Title

Essays in Econometrics: Causal Inference With Large Panel Data

Permalink

https://escholarship.org/uc/item/2wn079r5

Author

Banafti, Saman

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Essays in Econometrics: Causal Inference With Large Panel Data

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Saman Banafti

June 2022

Dissertation Committee:

Dr. Tae-Hwy Lee, Chairperson
Dr. Gloria Gonzalez-Rivera
Dr. Jean Helwege
Dr. Ruoyao Shi
Dr. Aman Ullah

The Dissertation of Saman Banafti is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, Tae-Hwy Lee, for the constant guidance and support throughout this journey. I learned a great deal from you; I am very fortunate to call you my advisor. I also owe a debt of gratitude to Richard Arnott, Michael Bates, Gloria-Gonzalez Rivera, Jang-Ting Guo, Jean Helwege, Ruoyao Shi, Aman Ullah, Gary Kuzas and other economics faculty and staff for their inspiration and guidance throughout the years. Lastly, thank you to my colleagues, I will miss our conversations together in Sproul Hall and I will never forget the bond we forged suffering through the first couple of years of the program. I can not thank you all enough for the impact you have impressed upon me.

To my amazing parents, sister, fiancé, family and friends for their unwavering patience, support and unconditional love. I would not have been here were it not for them; I am eternally indebted to them. I dedicate this dissertation to them.

ABSTRACT OF THE DISSERTATION

Essays in Econometrics: Causal Inference With Large Panel Data

by

Saman Banafti

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2022
Dr. Tae-Hwy Lee, Chairperson

This dissertation is concerned with econometric theory and its applications. More specifically, the research issue of interest is the classic topic of estimation and inference in econometric models where researchers wish to investigate causal relationships in the absence of a randomized control trial. To overcome the problem of confounding effects we propose extensions of the recent Granular Instrumental Variables (GIV) methodology.

In chapter 1 we provide a high-level introduction and motivate the topic in greater detail. In chapter 2, we extend the GIV methodology by relaxing several strong assumptions imposed on the error term and factor loadings and we further allow for asymptotic regimes where both the cross-sectional dimension and time series dimension diverge jointly to infinity. Additionally, we fully exploit the structure of the model and overidentify the parameters of interest. We illustrate our contributions with an empirical application to the global crude oil market.

In the 3rd chapter, the GIV methodology is further developed to accommodate large dynamic panels with unit specific endogenous variates, which require unit-specific GIVs. We develop a split-panel jackknife (SPJ) GMM-PCA iterative procedure to estimate the structural parameters of interest. Overidentification tests can be carried out to test model validity. We illustrate the SPJ GMM-PCA procedure in two applications: (1) estimation of demand for new automobiles and (2) estimation of the determinants of banks' capital adequacy ratios.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation is concerned with econometric theory and its applications. More specifically, the research issue of interest is the classic topic of estimation and inference in econometric models where researchers wish to investigate causal relationships in the absence of a randomized control trial. Investigation of causal relationships is conducted through estimation and inference on a range of causal model parameters. One such example of causal parameters of interest in economics is elasticities.

Many researchers in economics do not have the resources to carry out a randomized experiment. As such, we are left with economic theory, econometric theory and observational data sets, such as survey data. Unfortunately, estimating causal relationships with survey data is not readily feasible and many times it is never feasible. Why? Due to unobserved variables confounding observed variables. All econometric models entail an observed dependent variable, a set of observed independent variables and finally a set of unobserved variables. The reality of the matter is that unobserved variables are correlated with our observed variables. Put differently, since the unobserved variables confound the variables we do observe, any subsequent claim that our independent variables are causally related to our dependent variables is necessarily false, precisely due to confounding.

Typically, to try and break the problem of confounding effects one must find a so-called *instrumental* variable to provide an alternative identification mechanism for the causal parameters of interest. The identification relies on two conditions: variation induced by the instrument on the confounded independent variable; and another obvious condition that should be met is that the instrument itself must not be confounded! In principle, when these two conditions are met, instrumental variables estimation solve the problem of confounding effects of unobserved variables. But in practice, finding an instrument to satisfy both of the conditions above is no easy task. Moreover, many clever instruments are quite controversial within the economics literature regarding whether they satisfy the criteria or not. It would be quite nice to be able to systematically form instrumental variables without having to come up with one in heroic fashion. Fortunately, a recent procedure developed by [1] (hereafter GK) aims to solve that problem. GK remarkably invent a systematic way to construct instruments from observational datasets. Their contribution, most importantly, eliminates the need to find an instrument. Note that there are other methodologies which seek to eliminate the need to find an instrument, e.g., [2], [3], [4], [5], [6] and have been well studied in the literature.

What is GKs instrument? In short, it is a size-weighted average of idiosyncratic shocks. An example should make it clearer. Suppose a researcher has a monthly panel (longitudinal) dataset, consisting of the same 20 countries over a 30 year period. For simplicity, suppose the independent variable is the price of oil and the dependent variable is demand for oil. The price of oil is correlated with unobserved aggregate demand shocks, such as a stock market crash. Hence, aggregate shocks to demand are confounding our analysis here where we would like to actually infer the causal relationship between price and demand. However, if we consider idiosyncratic shocks to the individual countries, they are, presumably by definition, uncorrelated with aggregate demand shocks. Moreover, in this setting, idiosyncratic shocks to Saudi Arabia, Iran, Iraq, and so forth, do actually induce variation on the price of oil. We can conclude that idiosyncratic shocks can serve as valid instrumental variables as they satisfy the conditions stated earlier!

But idiosyncratic shocks are also unobserved? However, GK lay out a formulaic approach to extract them from the data. As their instrument is a weighted average of these individual shocks, they call it granular instrumental variables (GIV hereafter) as it is composed of the *incompressible grains* of the economy, see [7] for more details on this notion albeit in a macroeconomic model and not in an econometric framework. See also [8].

The econometric theory for classic instrumental variables (IVs) estimation is well established. However, with GIVs in progress there are many questions that come to mind. Do GIVs share the same statistical properties as IVs? The GIV extraction procedure initially involves estimation of multiple nuisance parameters via principal components analysis, and the GIV itself is a generated instrument. Considering this, the statistical properties of GIV should generally be different from the properties of IVs. These are some of the issues that are addressed in this dissertation.

Scholars within the field of economics, government officials and policy makers will find this work interesting and useful. Consider the example of elasticities and assume we have an accurately measured estimate for an important elasticity, such as the elasticity of demand for oil or for water. In principle, this means we know how consumer's demand will adjust to unexpected changes in the price of oil or water (say due to a geopolitical event in the case of oil or due to a drought in the case of water). These important examples, transcend the field of economics and academia in general.

With this high-level motivation in place, we turn to a brief summary of the specific content of the subsequent chapters:

As alluded to earlier, the GIV methodology exploits panels with factor error structures to construct instruments to estimate structural time series models with endogeneity even after controlling for latent factors. In chapter 2, we extend the GIV methodology in several dimensions. First, we extend the identification procedure to a large N and large T framework, which depends on the asymptotic Herfindahl index of the size distribution of N cross-sectional units. Second, we treat both the factors and loadings as unknown and show that the sampling error in the estimated instrument and factors is negligible when considering the limiting distribution of the structural parameters. Third, we show that the sampling error in the high-dimensional precision matrix is negligible in our estimation algorithm. Fourth, we overidentify the structural parameters with additional constructed instruments, which leads to efficiency gains. Monte Carlo evidence is presented to support our asymptotic theory and application to the global crude oil market leads to new results.

In the 3rd chapter, the GIV methodology is further developed to accommodate large dynamic panels with unit specific endogenous variates, which require unit-specific GIVs. We develop a split-panel jackknife (SPJ) GMM-PCA iterative procedure to estimate the structural parameters of interest. Overidentification tests can be carried out to test model validity. We illustrate the SPJ GMM-PCA procedure in two applications: (1) estimation of demand for new automobiles and (2) estimation of the determinants of banks' capital adequacy ratios.

Chapter 2

Inferential Theory for Granular Instrumental Variables in High Dimensions

2.1 Introduction

In the absence of randomized control trials, *finding* valid and strong instruments to circumvent unobserved confounders is a very challenging task. The Granular Instrumental Variables, hereafter GIV, methodology that [1] propose, establishes a systematic way to *construct* instruments from suitably weighted idiosyncratic shocks, from observational datasets and use them as instruments for aggregate endogenous variables.[1]

[1] We are grateful to Xu Cheng, Xavier Gabaix, Gloria Gonzalez-Rivera, Bruce Hansen, Jean Helwege, Bo Honoré, Guido Imbens, Ralph Koijen, Andrew Patton, Markus Pelger, Hashem Pesaran, Ekaterina Seregina, Ruoyao Shi, Aman Ullah, Qiankun Zhou, seminar participants at UC Riverside and JSM 2021 for helpful comments and suggestions. We thank Michael Bates for providing access to UCR's High Performance Computing Center to carry out simulations efficiently.

Constructing instruments. There are some existing methodologies which seek to eliminate the need to find an instrument. A leading example is the [2] framework in the context of estimating the speed of adjustment or state dependence parameters using dynamic panel data models with fixed effects, in which higher order lags of the dependent variable serve as instruments for the included lags of the dependent variable. The [3] methodology (aka shift-share estimators) where instruments are constructed from identities involving the (endogenous) explanatory variable whose shift component is interacted with shares. The [4] setting exploits the existence of structural breaks in the conditional heteroskedasticity regime, which is common place in many applications of interest. This allows one to bring a system with less equations than unknowns to a just identified system with as many equations as unknowns. The [5] methodology lays out a panel simultaneous equations model (similar to the model analyzed in this paper) where the estimated (strong) factors can be used as instrumental variables under certain conditions. We will return to the [5] methodology when we overidentify the structural parameters of interest as it is inspired by their framework. The vast methodological refinements cited within the papers referenced above are not listed here for brevity.

Microeconomic (granular) origins of aggregate fluctuations. How can idiosyncratic shocks be relevant for endogenous aggregate variables? The literature on "granularity" traces back to historic debates in macroeconomics; no attempt to fully catalog this debate is made here, rather a concise summary is offered. [9] demonstrate that in a multisector stochastic neoclassical growth model, sectoral shocks (as opposed to aggregate shocks) can potentially lead to GDP fluctuations. Intuitively, complex production processes form sectoral linkages which in turn provide a transmission mechanism of shocks across sectors. Subsequently, [10] and [11] debate whether sectoral shocks decay according to $\frac{1}{\sqrt{N}}$ as the central limit theorem would suggest. [7] provides an initial theoretical solution to the debate by showing that when the firm size distribution is heavy tailed, the central limit theorem does not apply and sectoral volatility decays much slower than $\frac{1}{\sqrt{N}}$. [7] coins this mechanism as the so-called "granular" hypothesis, in which the economy is composed of incompressible grains as opposed to infinitesimally small micro units. [12] formulates a network approach to demonstrate that sectoral idiosyncratic shocks generate non-negligible aggregate volatility when there exists sufficient asymmetry in the input-output relationships. [8] build off of the theoretical approach of [12] and develop econometric theory to measure the degree of network dominance and in their application they find some evidence of sector-specific shock propagation albeit not overwhelmingly strong for the US input-output accounts data over the period 1972-2002. See Figures 2.1 and 2.2 for empirical evidence of this notion of granularity. Figure 2.1 replicates Figure 1 of [7] with additional recent data and Figure 2.2 replicates Figure 1 of [13] with additional recent data as well. More empirical evidence for such propagation mechanism is presented in [14], [15], [16], [17], [18], [19], [7], [20], [21], [13], [22] and [23].

GIV, Gabaix and Koijen (2021). In an econometric framework, GK illustrate that when the market under consideration is sufficiently concentrated, then one can use the collection of idiosyncratic shocks to individual micro units, at each time period t, as an instrument for endogenous aggregate variables. The instrumental relevance follows heuristically from the paragraphs above. The exogeneity condition, as in any instrumental variables procedure, requires assumptions on unobserved random variables. However it should be noted that the exogeneity condition exploited in this framework is a relatively mild assumption that is often made in factor models (e.g. [24]) for identification purposes. The insight and contribution of GK opens the doors to a wide possibility of ways in which one can continue building on the promising new GIV methodology.

Contributions of this paper. Our contributions to the GIV methodology are primarily focused on the underlying econometric issues. First, we naturally extend GK's identification procedure to a large N and large T framework (GK formally introduced GIV for a fixed N and large T) by establishing and restricting the asymptotic behavior of the Herfindahl index for large N markets as a function of the tail index of the size distribution. Given the large N and large T framework, we treat both the factors and loadings as unknown and allow the idiosyncratic error term to be weakly cross-sectionally correlated.[2] As such, from our preliminary stage, we extract not only the estimated factors but also the estimated loadings via principal components analysis, PCA hereafter, or depending on the generality of the model ($k_x \neq 0$ in our notation from Section 2.2), we use the iterative OLS-PCA method of [25]. Second, we show that the sampling error in the estimated instrument and estimated factors is negligible when considering the limiting distribution of the structural parameters of interest; that is, the estimator is robust to the latent factor structure. Moreover, the exogeneity requirement for one of the structural parameters generally depends on a potentially high dimensional precision matrix (the inverse of the covariance matrix). Third, we show that the sampling error in the high dimensional precision matrix is negligible in our iterative estimation algorithm for said structural parameter. Fourth, we overidentify the structural parameters which leads to efficiency gains. This leads to new and improved results in our empirical application of GIV to the global crude oil markets. Monte Carlo evidence is presented to confirm the finite sample behavior of our estimators are well approximated by the asymptotic distributions. We label our refinement to the GIV methodology as *Feasible Granular Instrumental Variables* or FGIV for

[2] GK treat the factor loadings as known and extract the factors via period-by-period cross-sectional regressions. While they advocate extraction of latent factors via principal components analysis when loadings are unknown, they abstract away from the corresponding sampling error. We will show that the sampling error is indeed negligible.

short. Finally, an empirical application of the estimation methods to estimate demand and supply elasticities of the global crude oil markets are presented to demonstrate the estimation procedures.

Notation. We distinguish vectors and matrices from scalars by making an object bold. Let $\{X_{it}, i = 1, \dots, N; t = 1, \dots, T\}$ be a double index process of random variables where N denotes the number of cross-sectional units and T denotes the number of time periods. We frequently stack across i, in which we obtain $\underset{N\times 1}{\boldsymbol{X}_{\cdot t}} := \begin{pmatrix} X_{1t} & \dots & X_{Nt} \end{pmatrix}'$. Similarly, if we stack across t we obtain $\underset{T\times 1}{\boldsymbol{X}_{i\cdot}} := \begin{pmatrix} X_{i1} & \dots & X_{iT} \end{pmatrix}'$. When $\boldsymbol{X}_{it}$ is itself a vector, say of dimension k, then we obtain a matrix when we stack across i or t, e.g. $\underset{N\times k}{\boldsymbol{X}_{\cdot t}}$ or $\underset{T\times k}{\boldsymbol{X}_{i\cdot}}$. Define $X_{\boldsymbol{w}t}$ as the cross-sectionally weighted average of X_{it}, that is $X_{\boldsymbol{w}t} := \boldsymbol{w}'X_{\cdot t} = \sum_{i=1}^{N} w_i X_{it}$. Common weights, $\underset{N\times 1}{\boldsymbol{w}} = (w_i)$, used frequently throughout the paper are (1) the precision weights, $\boldsymbol{E} := \frac{\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\iota}}$ where $\underset{N\times N}{\boldsymbol{\Sigma}_u} := \mathbb{E}(\boldsymbol{u}_{\cdot t}\boldsymbol{u}'_{\cdot t})$ is the covariance matrix of the idiosyncratic error term, u_{it}, $\underset{N\times 1}{\boldsymbol{\iota}}$ is a vector of ones and (2) the share weights, which we simply refer to as size weights, $\boldsymbol{S} := \begin{pmatrix} S_1 & \dots & S_N \end{pmatrix}'$. Let $\widetilde{X}_{it} = X_{it} - \bar{X}_t$, where $\bar{X}_t = \frac{1}{N}\sum_{i=1}^{N} X_{it}$, denote a cross-sectionally demeaned variable. Unless otherwise specified, we denote the L^2-norm as $||\cdot||$ or sometimes explicitly as $||\cdot||_2$, the L^1-norm as $||\cdot||_1$ and the Frobenius norm as $||\cdot||_F$; if another norm is used, it will be explicitly noted. Given a square matrix $\boldsymbol{A}$, let $\gamma_{max}(\boldsymbol{A})$ denote the maximum eigenvalue of $\boldsymbol{A}$. Joint convergence of N and T will be denoted as $(N, T) \xrightarrow{j} \infty$ without any restriction on the relative rates; whenever restrictions on relative rates of convergence are imposed, it will be explicitly noted. The expression $\xrightarrow{p}$ denotes convergence in probability while $\xrightarrow{d}$ denotes convergence in distribution. The equation $\boldsymbol{y} = \mathcal{O}_p(\boldsymbol{x})$ states that the vector of random variables $\boldsymbol{y}$ is at most of order $\boldsymbol{x}$ in probability. The equation $a = \Theta_p(b)$ states that a is stochastically bounded by b and b is stochastically bounded by a, hence a and b rise jointly proportionally.

2.2 Model

A general formulation of the model examined in this paper is given in the following panel simultaneous equations model with factor error structure

$$\boldsymbol{y}_{it} = \boldsymbol{B}\boldsymbol{x}_{it} + \boldsymbol{C}\boldsymbol{a}_t + \boldsymbol{v}_{it},$$

$$\boldsymbol{v}_{it} = \boldsymbol{\Lambda}_i'\boldsymbol{F}_t + \boldsymbol{u}_{it},$$

where $\boldsymbol{y}_{it} = \begin{pmatrix} y_{1,it} & \dots & y_{G,it} \end{pmatrix}'$ is a $G \times 1$ vector of dependent variables, $\boldsymbol{x}_{it} = \begin{pmatrix} x_{1,it} & \dots & x_{k_x,it} \end{pmatrix}'$ is a $k_x \times 1$ vector of strictly exogenous variables (which can be arbitrarily correlated with the common factors, $\boldsymbol{F}_t$, and/or the loadings, $\boldsymbol{\Lambda}_i$), $\boldsymbol{a}_t = \begin{pmatrix} a_{1,t} & \dots & a_{k_a,t} \end{pmatrix}'$ is a $k_a \times 1$ vector of potentially endogenous aggregate variables, $\boldsymbol{v}_{it}$ is a $G \times 1$ vector of composite error terms which admit a low-rank plus sparse (factor structure) error decomposition, where $\boldsymbol{\Lambda}_i$ is an $r \times G$ matrix of latent factor loadings and $\boldsymbol{F}_t$ is an $r \times 1$ vector of latent factors.

In our exposition, we focus on the canonical setting of estimating the supply and demand elasticities in the global crude oil market, so we set the dimension of $G = 2$ for supply and demand variables respectively. We take $k_x = 0$ for ease of exposition but we present a general estimation algorithm for when $k_x \neq 0$. Moreover, we assume that only one of the $G = 2$ variables has a panel structure, whereas the other variable is an aggregate time series. That is, $\boldsymbol{y}_{it} = \begin{pmatrix} d_t & y_{it} \end{pmatrix}'$ where d_t is the log change of aggregate crude oil consumption and y_{it} is the log change of country i's crude oil production, $a_t = p_t$, with $k_a = 1$, is the log change of real crude oil price (where we deflate the nominal oil price with the U.S. general price index).[3],[4] Given our stylizations the coefficient matrix $\boldsymbol{C}$ and composite error, $\boldsymbol{v}_{it}$ becomes

$$\boldsymbol{C} = \begin{pmatrix} \phi^d & 0 \\ 0 & \phi^s \end{pmatrix},$$

$$\boldsymbol{v}_{it} = \boldsymbol{\Lambda}_i' \boldsymbol{F}_t + \boldsymbol{u}_{it} = \begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{\lambda}_i \end{pmatrix} \begin{pmatrix} \varepsilon_t \\ \boldsymbol{\eta}_t \end{pmatrix} + \begin{pmatrix} 0 \\ u_{it} \end{pmatrix},$$

where the coefficients ϕ^d and ϕ^s denote the crude oil demand and supply elasticities, respectively, and $\boldsymbol{\eta}_t, \boldsymbol{\lambda}_i$ are $r \times 1$ vectors of latent factors and latent loadings, respectively. Our stylized simultaneous equations model takes the simple form

$$d_t = \phi^d p_t + \varepsilon_t \tag{2.1}$$

$$y_{it} = \phi^s p_t + \boldsymbol{\lambda}_i' \boldsymbol{\eta}_t + u_{it}. \tag{2.2}$$

The global market clearing condition is given by $y_{St} = d_t$, where $y_{St} := \boldsymbol{S}' \boldsymbol{y}_{\cdot t} = \sum_{i=1}^N S_i y_{it}$, $\boldsymbol{S}$ is the $N \times 1$ vector of shares that are normalized such that $\sum_{i=1}^N S_i = 1$ and i and t take the values $i = 1, \ldots, N$ and $t = 1, \ldots, T$, respectively.[5] Making use of the global market clearing condition

[3] One may wonder why p_t is not disaggregated; in fact the p_t we use can be considered as the weighted average of country specific real oil prices (in changes). As shown in [26], for a proper global analysis, deflating the nominal oil price in U.S. dollars by the U.S. price index is generally theoretically invalid unless the law of one price holds universally. Namely, let P_{it} denote the general price index faced by country i, E_{it} denotes country i's exchange rate measured as units of country i's currency per U.S. dollar, p_{it} denote country specific log of real oil prices and $\tilde{p}_t$ denotes nominal oil prices in U.S. dollars, if $E_{it} P_{US,t} = P_{it}$ $\forall i$; then it follows that $\sum_{i=1}^N w_i p_{it} = \tilde{p}_t + \sum_{i=1}^N w_i \text{log}(E_{it}/P_{it}) = \tilde{p}_t + \sum_{i=1}^N w_i \text{log}(1/P_{US,t}) = \tilde{p}_t - p_{US,t} := p_t$. As it turns out, $p_t = \tilde{p}_t - p_{US,t}$ is an appropriate approximation as documented in [26] for their long run analysis, in the sense that it respects the long-run equilibrium relationships. We assume it is an appropriate approximation for our short-run analysis.

[4] The main results extend relatively naturally to the case where both variables have a panel model.

[5] As oil is a storable good, one could easily allow oil prices to adjust to the gap between supply and demand, e.g. as in [26]. This introduces more complex notations without adding any substance to the main points of the paper.

we see that

$$p_t = \frac{1}{\phi^d - \phi^s}\left(u_{St} + \boldsymbol{\lambda}_S'\boldsymbol{\eta}_t - \varepsilon_t\right), \tag{2.3}$$

which makes the simultaneity clear, e.g., that prices are composed of size-weighted idiosyncratic shocks, aggregate supply shocks and the demand shock. The objective of the GIV methodology is to extract the idiosyncratic shocks and use them as instruments for price.

Demand estimation in the case of uniform loadings ($\boldsymbol{\lambda}_i = \boldsymbol{\lambda}\,\forall\, i$). To momentarily fix ideas, it is helpful to consider a major simplification when constructing the instrument. Suppose that the loadings are uniform, $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}\,\forall i$. Then, the instrument, z_t, can be formed as

$$\begin{aligned} z_t &= y_{St} - \frac{1}{N}\sum_{i=1}^{N} y_{it} = (\phi^d p_t + \boldsymbol{\lambda}'\boldsymbol{\eta}_t + u_{St}) - (\phi^d p_t + \boldsymbol{\lambda}'\boldsymbol{\eta}_t + \frac{1}{N}\sum_{i=1}^{N} u_{it}), \\ &= u_{St} - \frac{1}{N}\sum_{i=1}^{N} u_{it} := u_{\Gamma t}, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\Gamma} := \boldsymbol{S} - \boldsymbol{\iota}/N$ is an $N \times 1$ random vector such that $\boldsymbol{\iota}'\boldsymbol{\Gamma} = \sum_{i=1}^{N} \Gamma_i = 0$, by construction. Γ_i is random because we assume the shares follow a fat-tailed distribution, see Assumption 4. Identification and estimation of demand by GIV requires that

$$\mathbb{E}(z_t\varepsilon_t) = \mathbb{E}\left(\sum_{i=1}^{N} \Gamma_i \mathbb{E}(u_{it}\varepsilon_t|\boldsymbol{\Gamma})\right) = 0. \tag{2.5}$$

(2.5) is our exogeneity condition and (2.3) gives $\mathbb{E}(z_t p_t) \neq 0$, relevance. A sufficient condition for the moment condition in (2.5) to be zero is $\mathbb{E}(u_{it}\varepsilon_t|\boldsymbol{\Gamma}) = 0$, which effectively requires that conditional on size, u_{it} and ε_t are uncorrelated. Given relevance, exogeneity implies the following demand elasticity estimator $\widehat{\phi}^d = \frac{\sum_t d_t z_t}{\sum_t p_t z_t}$. Intuitively, z_t places larger weights on the idiosyncratic shocks to larger oil producers, these granular shocks will shift the supply curve while keeping the aggregate demand curve fixed since demand responds to these shocks only through their affects on prices. This allows for consistent estimation of the demand elasticity. The uniform loadings assumption in this case tremendously facilitate the analysis. Uniform loadings allow one to construct the instrument, as in (2.4), from observables. In practice, uniform loadings are quite restrictive and we subsequently relax this assumption. However, before moving on to the general case, we also illustrate supply estimation under simplifying assumptions to fix ideas.

