which may exacerbate differences in test scores across districts.

We estimate equation (17) separately by subject for grades $j \in A = \{4, 5, 6\}$ (i.e., grades 6, 7 and 8) considering $6,099$ observations. The results are presented in Table 2. Districts are assumed to be independent, and the error term u_{gj} is assumed to be conditionally independent across grades and subjects. For the consistency of the GVE and WGVE estimators, the errors can be weakly dependent within district at the middle school level, but errors in the middle school and elementary school equations are assumed to be conditionally independent.

5.3 Empirical results

We employ several estimators discussed in this paper. IEE denotes Bai's (2009) estimator, GVE denotes the group variable estimator, and WGVE is the optimally weighted group variable estimator. OLS does not use IVs and does not cluster the standard errors and GVE uses IVs and does cluster the standard errors as in Theorem 1. We find that district-level educational performance is lower in districts with a higher percentage of minority students and also lower income students who rely on free or reduced lunch programs. At the same time, district-level educational performance is higher in districts with a more educated population. Average class size has a very small positive effect by comparison. The impact of these district-level variables is very similar for both subjects (Mathematics and Reading). The magnitudes of the slope coefficients are slightly higher for the GVE and WGVE estimators compared to the IEE estimator. As anticipated, OLS tends to produce noticeably different estimates than the other estimators that also account for latent heterogeneity.

When computing the GVE estimates, the practitioner can get different results depending on the chosen subset B_q. In Figure 3, we compare the impact of using different grades and/or subjects to proxy the latent district level loading in (3) on the estimated coefficients on the demographic explanatory variables measuring poverty, parental education, and class size. The figure presents results for mathematics only, because the evidence for reading is similar. While the choices of B_q do not change the qualitative interpretation of the results, the range of estimated coefficients varies substantially and can exceed 20%. Because the WGVE estimator uses optimal weights, we continue to see improvements on the precision of the proposed method in relation to the GVE estimator, see Figure 1 and Table 2.

Lastly, we show how the approach can be used to estimate district loadings, which are naturally

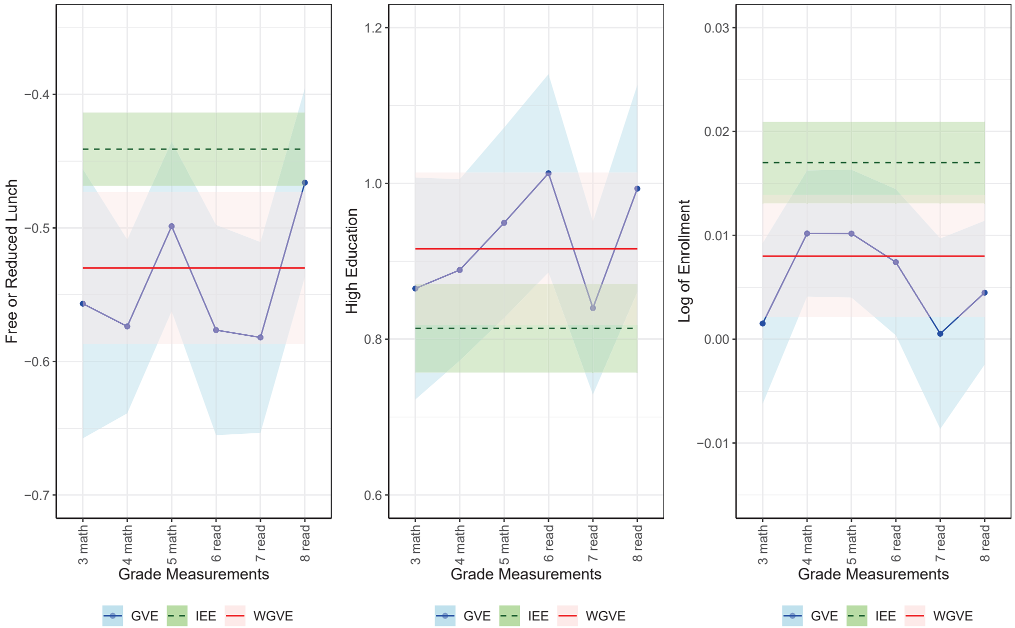

Figure 3: *Results for the slope parameters in a model for mathematics. The shaded areas are 95% point-wise confidence intervals. Third grade math is denoted by '3 math', fourth grade math by '4 math', etc. IEE denotes Bai's (2009) estimator, GVE denotes the group variable estimator, and WGVE is the optimally weighted group variable estimator.*