Supply estimation in the case of uniform loadings and u_{it} *i.i.d.* Continuing on with the uniform loadings case, remarkably, GK show that one can use the *same* instrument, z_t, to *also* estimate the supply elasticity using a cross-sectionally aggregated supply equation. Now, GK further assume that u_{it} are $i.i.d.$, $\mathbb{E}(\boldsymbol{u}_{\cdot t}\boldsymbol{u}'_{\cdot t}) := \boldsymbol{\Sigma}_u = \sigma_u^2 \boldsymbol{I}_N$, where $\boldsymbol{u}_{\cdot t} := \begin{pmatrix} u_{1t} & \dots & u_{Nt} \end{pmatrix}'$ and $\boldsymbol{I}_N$ is the identity matrix and define the $N \times 1$ precision weight vector $\boldsymbol{E} := \frac{\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\iota}}$ which reduces to $\boldsymbol{\iota}/N$ when u_{it} are $i.i.d.$ across i. Aggregation of the supply equation is performed using the vector $\boldsymbol{E}$, we have that $y_{Et} = \phi^s p_t + \boldsymbol{\lambda}'\boldsymbol{\eta}_t + u_{Et}$. Identification and estimation of supply by GIV requires that the instrument satisfies exogeneity with respect to the *composite* error term[6]

$$\mathbb{E}((\boldsymbol{\lambda}'\boldsymbol{\eta}_t + u_{Et})z_t) = 0. \tag{2.6}$$

The first term in (2.6) has similar interpretation as in (2.5), i.e., size-weighted idiosyncratic supply shocks are uncorrelated with the aggregate supply component, $\boldsymbol{\lambda}'\boldsymbol{\eta}_t$. Miraculously, the second term

[6] In the general case to follow, we estimate the factors and thus only exploit $\mathbb{E}(u_{Et}z_t) = 0$ to estimate ϕ^s.

is exactly zero

$$\mathbb{E}(u_{Et}z_t) = \mathbb{E}(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}\boldsymbol{u}_{\cdot t}'\boldsymbol{\Gamma}) = \mathbb{E}(\boldsymbol{E}'\mathbb{E}(\boldsymbol{u}_{\cdot t}\boldsymbol{u}_{\cdot t}'|\boldsymbol{\Gamma})\boldsymbol{\Gamma}) = \frac{\sigma_u^2}{N}\mathbb{E}(\boldsymbol{\iota}'\boldsymbol{\Gamma}) = 0.^{7} \tag{2.7}$$

The moment condition (2.7) is zero due to independence of Γ_i and u_{it} by assumption and the sum-to-zero property of $\boldsymbol{\Gamma}$. For identification with large N, we assume size to follow a power law in tail (see Assumption 4), thus Γ_i is stochastic and assumed to be independent of u_{it}.[8] So again, we have $\mathbb{E}(z_t p_t) \neq 0$ and for this simplified example, we avoid the need to estimate the factor structure since (i) due to uniform loadings, z_t is constructed from observables and (ii) z_t is uncorrelated with the *composite* error term. If either of (i) or (ii) fails to hold, estimation of the factor structure becomes a preliminary step, as in our general procedure. Nevertheless, (2.6) leads to the following simple supply elasticity estimator $\widehat{\phi}^s = \frac{\sum_t y_{Et}z_t}{\sum_t p_t z_t}$. The intuition here is that, again, z_t places larger weights on the idiosyncratic shocks to larger oil producers, these granular shocks keep the simple average (or more generally precision-weighted, i.e., weighted heavily towards more stable oil producers) supply curve fixed. That is, on average, precision-weighted supply responds to these granular shocks only through their effects on prices (due to $\mathbb{E}((\boldsymbol{\lambda}'\boldsymbol{\eta}_t + u_{Et})z_t) = 0$) and at the same time since smaller oil producers take as given price changes caused by these granular shocks, it will shift their supply curves which enables consistent estimation of the supply elasticity.

Discussion. In the case of uniform loadings and u_{it} $i.i.d.$, the vector $\boldsymbol{E}$ and the instrument are constructed from observables, the large sample properties of $\widehat{\phi}^s$ and $\widehat{\phi}^d$ only entail fixed

[7] One may wonder then why this particular form of $\boldsymbol{\Gamma}$ was selected. Appealing to Proposition 3 in GK, they establish that $\boldsymbol{\Gamma} = \boldsymbol{S} - \boldsymbol{E}$, for this example, turns out to be the optimal weight vector, amongst the class of weights which sum-to-zero. $\boldsymbol{\Gamma}$ is optimal in the sense that it minimizes the asymptotic variance of the structural parameters.

[8] For a fixed N, it is not required to assume independence of Γ_i and u_{it} because Γ_i can be treated as constant and (2.7) is zero solely by virtue of the fact that $\boldsymbol{\iota}'\boldsymbol{\Gamma} = 0$.

N, large T asymptotics for which GK have laid out. In general, however, the cross-section will need to be exploited to estimate $\boldsymbol{E}$ since one can not know if u_{it} are $i.i.d.$ across i. Indeed, the factors typically take care of a substantial portion of the cross-sectional correlations but it is prudent to allow for cross correlations in u_{it} since the exogeneity condition for estimation of the supply elasticity heavily exploits the structure of $\boldsymbol{\Sigma}_u$. Therefore, it will be important to generally allow for some weak cross correlations in $\boldsymbol{\Sigma}_u$, which our algorithm accommodates, as discussed in Section 2.3 and Section 2.4.

Moreover,although homogeneous loadings was only an abstraction to illustrate the instrument, GK advocate the use of $y_{\Gamma t} = y_{St} - \frac{1}{N}\sum_i y_{it}$ in practice even when the loadings are not uniform. In the general heterogeneous loadings case, their instrument becomes

$$Z_t := y_{\Gamma t} = u_{\Gamma t} + \boldsymbol{\lambda}'_{\Gamma}\boldsymbol{\eta}_t. \tag{2.8}$$

They label this instrument with a capital case convention, to distinguish it because it is no longer solely composed of weighted idiosyncratic shocks, $u_{\Gamma t}$, as the $\boldsymbol{\lambda}'_{\Gamma}\boldsymbol{\eta}_t$ term is contaminating the instrument. However, this clever formulation is possible because they advocate estimation of the factors in practice, which they augment to their structural equations, thereby controlling for the second term which can potentially make their moment conditions different from zero.

2.3 Feasible Granular Instrumental Variables

Homogeneous loadings are overly restrictive but relaxing this can be easily accommodated in practice via PCA or iterative OLS-PCA methods, e.g., [27] or [25] in a preliminary stage to construct an estimate of the instrument.[9] Although in GK's asymptotic theory they assume homogeneous loadings and that the instrument is exogenous with respect to the composite error, which circumvents the need to estimate the factor structure, they indeed advocate augmenting their structural equations with estimated factors either via period-by-period cross sectional regressions when the loadings are known or via PCA in the case of non-parametric (unknown) loadings. GK abstract away from the sampling error in suggesting the use of augmented factors, which only vanishes for both large N and T. [31] and [32] have developed the asymptotic distribution for structural parameters in factor augmented regressions in time series and panel models respectively. In this paper, a variant of their corresponding result is established in showing the sampling error from estimating the high dimensional precision matrix, the factors, as well as the instrument is negligible in the asymptotic distribution of the structural parameters.

The general heterogeneous loadings case and u_{it} non-$\boldsymbol{i.i.d.}$ Now we formulate the estimation approach in the general case, which makes much heavier use of the cross-section. When we cross-sectionally demean the supply equation and stack across i we obtain (recall $\widetilde{\boldsymbol{X}}$ denotes a generic demeaned variate)

$$\widetilde{\boldsymbol{y}}_{\cdot t} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\eta}_t + \widetilde{\boldsymbol{u}}_{\cdot t}, \tag{2.9}$$

which is estimable with vanilla PCA when the factor structure is strong.[10] Letting $\boldsymbol{Q} = (\boldsymbol{I}_N -$

[9] For our theory, we assume a balanced panel. However, in the case of unbalanced panels with data missing at random (which is beyond the scope of this paper) one can instead use the [28] method or [29] method to estimate the factor structure and the instrument. In the more realistic case where data are not missing at random, one can use the methods developed in [30].

[10] Strong factors in the sense that $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}/N \xrightarrow{p} \boldsymbol{\Sigma}_{\Lambda} > 0$; thus we assume the factors are strong/pervasive in the sense that a significant fraction of cross-sectional units are affected by their presence. Consistent estimation of weak factors is beyond the scope of this paper, see for example [33], [34] or [35] for suitable conditions for which it is possible. Even when estimable, their convergence rates are slower relative to estimates of strong factors, e.g., see [36]. This will generally require modifications to the limiting distributions we derive in this paper.

$\widetilde{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}})^{-1}\widetilde{\boldsymbol{\Lambda}}')$, then $\boldsymbol{Q}\widetilde{\boldsymbol{y}}_{\cdot t} = \boldsymbol{Q}\widetilde{\boldsymbol{u}}_{\cdot t}$, completely purges the process of the common factors through the loading space. Premultiplying the share weights gives the instrument

$$z_t := \boldsymbol{S}'\boldsymbol{Q}\widetilde{\boldsymbol{y}}_{\cdot t}, \tag{2.10}$$

$$= \boldsymbol{S}'\boldsymbol{Q}\widetilde{\boldsymbol{u}}_{\cdot t} := \boldsymbol{\Gamma}'\widetilde{\boldsymbol{u}}_{\cdot t}, \tag{2.11}$$

where $\boldsymbol{\Gamma} := \boldsymbol{Q}\boldsymbol{S}$ is unknown because $\boldsymbol{Q}$ is unknown, but $\boldsymbol{Q}$ is easily estimated from data. Once we have $\widehat{\boldsymbol{Q}}$, which just replaces $\widetilde{\Lambda}$ with $\widehat{\widetilde{\Lambda}}$, we form $\widehat{z}_t = \boldsymbol{S}'\widehat{\boldsymbol{Q}}\widetilde{\boldsymbol{y}}_{\cdot t}$ from observables. Importantly, when $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}\,\forall i$, then $\boldsymbol{\Gamma} = (\boldsymbol{I}_N - \widetilde{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}})^{-1}\widetilde{\boldsymbol{\Lambda}}')\boldsymbol{S} = \boldsymbol{S} - \boldsymbol{\iota}/N$ as in the previous case with homogenous loadings. This gives rise to a more general demand elasticity estimator

$$\widehat{\phi}^d = \widehat{\phi}^d(\widehat{\boldsymbol{z}}) = \frac{\sum_t d_t\widehat{z}_t}{\sum_t p_t\widehat{z}_t}. \tag{2.12}$$

In Section 2.6, we show that the demand elasticity can be estimated *as if* the infeasible instrument, z_t, is used.

In the case of the supply elasticity, the estimator will additionally depend on the estimated (potentially high dimensional) precision matrix. That is, $\widehat{\phi}^s = \widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$. This creates the need to jointly estimate $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ to form $\widehat{\boldsymbol{E}}$ in order to aggregate the panel to estimate $\widehat{\phi}^s$. We propose a simple iterative procedure and show that the supply elasticity can be estimated *as if* the infeasible precision matrix, $\boldsymbol{\Sigma}_u^{-1}$, and instrument, z_t, were used. More specifically, let $y_{Et} = \phi^s p_t + \boldsymbol{\lambda}_E'\boldsymbol{\eta}_t + u_{Et} := \boldsymbol{f}_t'\boldsymbol{\theta}^s + u_{Et}$, where $\boldsymbol{\theta}^s = \begin{pmatrix}\phi^s & \boldsymbol{\lambda}_E'\end{pmatrix}'$ and $\boldsymbol{f}_t = \begin{pmatrix}p_t & \boldsymbol{\eta}_t'\end{pmatrix}'$ are $(1+r)\times 1$ vectors. The remarkable result $\mathbb{E}(z_t u_{Et}) = 0$, shown in (2.7) for the previous simple example with homogeneous loadings, continues to hold in this setting as well, with $z_t = \boldsymbol{S}'\boldsymbol{Q}\tilde{\boldsymbol{y}}_{\cdot t} = \boldsymbol{S}'\boldsymbol{Q}\tilde{\boldsymbol{u}}_{\cdot t}$ and $\boldsymbol{\Gamma} = \boldsymbol{Q}\boldsymbol{S}$ (recall that

$\boldsymbol{\iota}'\boldsymbol{\Gamma} = 0$)

$$\begin{aligned}\mathbb{E}(u_{Et}z_t) &= \mathbb{E}\left(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}\tilde{\boldsymbol{u}}_{\cdot t}'\boldsymbol{\Gamma}\right) = \mathbb{E}(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}(\boldsymbol{u}_{\cdot t} - \bar{u}_t\boldsymbol{\iota})'\boldsymbol{\Gamma}) = \mathbb{E}(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}\boldsymbol{u}_{\cdot t}'\boldsymbol{\Gamma}) - \mathbb{E}(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}\bar{u}_t\boldsymbol{\iota}'\boldsymbol{\Gamma}) \\ &= \mathbb{E}(\boldsymbol{E}'\,\mathbb{E}(\boldsymbol{u}_{\cdot t}\boldsymbol{u}_{\cdot t}'|\boldsymbol{\Gamma})\boldsymbol{\Gamma}) - 0 = \frac{1}{\boldsymbol{\iota}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\iota}}\mathbb{E}\left(\boldsymbol{\iota}'\boldsymbol{\Gamma}\right) = 0. \end{aligned} \tag{2.13}$$

So we have that (where the estimated factors self-instrument)

$$\mathbb{E}\left[\begin{pmatrix} z_t \\ \boldsymbol{\eta}_t \end{pmatrix} \cdot u_{Et}\right] = \mathbb{E}\left[\begin{pmatrix} z_t \\ \boldsymbol{\eta}_t \end{pmatrix} \cdot \left(y_{Et} - \phi^s p_t - \boldsymbol{\lambda}_E'\boldsymbol{\eta}_t\right)\right] = \boldsymbol{0}. \tag{2.14}$$

However, given our interest lies in inference for ϕ^s, it is useful to stack over t, $\boldsymbol{y}_E = \boldsymbol{p}\,\phi^s + \boldsymbol{\eta}\,\boldsymbol{\lambda}_E + \boldsymbol{u}_E$, where $\boldsymbol{y}_E, \boldsymbol{p}$, and $\boldsymbol{u}_E$ are $T \times 1$ vectors and $\boldsymbol{y}_{\widehat{E}}$ is the feasible counterpart of $\boldsymbol{y}_E$. Let $\boldsymbol{M}_{\widehat{\eta}} = (\boldsymbol{I}_T - \widehat{\boldsymbol{\eta}}(\widehat{\boldsymbol{\eta}}'\widehat{\boldsymbol{\eta}})^{-\mathbf{1}}\widehat{\boldsymbol{\eta}}')$, then it follows from standard partitioned regression results that

$$\widehat{\phi}^s = \widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1}) = \frac{\widehat{\boldsymbol{z}}'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{y}_{\widehat{E}}}{\widehat{\boldsymbol{z}}'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{p}}. \tag{2.15}$$

As $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ depends on $\widehat{\phi}^s$, (2.15) generally requires an iterative estimation procedure. To that end, note that if ϕ^s were known, $y_{it} - p_t\phi^s = \boldsymbol{\lambda}_i'\boldsymbol{\eta}_t + u_{it}$ follows an approximate factor structure. Thus, a covariance estimator, $\widehat{\boldsymbol{\Sigma}}_u$, for the idiosyncratic part can be obtained following [37] by applying thresholding to the eigenvalue decomposition, $\frac{1}{T}\sum_{t=1}^T(\boldsymbol{y}_{\cdot t} - \boldsymbol{\iota}p_t\phi^s)(\boldsymbol{y}_{\cdot t} - \boldsymbol{\iota}p_t\phi^s)' = \sum_{i=1}^N \gamma_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i'$, where γ_i and $\boldsymbol{\xi}_i$ are the eigenvalues (sorted in decreasing order) and corresponding eigenvectors, respectively. More specifically, if ϕ^s were known, we have

$$\widehat{\boldsymbol{\Sigma}}_{(\boldsymbol{y}_{\cdot t} - \boldsymbol{\iota}p_t\phi^s)} := \sum_{i=1}^r \widehat{\gamma}_i\widehat{\boldsymbol{\xi}}_i\widehat{\boldsymbol{\xi}}_i' + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{\boldsymbol{\mathcal{T}}}, \tag{2.16}$$

where $\widehat{\boldsymbol{\Sigma}}_u^{\mathcal{T}} = \sum_{i=r+1}^N \widehat{\gamma}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i' = (\widehat{\sigma}_{u,ij}^{\mathcal{T}})_{N\times N}$,

$$\widehat{\sigma}_{u,ij}^{\mathcal{T}} = \begin{cases} \widehat{\sigma}_{u,ii}, & i = j, \\ \\ h_{ij}(\widehat{\sigma}_{u,ij}), & i \neq j, \end{cases} \tag{2.17}$$

and $h_{ij}(\cdot)$ is a generalized shrinkage function of [38].[11] Of course, ϕ^s can not be known as it requires an estimate of $\boldsymbol{\Sigma}_u^{-1}$. Thus, we now address joint estimation of ϕ^s and $\boldsymbol{\Sigma}_u^{-1}$ in what follows and subsequently establish that the sampling error in $\widehat{\boldsymbol{E}}$ is negligible given some regularity conditions. The iterative procedure is summarized in Algorithm 1 presented below.

Algorithm 1 FGIV for ϕ^s (when $k_x = 0$):

```
• Step 1: Run PCA on (2.9) and obtain $\widehat{z}_t = \boldsymbol{S}'\widehat{\boldsymbol{Q}}\widetilde{\boldsymbol{y}}_{\cdot t}$ as the sample counterpart of (2.10).
• Step 2: Initialize $\widehat{\boldsymbol{\Sigma}}_u^{-1} = \boldsymbol{I}_N$.
• Step 3: Obtain $\boldsymbol{y}_{\widehat{E}}(\widehat{\boldsymbol{\Sigma}}_u^{-1})$ and $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$ as in (2.15).
• Step 4: Update $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ by inverting $\widehat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}$ defined in (2.17), $\boldsymbol{y}_{\widehat{E}}(\widehat{\boldsymbol{\Sigma}}_u^{-1})$ and $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$.
• Step 5: Iterate Step 3 and Step 4 until convergence.
```

When r is unknown, one can augment *Step 1* and estimate r using a procedure as in [24], [39] or [40]; we use the ER and GR methods of [40] (hereafter AH). For more details of the ER and GR methods, see Section 2.12.3 of the Supplementary Appendix.

FGIV algorithm accommodating cross-section specific covariates. When $k_x \neq 0$ then the demeaning transformation from (2.9) results in $\widetilde{\boldsymbol{y}}_{\cdot t} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\eta}_t + \widetilde{\boldsymbol{x}}_{\cdot t}\boldsymbol{\beta} + \widetilde{\boldsymbol{u}}_{\cdot t}$, where $\widetilde{\boldsymbol{x}}_{\cdot t}$ is an $N \times k_x$ matrix, which leaves $\underset{k_x\times 1}{\boldsymbol{\beta}}$ as an additional parameter to estimate. $\boldsymbol{\beta}$ can be easily estimated

[11] Examples of $h_{ij}(\cdot)$ include hard thresholding $h_{ij}(x) = x\mathbb{1}(|x| \geq \tau_{ij})$ and soft thresholding $h_{ij}(x) = \text{sgn}(x)(|x| - \tau_{ij})_+$. The entry dependent threshold, $\tau_{ij} > 0$, can be defined as $C\omega_T\sqrt{\widehat{\alpha}_{ij}}$, where $\widehat{\alpha}_{ij} = \frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}\widehat{u}_{jt} - \widehat{\sigma}_{u,ij})^2$, $\widehat{\sigma}_{u,ij} = \frac{1}{T}\sum_{t=1}^T \widehat{u}_{it}\widehat{u}_{jt}$ and $\widehat{u}_{it} = y_{it} - \phi^s p_t - \widehat{\boldsymbol{\lambda}}_i'\widehat{\boldsymbol{\eta}}_t$ for some predetermined decreasing sequence $\omega_T > 0$ and $C > 0$. The choice of C can be data driven; [37] choose C through multifold cross-validation to maintain positive definiteness of $\widehat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}(C)$. In our algorithm below, we make use of the R package for POET, written by the authors [37].

by adapting the procedure of [41], which is generalizing [25], to handle endogeneity of prices even after controlling for latent common factors. More specifically,

$$\boldsymbol{\beta}(\widetilde{\boldsymbol{\Lambda}}, \boldsymbol{\eta}_t, \boldsymbol{\Sigma}_u^{-1}) = \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{x}}_{\cdot t}' \boldsymbol{\Sigma}_u^{-1} \widetilde{\boldsymbol{x}}_{\cdot t} \right)^{-1} \sum_{t=1}^{T} \widetilde{\boldsymbol{x}}_{\cdot t}' \boldsymbol{\Sigma}_u^{-1} (\widetilde{\boldsymbol{y}}_{\cdot t} - \widetilde{\boldsymbol{\Lambda}} \boldsymbol{\eta}_t), \tag{2.18}$$

$$(\widetilde{\boldsymbol{y}}_{\cdot t} - \widetilde{\boldsymbol{x}}_{\cdot t} \boldsymbol{\beta}) = \widetilde{\boldsymbol{\Lambda}} \boldsymbol{\eta}_t + \widetilde{\boldsymbol{u}}_{\cdot t}, \tag{2.19}$$

since (2.19) follows a factor structure, the $T \times r$ factor matrix, $\boldsymbol{\eta}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_u^{-1})$, can be estimated using the principal components estimator whose columns are the eigenvectors corresponding to the largest r eigenvalues of the $T \times T$ matrix $(\widetilde{\boldsymbol{y}}_{\cdot\cdot} - \widetilde{\boldsymbol{x}}_{\cdot\cdot}(\boldsymbol{\beta})) \boldsymbol{\Sigma}_u^{-1} (\widetilde{\boldsymbol{y}}_{\cdot\cdot} - \widetilde{\boldsymbol{x}}_{\cdot\cdot}(\boldsymbol{\beta}))'$, where the $T \times N$ matrix $\widetilde{\boldsymbol{x}}_{\cdot\cdot}(\boldsymbol{\beta}) := \begin{pmatrix} \widetilde{\boldsymbol{x}}_{1\cdot}\boldsymbol{\beta} & \ldots & \widetilde{\boldsymbol{x}}_{N\cdot}\boldsymbol{\beta} \end{pmatrix}$ and $\widetilde{\boldsymbol{\Lambda}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_u^{-1}) = \frac{1}{T} \sum_{t=1}^{T} (\widetilde{\boldsymbol{y}}_{\cdot t} - \widetilde{\boldsymbol{x}}_{\cdot t} \boldsymbol{\beta}) \boldsymbol{\eta}_t'(\boldsymbol{\beta}, \boldsymbol{\Sigma}_u^{-1})$. Thus, to deal with general (strictly exogenous) covariates, $\boldsymbol{x}_{it}$, Algorithm 2 can be applied.

Algorithm 2 FGIV for ϕ^s (when $k_x \neq 0$):

- *Step 1:* Initialize $\widehat{\boldsymbol{\beta}} = \mathbf{0}$, $\widehat{\boldsymbol{\Sigma}}_u^{-1} = \boldsymbol{I}_N$.
- *Step 2:* Run PCA on (2.19) to obtain $\widehat{\boldsymbol{\eta}}_t(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$ and $\widehat{\widetilde{\boldsymbol{\Lambda}}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$ as explained above.
- *Step 3:* Update $\widehat{\boldsymbol{\beta}}$ as the sample counterpart of (2.18).
- *Step 4:* Obtain $\widehat{z}_t = \boldsymbol{S}' \widehat{\boldsymbol{Q}} (\widetilde{\boldsymbol{y}}_{\cdot t} - \widetilde{\boldsymbol{x}}_{\cdot t} \widehat{\boldsymbol{\beta}})$.
- *Step 5:* Initialize $\boldsymbol{y}_{\widehat{E}}(\widehat{\boldsymbol{\Sigma}}_u^{-1})$ and $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1}) = \left(\widehat{\boldsymbol{z}}' \, \boldsymbol{M}_{\widehat{\eta}} \, \boldsymbol{p} \right)^{-1} \widehat{\boldsymbol{z}}' \, \boldsymbol{M}_{\widehat{\eta}} \, (\boldsymbol{y}_{\widehat{E}} - \boldsymbol{x}_{i\cdot} \boldsymbol{\beta})$.
- *Step 6:* Update $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ by inverting $\widehat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}$ defined in (2.17), where $\widehat{\gamma}_i$ and $\widehat{\boldsymbol{\xi}}_i$ are the eigenvalues and eigenvectors (sorted in decreasing order) corresponding to the sample analog of $\frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{y}_{\cdot t} - \boldsymbol{\iota} p_t \phi^s - \boldsymbol{x}_{\cdot t} \boldsymbol{\beta})(\boldsymbol{y}_{\cdot t} - \boldsymbol{\iota} p_t \phi^s - \boldsymbol{x}_{\cdot t} \boldsymbol{\beta})'$ respectively.
- *Step 7:* Iterate *Step 2* through *Step 6* until convergence.

When r is unknown, one can augment *Step 2* and iteratively estimate r using the ER and GR methods of [40].

The main takeaway is that when both (N, T) are large, one can generalize the GIV estimators proposed by GK along different dimensions; here we accommodate latent heterogeneous loadings, latent factors and latent precision matrix (e.g., u_{it} can be weakly cross-correlated and heteroskedastic). As mentioned earlier, we call the proposed estimators of the elasticities in (2.12), Algorithm 1 and Algorithm 2 as FGIV estimators.

Remark 1 *In principle, the theory for the estimators proposed in this paper allows for $N \gg T$. This case is relevant in many empirical settings (e.g., empirical industrial organization and finance). However, it may be beneficial to avoid estimating the precision matrix for cases where $N \ll T$ (e.g., empirical macro). But, as (2.7) and (2.13) show, to have a valid instrument for which the moment equation is exactly zero, we must specify $\boldsymbol{\Sigma}_u$ correctly. This is the primary motivation for estimating the general precision matrix in Algorithms 1 and 2. In order to avoid estimating the precision matrix, we must assume (potentially erroneously) u_{it} are cross-sectionally independent. We now analyze the consequences of making this assumption when in fact u_{it} are cross-sectionally correlated. Suppose we erroneously assume cross-sectional independence, then the vector $\boldsymbol{E}$ reduces to $\boldsymbol{\iota}/N$ and we end up with the following moment equation*

$$\begin{aligned}\mathbb{E}(u_{Et}z_t) &= \mathbb{E}\left(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}\tilde{\boldsymbol{u}}'_{\cdot t}\boldsymbol{\Gamma}\right) = \mathbb{E}(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}(\boldsymbol{u}_{\cdot t} - \bar{u}_t\boldsymbol{\iota})'\boldsymbol{\Gamma}) \\ &= \mathbb{E}(\boldsymbol{E}'\mathbb{E}(\boldsymbol{u}_{\cdot t}\boldsymbol{u}'_{\cdot t}|\boldsymbol{\Gamma})\boldsymbol{\Gamma}) - \mathbb{E}(\boldsymbol{E}'\boldsymbol{u}_{\cdot t}\boldsymbol{\iota}'\boldsymbol{\Gamma})\bar{u}_t = \frac{1}{N}\mathbb{E}(\boldsymbol{\iota}'\boldsymbol{\Sigma}_u\boldsymbol{\Gamma}) - 0 = o(1). \qquad (2.20)\end{aligned}$$

Hence, z_t is not a valid instrument in the traditional sense because we allow $\mathbb{E}(u_{Et}z_t) \neq 0$ for any given sample. Nevertheless, this moment converges to zero for large N. Indeed, the moment satisfies $\mathbb{E}(u_{Et}z_t) = o\left(1\right)$ under our regularity assumptions, and thus, z_t is asymptotically a valid instrument.[12] *This insight reveals that this moment is approaching zero, hence it may prove to*

[12] It can be shown that $\boldsymbol{\iota}'\boldsymbol{\Sigma}_u\boldsymbol{\Gamma} = \boldsymbol{\iota}'\boldsymbol{\Sigma}_u\boldsymbol{Q}\boldsymbol{S} \leq \boldsymbol{\iota}'\boldsymbol{\Sigma}_u\boldsymbol{S}\gamma_{max}(\boldsymbol{Q}) = \sum_{i,j}\sigma_{u,ij}S_j \leq \left(\sum_{i,j}\sigma^2_{u,ij}\right)^{1/2}||\boldsymbol{S}||_2^2 =$

be beneficial to aggregate the panel, y_{it}, using weights $\boldsymbol{\iota}/N$ regardless of the covariance structure. The immediate implication is that $\widehat{\phi}^s = \widehat{\phi}^s(\widehat{\boldsymbol{z}}, \boldsymbol{I}_N)$, so there is no need for an algorithmic estimation procedure, the simple analytical formula for the supply elasticity estimator with potentially misspecified covariance structure for u_{it} is given by $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \boldsymbol{I}_N) = \frac{\widehat{\boldsymbol{z}}' \boldsymbol{M}_{\widehat{\eta}} \bar{\boldsymbol{y}}}{\widehat{\boldsymbol{z}}' \boldsymbol{M}_{\widehat{\eta}} \boldsymbol{p}}$, where $\bar{\boldsymbol{y}}$ stacks $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$ for each $t = 1, \ldots, T$; this estimator is essentially Step 2 and Step 3 of Algorithm 1. Asymptotically, it holds that $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1}) = \widehat{\phi}^s(\widehat{\boldsymbol{z}}, \boldsymbol{I}_N) + o_p(1)$. However, regarding performance in finite samples, when u_{it} are not $i.i.d.$ and when $N \ll T$, it is not clear ex-ante if $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \boldsymbol{I}_N)$ will outperform $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$. When $N \gg T$ one would expect ex-ante that $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \boldsymbol{I}_N)$ will be less efficient than $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$ since the former is not optimally weighting the observations, whereas the latter is. When u_{it} are indeed $i.i.d.$ we would expect $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \boldsymbol{I}_N)$ to perform better.[13]

2.4 Efficient GMM Estimation: Factor-Augmented FGIV

We now proceed to overidentify the elasticities, which yields overidentified FGIV estimators. We will refer to the overidentified FGIV estimators simply as efficient GMM estimators and the just identified FGIV estimators simply as FGIV estimators. It will be of interest to practitioners to see if overidentification is possible for the supply and demand equations. In this section, we show that the system is indeed overidentified to varying degrees for the supply and demand equations.

Demand. It is common practice to assume uncorrelated aggregate supply and aggregate demand shocks, that is $\mathbb{E}(\boldsymbol{\eta}_t \varepsilon_t) = 0$. When we are willing to entertain this, then our supply factors, estimated via principal components, serve as valid instruments in estimation of the demand elas-

$||\boldsymbol{\Sigma}_u||_F ||\boldsymbol{S}||_2^2 \leq ||\boldsymbol{\Sigma}_u||_1 ||\boldsymbol{S}||_2^2 \leq \mathcal{O}(m_N)\Theta_p(1) = o(N)\Theta_p(1) = o(N)$, where m_N is defined in Assumption 3 and $m_N = o(N)$.

[13] In unreported simulations where $N \ll T$ and u_{it} are non-i.i.d., we find that $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\Sigma}}_u^{-1})$ typically has a smaller bias than $\widehat{\phi}^s(\widehat{\boldsymbol{z}}, \boldsymbol{I}_N)$ (in absolute terms, the bias of both estimators are very small) but with a slightly larger variance.

ticity, rendering an overidentified parameter. In fact, the theory for using principal components as instruments was laid out in [5] under strong instrument asymptotics, as well as [6] under many/weak instrument asymptotics. In the remainder of this section, we let the GIV be denoted as $z_{t,GIV} := z_t$ to distinguish it from the full instrument vector we introduce with upper case conventions. Our full instrument matrix for the demand equation is $\underset{T\times(1+r)}{\boldsymbol{Z}_d} := \begin{pmatrix}\boldsymbol{z}_{GIV} & \boldsymbol{\eta}\end{pmatrix}$ with $\mathbb{E}(\boldsymbol{Z}_{dt}\varepsilon_t) = 0$; $\boldsymbol{Z}_{dt}$ simply augments factors to be used as instruments. Making use of the $(1+r)\times 1$ dimensional moment condition, the efficient GMM demand elasticity estimator is defined as

$$\begin{aligned}\widehat{\phi}^d_{GMM} &= \underset{\phi^d}{\text{argmin}}\, \frac{\boldsymbol{\varepsilon}'\boldsymbol{Z}_d}{T}\boldsymbol{W}_d\frac{\boldsymbol{Z}_d'\boldsymbol{\varepsilon}}{T},\\ &= \left(\boldsymbol{p}'\,\widehat{\boldsymbol{Z}}_d\,\widehat{\boldsymbol{\Omega}}_d^{-1}\,\widehat{\boldsymbol{Z}}_d'\boldsymbol{p}\right)^{-1}\boldsymbol{p}'\,\widehat{\boldsymbol{Z}}_d\,\widehat{\boldsymbol{\Omega}}_d^{-1}\,\widehat{\boldsymbol{Z}}_d'\boldsymbol{d},\end{aligned} \tag{2.21}$$

where $\underset{(1+r)\times(1+r)}{\boldsymbol{W}_d}$ is an arbitrary positive definite weight matrix, but is optimally set as $\widehat{\boldsymbol{W}}_d = \widehat{\boldsymbol{\Omega}}_d^{-1}$, where $\widehat{\boldsymbol{\Omega}}_d = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{Z}}_{dt}\widehat{\boldsymbol{Z}}_{dt}'(d_t - p_t\widehat{\phi}^d_{2SLS})^2$. It is clear that (2.21) nests the FGIV estimator for the demand elasticity as a special case. In this sense, $\widehat{\phi}^d_{GMM}$ will be robust to scenarios where z_t is weaker.

Supply. In the same vein, the supply elasticity can *always* be overidentified given our identifying assumptions because $\mathbb{E}(\varepsilon_t u_{Et}) = 0$ and thus ε_t can serve as an additional instrument. To estimate the entire parameter vector for the supply equation, let $\underset{T\times(2+r)}{\boldsymbol{Z}_s} := \begin{pmatrix}\boldsymbol{z}_{GIV} & \boldsymbol{\varepsilon} & \boldsymbol{\eta}\end{pmatrix}$, where the augmented factors self-instrument as they are part of the supply equation. Then $\boldsymbol{y}_E = \boldsymbol{f}\boldsymbol{\theta}^s + \boldsymbol{u}_E$ and recall $\boldsymbol{\theta}^s = \begin{pmatrix}\phi^s & \boldsymbol{\lambda}_E'\end{pmatrix}'$ and $\boldsymbol{f}_t = \begin{pmatrix}p_t & \boldsymbol{\eta}_t'\end{pmatrix}'$ are $(1+r)\times 1$ vectors and the matrix $\boldsymbol{f}$ is $T\times(1+r)$, which stacks $\boldsymbol{f}_t$. We have $\mathbb{E}(\boldsymbol{Z}_{st}\boldsymbol{u}_{Et}) = 0$; hence, making use of the $(2+r)\times 1$

dimensional moment conditions, the efficient GMM supply elasticity estimator is defined as

$$\widehat{\boldsymbol{\theta}}^s_{GMM} = \underset{\boldsymbol{\theta}^s}{\text{argmin}}\, \frac{\boldsymbol{u}_E'\boldsymbol{Z}_s}{T}\, \boldsymbol{W}_s\, \frac{\boldsymbol{Z}_s'\boldsymbol{u}_E}{T},$$

$$= \left(\widehat{\boldsymbol{f}}'\,\widehat{\boldsymbol{Z}}_s\,\widehat{\boldsymbol{\Omega}}_s^{-1}\,\widehat{\boldsymbol{Z}}_s'\,\widehat{\boldsymbol{f}}\right)^{-1} \widehat{\boldsymbol{f}}'\,\widehat{\boldsymbol{Z}}_s\,\widehat{\boldsymbol{\Omega}}_s^{-1}\,\widehat{\boldsymbol{Z}}_s'\boldsymbol{y}_{\widehat{E}}. \tag{2.22}$$

where $\underset{(2+r)\times(2+r)}{\boldsymbol{W}_s}$ is an arbitrary positive definite weight matrix, but is also optimally set as $\widehat{\boldsymbol{W}}_s = \widehat{\boldsymbol{\Omega}}_s^{-1}$, where $\widehat{\boldsymbol{\Omega}}_s = \frac{1}{T}\sum_{t=1}^T \widehat{\boldsymbol{Z}}_{st}\widehat{\boldsymbol{Z}}_{st}'(y_{\widehat{E}t} - \widehat{\boldsymbol{f}}_t'\widehat{\boldsymbol{\theta}}^s_{GMM})^2$.[14] It is clear that (2.22) nests the FGIV estimator for the supply equation as a special case. As in the just identified case in (2.15), $\widehat{\boldsymbol{\theta}}^s_{GMM}$ in (2.22) depends on $\widehat{\boldsymbol{\Sigma}}_u^{-1}$, hence, will generally require an iterative estimation procedure. Algorithm 3 below generalizes Algorithm 1 by extending the joint estimation of the supply elasticity estimator and the precision matrix to the overidentified case for when $k_x = 0$. In view of Algorithm 2, Algorithm 3 can be further extended to the case when $k_x > 0$, but we omit the details for brevity.

Algorithm 3 Efficient GMM for ϕ^s (when $k_x = 0$):

- *Step 1:* Run PCA on (2.9) and obtain $\widehat{z}_t = \boldsymbol{S}'\widehat{\boldsymbol{Q}}\widetilde{\boldsymbol{y}}_{\cdot t}$ as the sample counterpart of (2.10).
- *Step 2:* Initialize $\widehat{\boldsymbol{\Sigma}}_u^{-1} = \boldsymbol{I}_N$.
- *Step 3:* Estimate (2.21) to obtain $\widehat{\varepsilon}$, initialize $\widehat{\boldsymbol{W}}_s = (\widehat{\boldsymbol{Z}}_s'\widehat{\boldsymbol{Z}}_s)^{-1}$ and obtain $\widehat{\boldsymbol{\theta}}^s_{2SLS}(\widehat{\boldsymbol{Z}}_s, \widehat{\boldsymbol{\Sigma}}_u^{-1})$.
- *Step 4:* Obtain $\boldsymbol{y}_{\widehat{E}}(\widehat{\boldsymbol{\Sigma}}_u^{-1})$.
- *Step 5:* Update $\widehat{\boldsymbol{W}}_s = \left(\frac{1}{T}\sum_{t=1}^T \widehat{\boldsymbol{Z}}_{st}\widehat{\boldsymbol{Z}}_{st}'\widehat{u}^2_{\widehat{E}t}\right)^{-1}$, where $\widehat{u}_{\widehat{E}t} = y_{\widehat{E}t} - \widehat{\boldsymbol{\theta}}^s_{GMM}(\widehat{\boldsymbol{Z}}_s, \widehat{\boldsymbol{\Sigma}}_u^{-1})'\widehat{\boldsymbol{f}}_t$ and construct $\widehat{\boldsymbol{\theta}}^s_{GMM}(\widehat{\boldsymbol{Z}}_s, \widehat{\boldsymbol{\Sigma}}_u^{-1})$ as the sample counterpart of (2.22).
- *Step 6:* Update $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ by inverting $\widehat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}$ defined in (2.17).
- *Step 7:* Iterate *Step 4* through *Step 6* until convergence.

In addition to efficiency gains, the efficient GMM estimators exhibit superior finite sample properties and are also robust to the GIV itself being a weak instrument. We illustrate these points in

[14] In the case of the demand elasticity estimator in (2.21) we use 2SLS residuals to construct $\widehat{\boldsymbol{\Omega}}_d$. However, we implement (2.22) via Algorithm 3 which, by iteration, renders the residuals used to construct $\widehat{\boldsymbol{\Omega}}_s$ to be GMM residuals.

greater detail in Remark 5 and Section 2.7.

The intuition for the overidentified estimators can be seen from observing the reduced form equation for (equilibrium) prices, $p_t = \frac{1}{\phi^d - \phi^s}(u_{St} + \boldsymbol{\lambda}_S' \boldsymbol{\eta}_t - \varepsilon_t)$. Clearly $\mathbb{E}(p_t \boldsymbol{\eta}_t) \neq 0$ and $\mathbb{E}(p_t \varepsilon_t) \neq 0$ and so instrumental relevancy is established. Thus, we are effectively back to the classical approach of finding exogenous supply shifters, in this case ε_t, to estimate the supply elasticity and finding exogenous demand shifters, in this case $\boldsymbol{\eta}_t$, to estimate the demand elasticity. With the exception that these shifters, $\boldsymbol{\eta}_t$ and ε_t are unobserved. In what follows, we show that estimating $\boldsymbol{\eta}_t$ and ε_t has a negligible effect on the limiting distributions of the estimators of demand and supply elasticities, respectively.

2.5 Assumptions

Below we lay out the assumptions needed to derive our main results. Assumption 1, Assumption 2 and Assumption 3 are standard in the literature; see, for example, [27], [37] and [41], but are relevant for a thorough understanding of the subsequent theorems. Whereas, Assumption 4 parts ii.) and iii.) are new so we provide more details.

Assumption 1 (Factor Error Structure) *The composite error term in (2.2) is assumed to admit an (approximate) factor structure representation $v_{it} := \boldsymbol{\lambda}_i' \boldsymbol{\eta}_t + u_{it}$, where $\boldsymbol{\eta}_t = \begin{pmatrix} \eta_{1t} & \dots & \eta_{rt} \end{pmatrix}'$ is an $r \times 1$ vector of latent common factors and $\boldsymbol{\lambda}_i = \begin{pmatrix} \lambda_{1i} & \dots & \lambda_{ri} \end{pmatrix}'$ is an $r \times 1$ vector of latent factor loadings. We assume the factors are pervasive in the sense that $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}/N$ converges to some $r \times r$ positive definite matrix.*

Assumption 2 (Strict Stationarity, Exponential Tails & Strong Mixing)

(A2i.) $\{\boldsymbol{\eta}_t, u_{it}, \varepsilon_t\}_{t \geq 1}$ *is strictly stationary and each with a zero mean.*

(A2ii.) *$\exists\, c_1, c_2 > 0$ with $\gamma_{\min}(\mathbf{\Sigma}_u) > c_2$, $\max_{j \leq N} ||\gamma_j|| < c_1$, $c_2 < \gamma_{\min}(\mathbb{cov}(\boldsymbol{\eta}_t)) \leq \gamma_{\max}(\mathbb{cov}(\boldsymbol{\eta}_t)) < c_1$.*

(A2iii.) *Exponential tail: $\exists\, r_1, r_2 > 0$ and $b_1, b_2 > 0$, such that for any $s > 0$, $i \leq N$ and $j \leq r$, $\mathbb{P}(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1})$, and $\mathbb{P}(|\eta_{t,j}| > s) \leq \exp(-(s/b_2)^{r_2})$.*

(A2iv.) *Strong Mixing: $\exists\, r_3,\ C > 0\ \forall\ T > 0,\ r_1^{-1} + r_2^{-1} + r_3^{-1} > 1,\ \sup_{A \in \mathcal{F}_{-\infty}^{0},\, B \in \mathcal{F}_T^{\infty}} |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(AB)| < \exp(-CT^{r_3})$, where $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_T^{\infty}$ denote the σ-algebras generated by $\{(\boldsymbol{\eta}_t, u_{it}, \varepsilon_t) : t < 0\}$ and $\{(\boldsymbol{\eta}_t, u_{it}, \varepsilon_t) : t > T\}$ respectively.*

Assumption 3 (Sparsity on $\mathbf{\Sigma}_u$) *Let $\mathbf{\Sigma}_u = (\sigma_{u,ij})$, for some $q \in [0, 1/2)$, define*

$$m_N = \max_{i \leq N} \sum_{j=1}^{N} |\sigma_{u,ij}|^q. \tag{2.23}$$

We require that there is $q \in [0, 1/2)$ such that $m_N \omega_{N,T}^{1-q} = o(1)$, where $\omega_{N,T} = \sqrt{\frac{\log(N)}{T}} + \frac{1}{\sqrt{N}}$.

Assumption 4 (Identification by GIV)

(A4i.) $\mathbb{E}(z_t u_{Et}) = \mathbb{E}(z_t \varepsilon_t) = \mathbb{E}(\boldsymbol{Z}_{st} u_{Et}) = \mathbb{E}(\boldsymbol{Z}_{dt} \varepsilon_t) = 0.$

(A4ii.) *The sizes $\mathcal{S}_1, \ldots, \mathcal{S}_N$ are drawn $i.i.d.$ from an arbitrary distribution for which the tail of the size distribution (i.e. above some threshold) follows a power law, with tail index, $\mu \in (0, 1)$*

$$\mathbb{P}(\mathcal{S} > s) = cs^{-\mu}.$$

The tail index μ determines the probability of observing extreme values. We assume that $\mathcal{S}_i$ is independent of u_{it}.

(A4iii.) *Suppose the sizes are ordered in decreasing fashion as such: $\mathcal{S}_{(1)} \geq \mathcal{S}_{(2)} \geq \cdots \geq \mathcal{S}_{(N-1)} \geq \mathcal{S}_{(N)}$, and we partition the cross-section as, $\mathcal{N}_{dominant} := \{1, \ldots, N_1\}$ and $\mathcal{N}_{fringe} :=$*

$\{N_1+1,\ldots,N\}$ *such that* $\mathcal{N}_{dominant} \cup \mathcal{N}_{fringe} := \mathcal{N}_{full}$*. Let* $S_i = \frac{\mathcal{S}_i}{\sum_{j=1}^{N} \mathcal{S}_j}$ *denote the normalized shares such that* $\sum_i S_i = 1$*. We assume* $\forall\, i \in \mathcal{N}_{fringe}$, $S_i = \mathcal{O}_p\left(\frac{1}{N}\right)$.

Remark 2 *The first condition gives us instrumental exogeneity for the FGIV and efficient GMM estimators. The second condition allows for instrumental relevance in the extension of a large N framework. An important implication of the second condition is that the Herfindahl index,* $h_{N,\mu}$*, has the following asymptotic property*

$$\sqrt{h_{N,\mu}} = ||\boldsymbol{S}||_2 = \begin{cases} \Theta_p(1) & \textit{for} \quad \mu \in (0,1), \\ \mathcal{O}_p(g_{N,\mu}) & \textit{for} \quad \mu \in [1,2), \end{cases}$$

with $\mathcal{O}_p(g_{N,\mu}) \gg 1/\sqrt{N}$*. The variance of the just identified estimators is inversely proportional to the Herfindahl index, that is* $\mathbb{V}(\widehat{\phi}^{j}_{FGIV}) = \mathcal{O}(h^{-1}_{N,\mu})$ *for* $j = s, d$*, reflecting the fact that the more concentrated the market, the more precise the GIV methodology will be and also reflecting the fact that if the Herfindahl converges to zero in the limit, the variance will diverge.*[15] *However, if* μ *is slightly greater than 1, theoretically identification breaks down for large* N *but in any finite sample the GIV could be relevant (precisely due to* $\mathcal{O}_p(g_{N,\mu}) \gg 1\sqrt{N}$*). Nevertheless, we rule this case out for the purpose of asymptotic inference.*[16] *The third condition is also a generalization of the so-called "granular" weights in the panel data literature, say* $\underset{N\times 1}{\boldsymbol{w}}$*, which are typically assumed to satisfy* $||\boldsymbol{w}||_2 = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ *and* $\frac{w_i}{||\boldsymbol{w}||_2} = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ $\forall\, i$*. The third condition allows the share vector to be partitioned into a dominant part and a fringe part. That is,* $\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_d' & \boldsymbol{S}_f' \end{pmatrix}'$ *where* $\boldsymbol{S}_d$ *is* $N_1 \times 1$*, is the dominant part and* $\boldsymbol{S}_f$ *is* $N_2 \times 1$*, is the fringe part; with* $N_1 + N_2 = N$*, the key*

[15] The derivation of the asymptotic behavior of $h_{N,\mu}$ can be found in Supplementary Appendix 2.12.2.

[16] For more details on instrumental relevance for large N, see Section 2.7.

being that one can easily show that $N_1(N) = N_1$ is fixed while $N_2(N) \to \infty$ as $N \to \infty$. This assumption can be empirically justified in concentrated markets, see Section 2.9 as an example; as well as mathematically justified, see [42].

Remark 3 *Taking the variance of the equilibrium price process (assuming the covariances to be zero for simplicity) we obtain $\mathbb{V}(p_t) = \frac{1}{(\phi^d - \phi^s)^2}(\mathbb{V}(u_{St}) + \mathbb{V}(\boldsymbol{\lambda}'_S \boldsymbol{\eta}_t) + \mathbb{V}(\varepsilon_t)) = \Theta(1)$, where the last equality follows by the second and the third conditions in Assumption 4, details can be found in Lemma 1 in the Appendix. Without these conditions, one would obtain the unsatisfactory result that $\mathbb{V}(p_t) = \mathcal{O}(N)$, that is, the variance of the price process is unbounded for each t as $N \to \infty$. Effectively, Assumption 4 allows the coexistence of a finite number of dominant units, in terms of size, whose cardinality can not grow with N, while at the same time allowing for a bounded variance for the aggregate endogenous variable p_t.*

2.6 Limiting Distributions

In this section, we first present the limiting distributions of the FGIV elasticity estimators, corresponding to (2.12) and (2.15) with Algorithm 1. We then move on to the limiting distributions of the efficient GMM elasticity estimators, corresponding to (2.21) and (2.22) with Algorithm 3.

Just identified demand elasticity. The just identified demand elasticity estimator in (2.12) is given by

$$\widehat{\phi}^d(\widehat{\boldsymbol{z}}) = \frac{\sum_{t=1}^T \widehat{z}_t d_t}{\sum_{t=1}^T \widehat{z}_t p_t} = \frac{\sum_{t=1}^T \sum_{i,j} S_i \widehat{Q}_{ij} \tilde{y}_{jt} d_t}{\sum_{t=1}^T \sum_{i,j} S_i \widehat{Q}_{ij} \tilde{y}_{jt} p_t}.$$

Hence,

$$\begin{aligned}\widehat{\phi}^d - \phi^d &= \left(\sum_t \widehat{z}_t p_t\right)^{-1}\left(\sum_t \widehat{z}_t \varepsilon_t\right), \\ &= \left(T^{-1}\sum_t z_t p_t + T^{-1}\sum_t (\widehat{z}_t - z_t) p_t\right)^{-1}\left(T^{-1}\sum_t z_t \varepsilon_t + T^{-1}\sum_t (\widehat{z}_t - z_t)\varepsilon_t\right).\end{aligned}$$

From above, it is apparent we need to show $\frac{1}{T}\sum_{t=1}^{T}(\widehat{z}_t - z_t)\varepsilon_t = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{Q}} - \boldsymbol{Q})\tilde{\boldsymbol{y}}_{\boldsymbol{\cdot t}}\varepsilon_t = o_p(1)$ and $\frac{1}{T}\sum_{t=1}^{T}(\widehat{z}_t - z_t)p_t = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{Q}} - \boldsymbol{Q})\tilde{\boldsymbol{y}}_{\cdot t}p_t = o_p(1)$. Indeed, we show in Lemma 2, in the Appendix, that

$$T^{-1}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{Q}} - \boldsymbol{Q})\tilde{\boldsymbol{y}}_{\cdot t}\varepsilon_t = \mathcal{O}_p\left(\frac{1}{\min\{N,T\}}\right) + \mathcal{O}_p\left(\frac{1}{\min\{N,\sqrt{NT}\}}\right) = o_p(1), \qquad (2.24)$$

$$T^{-1}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{Q}} - \boldsymbol{Q})\tilde{\boldsymbol{y}}_{\cdot t}p_t = \mathcal{O}_p\left(\frac{1}{\min\{N,T\}}\right) + \mathcal{O}_p\left(\frac{1}{\min\{N,\sqrt{NT}\}}\right) = o_p(1). \qquad (2.25)$$

Thus, making use of (2.24) and (2.25) we obtain

$$\widehat{\phi}^d - \phi^d = \left(\sum_t \widehat{z}_t p_t\right)^{-1}\left(\sum_t \widehat{z}_t \varepsilon_t\right) = \left(T^{-1}\sum_t z_t p_t\right)^{-1} T^{-1}\sum_t z_t\varepsilon_t + o_p(1). \qquad (2.26)$$

The order of the sampling error generally relies, in part, on the order of the Herfindahl. The order of the Herfindahl, in turn, critically depends on μ, the tail index of the size distribution. Results on the order of the Herfindahl as a function of the tail index parameter μ entails a total of six possible cases. The results can be found in Table 2.6 of Supplementary Appendix 2.12.2. However, for inference, we require $\mu \in (0,1)$ (regularly varying tails) or $\mu \to 0$ (slowly varying tails) as discussed in detail in the previous section's remarks. Given this, even after pinning down the order of the Herfindahl, the panel dimensions can distinguish more cases as seen above. Nevertheless, as (2.24), (2.25) and

(2.26) indicate, for consistency we have the following result:

Theorem 1 (Consistency of $\widehat{\phi}^d$) *Under Assumptions 1-4, as* $(N,T)\xrightarrow{j}\infty$*, we have*

$$\widehat{\phi}^d-\phi^d\xrightarrow{p}0. \tag{2.27}$$

Now, multiplying (2.26) by $\sqrt{T}$

$$\sqrt{T}(\widehat{\phi}^d-\phi^d)=\left(T^{-1}\sum_t z_t p_t\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_t z_t\varepsilon_t+\mathcal{O}_p\left(\frac{\sqrt{T}}{\min\{N,T\}}\right)+\mathcal{O}_p\left(\frac{\sqrt{T}}{\min\{N,\sqrt{NT}\}}\right)\right), \tag{2.28}$$

we can state the following result for the limiting distribution:

Theorem 2 (Limiting distribution for $\widehat{\phi}^d$) *Under Assumptions 1-4 as* $(N,T)\xrightarrow{j}\infty$*, we have that when* $N\geq T$ *or* $N<T$ *and* $\sqrt{T}/N\rightarrow 0$

$$\sqrt{T}(\widehat{\phi}^d-\phi^d)\xrightarrow{d}\mathcal{N}\left(0,\mathbb{v}_d\right), \tag{2.29}$$

where $\mathbb{v}_d:=\mathbb{m}_{zp}^{-2}\,\mathbb{v}_{z\varepsilon}$*,* $\mathbb{v}_{z\varepsilon}:=\mathbb{E}(z_t^2\varepsilon_t^2)$ *and* $\mathbb{m}_{zp}:=\mathbb{E}(z_tp_t)$*.*

$\mathbb{v}_{z\varepsilon}$ can be consistently estimated with

$$\widehat{\mathbb{v}}_{z\varepsilon}=\begin{cases}T^{-1}\sum_{t=1}^{T}\widehat{z}_t^2\,\widehat{\varepsilon}_t^2 & \text{HC,}\\ T^{-1}\sum_{t=1}^{T}\widehat{z}_t^2\,\widehat{\varepsilon}_t^2+2\cdot T^{-1}\sum_{j=1}^{m}\left(1-\dfrac{j}{m+1}\right)\sum_{t=j+1}^{T}\widehat{z}_t\widehat{\varepsilon}_t\widehat{z}_{t-j}\widehat{\varepsilon}_{t-j} & \text{HAC,}\end{cases} \tag{2.30}$$

where HC and HAC denote heteroskedasticity-consistent and heteroskedasticity and autocorrelation consistent estimators, respectively. Hence $\frac{\sqrt{T}(\widehat{\phi}^d - \phi^d)}{\widehat{\mathbb{v}}_d^{1/2}} \sim \mathrm{t}_{df} \xrightarrow{d} \mathcal{N}(0,1)$, where $\widehat{\mathbb{v}}_d^{1/2} = \widehat{\mathbb{m}}_{zp}^{-2} \widehat{\mathbb{v}}_{z\varepsilon}^{1/2}$, with $\widehat{\mathbb{m}}_{zp} = T^{-1}\sum_{t=1}^{T} \widehat{z}_t p_t$ also consistent for $\mathbb{m}_{zp}$. We will see in Section 2.8 that the asymptotic theory provides good approximations to the finite sample distribution.