of interest from a policy point of view. Using the estimates in Table 2, we obtain residuals $\hat{e}_{gj} = y_{gj} - \boldsymbol{x}'_{gj}\widehat{\boldsymbol{\gamma}}^*_W$. We then apply the method of principal components with a PC3-type restriction to identify $(\lambda_1, \lambda_2, \ldots, \lambda_G)$. In Figure 4, we investigate the loadings capturing the district level heterogeneity further by displaying the distributions by state, subject to the caveats that the consistency of the estimator requires large J, and the data available to us cannot capture the full national distribution due to a limited number of observations in some states. Nevertheless, it is particularly striking how wide the range of the estimated loadings is by state. In mathematics, the worst performing school districts as measured by the magnitude of the loading λ_g are Grenada (MS), Marengo (AL), Montgomery (AL), Glynn (GA) and Bryan (OK).

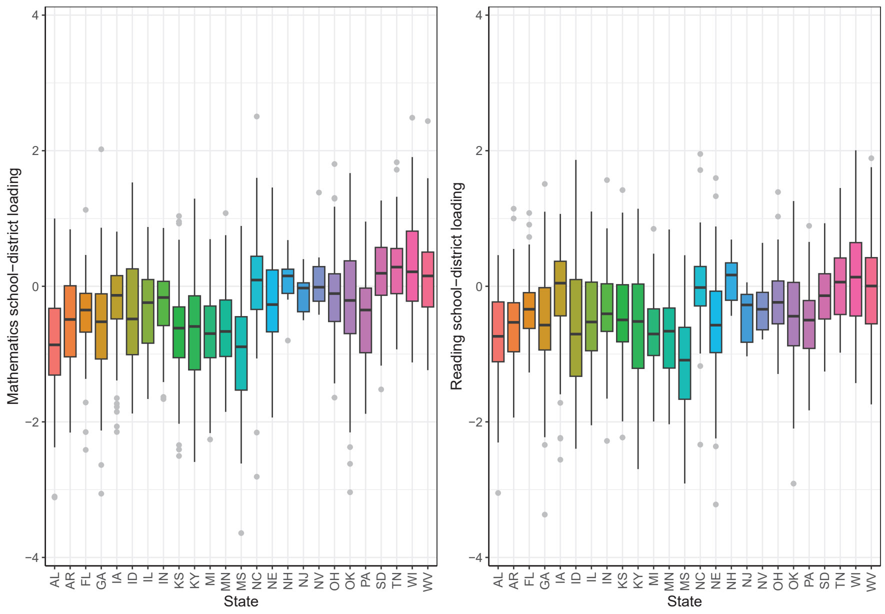

Figure 4: *Geographical disparities and district educational attainment.*

6 Conclusions

In this paper, we present a novel solution using internally generated instruments for the estimation of linear factor models. We propose a new class of estimators and establish large sample results. We demonstrate that there are theoretical and practical advantages of creating weighted linear combinations of instrumental variables estimators, which can lead to efficiency improvements. While the proposed approach is computationally intensive and identification relies on correctly specifying the dependence of the error term across partitions, it nevertheless leads to a simple approach to estimating linear models. Further research may involve relaxing the identification assumptions to more general cases and to the extension of approximate factor models.

References

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

Ando, T. & Bai, J. (2016), 'Panel data models with grouped factor structure under unknown group membership', *Journal of Applied Econometrics* **31**(1), 163–191.

Ando, T. & Bai, J. (2017), 'Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures', *Journal of the American Statistical Association* **112**(519), 1182–1198.

Attanasio, O., Meghir, C. & Nix, E. (2020), 'Human Capital Development and Parental Investment in India', *The Review of Economic Studies* **87**(6), 2511–2541.

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77(4)**, 1229–1279.

Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Bai, J. & Ng, S. (2010), 'Instrumental variable estimation in a data rich environment', *Econometric Theory* **26**(6), 1577–1606.

Bai, J. & Ng, S. (2013), 'Principal components estimation and identification of static factors', *Journal of Econometrics* **176**(1), 18 – 29.

Bai, J. & Wang, P. (2016), 'Econometric analysis of large factor models', *Annual Review of Economics* **8**(1), 53–80.

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012), 'Sparse models and methods for optimal instruments with an application to eminent domain', *Econometrica* **80**(6), 2369–2429.

Bernanke, B. S., Boivin, J. & Eliasz, P. (2005), 'Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach', *The Quarterly Journal of Economics* **120**(1), 387–422.

Borghans, L., Duckworth, A. L., Heckman, J. J. & Weel, B. t. (2008), 'The economics and psychology of personality traits', *Journal of Human Resources* **43**(4), 972–1059.

Brown, N. L., Schmidt, P. & Wooldridge, J. M. (2021), ‘Simple alternatives to the common correlated effects model’, arXiv:2112.01486 [econ.EM].