Remark 4 *As in GK, we express $\mathbb{v}_d$ as inversely related to the Herfindahl, $h_{N,\mu}$, as claimed in Remark 2, for insights on the role of market concentration on precision of the GIV. Assuming conditional homoskedasticity of ε_t and homoskedasticity of u_{it}, we have that*

$$\mathbb{v}_{z\varepsilon} = \mathbb{E}(z_t^2 \varepsilon_t^2) = \sigma_\varepsilon^2 \cdot \sigma_{\tilde{u}}^2 \cdot \mathbb{E}(\boldsymbol{S}'\boldsymbol{Q}\boldsymbol{S}). \tag{2.31}$$

If $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}\ \forall i$, then there is no need to purge the factor structure through the loading space. That is, a simple cross-sectional demeaning transformation will suffice, $\boldsymbol{Q} = (\boldsymbol{I}_N - \tilde{\boldsymbol{\Lambda}}(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}})^{-1}\tilde{\boldsymbol{\Lambda}}') = (\boldsymbol{I}_N - \frac{\boldsymbol{\iota\iota}'}{\boldsymbol{N}})$. We can simplify equation (2.31) to (where we make use of the normalization that $\boldsymbol{S}'\boldsymbol{\iota} = 1$)

$$\mathbb{v}_{z\varepsilon} = \sigma_\varepsilon^2 \cdot \sigma_{\tilde{u}}^2 \cdot \left(\mathbb{E}(\boldsymbol{S}'\boldsymbol{S}) - \frac{1}{N}\right) = \sigma_\varepsilon^2 \cdot \underbrace{\sigma_{\tilde{u}}^2 \cdot \left(\mathbb{E}(h_{N,\mu}) - \frac{1}{N}\right)}_{\mathbb{E}(z_t^2)},$$

whereas, $\mathbb{m}_{zp} = \mathbb{E}(p_t z_t) \propto \mathbb{E}(z_t^2)$. Hence,

$$\mathbb{v}_d \propto \frac{\sigma_\varepsilon^2 \cdot \sigma_{\tilde{u}}^2 \cdot \left(\mathbb{E}(h_{N,\mu}) - \frac{1}{N}\right)}{\left[\sigma_{\tilde{u}}^2 \cdot \left(\mathbb{E}(h_{N,\mu}) - \frac{1}{N}\right)\right]^2} = \frac{\sigma_\varepsilon^2}{\sigma_{\tilde{u}}^2 \cdot \left(\mathbb{E}(h_{N,\mu}) - \frac{1}{N}\right)}. \tag{2.32}$$

Thus, the more concentrated the market, the more precise the estimator. See Section 2.7 for a more general treatment.

Just identified supply elasticity. For the just identified supply elasticity estimator in (2.15), upon convergence of Algorithm 1, we have that

$$\widehat{\phi}^s - \phi^s = \underbrace{\left(T^{-1}\widehat{\boldsymbol{z}}'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{p}\right)^{-1}}_{\widehat{A}^{-1}} \underbrace{T^{-1}\widehat{\boldsymbol{z}}'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{\eta}\cdot\boldsymbol{\lambda}_{\widehat{E}}}_{\widehat{B}} + \left(T^{-1}\widehat{\boldsymbol{z}}'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{p}\right)^{-1} \underbrace{T^{-1}\widehat{\boldsymbol{z}}'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{u}_{\widehat{E}}}_{\widehat{C}}. \tag{2.33}$$

We can write the scalars $\widehat{A}$, $\widehat{B}$ and $\widehat{C}$ as follows

$$\widehat{A} = T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_{\eta}\,\boldsymbol{p} + \underbrace{T^{-1}(\widehat{\boldsymbol{z}} - \boldsymbol{z})'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{p}}_{a_1} + \underbrace{T^{-1}\boldsymbol{z}'\,(\boldsymbol{M}_{\widehat{\eta}} - \boldsymbol{M}_{\eta})\,\boldsymbol{p}}_{a_2}, \tag{2.34}$$

$$\widehat{B} = \underbrace{T^{-1}(\widehat{\boldsymbol{z}} - \boldsymbol{z})'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{\eta}\lambda_{\widehat{E}}}_{b_1} + \underbrace{T^{-1}\boldsymbol{z}'\,(\boldsymbol{M}_{\widehat{\eta}} - \boldsymbol{M}_{\eta})\,\boldsymbol{\eta}\boldsymbol{\lambda}_{\widehat{E}}}_{b_2}, \tag{2.35}$$

$$\widehat{C} = T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_{\eta}\,\boldsymbol{u}_E + \underbrace{T^{-1}(\widehat{\boldsymbol{z}} - \boldsymbol{z})'\,\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{u}_{\widehat{E}}}_{c_1} + \underbrace{T^{-1}\boldsymbol{z}'\,(\boldsymbol{M}_{\widehat{\eta}} - \boldsymbol{M}_{\eta})\,\boldsymbol{u}_{\widehat{E}}}_{c_2} + \underbrace{T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_{\eta}\,(\boldsymbol{u}_{\widehat{E}} - \boldsymbol{u}_E)}_{c_3}. \tag{2.36}$$

It is shown in Lemma 3 of the Appendix that the terms a_i, b_i, c_j are $o_p(1)$ for $i = 1, 2$; $j = 1, 2, 3$, such that

$$\widehat{\phi}^s - \phi^s = \left(T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_{\eta}\,\boldsymbol{p}\right)^{-1}\left(T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_{\eta}\,\boldsymbol{u}_E\right) + o_p(1). \tag{2.37}$$

We can now state the following result:

Theorem 3 (Consistency of $\widehat{\phi}^s$) *Under Assumptions 1-4, as $(N, T) \xrightarrow{j} \infty$, we have*

$$\widehat{\phi}^s - \phi^s \xrightarrow{p} 0. \tag{2.38}$$

Now, multiplying (2.37) by $\sqrt{T}$

$$\sqrt{T}(\widehat{\phi}^s - \phi^s) = \left(T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_\eta\,\boldsymbol{p}\right)^{-1}\left(\frac{\boldsymbol{z}'\,\boldsymbol{M}_\eta\,\boldsymbol{u}_E}{\sqrt{T}} + \mathcal{O}_p\left(\frac{\sqrt{T}}{\min\{N,T\}}\right) + \mathcal{O}_p\left(\frac{\sqrt{T}}{\min\{N,\sqrt{NT}\}}\right) + \dots \right.$$
$$\left. \dots + \mathcal{O}_p\left(\frac{m_N\omega_{N,T}^{1-q}}{\sqrt{T}}\right)\right), \tag{2.39}$$

we can state the following result for the limiting distribution:

Theorem 4 (Limiting distribution for $\widehat{\phi}^s$) *Under Assumptions 1-4, as $(N,T)\xrightarrow{j}\infty$, we have that when $N \geq T$ or $N < T$ and $\sqrt{T}/N \to 0$*

$$\sqrt{T}(\widehat{\phi}^s - \phi^s) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{v}_s\right), \tag{2.40}$$

where $\mathbb{v}_s := \mathbb{m}_{z\tilde{p}}^{-2}\,\mathbb{v}_{zu}$, $\mathbb{v}_{zu} := \mathbb{E}(z_t^2\,(\boldsymbol{M}_\eta\,\boldsymbol{u}_E)_t^2)$ and $\mathbb{m}_{z\tilde{p}} = \mathbb{E}(z_t(\boldsymbol{M}_\eta\,\boldsymbol{p})_t)$.

$\mathbb{v}_{zu}$ can be consistently estimated with

$$\widehat{\mathbb{v}}_{zu} = \begin{cases} T^{-1}\sum_{t=1}^T \widehat{z}_t^2\,(\boldsymbol{M}_{\widehat{\eta}}\,\widehat{\boldsymbol{u}}_E)_t^2 & \text{HC,} \\ T^{-1}\sum_{t=1}^T \widehat{z}_t^2(\boldsymbol{M}_{\widehat{\eta}}\,\widehat{\boldsymbol{u}}_{\widehat{E}})_t^2 + 2\cdot T^{-1}\sum_{j=1}^m \left(1-\dfrac{j}{m+1}\right)\sum_{t=j+1}^T \widehat{z}_t(\boldsymbol{M}_{\widehat{\eta}}\,\widehat{\boldsymbol{u}}_{\widehat{E}})_t\widehat{z}_{t-j}(\boldsymbol{M}_{\widehat{\eta}}\,\widehat{\boldsymbol{u}}_{\widehat{E}})_{t-j} & \text{HAC.} \end{cases} \tag{2.41}$$

Hence $\frac{\sqrt{T}(\widehat{\phi}^s-\phi^s)}{\widehat{\mathbb{v}}_s^{1/2}} \sim \mathrm{t}_{df} \xrightarrow{d} \mathcal{N}(0,1)$, where $\widehat{\mathbb{v}}_s^{1/2} = \widehat{\mathbb{m}}_{z\tilde{p}}^{-2}\widehat{\mathbb{v}}_{zu}$, with $\widehat{\mathbb{m}}_{z\tilde{p}} = T^{-1}\sum_{t=1}^T \widehat{z}_t(\boldsymbol{M}_{\widehat{\eta}}\,\boldsymbol{p})_t$ consistent for $\mathbb{m}_{z\tilde{p}}$. We will see in Section 2.8 that asymptotic theory provides good approximations to the finite sample distribution.

Overidentified demand elasticity. For the overidentified demand elasticity estimator in (2.21), recall the estimated instrument matrix consists of $\widehat{\boldsymbol{Z}}_d = \begin{pmatrix} \widehat{\boldsymbol{z}}_{GIV} & \widehat{\boldsymbol{\eta}} \end{pmatrix}$. For the strong factors, $\widehat{\boldsymbol{\eta}}$, estimated via PCA, [5] showed that the generated regressors problem of [43] does not arise when both N and T are large. Thus, the sampling error in $\widehat{\boldsymbol{\eta}}$ is negligible in consideration of the limiting distribution of the overidentified demand elasticity estimate. In the previous section, we established that estimation of $\widehat{\boldsymbol{z}}_{GIV}$ is also negligible under regularity, we can then use standard asymptotic theory to also obtain asymptotic normality of the efficient GMM estimator in the case of demand since

$$\begin{aligned}\widehat{\phi}^d_{GMM} - \phi^d &= \left(\boldsymbol{p}'\, \widehat{\boldsymbol{Z}}_d\, \widehat{\boldsymbol{\Omega}}_d^{-1}\, \widehat{\boldsymbol{Z}}_d'\, \boldsymbol{p}\right)^{-1} \boldsymbol{p}'\, \widehat{\boldsymbol{Z}}_d\, \widehat{\boldsymbol{\Omega}}_d^{-1}\, \widehat{\boldsymbol{Z}}_d^{'}\, \boldsymbol{\varepsilon}, \\ &= \left(\boldsymbol{p}'\, \boldsymbol{Z}_d\, \boldsymbol{\Omega}_d^{-1}\, \boldsymbol{Z}_d^{'}\, \boldsymbol{p}\right)^{-1} \boldsymbol{p}'\, \boldsymbol{Z}_d\, \boldsymbol{\Omega}_d^{-1}\, \boldsymbol{Z}_d^{'}\, \boldsymbol{\varepsilon} + o_p(1). \end{aligned} \tag{2.42}$$

We can now state the following theorem:

Theorem 5 (Limiting distribution for $\widehat{\phi}^d_{GMM}$) *Under Assumptions 1-4, with $\mathbb{E}(\varepsilon_t \boldsymbol{\eta}_t) = 0\,\forall t$, as $(N,T) \xrightarrow{j} \infty$, we have that when $N \geq T$ or $N < T$ and $\sqrt{T}/N \to 0$*

$$\sqrt{T}(\widehat{\phi}^d_{GMM} - \phi^d) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}(\widehat{\phi}^d_{GMM})\right), \tag{2.43}$$

where

$$\mathbb{V}(\widehat{\phi}^d_{GMM}) = \left(\boldsymbol{m}'_{Zdp}\, \boldsymbol{\Omega}_d^{-1} \boldsymbol{m}_{Zdp}\right)^{-1}, \tag{2.44}$$

with $\boldsymbol{m}_{Zdp} = \mathbb{E}(\boldsymbol{Z}_{dt}\, p_t)$ and $\boldsymbol{\Omega}_d = \operatorname{plim} T^{-1} \sum_{t=1}^T \widehat{\boldsymbol{Z}}_{dt} \widehat{\boldsymbol{Z}}_{dt}' (d_t - p_t \widehat{\phi}^d_{2SLS})^2$.

$\mathbb{V}(\widehat{\phi}^d_{GMM})$ can be consistently estimated using $2SLS$ residuals with

$$\widehat{\mathbb{V}}(\widehat{\phi}^d_{GMM}) = \left(\frac{\boldsymbol{p}'\widehat{\boldsymbol{Z}}_d}{T} \, \widehat{\boldsymbol{\Omega}}_d^{-1} \, \frac{\widehat{\boldsymbol{Z}}_d'\boldsymbol{p}}{T} \right)^{-1}, \tag{2.45}$$

where $\widehat{\boldsymbol{\Omega}}_d = T^{-1}\sum_{t=1}^{T} \widehat{\boldsymbol{Z}}_{dt}\widehat{\boldsymbol{Z}}_{dt}' \, (d_t - p_t\widehat{\phi}^d_{2SLS})^2$. It is well known that $\mathbb{V}(\widehat{\phi}^d_{GMM})$ attains the semiparametric efficiency bound, as shown by [44], which reduces to (2.44) in the linear model. Standard overidentification tests can be carried out since

$$J_d = T \cdot \left(T^{-1} \sum_{t=1}^{T} \boldsymbol{Z}_{dt} \, \varepsilon_t(\widehat{\phi}^d_{GMM}) \right)' \widehat{\boldsymbol{\Omega}}_d^{-1} \left(T^{-1} \sum_{t=1}^{T} \boldsymbol{Z}_{dt} \, \varepsilon_t(\widehat{\phi}^d_{GMM}) \right) \xrightarrow{d} \chi^2_{df_d}, \tag{2.46}$$

where the degrees of freedom is given by $df_d = (1+r) - k_d = r$ and $k_d = 1$ is the number of endogenous regressors. We will see that simulation evidence shows that the size of the J-test is near the nominal size when the true r is used and when $rmax > r$ factors are used; which is important in empirical work when r is typically estimated and it is generally known that an overestimate of r is preferred in order to prevent an effect akin to omitted variable bias, see [45] who formalize this notion.

Overidentified supply elasticity. The full instrument matrix for the overidentified supply elasticity estimator in (2.22) consists of $\widehat{\boldsymbol{Z}}_s = \begin{pmatrix} \widehat{\boldsymbol{z}}_{GIV} & \widehat{\boldsymbol{\varepsilon}} & \widehat{\boldsymbol{\eta}} \end{pmatrix}$, (recall the factors self instrument here, as they are part of the supply equation). We show that the sampling error in $\widehat{\boldsymbol{Z}}_s$ is indeed negligible. This is again due to both large N and T. As a result, $\underset{(2+r)\times(2+r)}{\boldsymbol{\Omega}_s} =$ plim $T^{-1}\sum_{t=1}^{T} \widehat{\boldsymbol{Z}}_{st}\widehat{\boldsymbol{Z}}_{st}'(y_{\widehat{E}t} - \widehat{\boldsymbol{f}}_t'\widehat{\boldsymbol{\theta}}^s_{GMM})^2$ is sufficient when constructing the efficient weighting matrix, even though it does not take the sampling error in our estimate of ϕ^d into account (since

$\widehat{\varepsilon} = \varepsilon(\widehat{\phi}^d_{GMM})$). That is

$$\widehat{\boldsymbol{\theta}}^s_{GMM} - \boldsymbol{\theta}^s = \left(\boldsymbol{f}'\,\widehat{\boldsymbol{Z}}_s\,\widehat{\boldsymbol{\Omega}}_s^{-1}\,\widehat{\boldsymbol{Z}}_s'\boldsymbol{f}\right)^{-1}\boldsymbol{f}'\,\widehat{\boldsymbol{Z}}_s\,\widehat{\boldsymbol{\Omega}}_s^{-1}\,\widehat{\boldsymbol{Z}}_s'\boldsymbol{u}_E,$$
$$= \left(\boldsymbol{f}'\,\boldsymbol{Z}_s\,\boldsymbol{\Omega}_s^{-1}\,\boldsymbol{Z}_s'\boldsymbol{f}\right)^{-1}\boldsymbol{f}'\,\boldsymbol{Z}_s\,\boldsymbol{\Omega}_s^{-1}\,\boldsymbol{Z}_s'\boldsymbol{u}_E + o_p(1). \tag{2.47}$$

We can now state the following theorem:

Theorem 6 (Limiting distribution for $\widehat{\boldsymbol{\theta}}^s_{GMM}$) *Under Assumptions 1-4, as $(N,T)\xrightarrow{j}\infty$, we have that when $N \geq T$ or $N < T$ and $\sqrt{T}/N \to 0$*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}^s_{GMM} - \boldsymbol{\theta}^s) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{V}(\widehat{\boldsymbol{\theta}}^s_{GMM})\right), \tag{2.48}$$

where

$$\mathbb{V}(\widehat{\boldsymbol{\theta}}^s_{GMM}) = \left(\boldsymbol{m}'_{Z_sf}\,\boldsymbol{\Omega}_s^{-1}\boldsymbol{m}_{Z_sf}\right)^{-1}, \tag{2.49}$$

with $\boldsymbol{m}_{Z_sf} = \mathbb{E}(\boldsymbol{Z}_{st}\boldsymbol{f}_t')$ and $\boldsymbol{\Omega}_s = \operatorname{plim} T^{-1}\sum_{t=1}^T \widehat{\boldsymbol{Z}}_{st}\widehat{\boldsymbol{Z}}_{st}'(y_{\widehat{E}t} - \widehat{\boldsymbol{f}}_t'\widehat{\boldsymbol{\theta}}^s_{GMM})^2$.

$\mathbb{V}(\widehat{\boldsymbol{\theta}}^s_{GMM})$ can be consistently estimated using GMM residuals with

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\theta}}^s_{GMM}) = \left(\frac{\widehat{\boldsymbol{f}}'\,\widehat{\boldsymbol{Z}}_s}{T}\,\widehat{\boldsymbol{\Omega}}_s^{-1}\,\frac{\widehat{\boldsymbol{Z}}_s'\widehat{\boldsymbol{f}}}{T}\right)^{-1}, \tag{2.50}$$

where $\widehat{\boldsymbol{\Omega}}_s = T^{-1}\sum_{t=1}^T \widehat{\boldsymbol{Z}}_{st}\widehat{\boldsymbol{Z}}_{st}'\,(y_{\widehat{E}t} - \widehat{\boldsymbol{f}}_t'\widehat{\boldsymbol{\theta}}^s_{GMM})^2$. Just as in the case of the overidentified demand elasticity estimator, (2.49) achieves the semiparametric efficiency bound. Overidentification tests can be carried out since

$$J_s = T \cdot \left(T^{-1} \sum_{t=1}^{T} \boldsymbol{Z}_{st}\, u_{\widehat{E}t}(\widehat{\boldsymbol{\theta}}^s_{GMM}) \right)' \, \widehat{\boldsymbol{\Omega}}_s^{-1} \, \left(T^{-1} \sum_{t=1}^{T} \boldsymbol{Z}_{st}\, u_{\widehat{E}t}(\widehat{\boldsymbol{\theta}}^s_{GMM}) \right) \xrightarrow{d} \chi^2_{df_s}, \tag{2.51}$$

where the degrees of freedom are given by $df_s = 2 - k_s = 2 - 1 = 1$ and $k_s = 1$ is the number of endogenous regressors for the supply equation.

Remark 5 *The asymptotic distribution of the FGIV and efficient GMM estimators of demand and supply elasticities are established. However, the finite sample moments of these estimators, are unbounded to different degrees. The extensive literature on the classic simultaneous equations model has documented this result in many forms, see [46], [47], [48], [49], [50], [51], [52] and [53]. A complete representation of the above results was given by [54]. Kinal's result for 2SLS states that, if the dependent variable, explanatory variables and instruments are jointly normal, then* $\mathbb{E}||\widehat{\phi}^j_{2SLS}||^m < \infty$ *for* $m < \ell_j - k_j + 1$*,* $j = d, s$*, where* ℓ_j *is the number of instruments and* k_j *is the number of endogenous regressors.*

Thus, the FGIV estimators for supply and demand exhibit no bounded absolute moments since $\ell_j = k_j = 1$*,* $j = d, s$*. Whereas, the efficient GMM estimators exhibits* $\ell_d - k_d = (1+r) - 1 = r$ *bounded absolute moments in finite samples for the case of demand and* $\ell_s - k_s = 2 - 1 = 1$ *bounded absolute moment in finite samples for the case of supply. Hence, the efficient GMM elasticity estimators (overidentified FGIV estimators) exhibit superior finite sample properties relative to their (just identified) FGIV counterparts. Of course, in general, just identified instrumental variables estimators (with strong instruments) exhibit nice properties asymptotically.*

2.7 Weak Instruments

The classical weak instruments framework introduced by [55] has its analog in this framework. Interestingly, here the "weak" aspect is partially linked to the Herfindahl index without making the usual *local-to-zero* assumption as in [55]. Moreover, the traditional notion of local-to-zero with $\frac{1}{\sqrt{T}}$ scaling which matches the rate of convergence of the estimator need not necessarily apply here for weak instruments to arise. More specifically, the locality to zero can be expressed as decaying functions of N, except in the case of $\mu \in (0,1)$, which we require for inference under our maintained strong instruments assumption; whereas the rate of convergence is at the $\sqrt{T}$ rate. To make things more clear, it is useful to see the reduced form, equilibrium price equation again. Recall from (2.3), we have that $p_t = \dfrac{1}{\phi^d - \phi^s}\left(u_{St} + \boldsymbol{\lambda}_S' \boldsymbol{\eta}_t - \varepsilon_t\right)$. Thus, it is clear that for finite N, $\mathbb{Cov}(p_t, z_t) > 0$, which automatically renders the GIV as relevant. However, for large N, writing $z_t = \boldsymbol{S}'\boldsymbol{Q}\tilde{\boldsymbol{u}}_{\cdot t}$ we observe that

$$\mathbb{V}(\boldsymbol{S}'\boldsymbol{Q}\tilde{\boldsymbol{u}}_{\cdot t}) = \mathbb{E}(\boldsymbol{S}'\boldsymbol{Q}\mathbb{E}(\tilde{\boldsymbol{u}}_{\cdot t}\tilde{\boldsymbol{u}}_{\cdot t}'|\boldsymbol{\Gamma})\boldsymbol{Q}\boldsymbol{S}) = \mathbb{E}(\boldsymbol{S}'\boldsymbol{Q}\,\boldsymbol{\Sigma}_{\tilde{u}}\,\boldsymbol{Q}\boldsymbol{S}), \tag{2.52}$$

where $\boldsymbol{\Sigma}_{\tilde{u}} := \mathbb{E}(\tilde{\boldsymbol{u}}_{\cdot t}\tilde{\boldsymbol{u}}_{\cdot t}')$. The term inside the expectation can be simplified to

$$\begin{aligned}\boldsymbol{S}'\boldsymbol{Q}\,\boldsymbol{\Sigma}_{\tilde{u}}\,\boldsymbol{Q}\boldsymbol{S} &= \boldsymbol{S}'\boldsymbol{Q}\,\boldsymbol{\Sigma}_{u}\,\boldsymbol{Q}\boldsymbol{S} + \mathcal{O}_p\left(N^{-1}\right) \\ &\leq \boldsymbol{S}'\boldsymbol{Q}\boldsymbol{S}\,\gamma_{max}(\boldsymbol{\Sigma}_u) + \mathcal{O}_p\left(N^{-1}\right) \leq \boldsymbol{S}'\boldsymbol{S}\cdot\gamma_{max}(\boldsymbol{\Sigma}_u)\cdot\gamma_{max}(\boldsymbol{Q}) + \mathcal{O}_p\left(N^{-1}\right) \\ &= \mathcal{O}\left(1\right)h_{N,\mu} + \mathcal{O}_p\left(N^{-1}\right),\end{aligned} \tag{2.53}$$

where we make use of $\gamma_{max}(\mathbf{\Sigma}_u) = \mathcal{O}(1)$ and the fact that a symmetric idempotent matrix, such as $\boldsymbol{Q}$, has eigenvalues of 0 or 1 and so $\gamma_{max}(\boldsymbol{Q}) = 1$. Taken together, (2.52) and (2.53) imply

$$\mathbb{V}(z_t) \leq \mathcal{O}(1)\, \mathbb{E}(h_{N,\mu}) + \mathcal{O}\left(N^{-1}\right). \tag{2.54}$$

As such, only when we are in a tail regime indexed by $\mu \in (0,1)$ do we avoid the locality to zero. For example, when $\mu > 2$, we have that $\mathbb{V}(\boldsymbol{S}'\boldsymbol{Q}\tilde{\boldsymbol{u}}_{\cdot t}) \leq \mathcal{O}(1)\, \mathbb{E}(h_{N,\,\mu>2}) = \mathcal{O}\left(\frac{1}{N}\right)$. As a result, when $\mu > 2$, we have that $z_t = \boldsymbol{S}'\boldsymbol{Q}\tilde{\boldsymbol{u}}_{\cdot t} = \mathcal{O}_p\left(\frac{1}{\sqrt{N}}\right)$, so our equilibrium price equation simplifies to the following large N representation

$$p_t = \frac{1}{\phi^d - \phi^s}\left(\boldsymbol{\lambda}_S'\boldsymbol{\eta}_t - \varepsilon_t\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N}}\right). \tag{2.55}$$

This would render the GIV as very weak since $\mathbb{C}\mathrm{ov}(p_t, u_{St}) = \mathbb{C}\mathrm{ov}(p_t, z_t) = \mathcal{O}\left(\frac{1}{N}\right)$. Note that the $\mathbb{C}\mathrm{ov}(p_t, u_{St})$ and $\mathbb{C}\mathrm{ov}(p_t, z_t)$ are of the same order precisely because $\gamma_{max}(\boldsymbol{Q}) = 1$. (2.55) is effectively the relationship that was exploited by [26], who assumed the so-called "granular" weights of order $\mathcal{O}\left(\frac{1}{N}\right)$ and used this weak correlation for large N to ultimately deduce that prices can be treated as weakly exogenous.[17]

Consider the well documented and empirically relevant case where μ is just above 1 (Zipf's law corresponds to $\mu = 1$); when $\mu \in (1,2)$ we have $h_{N,\,\mu\in(1,2)} = \mathcal{O}_p\left(1/(N^{2-\frac{2}{\mu}})\right)$.

[17] "Granular" has a different definition in the panel data literature, which is referring to properties of weights and heuristically, rules out the existence of dominant units, see e.g., [26] and Remark 2. On the contrary, our usage of the term "granular" follows [7] and is essentially referring to the existence of dominant cross-sectional units, see Section 2.1.

So,

$$p_t = \frac{1}{\phi^d - \phi^s} \left(\boldsymbol{\lambda}'_S \boldsymbol{\eta_t} - \varepsilon_t \right) + \mathcal{O}_p \left(\frac{1}{N^{1-\frac{1}{\mu}}} \right). \tag{2.56}$$

Therefore, even though $\mathbb{Cov}(p_t, z_t) = \mathcal{O}\left(1/(N^{2-\frac{2}{\mu}})\right)$, it is in fact decaying to zero so slowly for μ near 1, that this potentially corresponds to a highly relevant instrument in any finite sample. That is, $\mathbb{Cov}(p_t, z_t) = \mathcal{O}\left(1/(N^{2-\frac{2}{\mu}})\right)$, is potentially consistent with z_t accounting for large fractions of aggregate variation, see [7].

However, the case we theoretically entertain, for consistency and valid asymptotic inference, requires $\mu \in (0, 1)$, which in conjunction with the additional regularity assumptions, renders $\mathbb{Cov}(p_t, z_t) = \Theta(1)$ even as $N \to \infty$.

Rothemberg Representations. Moreover, to further assess the likelihood of weak instruments, we can analyze the efficient GMM estimator of the demand elasticity which uses both the GIV and the factors as instruments and for comparison we can analyze the just identified demand elasticity estimator which uses only the GIV as an instrument. We analyze the overidentified case with conditional homoskedasticity (assuming only for remainder of this section). Define the projection matrix $\boldsymbol{P}_{Z_d} = \boldsymbol{Z}_d(\boldsymbol{Z}'_d\boldsymbol{Z}_d)^{-1}\boldsymbol{Z}'_d$, the 2SLS estimator takes the form

$$\widehat{\phi}^d_{GMM} - \phi^d = \frac{\boldsymbol{p}' \, \boldsymbol{P}_{Z_d} \, \boldsymbol{\varepsilon}}{\boldsymbol{p}' \, \boldsymbol{P}_{Z_d} \, \boldsymbol{p}}. \tag{2.57}$$

Write the structural and reduced form equations as

$$\begin{aligned} \boldsymbol{d} &= \boldsymbol{p}\,\phi^d + \boldsymbol{\varepsilon} \\ \boldsymbol{p} &= \boldsymbol{z}\boldsymbol{\pi}' + \boldsymbol{v}, \end{aligned} \tag{2.58}$$

where $\underset{T\times(1+r)}{\boldsymbol{z}} = \begin{pmatrix} \boldsymbol{u}_S & \boldsymbol{\eta} \end{pmatrix}$, $\underset{(1+r)\times 1}{\boldsymbol{\pi}} = \begin{pmatrix} \frac{1}{\phi^d-\phi^s} & \frac{1}{\phi^d-\phi^s}\cdot\boldsymbol{\lambda}_S' \end{pmatrix}'$ and $\underset{T\times 1}{\boldsymbol{v}} = \frac{1}{\phi^d-\phi^s}\cdot\boldsymbol{\varepsilon}$.

Remark 6 *The difference between $\boldsymbol{z}$ in (2.58) and our actual instrument, $\boldsymbol{Z}_d$, boils down to the difference between their first columns, $\boldsymbol{Z}_d[\cdot,1] = \boldsymbol{z}_{GIV} = \tilde{\boldsymbol{u}}_{\cdot\cdot}\boldsymbol{Q}\boldsymbol{S}$ and $\boldsymbol{z}[\cdot,1] = \boldsymbol{u}_S = \boldsymbol{u}_{\cdot\cdot}\boldsymbol{S}$. In the case of the demand equation, $\boldsymbol{u}_S$ is ideal, whereas $\boldsymbol{z}_{GIV}$ is a proxy. The reason the proxy is used is simply due to a simpler theoretical exposition than a direct estimate for the ideal. Indeed $\boldsymbol{z}_{GIV}$ is in fact a good proxy. For example, the correlation between $\boldsymbol{u}_S$ and $\boldsymbol{z}_{GIV}$ is over 90% regardless of the complexity of our DGP in Monte Carlo simulations even for small configurations of (N,T). Moreover, in the case of the supply equation, $\boldsymbol{u}_S$ is no longer valid, whereas $\boldsymbol{z}_{GIV}$ is; see (2.13). Mathematically, $\boldsymbol{S}'\boldsymbol{u}_{\cdot t} - \boldsymbol{S}'\boldsymbol{Q}\tilde{\boldsymbol{u}}_{\cdot t} = \boldsymbol{S}'\boldsymbol{P}_{\tilde{\Lambda}}\boldsymbol{u}_{\cdot t} = \boldsymbol{S}'\boldsymbol{P}_{\tilde{\Lambda}}\boldsymbol{P}_{\tilde{\Lambda}}\boldsymbol{u}_{\cdot t}$ for each t, where $\boldsymbol{P}_{\tilde{\Lambda}}$ is the symmetric and idempotent projection matrix in the demeaned loading space. Hence, $\boldsymbol{S}'\boldsymbol{P}_{\tilde{\Lambda}}\boldsymbol{P}_{\tilde{\Lambda}}\boldsymbol{u}_{\cdot t}$ is zero when the loadings and the share vector are asymptotically uncorrelated and/or the loadings and idiosyncratic errors are asymptotically uncorrelated; which explains why our simulations exhibit near perfect correlation.*

Then, it follows from [56] that (2.57) has the following illustrative representation

$$\mu_{d,GMM}(\widehat{\phi}^d_{GMM} - \phi^d) = \left(\frac{\sigma^2_\varepsilon}{\sigma^2_v}\right)^{\frac{1}{2}} \frac{X + (\omega_1/\mu_{d,GMM})}{1 + 2Y/\mu_{d,GMM} + (\omega_2/\mu^2_{d,GMM})}, \tag{2.59}$$

where $X = \boldsymbol{\pi}'\boldsymbol{z}'\boldsymbol{P}_{Z_d}\boldsymbol{\varepsilon}/(\sigma^2_\varepsilon\boldsymbol{\pi}'\boldsymbol{z}'\boldsymbol{z}\boldsymbol{\pi})^{\frac{1}{2}}$ and $Y = \boldsymbol{\pi}'\boldsymbol{z}'\boldsymbol{P}_{Z_d}\boldsymbol{v}/(\sigma^2_v\boldsymbol{\pi}'\boldsymbol{z}'\boldsymbol{z}\boldsymbol{\pi})^{\frac{1}{2}}$ are bivariate standard normal variates with correlation coefficient ρ. The random variable $\omega_1 = \boldsymbol{v}'\boldsymbol{P}_{Z_d}\boldsymbol{\varepsilon}/(\sigma_\varepsilon^2\sigma_v^2)^{\frac{1}{2}}$ has mean equal to rank$(P_{Z_d})\rho = (r+1)\rho$ and variance equal to $(r+1)(1+\rho^2)$. The random variable $\omega_2 = \boldsymbol{v}'\boldsymbol{P}_{Z_d}\boldsymbol{v}/\sigma_v^2$ has mean equal to rank$(\boldsymbol{P}_{Z_d}) = (r+1)$ and variance equal to $2(r+1)$. Finally, $\mu_{d,GMM}$ is the square root of the so-called concentration parameter $\mu_{d,GMM}^2 = \boldsymbol{\pi}'\boldsymbol{z}'\boldsymbol{z}\boldsymbol{\pi}/\sigma_v^2$ for the demand equation. $\mu_{d,GMM}$ plays the role of $\sqrt{T}$, that is, when $\mu_{d,GMM}$ is large, $\mu_{d,GMM}(\widehat{\phi}_{GMM}^d - \phi^d)$ is well approximated by a $\mathcal{N}(0,1)$ variate. Large values of $\mu_{d,GMM}$ are consistent with large values of T, i.e., our typical large sample approximations. However, large values of $\mu_{d,GMM}$ are also consistent with small values of σ_v^2, regardless of the value of T, i.e., small-σ asymptotics, as introduced originally by [57]. More insights can be gained by simplifying the concentration parameter for the demand elasticity

$$\mu_{d,GMM}^2 = \boldsymbol{\pi}'\boldsymbol{z}'\boldsymbol{z}\boldsymbol{\pi}/\sigma_v^2 = \frac{1}{\sigma_v^2}\cdot\begin{pmatrix}\pi_1 & \boldsymbol{\pi_2'}\end{pmatrix}\begin{pmatrix}\boldsymbol{u}_S'\boldsymbol{u}_S & \boldsymbol{u}_S'\boldsymbol{\eta} \\ \boldsymbol{\eta}'\boldsymbol{u}_S & \boldsymbol{\eta}'\boldsymbol{\eta}\end{pmatrix}\begin{pmatrix}\pi_1 \\ \boldsymbol{\pi_2}\end{pmatrix} \approx \frac{\boldsymbol{u}_S'\,\boldsymbol{u}_S + \boldsymbol{\lambda}_S'\boldsymbol{\lambda}_S}{\sigma_\varepsilon^2}, \quad (2.60)$$

where the approximation is due to ignoring the terms involving $\boldsymbol{\eta}'\boldsymbol{u}_S$, which are zero only in expectation. (2.60) is very intuitive, if the proportion of the volatility in the GIV *and* size-weighted common components dominate the volatility of the demand shocks, so that the ratio in (2.60) is large, then the concentration parameter $\mu_{d,GMM}$ will be large and one should expect good approximations to the finite sampling distributions.

On the other hand, when only the GIV is used as an instrument, if we redefine $\underset{T\times 1}{\boldsymbol{z}} = \boldsymbol{u}_S$, $\underset{1\times 1}{\boldsymbol{\pi}} = \frac{1}{\phi^d - \phi^s}$ and $\underset{T\times 1}{\boldsymbol{v}} = \frac{1}{\phi^d - \phi^s}\cdot(\boldsymbol{\varepsilon} + \boldsymbol{\eta}\boldsymbol{\lambda}_S)$ from (2.58) and simply follow the logic above through (2.60), we arrive at the following concentration parameter for the FGIV estimator

$$\mu_{d,FGIV}^2 \approx \frac{\boldsymbol{u}_S'\boldsymbol{u}_S}{\boldsymbol{\lambda}_S'\boldsymbol{\lambda}_S + \sigma_\varepsilon^2}. \quad (2.61)$$

Thus, by inspection of (2.60) against (2.61) we can see that in the case of the just identified FGIV estimator, we would need the volatility of just the GIV to drive up the ratio of the concentration parameter, $\mu^2_{d,GIV}$, and the size-weighted common component would be working against us (in the denominator), in this case, instead of working for us as in $\mu^2_{d,GMM}$.

Although the literature on granularity has demonstrated that idiosyncratic shocks alone can be quite volatile, in this context, we advocate starting with the efficient GMM estimators, since the J-test is well sized as illustrated with simulation evidence, because the efficient GMM estimators can exhibit substantially improved finite sample properties relative to the just identified estimators and are less likely to suffer from weak instrument issues as well.

2.8 Monte Carlo

We simulate the following panel simultaneous equations system with latent factor structure that was analyzed in the theoretical sections:

$$d_t = \phi^d p_t + \varepsilon_t,\ y_{it} = \phi^s p_t + \lambda_{1i}\eta_{1t} + \lambda_{2i}\eta_{2t} + u_{it},\ y_{\boldsymbol{S}t} = d_t,$$
$$p_t = \frac{1}{\phi^d - \phi^s}\left(u_{St} + \boldsymbol{\lambda}'_S \boldsymbol{\eta_t} - \varepsilon_t\right),\ \mathcal{S}_i = \left(\frac{i}{N}\right)^{-\frac{1}{\mu}},\ S_i = \frac{\mathcal{S}_i}{\sum_{j=1}^N \mathcal{S}_j}.$$

We consider two sets of simulated experiments. In Design 1, we let u_{it} be $i.i.d.$ to establish a set of baseline results. In Design 2, we allow for sparse cross-sectional dependence in u_{it}. In addition, in unreported simulations, we simulate $z_{t,GIV}$ to be a weak instrument to illustrate that the efficient GMM estimators are robust to this as they optimally shift their weights away from this point of weakness, whereas the just identified estimators of GK and our FGIV will substantially deteriorate in their performances.

Design 1 - u_{it} *i.i.d.* case. We set $\phi^s = 0.1$ and $\phi^d = -0.3$. We draw the supply factors and loadings as, $\underset{T\times r}{\boldsymbol{\eta}} \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_r)$ and $\underset{N\times r}{\boldsymbol{\Lambda}} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\Lambda^2 \boldsymbol{I}_N)$, respectively, with $r = 2$.[18] We draw the idiosyncratic supply shocks as $\underset{T\times N}{\boldsymbol{u}} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2 \boldsymbol{I}_N \otimes \boldsymbol{I}_T)$ and aggregate demand shocks as $\underset{T\times 1}{\boldsymbol{\varepsilon}} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 \boldsymbol{I}_T)$.

Design 2 - u_{it} non-*i.i.d.* case. Everything is identical to Design 1, except that we no longer set $\boldsymbol{\Sigma}_u = \sigma_u^2 \boldsymbol{I}_N$ for each $t = 1, \ldots, T$. We generate a non-diagonal banded covariance matrix; as such, it satisfies the sparsity requirement from Assumption 3.[19] We consider the following banded idiosyncratic covariance matrix with cross-sectional dependence and heteroskedasticity

$$\sigma_{u,ij} = \begin{cases} \tau^{|i-j|}\sqrt{\sigma_{u,i}\sigma_{u,j}} & |i-j| \leq k; k \geq 0, \\ 0 & |i-j| > k, \end{cases} \tag{2.62}$$

with bandwidth $k = 3$ and $\sigma_{u,i}^2$ are drawn from $\mathcal{U}[0.5, 1]$.

Target parameterizations. The variance of the price process takes the following form $\mathbb{V}(p_t) = c \cdot (\mathbb{V}(u_{St}) + \mathbb{V}(\boldsymbol{\lambda}_S' \boldsymbol{\eta}_t) + \mathbb{V}(\varepsilon_t))$, where $c = \dfrac{1}{(\phi^d - \phi^s)^2}$. This conveniently allows us to parameterize the relative volatilities of the various components of equilibrium prices. We parameterize the individual variances, σ_u^2 (for Design 1), σ_Λ^2 and σ_ε^2 such that $\psi_u := \dfrac{\mathbb{V}\left(\sqrt{c} \cdot u_{St}\right)}{\mathbb{V}(p_t)} \in (0.15, 0.35)$, $\psi_{u+\eta} := \dfrac{\mathbb{V}\left(\sqrt{c} \cdot (u_{St} + \boldsymbol{\lambda}_S' \boldsymbol{\eta}_t)\right)}{\mathbb{V}(p_t)} \in (0.45, 0.65)$, and $\psi_{u+\varepsilon} := \dfrac{\mathbb{V}\left(\sqrt{c} \cdot (u_{St} + \varepsilon_t)\right)}{\mathbb{V}(p_t)} \in (0.45, 0.65)$.[20] In Design 1, we achieve an average across simulations of $\bar{\psi}_u \approx 0.23$, $\bar{\psi}_{u+\eta} \approx 0.58$

[18] The results do not change significantly if we draw the loadings from a Uniform distribution with non-zero mean.

[19] In unreported simulations, we find that the results do not change significantly if we generate a dense, non-diagonal $\boldsymbol{\Sigma}_u$, such as one arising from a cross-sectional AR(p) process. This is because although the cross-sectional AR(p) generates a dense matrix, it is not too dense since the off-diagonals decay exponentially fast to 0 as $|i - j| \to \infty$.

[20] The interval for ψ_u is consistent with the literature on granularity, which has documented the proportion of aggregate fluctuations traced back to idiosyncratic shocks falling in this specified range.

and $\bar{\psi}_{u+\varepsilon} \approx 0.65$ which implies that $\bar{\psi}_{\eta} = 0.35$ and $\bar{\psi}_{\varepsilon} = 0.42$. That is, the idiosyncratic shocks are not the dominating force in terms of observed price volatility; however, their granular role is still substantial enough to draw inferences from when used as instruments. In Design 2, we achieve an average across simulations of $\bar{\psi}_u \approx 0.27$, $\bar{\psi}_{u+\eta} \approx 0.64$ and $\bar{\psi}_{u+\varepsilon} \approx 0.63$.

Let $\widehat{\phi}^j(m)$, $j = d, s$, denote the estimate in the m^{th} monte carlo repetition, $m = 1, ..., M$. We report the monte carlo bias: Bias$(\widehat{\phi}^j) = \left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\phi}^j(m) - \phi^j\right)$ for $j = d, s$; and square root of the monte carlo MSE: RMSE$(\widehat{\phi}^j) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\phi}^j - \phi^j)^2}$ for $j = s, d$. Additionally we report the size of the t-test for all estimators and size of the J-test for the efficient GMM estimators. The results are reported in Table 2.1 (Design 1) and Table 2.2 (Design 2). In Table 2.1, we multiply the bias by 100 because all the estimators perform quite well in this ideal setting. For nearly each configuration of (N, T) the GMM estimators perform the best, in terms of bias and RMSE, as the theory suggests. Importantly, the t-test and J-test is well sized even when $rmax = 3 > r = 2$ factors are used. In Table 2.2, we report the bias as is, and we find that for a given configuration of (N, T) the bias is two orders of magnitude larger in Design 2 relative to Design 1. Nevertheless, as the theory would suggest, the efficient GMM estimators perform the best in terms of bias and RMSE. There are some size distortions for the supply side estimators but the distortions are decreasing in N for a given T.

2.9 Empirical Application to Global Crude Oil Markets

The data construction follows the recent literature: [58], [59], [60] (hereafter BH) and GK. The following is a breakdown of the raw variables collected for Jan. 1985 - Dec. 2015 ($T =$ 372 months): monthly oil production for $N = 22$ countries from the U.S. Energy Information

Administration (hereafter EIA); world oil production from the U.S. EIA; monthly oil prices based on the refiner acquisition cost of imported crude oil from the U.S. EIA; U.S. CPI from the St. Louis FRED database; monthly change in inventories from BH; monthly industrial production index from BH. The CPI is used to deflate nominal oil prices to arrive at the real price of oil, which is highly non-stationary. Following the aforementioned literature, we take the logarithm of the real price of oil series and then take first differences. We apply the same transformation to the monthly oil production for each country. These transformations render the production and price series stationary as confirmed by a host of Dickey-Fuller tests. For ensuring the tail index of the size-distribution, μ, is in the region the theory requires, we provide visual evidence along with 6 estimates of μ that all fall beneath 1, see Table 2.3. Also, see Figure 2.3, Figure 2.4 & Figure 2.5.

Let y_{it} denote the log difference of the oil supply for country i at time t and p_t denote the log difference of the real price of oil. Following GK, we estimate an OPEC factor using information on the cross-section of countries (i.e., known loadings). To that end, let o_{it} denote a dummy variable equal to 1 if country i is an OPEC member at time t and note that $o_{it} = o_i$ for most i, with the exception of Gabon and Ecuador in our sample. Finally, $\boldsymbol{c}_{t-1}$ denotes a 4×1 vector containing: lagged p_t, lagged world supply growth, lagged change in inventories, and lagged growth in industrial production. The system is given as follows

$$y_{it} = \phi^s p_t + \boldsymbol{\gamma}_{\boldsymbol{s}}' \boldsymbol{c}_{t-1} + o_{it}\eta_{OPEC,t} + \boldsymbol{\lambda}_i' \boldsymbol{\eta}_t + u_{it}, \tag{2.63}$$

$$d_t = \phi^d p_t + \boldsymbol{\gamma}_d' \boldsymbol{c}_{t-1} + \lambda_d \eta_{OPEC,t} + \varepsilon_t, \tag{2.64}$$

$$\sum_{i=1}^{N} S_{it}\, y_{it} = d_t; \tag{2.65}$$

where we lose the observation $t = 1$ due to differencing. The cross-sectionally demeaned supply equation is given by the approximate factor model,

$$\tilde{y}_{it} = y_{it} - \frac{1}{N}\sum_i y_{it} = \tilde{o}_{it}\eta_{OPEC,t} + \tilde{\boldsymbol{\lambda}}_i'\boldsymbol{\eta}_t + \tilde{u}_{it} = \tilde{o}_{it}\eta_{OPEC,t} + \tilde{e}_{it}, \tag{2.66}$$

where $\tilde{e}_{it} := \tilde{\boldsymbol{\lambda}}_i'\boldsymbol{\eta}_t + \tilde{u}_{it}$. Note that (2.66) implies we can obtain the OPEC factor, $\eta_{OPEC,t}$, via cross-sectional regression, for each $t > 1$, that is $\widehat{\eta}_{OPEC,t} = (\tilde{\boldsymbol{o}}_{\cdot t}'\tilde{\boldsymbol{o}}_{\cdot t})^{-1}\tilde{\boldsymbol{o}}_{\cdot t}'\tilde{\boldsymbol{y}}_{\cdot t}$. Hence, in our preliminary stage, we extract $\widehat{\eta}_{OPEC,t}$ and then run PCA on $\widehat{\tilde{e}}_{it} = \tilde{y}_{it} - \widehat{\eta}_{OPEC,t}$ to extract the latent demeaned loadings and latent factors. Define $\boldsymbol{y}_{\cdot t}^* := \tilde{\boldsymbol{y}}_{\cdot t} - \tilde{\boldsymbol{x}}_{\cdot t}\widehat{\eta}_{OPEC,t}$, then we purge the latent factors via $\boldsymbol{Q}$ as in the main text: $\boldsymbol{Q}\,\boldsymbol{y}_{\cdot t}^* = \boldsymbol{Q}\,\tilde{\boldsymbol{u}}_{\cdot t}$. However, when forming the GIV, there is a minor difference that we have time-varying size-weights, so we no longer construct $\boldsymbol{z}_{GIV}$ with a time-invariant share vector S_i, but rather we weight each idiosyncratic component at time t with its corresponding share from time $t-1$ to avoid endogeneity issues arising from contemporaneous weighting

$$\underset{(T-1)\times 1}{\boldsymbol{z}_{GIV}} = \begin{pmatrix} \boldsymbol{S}_{\cdot 1}'\boldsymbol{Q}\,\tilde{\boldsymbol{y}}_{\cdot 2}^* \\ \boldsymbol{S}_{\cdot 2}'\boldsymbol{Q}\,\tilde{\boldsymbol{y}}_{\cdot 3}^* \\ \vdots \\ \boldsymbol{S}_{\cdot T-1}'\boldsymbol{Q}\,\tilde{\boldsymbol{y}}_{\cdot T}^* \end{pmatrix}. \tag{2.67}$$

Besides these modifications from the stylized model in the theory, we estimate the elasticities using the estimators outlined in the main text. The number of factors, r, is estimated via the AH procedures (as outlined in the Supplementary Appendix 2.12.3). The ER method of AH estimated $\widehat{r}_{ER} = 1$, while the GR method estimated $\widehat{r}_{GR} = 3$; with $k_{max} = 10$. To be safe, we take $\widehat{r} = \widehat{r}_{GR} + 1$.

Supply results. The results for the supply elasticity are presented in Table 2.4. In Table 2.4, the 2nd column displays GK's results. The instrument GK use is given by (2.8) and their dependent variable is simply the cross-sectional average of the log difference of oil supply (i.e., $\boldsymbol{E} = \boldsymbol{\iota}/N$). Our results are in columns 3 and 4. In contrast to (2.8), the instrument we use in column 3 purges the common factors through the loading space. The instrument we use in column 4 also adds an estimate of the unobserved aggregate demand shocks, $\widehat{\varepsilon}_t$, to our FGIV. Moreover, the dependent variable we use is weighted using the estimated precision vector $\widehat{\boldsymbol{E}}$, which allows for cross-sectional correlations and heteroskedasticity in u_{it}. These differences lead to significantly different results. Columns 2 and 3, which attempt to use *only* GIVs as instruments, both lead to weak instruments as indicated by the first-stage F-statistics less than the rule of thumb, 10. Nevertheless, the FGIV supply elasticity estimate (0.016) from column 3 (estimated via Algorithm 1) is roughly one third that of GK's (0.044). Whereas, our efficient GMM supply elasticity estimate (0.005) from column 4 (estimated via Algorithm 3) is highly significant at the 1% level. Additionally, our results reveal that using estimates of unobserved aggregate demand shocks as supply instruments indeed renders a strong instrument as indicated by the first-stage F-stat of 14.33 in column 4. Moreover, the p-value for the J-statistic (0.11) fails to reject the null hypothesis of a valid model. An F-stat greater than 10, coupled with a small J-statistic provides statistical evidence in favor of our efficient GMM point estimate for the supply elasticity.

Demand results. Turning now to the demand elasticity in Table 2.5, the dependent variable GK use in this case is the same as the one we use. However, the instruments are different. Column 2 displays GK's demand elasticity (-0.463), again using the instrument as in (2.8). Column 3 presents our result when using only the FGIV as an instrument (-.0009), which is roughly 400 times smaller than GK's estimate. Columns 4 through 7 sequentially add principal components to the instrument vector for our efficient GMM estimator from column 3 until 4 principal components are used. Here we find that none of the models yield first-stage F-statistics greater than 10. It is reassuring, however, that the J-statistic for columns 4 through 7 all fail to reject the null of a valid model. Lastly, column 8 presents the [5] estimator which only includes the four principal components but not the FGIV as instruments, nearly all statistics remain unchanged except that inclusion of the FGIV increases the t-stat by about 25%.

Taken together, our empirical results suggest that supply shocks, whether they be aggregate or idiosyncratic supply shocks, albeit valid, do not serve as strong instruments for estimation of the demand elasticity. Whereas, aggregate demand shocks indeed seem to be a strong source of exogenous variation to tease out the supply elasticity.