Chalfin, A. & McCrary, J. (2018), ‘Are U.S. Cities Underpoliced? Theory and Evidence’, *The Review of Economics and Statistics* **100**(1), 167–186.

Chen, X., Jacho-Chávez, D. T. & Linton, O. (2016), ‘Averaging of an increasing number of moment condition estimators’, *Econometric Theory* **32**(1), 30–70.

Chernozhukov, V., Hansen, C. & Spindler, M. (2016), ‘High-dimensional metrics in r’, *arXiv preprint arXiv:1603.01700* .

Chudik, A. & Pesaran, M. H. (2015), ‘Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors’, *Journal of Econometrics* **188**(2), 393 – 420.

Cunha, F. & Heckman, J. (2007), ‘The technology of skill formation’, *American Economic Review* **97**(2), 31–47.

Cunha, F. & Heckman, J. J. (2008), ‘Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation’, *Journal of Human Resources* **43**(4), 738–782.

Cunha, F., Nielsen, E. & Williams, B. (2021), ‘The econometrics of early childhood human capital and investments’, *Annual Review of Economics* **13**, 487–513.

Del Bono, E., Kinsler, J. & Pavan, R. (2020), ‘A note on the importance of normalizations in dynamic latent factor models of skill formation’, IZA DP No. 13714.

Fahle, E. M., Chavez, B., Kalogrides, D., Shear, B. R., Reardon, S. F. & Ho, A. D. (2021), Stanford education data archive 4.0, Technical report, Stanford University.

Freyberger, J. (2021), ‘Normalizations and misspecification in skill formation models’.
URL: *https://arxiv.org/abs/2104.00473*

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

Hansen, B. E. & Lee, S. (2019), 'Asymptotic theory for clustered samples', *Journal of Econometrics* **210**(2), 268–290.

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

Harding, M. & Lamarche, C. (2014), 'Estimating and testing a quantile regression model with interactive effects', *Journal of Econometrics* **178**, 101–113.

Harding, M., Lamarche, C. & Pesaran, M. H. (2020), 'Common correlated effects estimation of heterogeneous dynamic panel quantile regression models', *Journal of Applied Econometrics* **35**(3), 294–314.

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

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American Economic Review* **103**(6), 2052–2086.

Heckman, J. & Scheinkman, J. (1987), 'The Importance of Bundling in a Gorman-Lancaster Model of Earnings', *The Review of Economic Studies* **54**(2), 243–255.

Holtz-Eakin, D., Newey, W. & Rosen, H. S. (1988), 'Estimating vector autoregressions with panel data', *Econometrica* **56**(6), 1371–1395.

Juodis, A. & Sarafidis, V. (2018), 'Fixed t dynamic panel data estimators with multifactor errors', *Econometric Reviews* **37**(8), 893–929.

Juodis, A. & Sarafidis, V. (2020), ‘A linear estimator for factor-augmented fixed-t panels with endogenous regressors’, *Journal of Business & Economic Statistics* **0**(0), 1–15.

Juodis, A. & Sarafidis, V. (2022), ‘An incidental parameters free inference approach for panels with common shocks’, *Journal of Econometrics* **229**(1), 19–54.

Kim, D. & Oka, T. (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.

Moon, H. R. & Weidner, M. (2015), ‘Linear regression for panel with unknown number of factors as interactive fixed effects’, *Econometrica* **83**(4), 1543–1579.

Moon, H. R. & Weidner, M. (2017), ‘Dynamic linear panel regression models with interactive effects’, *Econometric Theory* pp. 158–195.

Norkutė, M., Sarafidis, V., Yamagata, T. & Cui, G. (2021), ‘Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure’, *Journal of Econometrics* **220**(2), 416–446.

Onatski, A. (2010), ‘Determining the number of factors from empirical distribution of eigenvalues’, *The Review of Economics and Statistics* **92**(4), 1004–1016.

Pesaran, M. H. (2006), ‘Estimation and inference in large heterogeneous panels with a multifactor error structure’, *Econometrica* **74(4)**, 967–1012.

Reardon, S. F. (2019), ‘Educational opportunity in early and middle childhood’, *RSF: The Russell Sage Foundation Journal of the Social Sciences* **5**(2), 40–68.

Robertson, D. & Sarafidis, V. (2015), ‘Iv estimation of panels with factor residuals’, *Journal of Econometrics* **185**(2), 526 – 541.

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

Stock, J. H. & Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**(460), 1167–1179.

Trapani, L. (2018), 'A randomized sequential procedure to determine the number of factors', *Journal of the American Statistical Association* **113**(523), 1341–1349.