2.10 Concluding Remarks

In this paper, we have further developed the GIV methodology introduced by [1], which takes advantage of panel data to construct instruments for estimation of structural time series regression models that involve endogenous regressors. This paper focuses on the underlying econometric issues involved in developing FGIV in a large N and large T framework where the loadings are treated as unknown parameters to be estimated before constructing the FGIV instrument. We further demonstrate that the sampling error arising from estimating the instrument, factors and a high dimensional precision matrix does not affect the limiting distribution for the structural parameters of interest. We also overidentify the structural parameters, which leads to new and improved results in the crude oil markets application and demonstrate that the J-test is well sized with simulation evidence. Our Monte Carlo study illustrates that our estimators and algorithms exhibit desirable performance with the finite sample distributions being well approximated by the asymptotic distributions.

More fruitful areas of research would be empirical applications of the theoretical results derived in this paper. Interesting theoretical extensions would be to allow for random slope coefficients with correlated heterogeneity, the presence of weak factors and unbalanced panels with data not missing at random. We are currently pursuing the dynamic panel data extension, as well as adapting the GIV methodology for unit-specific endogenous variables.

Table 2.1: Bias×100, RMSE, size of t-test and size of J-test for design 1.

Finite sample properties for Design 1.

	N	T	μ	$\widehat{\phi}^{s}_{FGIV}$	$\widehat{\phi}^{s}_{GK}$	$\widehat{\phi}^{s}_{GMM}$	$\widehat{\phi}^{s,rmax}_{GMM}$	$\widehat{\phi}^{d}_{FGIV}$	$\widehat{\phi}^{d}_{GK}$	$\widehat{\phi}^{d}_{GMM}$	$\widehat{\phi}^{d,rmax}_{GMM}$
1	30	400	0.92	0.0612	0.0910	0.0164	0.0273	-0.2760	-0.2723	0.1100	0.2100
				(0.0344)	(0.0330)	(0.0204)	(0.0204)	(0.0152)	(0.0173)	(0.0079)	(0.0080)
				[0.0635]	[0.1205]	[0.0830]	[0.0830]	[0.0570]	[0.0510]	[0.0685]	[0.0735]
				{N.A.}	{N.A.}	{0.0540}	{0.0625}	{N.A.}	{N.A.}	{0.0490}	{0.0465}
2	50	400	0.85	0.0071	0.0774	0.0174	0.0224	-0.1803	-0.1649	0.1462	0.2306
				(0.0313)	(0.0292)	(0.0200)	(0.0200)	(0.0106)	(0.0186)	(0.0058)	(0.0059)
				[0.0650]	[0.2685]	[0.0710]	[0.0715]	[0.0555]	[0.0555]	[0.0700]	[0.0740]
				{N.A.}	{N.A.}	{0.0520}	{0.0510}	{N.A.}	{N.A.}	{0.0480}	{0.0550}
3	100	400	0.80	0.0059	0.0159	0.0070	0.0102	-0.1886	-0.1558	0.1449	0.2330
				(0.0287)	(0.0271)	(0.0196)	(0.0197)	(0.0068)	(0.0112)	(0.0039)	(0.0040)
				[0.0610]	[0.2505]	[0.0585]	[0.0615]	[0.0515]	[0.0540]	[0.0705]	[0.0790]
				{N.A.}	{N.A.}	{0.0495}	{0.0485}	{N.A.}	{N.A.}	{0.0440}	{0.0425}
4	200	400	0.77	0.0192	0.0246	-0.006	0.0010	-0.1896	-0.1713	0.0524	0.1409
				(0.0276)	(0.0263)	(0.0188)	(0.0188)	(0.0046)	(0.0080)	(0.0027)	(0.0027)
				[0.0600]	[0.2660]	[0.0545]	[0.0535]	[0.0410]	[0.0450]	[0.0625]	[0.0635]
				{N.A.}	{N.A.}	{0.0590}	{0.0620}	{N.A.}	{N.A.}	{0.0495}	{0.0425}
5	500	400	0.75	-0.0013	-0.0078	-0.0097	-0.0102	0.2501	0.2255	-0.0893	-0.1747
				(0.0287)	(0.0280)	(0.0188)	(0.0187)	(0.0028)	(0.0030)	(0.0016)	(0.0016)
				[0.0545]	[0.0840]	[0.062]	[0.0625]	[0.0540]	[0.0590]	[0.0680]	[0.0730]
				{N.A.}	{N.A.}	{0.0560}	{0.0535}	{N.A.}	{N.A.}	{0.054}	{0.051}

Notes: We report Bias×100, (RMSE), [t-test],and $\{J$-test$\}$ (if applicable) with a nominal size of 5%. $\widehat{\phi}^{s}_{FGIV}$, $\widehat{\phi}^{s}_{GMM}$, $\widehat{\phi}^{s,rmax}_{GMM}$ estimated with Algorithm 1, 3, 3 with $r = 2$ and $rmax = 3$, respectively and $\widehat{\phi}^{s}_{GK}$, $\widehat{\phi}^{d}_{GK}$ both use (2.8) as an instrument. $\widehat{\phi}^{d}_{FGIV}$, $\widehat{\phi}^{d}_{GMM}$, $\widehat{\phi}^{d,rmax}_{GMM}$ estimated with (2.12), (2.21), (2.21) with $r = 2$ and $rmax = 3$, respectively. μ set to maintain $h_{N,\mu} = 0.12$ across all configurations of (N, T). $\bar{\psi}_u = 0.23$, $\bar{\psi}_{u+\eta} = 0.58$, $\bar{\psi}_{u+\varepsilon} = 0.65$.

Table 2.2: Bias, RMSE, size of t-test and size of J-test for design 2.

Finite sample properties for Design 2.

	N	T	μ	$\widehat{\phi}^s_{FGIV}$	$\widehat{\phi}^s_{GK}$	$\widehat{\phi}^s_{GMM}$	$\widehat{\phi}^{s,rmax}_{GMM}$	$\widehat{\phi}^d_{FGIV}$	$\widehat{\phi}^d_{GK}$	$\widehat{\phi}^d_{GMM}$	$\widehat{\phi}^{d,rmax}_{GMM}$
1	30	400	0.92	0.0263	0.0225	0.0080	0.0089	-0.0021	-0.0016	0.0003	0.0009
				(0.0365)	(0.0259)	(0.0150)	(0.0151)	(0.0371)	(0.0368)	(0.0148)	(0.0161)
				[0.2730]	[0.2895]	[0.12450]	[0.1410]	[0.0500]	[0.0510]	[0.0560]	[0.0600]
				{N.A.}	{N.A.}	{0.1895}	{0.3035}	{N.A.}	{N.A.}	{0.05200}	{0.0465}
2	50	400	0.85	0.0144	0.0136	0.0058	0.0063	-0.0009	-0.0007	0.0007	0.0013
				(0.0245)	(0.0219)	(0.0154)	(0.0155)	(0.0381)	(0.0327)	(0.0112)	(0.0117)
				[0.4030]	[0.4090]	[0.1245]	[0.1410]	[0.0465]	[0.0490]	[0.0705]	[0.0735]
				{N.A.}	{N.A.}	{0.1185}	{0.1615}	{N.A.}	{N.A.}	{0.0425}	{0.0405}
3	100	400	0.80	0.0070	0.0065	0.0029	0.0031	-0.0006	-0.0004	0.0009	0.0014
				(0.0221)	(0.0196)	(0.0137)	(0.0139)	(0.0117)	(0.0196)	(0.0068)	(0.0069)
				[0.1365]	[0.3865]	[0.0925]	[0.0915]	[0.0465]	[0.0455]	[0.0530]	[0.0590]
				{N.A.}	{N.A.}	{0.0895}	{0.0955}	{N.A.}	{N.A.}	{0.0495}	{0.0510}
4	200	400	0.77	0.0031	0.0029	0.0015	0.0016	-0.0011	-0.0009	0.0006	0.0010
				(0.0216)	(0.0197)	(0.0139)	(0.0139)	(0.0071)	(0.0130)	(0.0044)	(0.0045)
				[0.0950]	[0.3645]	[0.0745]	[0.0760]	[0.0400]	[0.0455]	[0.0640]	[0.0665]
				{N.A.}	{N.A.}	{0.0590}	{0.0615}	{N.A.}	{N.A.}	{0.0510}	{0.0515}
5	500	400	0.75	0.0014	0.0014	0.0006	0.0006	-0.0011	-0.0009	0.0004	0.0008
				(0.0218)	(0.0209)	(0.0141)	(0.0141)	(0.0037)	(0.0042)	(0.0024)	(0.0024)
				[0.0690]	[0.1125]	[0.0670]	[0.0690]	[0.0515]	[0.0570]	[0.0605]	[0.0675]
				{N.A.}	{N.A.}	{0.0535}	{0.0520}	{N.A.}	{N.A.}	{0.0575}	{0.0555}

Notes: We report Bias, (RMSE), [t-test],and $\{J$-test$\}$ (if applicable) with a nominal size of 5%. $\widehat{\phi}^s_{FGIV}$, $\widehat{\phi}^s_{GMM}$, $\widehat{\phi}^{s,rmax}_{GMM}$ estimated with Algorithm 1, 3, 3 with $r = 2$ and $rmax = 3$, respectively and $\widehat{\phi}^s_{GK}$, $\widehat{\phi}^d_{GK}$ both use (2.8) as an instrument. $\widehat{\phi}^d_{FGIV}$, $\widehat{\phi}^d_{GMM}$, $\widehat{\phi}^{d,rmax}_{GMM}$ estimated with (2.12), (2.21), (2.21) with $r = 2$ and $rmax = 3$, respectively. μ set to maintain $h_{N,\mu} = 0.12$ across all configurations of (N, T). $\bar{\psi}_u = 0.23$, $\bar{\psi}_{u+\eta} = 0.58$, $\bar{\psi}_{u+\varepsilon} = 0.65$.

Table 2.3: Tail index estimates by various methods

Tail index estimator	$\widehat{\mu}$
MLE	0.4216
OLS	0.5095
Percentiles Method	0.8987
Modified Percentiles Method	0.9000
Geometric Percentiles Method	0.5208
Weighted Least Squares	0.3725

Notes: The estimates are for a month selected at random. However, the estimates do not change significantly if we estimate $\widehat{\mu}$ for each month and average across months.

Table 2.4: Global crude oil market: supply elasticity

	$\widehat{\phi}^s_{GK}$	$\widehat{\phi}^s_{FGIV}$	$\widehat{\phi}^s_{GMM}$
Supply instruments	$\boldsymbol{Z}_{GIV}$	$\boldsymbol{z}_{GIV}$	$\boldsymbol{Z}_s = (\boldsymbol{z}_{GIV}, \boldsymbol{\varepsilon})$
Dep. variable	$\bar{\boldsymbol{y}}$	$\boldsymbol{y}_{\widehat{E}}$	$\boldsymbol{y}_{\widehat{E}}$
$\boldsymbol{p}$	0.044	0.016	0.005
t-stat	(1.43)	(1.35)	(4.32)
(N, T)	$(21, 370)$	$(21, 370)$	$(21, 370)$
J-stat p-value	{N.A.}	{N.A.}	0.11
First stage F-stat	< 10	< 10	14.33
First stage R^2	0.26	0.14	0.21

Notes: $\widehat{\phi}^s_{GK}$ is estimated using (2.8) as the instrument; whereas $\widehat{\phi}^s_{FGIV}$, and $\widehat{\phi}^s_{GMM}$ are estimated using Algorithm 1 and 3 respectively. See Section 2.9 for more details. The t-stat is reported in parenthesis below coefficient estimates. The coefficient estimates on $\widehat{\boldsymbol{\eta}}_{\boldsymbol{t}}$,$\widehat{\eta}_{OPEC,t}$ and $\boldsymbol{c}_{t-1}$ are omitted for brevity.

Table 2.5: Global crude oil market: demand elasticity

	$\widehat{\phi}^d_{GK}$	$\widehat{\phi}^d_{FGIV}$	$\widehat{\phi}^d_{GMM}(r)$				FGMM, BN
Demand instruments	$\boldsymbol{Z}_{GIV}$	$\boldsymbol{z}_{GIV}$	$(\boldsymbol{z}_{GIV}, \boldsymbol{\eta}[, 1])$	$(\boldsymbol{z}_{GIV}, \boldsymbol{\eta}[, 1:2])$	$(\boldsymbol{z}_{GIV}, \boldsymbol{\eta}[, 1:3])$	$(\boldsymbol{z}_{GIV}, \boldsymbol{\eta}[, 1:4])$	$\boldsymbol{\eta}[, 1:4]$
Dep. variable	$\boldsymbol{d}$	$\boldsymbol{d}$	$\boldsymbol{d}$	$\boldsymbol{d}$	$\boldsymbol{d}$	$\boldsymbol{d}$	$\boldsymbol{d}$
$\boldsymbol{p}$	-0.463	-0.0009	-0.0009	-0.0003	-0.0003	-0.0003	-0.0003
t-stat	(-3.54)	(-0.88)	(-0.89)	(-0.87)	(-0.93)	(-1.01)	(-0.80)
(N, T)	$(21, 370)$	$(21, 370)$	$(21, 370)$	$(21, 370)$	$(21, 370)$	$(21, 370)$	$(21, 370)$
J-stat p-value	{N.A.}	{N.A.}	0.83	0.67	0.85	0.93	0.98
First stage F-stat	< 10	< 10	< 10	< 10	< 10	< 10	< 10
First stage R^2	0.58	0.12	0.12	0.15	0.15	0.16	0.16

Note: $\widehat{\phi}^d_{GK}$ is estimated using (2.8) as the instrument; whereas $\widehat{\phi}^d_{FGIV}$, and $\widehat{\phi}^d_{GMM}$ are estimated using (2.12) and (2.21) respectively. See Section 2.9 for more details. The t-stat is reported in parenthesis below coefficient estimates. The coefficient estimates on $\widehat{\eta}_{OPEC,t}$ and $\boldsymbol{c}_{t-1}$ are omitted for brevity. The final column represents the [5] factor GMM estimator (FGMM).

2.11 Appendix

In this appendix, we prove Theorems 1-6, which require 4 lemmas. Lemmas 1, 2 and 3 are included in this appendix while Lemma 4 is deferred to Appendix 2.12.2 of the Supplementary Appendix.

Lemma 1 *Under Assumptions 1-4, we have that*

$$(i.)\ \left(\frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{it}\varepsilon_t\right)^2 = o_p\left(1\right) \tag{2.68}$$

$$(ii.)\ \mathbb{V}(p_t) = \Theta(1) \tag{2.69}$$

$$(iii.)\ \left(\frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{it}p_t\right)^2 = \mathcal{O}_p\left(1\right) \tag{2.70}$$

$$(iv.)\ \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}S_i\tilde{y}_{jt}\varepsilon_t\right)^2 = o_p(N). \tag{2.71}$$

Proof of Lemma 1: For $(i.)$, we have that for large T, the sum converges to its expected value of zero. For $(ii.)$, note that by Assumption 3 we can decompose our share vector into a dominant and a fringe part: $\boldsymbol{S} = \begin{pmatrix}\boldsymbol{S}_d' & \boldsymbol{S}_f'\end{pmatrix}'$ where $\boldsymbol{S}_d$ is $N_1 \times 1$, is the dominant part and $\boldsymbol{S}_f$ is $N_2 \times 1$, is the fringe part; with $N_1 + N_2 = N$. The key being that $N_1(N) = N_1$ is fixed while $N_2(N) \to \infty$ as $N \to \infty$. Recall that prices are given by $p_t = \dfrac{1}{\phi^d - \phi^s}\left(u_{St} + \boldsymbol{\lambda}_S'\boldsymbol{\eta}_t - \varepsilon_t\right)$. For simplicity, suppose that supply and demand shocks are uncorrelated, so that (ignoring the squared constant term)

$$\mathbb{V}(p_t) = \mathbb{E}(\boldsymbol{S}'\boldsymbol{\Sigma}_u\boldsymbol{S}) + \mathbb{E}(\boldsymbol{S}'\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{S}) + \mathbb{V}(\varepsilon_t) = \mathbb{E}(\boldsymbol{S}'\boldsymbol{\Sigma}_u\boldsymbol{S}) + \mathbb{E}(\boldsymbol{S}_d'\boldsymbol{\Lambda}_d\boldsymbol{\Lambda}_d'\boldsymbol{S}_d) + \mathbb{E}(\boldsymbol{S}_f'\boldsymbol{\Lambda}_f\boldsymbol{\Lambda}_f'\boldsymbol{S}_f) + \mathbb{V}(\varepsilon_t)$$

$$\leq \mathbb{E}(||\boldsymbol{S}||_2^2\gamma_{max}(\boldsymbol{\Sigma}_u)) + \mathbb{E}(||\boldsymbol{S}_d||_2^2\gamma_{max}(\boldsymbol{\Lambda}_d\boldsymbol{\Lambda}_d')) + \mathbb{E}(||\boldsymbol{S}_f||_2^2\gamma_{max}(\boldsymbol{\Lambda}_f\boldsymbol{\Lambda}_f')) + \mathcal{O}(1),$$

by Assumption 3 and Assumption 4; the first term consists of $||\boldsymbol{S}||^2$, which is $\Theta_p(1)$ for $\mu \in (0,1)$, see Lemma 1 in Appendix 2.12.2 of the Supplementary Appendix, and $\gamma_{max}(\boldsymbol{\Sigma}_u) = \mathcal{O}(1)$ by assumption, the second term is $\mathcal{O}(1)$ by Assumption 4 and the third term is $\mathcal{O}(\frac{1}{N}) \cdot \mathcal{O}(N) = \mathcal{O}(1)$ by Assumption 4. For part (iii.),

$$\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{it}p_t\right) \leq \left(\frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{it}^2\right)^{\frac{1}{2}} \cdot \left(\frac{1}{T}\sum_{t=1}^{T}p_t^2\right)^{\frac{1}{2}} = \left(\frac{1}{\sqrt{T}}||\tilde{\boldsymbol{y}}_{i\cdot}||\right) \cdot \left(\frac{1}{\sqrt{T}}||\boldsymbol{p}||\right) = \mathcal{O}_p(1).$$

For part $(iv.)$, we have

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}S_i\tilde{y}_{jt}\varepsilon_t\right)^2 = I + II + III + IV,$$

where $I = o_p(1)$, $II = o_p(1)$, $III = o_p\left(\frac{1}{N}\right)$ and $IV = o_p(N)$ are show below

$$\begin{aligned}
I &= \sum_{i=1}^{N_1} S_i^2 \sum_{j=1}^{N_1} \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{jt} \varepsilon_t \right)^2 \\
&= ||\boldsymbol{S}_d||^2 \cdot o_p(N_1) = o_p(1) \\
II &= \sum_{i=N_1+1}^{N} S_i^2 \sum_{j=N_1+1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{jt} \varepsilon_t \right)^2 \\
&= ||\boldsymbol{S}_f||^2 \cdot o_p(N_2) = \mathcal{O}_p\left(\frac{1}{N}\right) o_p(N) = o_p(1) \\
III &= \sum_{i=N_1+1}^{N} S_i^2 \sum_{j=1}^{N_1} \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{jt} \varepsilon_t \right)^2 \\
&= ||\boldsymbol{S}_f||^2 \cdot o_p\left(N_1\right) = o_p\left(\frac{1}{N}\right) \\
IV &= \sum_{i=1}^{N_1} S_i^2 \sum_{j=N_1+1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{jt} \varepsilon_t \right)^2 \\
&= o_p\left(N\right),
\end{aligned}$$

by Assumption 4 and Lemma 1 part (i.) ■

Lemma 2 *Under Assumptions 1-4, we have that*

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{z}_t - z_t)\varepsilon_t = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{S}'(\widehat{\boldsymbol{Q}} - \boldsymbol{Q})\tilde{\boldsymbol{y}}_{\cdot t}\varepsilon_t = o_p\left(1\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N}} \cdot C_{NT}^{-1}\right) = o_p(1) \qquad (2.72)$$

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{z}_t - z_t)p_t = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{S}'(\widehat{\boldsymbol{Q}} - \boldsymbol{Q})\tilde{\boldsymbol{y}}_{\cdot t}p_t = o_p\left(1\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N}} \cdot C_{NT}^{-1}\right) = o_p(1), \qquad (2.73)$$

where $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$.

Proof of Lemma 2: For the first term, it is well known that the loadings are only identified up to scale, so the usual notion of consistency is altered to consider consistency up to a rotation instead. For notational ease we will let $\tilde{\boldsymbol{\Lambda}}$ be denoted by $\boldsymbol{\Lambda}$. Recall, $\boldsymbol{Q} = \boldsymbol{I}_N - \boldsymbol{P}_{\Lambda H^{-1}}$ is an idempotent matrix spanned by the null space of $\boldsymbol{\Lambda H^{-1}}$ and is invariant to an orthogonal transformation. Let $\widehat{\boldsymbol{D}} = \dfrac{\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Lambda}}}{N} = \dfrac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\lambda}}_i\widehat{\boldsymbol{\lambda}}_i'$ and $\boldsymbol{D} = \dfrac{\boldsymbol{H^{-1'}}(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})\boldsymbol{H^{-1}}}{N} = \dfrac{1}{N}\boldsymbol{H^{-1'}}\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'\boldsymbol{H^{-1}}$, then we have (omitting subscripts on $\boldsymbol{P}$)

$$
\begin{aligned}
\widehat{\boldsymbol{Q}} - \boldsymbol{Q} &= \widehat{\boldsymbol{P}} - \boldsymbol{P} \\
&= N^{-1}\widehat{\boldsymbol{\Lambda}}\left(\frac{\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Lambda}}}{N}\right)^{-1}\widehat{\boldsymbol{\Lambda}}' - N^{-1}\boldsymbol{\Lambda}\boldsymbol{H^{-1}}\left(\frac{\boldsymbol{H^{-1'}}(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})\boldsymbol{H^{-1}}}{N}\right)^{-1}\boldsymbol{H^{-1'}}\boldsymbol{\Lambda}' \\
&= N^{-1}\left[\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}\widehat{\boldsymbol{\Lambda}}' - \boldsymbol{\Lambda}\boldsymbol{H^{-1}}\boldsymbol{D^{-1}}\boldsymbol{H^{-1'}}\boldsymbol{\Lambda}'\right] \\
&= N^{-1}\left[(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\boldsymbol{H^{-1}} + \boldsymbol{\Lambda}\boldsymbol{H^{-1}})\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\boldsymbol{H^{-1}} + \boldsymbol{\Lambda}\boldsymbol{H^{-1}})' - \boldsymbol{\Lambda}\boldsymbol{H^{-1}}\boldsymbol{D^{-1}}\boldsymbol{H^{-1'}}\boldsymbol{\Lambda}'\right] \\
&= N^{-1}\,[\,(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\boldsymbol{H^{-1}})\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\boldsymbol{H^{-1}})' + (\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\boldsymbol{H^{-1}})\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}\boldsymbol{H^{-1'}}\boldsymbol{\Lambda}' + \dots \\
&\qquad \dots + \boldsymbol{\Lambda}\boldsymbol{H^{-1}}\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\boldsymbol{H^{-1}}) + \boldsymbol{\Lambda}\boldsymbol{H^{-1}}(\widehat{\boldsymbol{D}}^{\boldsymbol{-1}} - \boldsymbol{D^{-1}})\boldsymbol{H^{-1'}}\boldsymbol{\Lambda}'\,].
\end{aligned}
$$

Therefore, $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{Q}} - \boldsymbol{Q})\tilde{\boldsymbol{y}}_{\cdot t}\varepsilon_t = I + II + III + IV$. Each term is analyzed below in order.[21]

$$\begin{aligned}
I &= \frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\boldsymbol{H}^{-1})\widehat{\boldsymbol{D}}^{-1}(\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\boldsymbol{H}^{-1})'\tilde{\boldsymbol{y}}_{\cdot t}\varepsilon_t \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(\widehat{\boldsymbol{\lambda}}_i-\boldsymbol{H}^{-1}\boldsymbol{\lambda}_i)'\widehat{\boldsymbol{D}}^{-1}(\widehat{\boldsymbol{\lambda}}_j-\boldsymbol{H}^{-1}\boldsymbol{\lambda}_j)\cdot\frac{1}{T}\sum_{t=1}^{T}S_i\tilde{y}_{jt}\varepsilon_t \\
&\leq \left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[(\widehat{\boldsymbol{\lambda}}_i-\boldsymbol{H}^{-1}\boldsymbol{\lambda}_i)'\widehat{\boldsymbol{D}}^{-1}(\widehat{\boldsymbol{\lambda}}_j-\boldsymbol{H}^{-1}\boldsymbol{\lambda}_j)\right]^2\right)^{\frac{1}{2}}\cdot\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}S_i\tilde{y}_{jt}\varepsilon_t\right)^2\right]^{\frac{1}{2}}
\end{aligned} \tag{2.74}$$

$$\begin{aligned}
&\leq \left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}||(\widehat{\boldsymbol{\lambda}}_i-\boldsymbol{H}^{-1}\boldsymbol{\lambda}_i)||^2\cdot||\widehat{\boldsymbol{D}}^{-1}||^2\cdot||(\widehat{\boldsymbol{\lambda}}_j-\boldsymbol{H}^{-1}\boldsymbol{\lambda}_j)||^2\right)^{\frac{1}{2}}\cdot o_p(N) \\
&= ||\widehat{\boldsymbol{D}}^{-1}||\cdot\left(\left(\frac{1}{N}\sum_{i=1}^{N}||(\widehat{\boldsymbol{\lambda}}_i-\boldsymbol{H}^{-1}\boldsymbol{\lambda}_i)||^2\right)^2\right)^{\frac{1}{2}}\cdot o_p(N) = \mathcal{O}_p(1)\cdot\mathcal{O}_p(C_{NT}^{-2})\cdot o_p(N) = o_p(1)\,,
\end{aligned}$$

where $\mathcal{O}_p(C_{NT}^{-2})$ follows from symmetry of Theorem 1 in [24], who show (while proving their Lemma 2) $||\widehat{D}^{-1}||$ is $\mathcal{O}_p(1)$ which again follows symmetrically here. Note, the first inequality follows from Cauchy-Schwarz applied to the summation in (i,j); the second inequality follows from Cauchy-Schwarz applied again but to the left most term's inner term in brackets being squared,

[21] In this Appendix, we use the Frobenius norm of a matrix $\boldsymbol{A}$ is $||\boldsymbol{A}||_F = [\mathrm{tr}(\boldsymbol{A}'\boldsymbol{A})]^{\frac{1}{2}} = \left[\sum_i\sum_j |a_{ij}|^2\right]^{\frac{1}{2}}$, but omit the subscript F for notational ease.

in equation (2.74).

$$\begin{aligned} II &= \frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\boldsymbol{H}^{-1})\widehat{\boldsymbol{D}}^{-1}\boldsymbol{H}^{-1\prime}\boldsymbol{\Lambda}'\tilde{\boldsymbol{y}}_{\cdot t}\varepsilon_t \\ &= \frac{1}{N}\sum_{i=1}^{N}(\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}^{-1}\boldsymbol{\lambda}_i)'\underbrace{\widehat{\boldsymbol{D}}^{-1}\sum_{j=1}^{N}\boldsymbol{H}^{-1\prime}\boldsymbol{\lambda}_j \cdot \frac{1}{T}\sum_{t=1}^{T}S_j\tilde{y}_{it}\varepsilon_t}_{\underset{r\times 1}{\boldsymbol{a_i}}} \\ &= \mathcal{O}_p\left(C_{NT}^{-2}\right), \end{aligned} \tag{2.75}$$

which is the same rate as I. In (2.75), we make use of the fact that $N^{-1}\sum_i(\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}^{-1}\boldsymbol{\lambda}_i)'\boldsymbol{a}_i = \mathcal{O}_p\left(C_{NT}^{-2}\right)$, which follows symmetrically from Lemma B.1 of [27]. The same logic leads to $III =$

$\mathcal{O}_p\left(C_{NT}^{-2}\right)$.

$$
\begin{aligned}
IV &= \frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{S}'\boldsymbol{\Lambda}\boldsymbol{H}^{\boldsymbol{-1}}(\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}-\boldsymbol{D}^{\boldsymbol{-1}})\boldsymbol{H}^{\boldsymbol{-1}\prime}\boldsymbol{\Lambda}'\tilde{\boldsymbol{y}}_{\cdot t}\varepsilon_t \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}S_i\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}(\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}-\boldsymbol{D}^{\boldsymbol{-1}})\boldsymbol{H}^{\boldsymbol{-1}\prime}\boldsymbol{\lambda}_j\cdot\frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{jt}\varepsilon_t \\
&\leq \left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}S_i^2\left[\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}(\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}-\boldsymbol{D}^{\boldsymbol{-1}})\boldsymbol{H}^{\boldsymbol{-1}\prime}\boldsymbol{\lambda}_j\right]^2\right)^{\frac{1}{2}}\cdot\mathcal{O}_p\left(1\right) \\
&\leq \left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}S_i^2\cdot||\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}||^2\cdot||\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}-\boldsymbol{D}^{\boldsymbol{-1}}||^2\cdot||\boldsymbol{H}^{\boldsymbol{-1}\prime}\boldsymbol{\lambda}_j||^2\right)^{\frac{1}{2}}\cdot\mathcal{O}_p\left(1\right) \\
&= \mathcal{O}_p(C_{NT}^{-1})\cdot\left(\frac{1}{N}\sum_{i=1}^{N}S_i^2\cdot||\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}||^2\right)^{\frac{1}{2}}\cdot\mathcal{O}_p(1)\cdot\mathcal{O}_p\left(1\right) \\
&= \mathcal{O}_p(C_{NT}^{-1})\cdot\left(\frac{1}{N}\left[\sum_{i=1}^{N_1}S_i^2\cdot||\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}||^2+\sum_{i=N_1+1}^{N}S_i^2\cdot||\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}||^2\right]\right)^{\frac{1}{2}}\cdot\mathcal{O}_p(1)\cdot\mathcal{O}_p\left(1\right) \\
&\leq \mathcal{O}_p(C_{NT}^{-1})\cdot\left(\frac{1}{N}\left(\sum_{i=1}^{N_1}S_i^4\right)^{\frac{1}{2}}\cdot\left(\sum_{i=1}^{N_1}||\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}||^4\right)^{\frac{1}{2}}+\frac{1}{N}\left(\sum_{i=N_1+1}^{N}S_i^4\right)^{\frac{1}{2}}\cdot\left(\sum_{i=N_1+1}^{N}||\boldsymbol{\lambda}_i'\boldsymbol{H}^{\boldsymbol{-1}}||^4\right)^{\frac{1}{2}}\right)^{\frac{1}{2}} \\
&= \mathcal{O}_p\left(C_{NT}^{-1}\right)\cdot\left(\mathcal{O}_p\left(\frac{N_1}{N}\right)+\mathcal{O}_p\left(\frac{1}{N}\right)\right)^{\frac{1}{2}} = \mathcal{O}_p\left(C_{NT}^{-1}\right)\cdot\mathcal{O}_p\left(\frac{1}{\sqrt{N}}\right),
\end{aligned}
$$

where $||\widehat{\boldsymbol{D}}^{\boldsymbol{-1}}-\boldsymbol{D}^{\boldsymbol{-1}}|| = \mathcal{O}_p(C_{NT}^{-1})$ again follows symmetrically from [24]. All in all, we have that

$$
\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}'(\widehat{\boldsymbol{Q}}-\boldsymbol{Q})\tilde{\boldsymbol{y}}_{\cdot t}\varepsilon_t = \mathcal{O}_p\left(C_{NT}^{-2}\right)+\mathcal{O}_p\left(\frac{1}{\sqrt{N}}\cdot C_{NT}^{-1}\right) = o_p(1),
$$

which is as the Lemma claimed. ■

Proof of Theorem 1: In light of Lemma 2, the result follows immediately from (2.26) by observing that $\frac{1}{T}\sum_t p_t z_t \xrightarrow{p} \mathbb{E}(p_t z_t) > 0$ for $\mu \in (0,1)$. Similarly, $\frac{1}{T}\sum_t z_t\varepsilon_t \xrightarrow{p} \mathbb{E}(z_t\varepsilon_t) = 0$ by Assumption 4. ■

Proof of Theorem 2: Under Assumptions 1-4 and Lemma 2, the result follows immediately. ■

As in the case of the demand elasticity, we need Lemma 3 before consistency and the limiting distribution of the supply elasticity can be established.

Lemma 3 *Under Assumptions 1-4, we have that the terms a_i, b_i, c_j for $i = 1, 2$; and $j = 1, 2$; defined in equations (2.34), (2.35) and (2.36) are $o_p(1) + \mathcal{O}_p\left(\frac{1}{\sqrt{N}} \cdot C_{NT}^{-1}\right) = o_p(1)$. While for $j = 3$, c_3 is $\mathcal{O}_p\left(\frac{m_N \omega_{N,T}^{1-q}}{T}\right) = o_p(1)$.*

Proof of Lemma 3: The terms, a_1, b_1, and c_1 follow very similarly as the proof for Lemma 2 and hence are omitted and the terms a_2, b_2 and c_2 follow symmetrically to the proof of Lemma 2 and thus, they are also omitted. The term c_3 is novel and warrants some further analysis. Recall c_3 is given by $c_3 = T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_\eta\,(\boldsymbol{u}_{\widehat{E}} - \boldsymbol{u}_E)$. Let us focus on $(\boldsymbol{u}_{\widehat{E}} - \boldsymbol{u}_E) = \boldsymbol{u}_{\cdot\cdot}(\widehat{\boldsymbol{E}} - \boldsymbol{E})$ momentarily, where $\boldsymbol{u}_{\cdot\cdot}$ is the $T \times N$ matrix of idiosyncratic errors and $\widehat{\boldsymbol{E}} - \boldsymbol{E} = \frac{\widehat{\boldsymbol{\Sigma}}_u^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\boldsymbol{\iota}} - \frac{\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\iota}}$. Let $\boldsymbol{\Theta}_u := \boldsymbol{\Sigma}_u^{-1}$, $C := \frac{\boldsymbol{\iota}'\boldsymbol{\Theta}_u\boldsymbol{\iota}}{N}$, then $\boldsymbol{E} = \frac{\boldsymbol{\Theta}_u\boldsymbol{\iota}/N}{C}$. We have that

$$\begin{aligned}
\widehat{\boldsymbol{E}} - \boldsymbol{E} &= \frac{\left[C\widehat{\boldsymbol{\Theta}}_u\boldsymbol{\iota} - \widehat{C}\boldsymbol{\Theta}_u\boldsymbol{\iota}\right]/N}{\widehat{C}C} = \frac{\left[C\widehat{\boldsymbol{\Theta}}_u\boldsymbol{\iota} - C\boldsymbol{\Theta}_u\boldsymbol{\iota} + C\boldsymbol{\Theta}_u\boldsymbol{\iota} - \widehat{C}\boldsymbol{\Theta}_u\boldsymbol{\iota}\right]/N}{\widehat{C}C}, \\
&= \frac{\left[C(\widehat{\boldsymbol{\Theta}}_u - \boldsymbol{\Theta}_u)\boldsymbol{\iota} + (C - \widehat{C})\boldsymbol{\Theta}_u\boldsymbol{\iota}\right]/N}{\widehat{C}C}, \\
\implies ||\widehat{\boldsymbol{E}} - \boldsymbol{E}||_1 &\leq \frac{\left[C \cdot ||(\widehat{\boldsymbol{\Theta}}_u - \boldsymbol{\Theta}_u)\boldsymbol{\iota}||_1 + |C - \widehat{C}| \cdot ||\boldsymbol{\Theta}_u\boldsymbol{\iota}||_1\right]/N}{|\widehat{C}|C},
\end{aligned} \tag{2.76}$$

where (2.76) follows from [61]. Let $\boldsymbol{\Theta}_{u,j}$ denote the j^{th} row of $\boldsymbol{\Theta}_u$ written as a column vector. Using Hölder's inequality we have

$$|\widehat{C} - C| = \left|\frac{\boldsymbol{\iota}'(\widehat{\boldsymbol{\Theta}}_u - \boldsymbol{\Theta}_u)\boldsymbol{\iota}}{N}\right| \leq \frac{||(\widehat{\boldsymbol{\Theta}}_u - \boldsymbol{\Theta}_u)\boldsymbol{\iota}||_1 \cdot ||\boldsymbol{\iota}||_{max}}{N} \leq \max_{1\leq j\leq N}||\widehat{\boldsymbol{\Theta}}_{u,j} - \boldsymbol{\Theta}_{u,j}||_1 = ||\widehat{\boldsymbol{\Theta}}_u - \boldsymbol{\Theta}_u||_1.$$

Thus,

$$\begin{aligned}
||\widehat{\boldsymbol{E}} - \boldsymbol{E}||_1 &\leq \frac{\left[C \cdot ||(\widehat{\boldsymbol{\Theta}}_u - \boldsymbol{\Theta}_u)\boldsymbol{\iota}||_1 + |C - \widehat{C}| \cdot ||\boldsymbol{\Theta}_u\boldsymbol{\iota}||_1\right]/N}{|\widehat{C}|C}, \\
&\leq \frac{\left[C \cdot \max\limits_{1\leq j\leq N}||\widehat{\boldsymbol{\Theta}}_{u,j} - \boldsymbol{\Theta}_{u,j}||_1 + \max\limits_{1\leq j\leq N}||\widehat{\boldsymbol{\Theta}}_{u,j} - \boldsymbol{\Theta}_{u,j}||_1 \cdot ||\boldsymbol{\Theta}_u\boldsymbol{\iota}||_1/N\right]}{|\widehat{C}|C}, \\
&= \frac{\max\limits_{1\leq j\leq N}||\widehat{\boldsymbol{\Theta}}_{u,j} - \boldsymbol{\Theta}_{u,j}||_1 \left[C \cdot + ||\boldsymbol{\Theta}_u\boldsymbol{\iota}||_1/N\right]}{|\widehat{C}|C} \leq \frac{\max\limits_{1\leq j\leq N}||\widehat{\boldsymbol{\Theta}}_{u,j} - \boldsymbol{\Theta}_{u,j}||_1 \left[C + \max\limits_{1\leq j\leq N}||\boldsymbol{\Theta}_{u,j}||_1\right]}{|\widehat{C}|C}, \\
&\leq \frac{||\widehat{\boldsymbol{\Theta}}_u - \boldsymbol{\Theta}_u||_1 \left[C + ||\boldsymbol{\Theta}_u||_1\right]}{|C + o_p(1)|C} = \frac{\mathcal{O}_p(m_N\omega_{N,T}^{1-q})\left[\mathcal{O}_p(1) + \mathcal{O}_p(1)\right]}{(\mathcal{O}_p(1) + o_p(1))\mathcal{O}_p(1)} = \mathcal{O}_p(m_N\omega_{N,T}^{1-q}),
\end{aligned} \tag{2.77}$$

where $C \geq \gamma_{min}(\boldsymbol{\Theta}_u) > 0$. Putting it all together for c_3, we have that

$$\begin{aligned}
c_3 &= T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_\eta(\boldsymbol{u}_{\widehat{E}} - \boldsymbol{u}_E) = T^{-1}\boldsymbol{z}'\,\boldsymbol{M}_\eta\boldsymbol{u}_{..}(\widehat{\boldsymbol{E}} - \boldsymbol{E}) \leq \gamma_{max}(\boldsymbol{M}_\eta) \cdot T^{-1}\boldsymbol{z}'\boldsymbol{u}_{..}(\widehat{\boldsymbol{E}} - \boldsymbol{E}), \\
&= T^{-1}\boldsymbol{z}'\boldsymbol{u}_{..}(\widehat{\boldsymbol{E}} - \boldsymbol{E}) \leq T^{-1}||\boldsymbol{u}'_{..}\boldsymbol{z}||_1 \cdot ||\widehat{\boldsymbol{E}} - \boldsymbol{E}||_1 \leq T^{-1}||\boldsymbol{u}'_{..}\boldsymbol{z}||_1 \cdot \mathcal{O}_p(m_N\omega_{N,T}^{1-q}), \\
&= T^{-1}\sum_{i=1}^{N}|S_i(\boldsymbol{u}'_{..}\boldsymbol{y}_{..}\boldsymbol{Q})_i| \cdot \mathcal{O}_p(m_N\omega_{N,T}^{1-q}) = \mathcal{O}_p\left(\frac{1}{T}\right) \cdot \mathcal{O}_p(m_N\omega_{N,T}^{1-q}) = o_p(1),
\end{aligned}$$

which concludes the proof. ■

Proof of Theorem 3: In light of Lemma 3, the result follows immediately from (2.37). ■

Proof of Theorem 4: Under Assumptions 1-4 and Lemma 3, the result follows immediately. ■

Proof of Theorem 5: In light of Theorem 2 and [5], the result follows immediately. ■

Proof of Theorem 6: In light of the theorems in the just identified case and standard GMM theory, we know that $\sqrt{T}(\widehat{\boldsymbol{\theta}}^s_{GMM} - \boldsymbol{\theta}^s)$ is asymptotically, a normal variate. The question remains whether using $\widehat{\varepsilon} = \varepsilon(\widehat{\phi}^d_{GMM})$ introduces sampling error that will effect the standard error of $\widehat{\boldsymbol{\theta}}^s_{GMM}$. To that end, let $\underset{(2+r)\times 1}{\boldsymbol{g}_{st}} := \boldsymbol{Z}_{st}u_{Et}$, we have that $\widehat{\boldsymbol{\theta}}^s_{GMM}$ solves the following first-order condition with probability approaching 1

$$0 = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\theta}^s}\boldsymbol{g}_{st}(\widehat{\boldsymbol{\theta}}^s_{GMM};\widehat{\phi}^d)\right)'\widehat{\boldsymbol{\Omega}}_s^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{g}_{st}(\widehat{\boldsymbol{\theta}}^s_{GMM};\widehat{\phi}^d)\right) \tag{2.78}$$

$$= \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\theta}^s}\boldsymbol{g}_{st}(\widehat{\boldsymbol{\theta}}^s_{GMM};\widehat{\phi}^d)\right)'\widehat{\boldsymbol{\Omega}}_s^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{g}_{st}(\boldsymbol{\theta}^s;\widehat{\phi}^d)\right)$$

$$+ \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\theta}^s}\boldsymbol{g}_{st}(\widehat{\boldsymbol{\theta}}^s_{GMM};\widehat{\phi}^d)\right)'\widehat{\boldsymbol{\Omega}}_s^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\theta}^s}\boldsymbol{g}_{st}(\bar{\boldsymbol{\theta}}^s;\widehat{\phi}^d)\right)(\widehat{\boldsymbol{\theta}}^s_{GMM}-\boldsymbol{\theta}^s) \tag{2.79}$$

$$= \boldsymbol{G}_s'\,\widehat{\boldsymbol{\Omega}}_s^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{g}_{st}(\boldsymbol{\theta}^s;\widehat{\phi}^d)\right) + \boldsymbol{G}_s'\,\widehat{\boldsymbol{\Omega}}_s^{-1}\,\boldsymbol{G}_s\sqrt{T}(\widehat{\boldsymbol{\theta}}^s_{GMM}-\boldsymbol{\theta}^s) \tag{2.80}$$

The basic idea of whether the sampling error from estimating $\widehat{\phi}^d$ can be ignored, boils down to whether the following expression holds: $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{g}_{st}(\boldsymbol{\theta}^s;\widehat{\phi}^d) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{g}_{st}(\boldsymbol{\theta}^s;\phi^d) + o_p(1)$; when this equation holds, then $\sqrt{T}(\widehat{\boldsymbol{\theta}}^s_{GMM}-\boldsymbol{\theta}^s)$ will not asymptotically depend on $\sqrt{T}(\widehat{\phi}^d-\phi^d)$. This can be easily seen if we take a mean value expansion of the left hand side of the expression above around ϕ^d, we obtain

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{g}_{st}(\boldsymbol{\theta}^s;\widehat{\phi}^d) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{g}_{st}(\boldsymbol{\theta}^s;\phi^d) + \boldsymbol{F}\sqrt{T}(\widehat{\phi}^d-\phi^d) + o_p(1), \tag{2.81}$$

where $\underset{(2+r)\times 1}{\boldsymbol{F}} := \mathbb{E}\left[\nabla_{\phi^d}\boldsymbol{g}_{st}(\boldsymbol{\theta}^s;\widehat{\phi}^d)\right]$, is generally different from zero, but here we have $\boldsymbol{F} = \mathcal{O}_p(\frac{1}{\sqrt{N}})$. This implies the asymptotic variance of $\sqrt{T}(\widehat{\boldsymbol{\theta}}^s_{GMM}-\boldsymbol{\theta}^s)$ need not take into account the sampling error induced by $\widehat{\phi}^d$. To see why, we need to get an expression for $\sqrt{T}(\widehat{\phi}^d-\phi^d)$, let $\underset{(1+r)\times 1}{\boldsymbol{g}_{dt}} = \boldsymbol{Z}_{dt}\varepsilon_t$; then taking a similar mean value expansion (as above) of the first-order conditions that $\widehat{\phi}^d$ solves with probability approaching 1

$$0 = \boldsymbol{G}_d' \widehat{\boldsymbol{\Omega}}_d^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{g}_{dt}(\phi^d) \right) + \boldsymbol{G}_d' \widehat{\boldsymbol{\Omega}}_d^{-1} \boldsymbol{G}_d \sqrt{T} (\widehat{\phi}^d - \phi^d), \tag{2.82}$$

hence, we obtain the usual influence function representation

$$\sqrt{T}(\widehat{\phi}^d - \phi^d) = -\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (\boldsymbol{G}_d' \boldsymbol{\Omega}_d^{-1} \boldsymbol{G}_d)^{-1} \boldsymbol{G}_d' \boldsymbol{\Omega}_d^{-1} \boldsymbol{g}_{dt}(\phi^d) := \frac{1}{\sqrt{T}} \sum_{t=1}^{T} r_{dt}(\phi^d). \tag{2.83}$$

Making use of (2.83) in (2.81) we obtain

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{g}_{st}(\boldsymbol{\theta}^s; \widehat{\phi}^d) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \breve{\boldsymbol{g}}_{st}(\boldsymbol{\theta}^s; \phi^d) + o_p(1), \tag{2.84}$$

where $\breve{\boldsymbol{g}}_{st}(\boldsymbol{\theta}^s; \phi^d) := \boldsymbol{g}_{st}(\boldsymbol{\theta}^s; \phi^d) + \boldsymbol{F}\, r_{dt}(\phi^d)$. Putting (2.84) and (2.80) together and solving for $\sqrt{T}(\widehat{\boldsymbol{\theta}}^s_{GMM} - \boldsymbol{\theta}^s)$ gives

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}^s_{GMM} - \boldsymbol{\theta}^s) \overset{d}{\to} -(\boldsymbol{G}_s'\, \boldsymbol{\Omega}_s^{-1}\, \boldsymbol{G}_s)^{-1} \boldsymbol{G}_s'\, \boldsymbol{\Omega}_s^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{g}_{st}(\boldsymbol{\theta}^s; \phi^d) + \boldsymbol{F}\, r_{dt}(\phi^d) \right) + o_p(1), \tag{2.85}$$

$$= -(\boldsymbol{G}_s'\, \boldsymbol{\Omega}_s^{-1}\, \boldsymbol{G}_s)^{-1} \boldsymbol{G}_s'\, \boldsymbol{\Omega}_s^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{g}_{st}(\boldsymbol{\theta}^s; \phi^d) + \mathcal{O}_p \left(\frac{1}{\sqrt{N}} \right) \mathcal{O}_p \left(1 \right) \right) + o_p(1), \tag{2.86}$$

which gives the result. ■

2.12 Supplementary Appendices

Section 2.12.1 contain figures pertaining to the empirical work from Section 2.9. Section 2.12.2 contains theoretical results for Herfindahl's in large N markets along with Lemma 1. Section 2.12.3 contains the estimation methods we use when r is unknown. Section 2.12.4 contains Algorithm $3'$ which employs an alternative estimator for the precision matrix, which is a hybrid of the factor approach and graphical models.

2.12.1 Figures

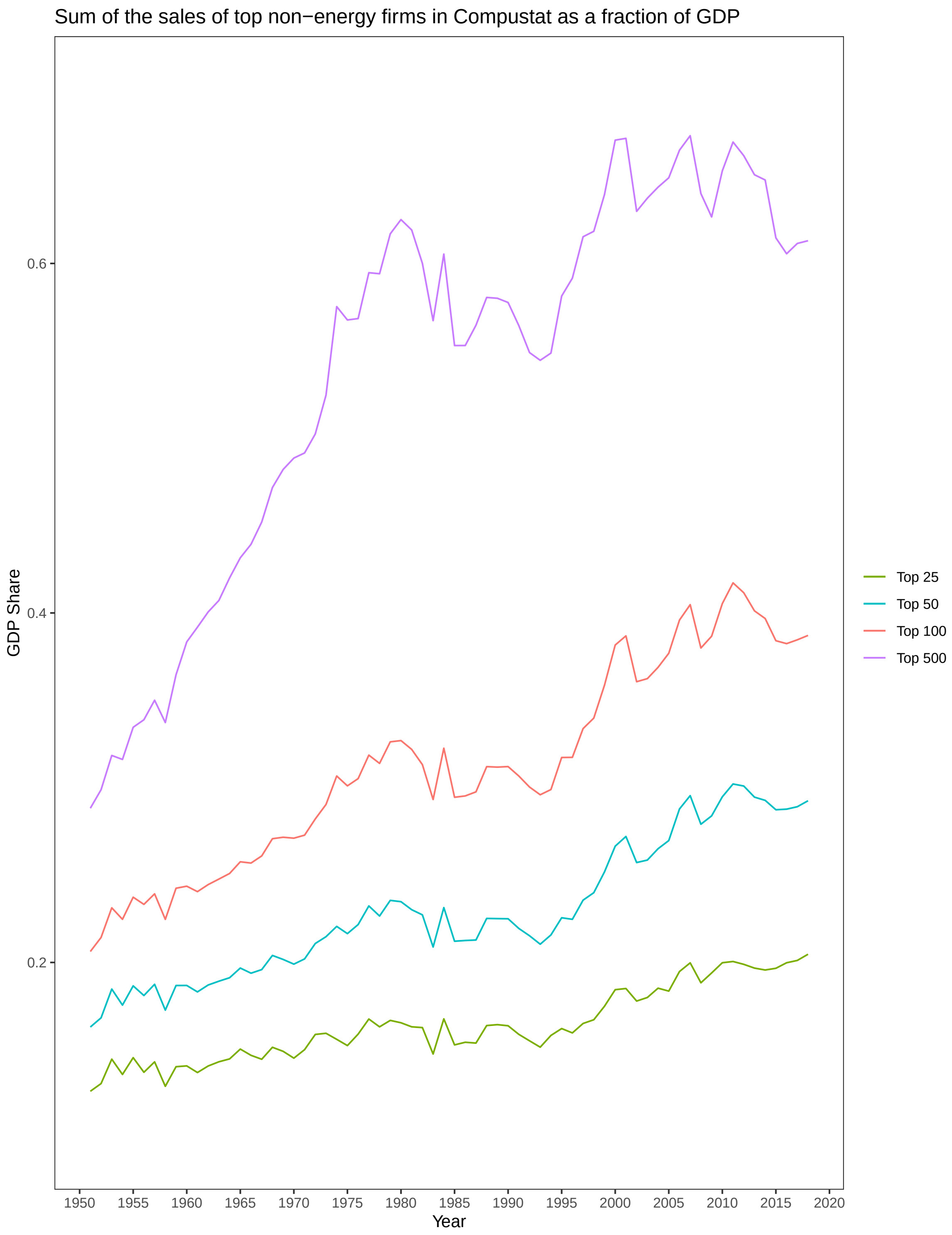

Figure 2.1: Sum of the sales of top non-energy firms in Compustat as a fraction of GDP

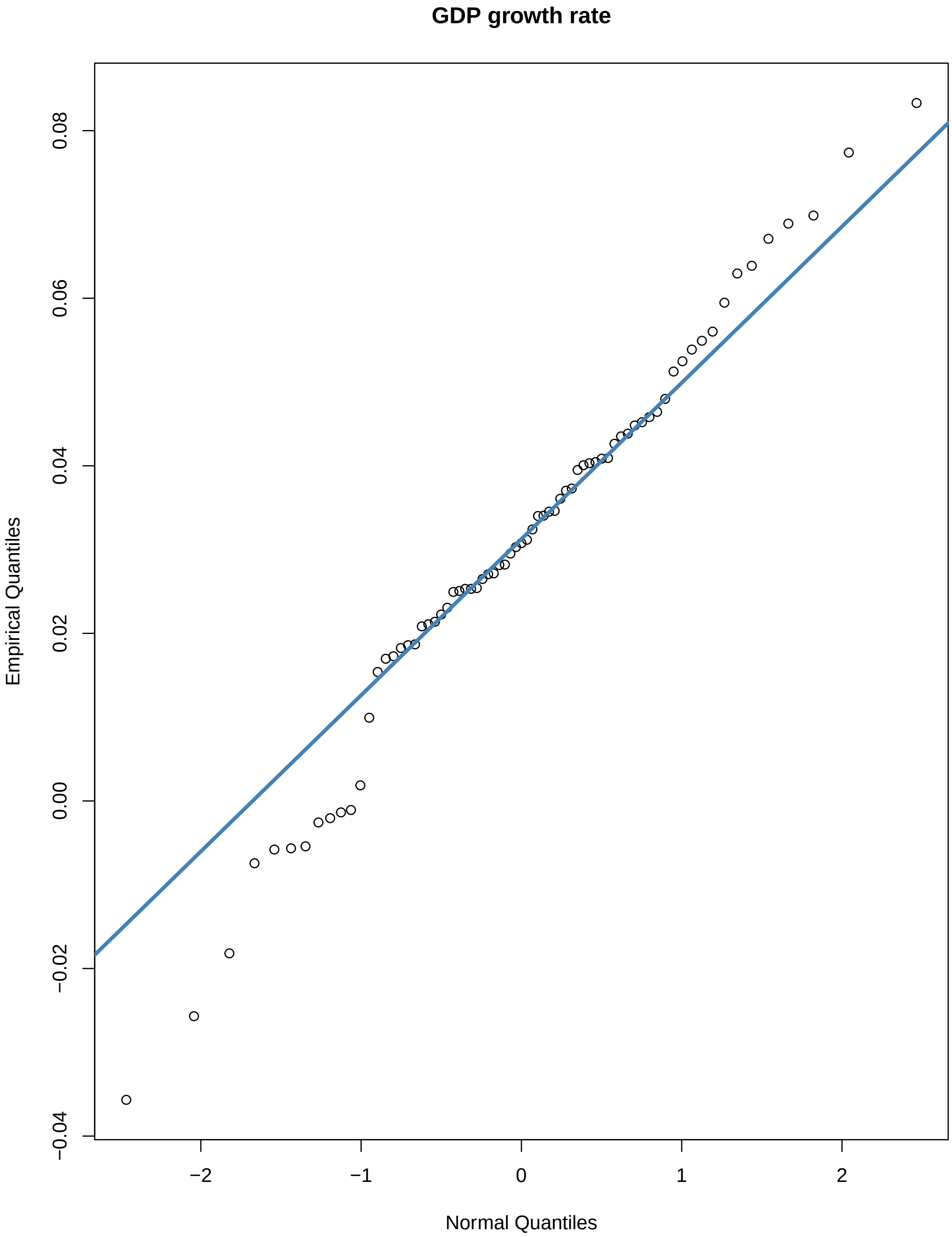

Figure 2.2: Quantile-quantile plots of US output fluctuations

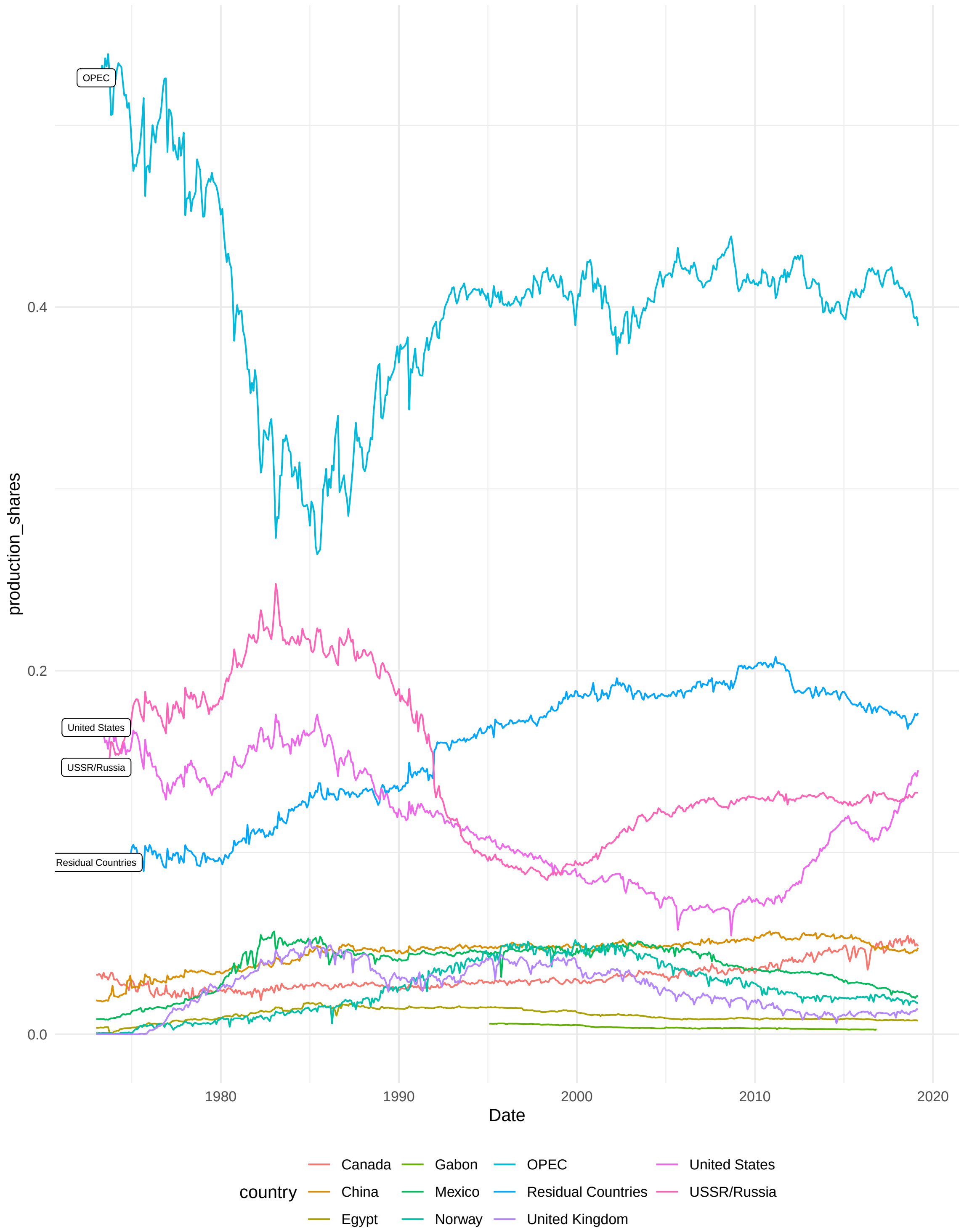

Figure 2.3: Temporal variation of production shares in global crude oil market

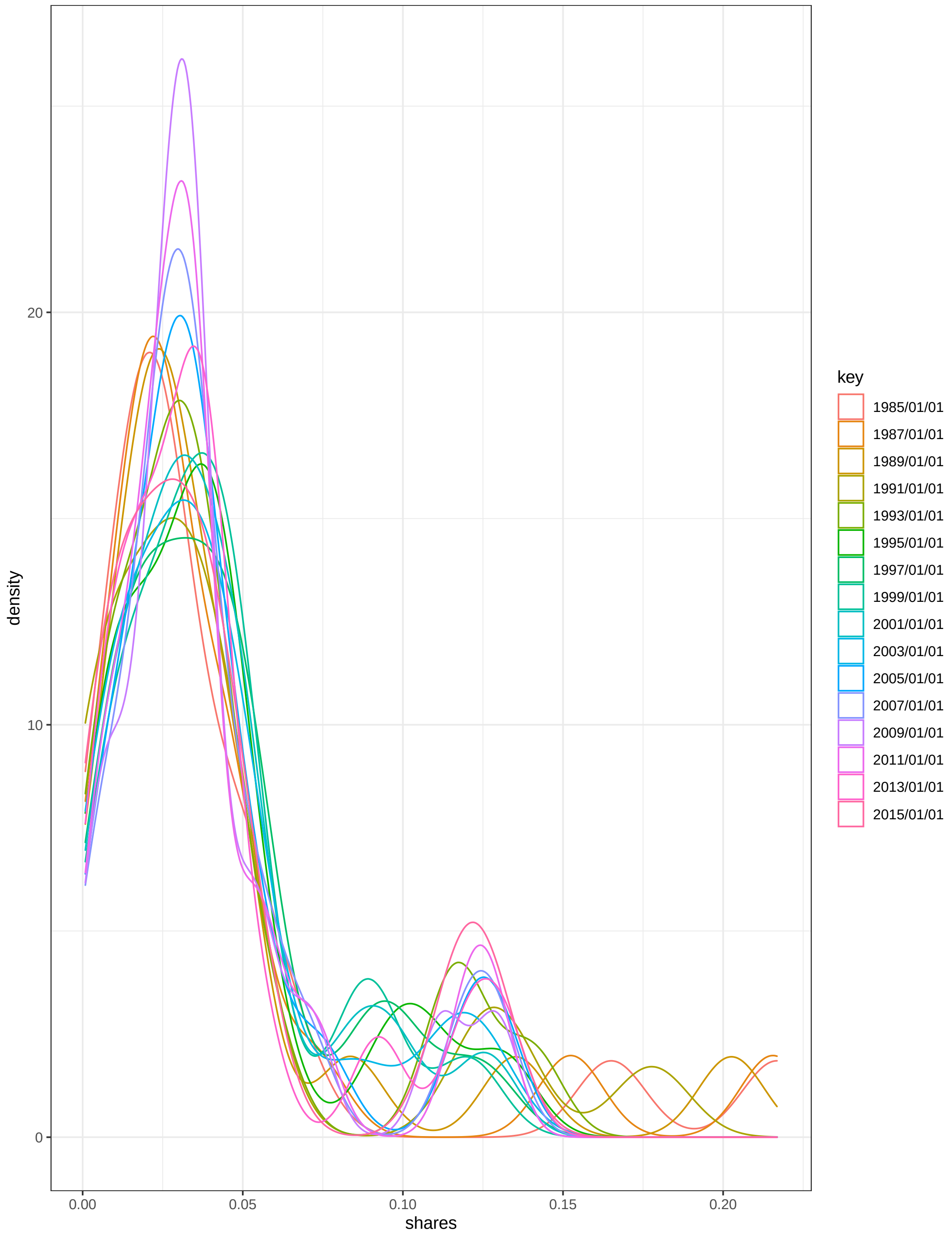

Figure 2.4: Distribution of production shares in global crude oil market

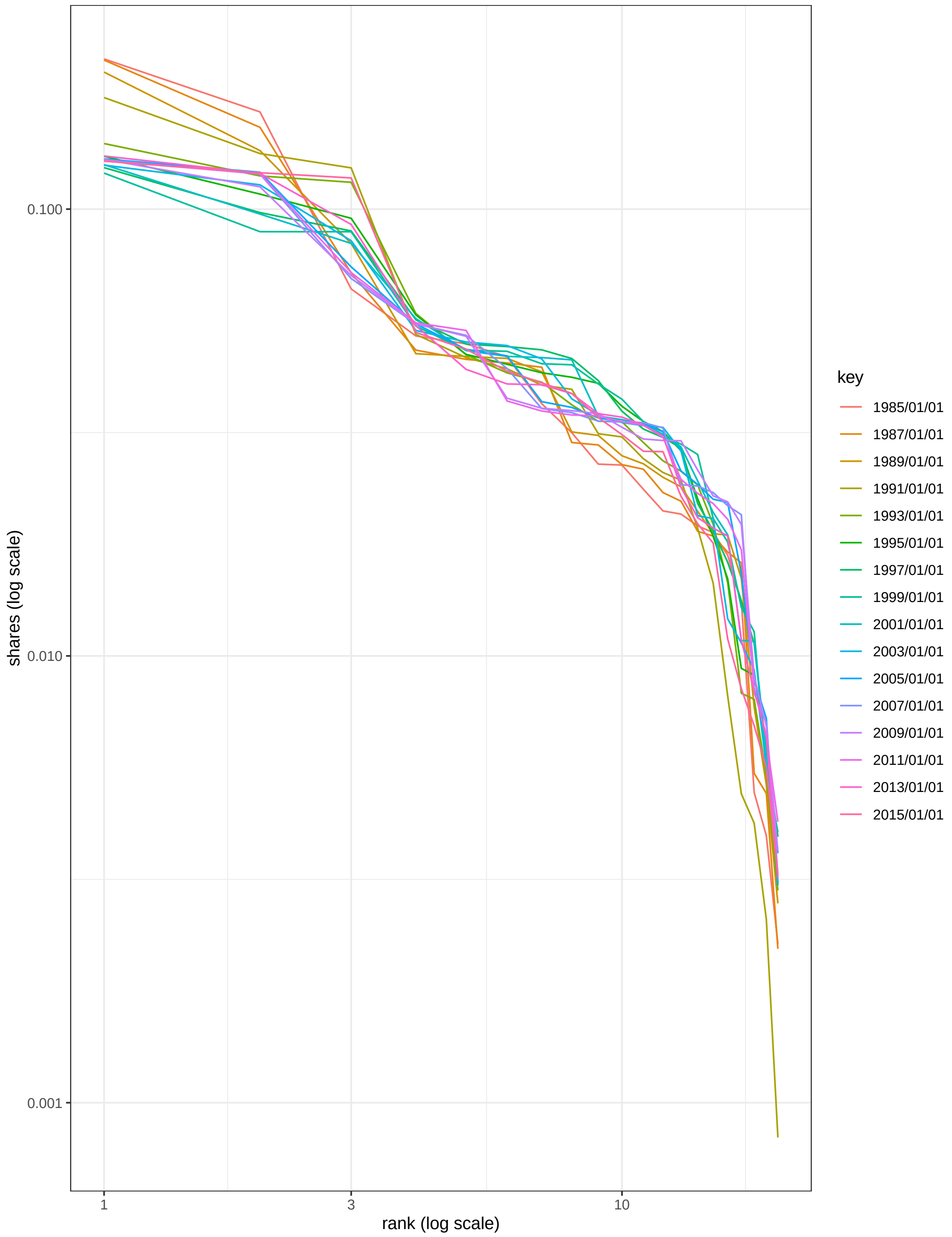

Figure 2.5: Size-Rank plot of crude oil production in global crude oil market (log-log scale)

2.12.2 Herfindahl's in Large *N* Markets

In this appendix we provide some basic information about properties of random variables that follow a power law and then we conclude with the statement of Lemma 1 and its proof. The following draws on [62], [63], [64], [42], [18], [7] and [65] for the $i.i.d.$ case presented below and can be adapted to allow for dependence over t using results from [66], [67], [68], and [69]. Recall, the sizes $\mathcal{S}_1, \ldots, \mathcal{S}_N$ are drawn $i.i.d.$ from a distribution for which the tail follows a power law, with tail index, $\mu > 0$. Note that the first and second moments can potentially diverge

$$\mathbb{E}(\mathcal{S}) = \int_1^\infty s\mu s^{-\mu-1} \mathrm{d}s = \int_1^\infty \mu s^{-\mu} \mathrm{d}s = \frac{\mu}{1-\mu} s^{1-\mu}\Big|_1^\infty = \begin{cases} \infty & \text{for } \mu \in (0,1] \\ -\dfrac{\mu}{1-\mu} & \text{for } \mu \in (1,\infty) \end{cases} \tag{2.87}$$

$$\mathbb{E}(\mathcal{S}^2) = \int_1^\infty s^2\mu s^{-\mu-1} \mathrm{d}s = \int_1^\infty \mu s^{1-\mu} \mathrm{d}s = \frac{\mu}{2-\mu} s^{2-\mu}\Big|_1^\infty = \begin{cases} \infty & \text{for } \mu \in (0,2] \\ -\dfrac{\mu}{2-\mu} & \text{for } \mu \in (2,\infty) \end{cases} \tag{2.88}$$

as a result of (2.87) and (2.88), $\mathbb{E}(\mathcal{S})$ is bounded for $\mu > 1$, while $\mathbb{V}(\mathcal{S})$ is bounded for $\mu > 2$. The literature refers to the cases $\mu \leq 2$ as thick tail regimes since the variance is infinite, rendering extreme tail events more likely. In light of this, there are some important cases to distinguish from one another when considering the limiting behavior of $h_{N,\mu}$, which are outlined in Table 2.6 below.

Table 2.6: Limiting behavior of the asymptotic Herfindahl index

	Tail index regime	Tail variation	First moment	Variance	$\mathcal{O}_p\left(g_{N,\mu}\right)$
Case I	$\mu > 2$	Exponential	$\mathbb{E}(\mathcal{S}) < \infty$	$\mathbb{V}(\mathcal{S}) < \infty$	$\frac{1}{\sqrt{N}}$
Case II	$\mu = 2$	Regularly varying	$\mathbb{E}(\mathcal{S}) < \infty$	$\mathbb{V}(\mathcal{S}) = \infty$	$\sqrt{\frac{\log(N)}{N}}$
Case III	$\mu \in (1,2)$	Regularly varying	$\mathbb{E}(\mathcal{S}) < \infty$	$\mathbb{V}(\mathcal{S}) = \infty$	$\frac{1}{N^{1-\frac{1}{\mu}}}$
Case IV	$\mu = 1$	Regularly varying	$\mathbb{E}(\mathcal{S}) = \infty$	$\mathbb{V}(\mathcal{S}) = \infty$	$\frac{1}{\log(N)}$
Case V	$\mu \in (0,1)$	Regularly varying	$\mathbb{E}(\mathcal{S}) = \infty$	$\mathbb{V}(\mathcal{S}) = \infty$	$\Theta_p(1)$
Case VI	$\mu \to 0$	Slowly varying	$\mathbb{E}(\mathcal{S}) = \infty$	$\mathbb{V}(\mathcal{S}) = \infty$	$\Theta_p(1)$

The Herfindahl is given by

$$h_{N,\mu} = \sum_{i=1}^{N} S_i^2 = \sum_{i=1}^{N} \left(\frac{\mathcal{S}_i}{\sum_{j=1}^{N} \mathcal{S}_j} \right)^2, \tag{2.89}$$

and the object of interest is the asymptotic Herfindahl and from (2.89), it can readily be written as

$$h_\mu := \lim_{N\to\infty} h_{N,\mu} = \lim_{N\to\infty} \sum_{i=1}^{N} \left(\frac{\mathcal{S}_i}{\sum_{j=1}^{N} \mathcal{S}_j} \right)^2 = \lim_{N\to\infty} \frac{1}{N} \frac{N^{-1}\sum_{i=1}^{N} \mathcal{S}_i^2}{\left(N^{-1}\sum_{j=1}^{N} \mathcal{S}_j\right)^2} := \lim_{N\to\infty} \frac{1}{N} \frac{a_N}{b_N}. \tag{2.90}$$

In **Case I**, when $\mathbb{E}(\mathcal{S}), \mathbb{E}(\mathcal{S}^2) < \infty$, the usual LLN and continuous mapping theorem gives us $a_N \xrightarrow{p} a = \mathbb{E}(\mathcal{S}^2)$ and $b_N \xrightarrow{p} b = (\mathbb{E}(\mathcal{S}))^2$. Therefore, $h_{N,\mu} \to h_\mu = 0$, for thin tailed regimes.[22] We will skip further details regarding **Cases II-IV** and state Lemma 1 which is pertaining to **Cases V** and **VI**.

[22] As we saw in (unreported) simulation evidence, a Herfindahl converging to zero in the large N limit, should not rule out identification by GIV; although theoretically it does. As illustrated in Remark 4, the variance of the elasticities diverges.

Lemma 1 *Under Assumption 4 (ii.), we have that*

$$\sqrt{h_{N,\mu}} = ||\boldsymbol{S}||_2 = \Theta_p(1) \quad \mu \in (0,1). \tag{2.91}$$

Proof of Lemma 1: Note that $\mathbb{P}(\mathcal{S} > s) = s^{-\mu}$ and hence, $\mathbb{P}(\mathcal{S}^{-\mu} > s) = s$. Or, put differently $\mathbb{P}(\mathcal{S}^{-\mu} > s) \sim \mathrm{U}[0,1]$; which is equivalently denoted as $U_i := 1 - F_{\mathcal{S}}(s_i) = s_i^{-\mu}$, where $F_{\mathcal{S}}(s_i)$ denotes the CDF of s_i. As a result, $U_1, \ldots, U_N$ are an $i.i.d.$ sample from $\mathrm{U}[0,1]$. It is well known that order statistics, denoted as $U_{(i)} = 1 - F_{\mathcal{S}}(s_{(i)})$, of the uniform distribution on the unit interval have marginal distributions belonging to the Beta distribution family. Hence, the PDF of the i^{th} order statistic, $U_{(i),N} \sim$ Beta(α, β), with $\alpha = i$ and $\beta = N - i + 1$. Finally, the size of the i^{th} largest unit out of N can be found by manipulating the expected value of the i^{th} order statistic, given by

$$\mathbb{E}(U_{(i),N}) = \mathbb{E}(\mathcal{S}_{(i),N}^{-\mu}) = \frac{\alpha}{\alpha+\beta} = \frac{i}{N+1}, \tag{2.92}$$

hence, the size of the i^{th} largest unit is

$$\mathcal{S}_{(i),N} = \left(\frac{i}{N+1}\right)^{-\frac{1}{\mu}} \simeq \left(\frac{i}{N}\right)^{-\frac{1}{\mu}}, \tag{2.93}$$

where the approximation is negligible for large N. Furthermore, $\mathcal{S}_i^2$ has tail index $\frac{\mu}{2} \leq 1$, since

$\mathbb{P}(\mathcal{S}^2 > s) = \mathbb{P}(\mathcal{S} > s^{\frac{1}{2}}) = s^{-\frac{\mu}{2}}$. Plugging (2.93) into (2.90), we obtain

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N} \frac{N^{-1}\sum_{i=1}^{N} \mathcal{S}_i^2}{\left(N^{-1}\sum_{j=1}^{N}\mathcal{S}_j\right)^2} = \lim_{N\to\infty} \frac{1}{N} \frac{N^{-1}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{-\frac{2}{\mu}}}{\left(N^{-1}\sum_{j=1}^{N}\mathcal{S}_j\right)^2} = \lim_{N\to\infty} \frac{1}{N^2}\frac{1}{N^{-\frac{2}{\mu}}} \frac{\sum_{i=1}^{N} i^{-\frac{2}{\mu}}}{\left(N^{-1}\sum_{j=1}^{N}\mathcal{S}_j\right)^2} \\
&= \lim_{N\to\infty} \frac{1}{N^{2-\frac{2}{\mu}}} \frac{\sum_{i=1}^{\infty} i^{-\frac{2}{\mu}}}{\left(\frac{1}{N^{1-\frac{1}{\mu}}}\sum_{j=1}^{\infty} j^{-\frac{1}{\mu}}\right)^2} = \lim_{N\to\infty} \frac{1}{N^{2-\frac{2}{\mu}}} \frac{\sum_{i=1}^{\infty} i^{-\frac{2}{\mu}}}{\frac{1}{N^{2-\frac{2}{\mu}}}\left(\sum_{j=1}^{\infty} j^{-\frac{1}{\mu}}\right)^2} = \lim_{N\to\infty} \frac{\sum_{i=1}^{\infty} i^{-\frac{2}{\mu}}}{\left(\sum_{j=1}^{\infty} j^{-\frac{1}{\mu}}\right)^2} \\
&= \frac{\zeta(\frac{2}{\mu})}{\left(\zeta(\frac{1}{\mu})\right)^2} \qquad (2.94)
\end{aligned}$$

Where $\zeta(\cdot)$ denotes the Riemann-zeta function. Therefore, we have just showed that for $\mu \in (0,1)$, $h_{N,\mu} \to h_\mu > 0$. ■

2.12.3 Estimating the Number of Factors

Generally, since the number of factors, r, is unknown we must estimate it. There are many estimators for r in static approximate factor models. Some examples are [24], [39] and [40]. We make use of the $ER(k)$ and $GR(k)$ estimators proposed by [40] (hereafter AH), which have been shown to outperform the existing estimators in the literature, particularly when the idiosyncratic errors are not $i.i.d.$, which is likely to be the more relevant case.

The estimators below can be used in (2.12) for estimation of the demand elasticity without affecting inference, and in Algorithm 1, 2, or 3 for estimation of the supply elasticity; again without affecting inference. The AH estimators are given by maximizing the following criteria

$$ER(k) = \frac{\tilde{\mu}_{NT,k}}{\tilde{\mu}_{NT,k+1}} \qquad k = 1, \dots, kmax \tag{2.95}$$

$$\begin{aligned} GR(k) &= \frac{\ln\left[V(k-1)/V(k)\right]}{\ln\left[V(k)/V(k+1)\right]} \\ &= \frac{\ln(1+\tilde{\mu}^{*}_{NT,k})}{\ln(1+\tilde{\mu}_{NT,k+1})} \qquad k = 1, \dots, kmax, \end{aligned} \tag{2.96}$$

where $\tilde{\mu}_{NT,k} := \psi_k\left[\boldsymbol{X}\boldsymbol{X}'/(NT)\right] = \psi_k\left[\boldsymbol{X}'\boldsymbol{X}/(NT)\right]$, $\boldsymbol{X}$ denotes a $T \times N$ matrix and $\psi_k(\boldsymbol{A})$ denotes the k^{th} largest eigenvalue of a positive semidefinite matrix $\boldsymbol{A}$. $V(k) = \sum_{j=k+1}^{m} \tilde{\mu}_{NT,j}$ and $\tilde{\mu}^{*}_{NT,k} = \tilde{\mu}_{NT,k}/V(k)$. Where $V(k)$ is the sample mean of the squared residuals from the time series regressions of individual response variables on the first k principal components of $\boldsymbol{X}\boldsymbol{X}'/(TN)$. Hence, the estimators are

$$\widehat{r}_{ER} = \underset{1 \le k \le kmax}{\text{argmax}} \; ER(\text{k}), \tag{2.97}$$

$$\widehat{r}_{GR} = \underset{1 \le k \le kmax}{\text{argmax}} \; GR(\text{k}). \tag{2.98}$$

The basic idea behind maximizing the $ER(k)$ and $GR(k)$ criteria is that $\dfrac{\tilde{\mu}_{NT,j}}{\tilde{\mu}_{NT,j+1}} = \mathcal{O}_p(1)$ for $j \neq r$, while $\dfrac{\tilde{\mu}_{NT,r}}{\tilde{\mu}_{NT,r+1}} = \mathcal{O}_p(C_{NT})$. This effective idea stems from the seminal paper of [70], who, among other things, demonstrate that only the r eigenvalues arising from the common component remain unbounded as the sample size tends to infinity, while those from the idiosyncratic part remain bounded. In particular, see their Theorem 4. Lastly, some recommendations on the choice of $kmax$ are provided by AH to avoid choosing $\widehat{r} < r$ wpa 1.

The important fact is that the limiting distribution of the elasticities remain unchanged so long as we use a consistent estimator for r. Let $\widehat{\phi}^j_{\widehat{r}}$, for $j = s, d$, denote the FGIV or efficient GMM estimator, using a consistent estimator for r, such as the AH estimator. It is easy to show that $\widehat{\phi}^j_{\widehat{r}}$ has the same limiting distribution as $\widehat{\phi}^j_r$, for $j = d, s$:

$$
\begin{align}
\mathbb{P}\left(\sqrt{T}(\widehat{\phi}^j_{\widehat{r}} - \phi^j) \leq x\right) &= \mathbb{P}\left(\sqrt{T}(\widehat{\phi}^j_{\widehat{r}} - \phi^j) \leq x | \widehat{r} = r\right) \cdot \mathbb{P}(\widehat{r} = r) \notag \\
&+ \mathbb{P}\left(\sqrt{T}(\widehat{\phi}^j_{\widehat{r}} - \phi^j) \leq x | \widehat{r} \neq r\right) \cdot \mathbb{P}(\widehat{r} \neq r) \tag{2.99} \\
&\rightarrow \mathbb{P}\left(\sqrt{T}(\widehat{\phi}^j_{\widehat{r}} - \phi^j) \leq x | \widehat{r} = r\right) \tag{2.100} \\
&= \mathbb{P}\left(\sqrt{T}(\widehat{\phi}^j_{r} - \phi^j) \leq x\right). \tag{2.101}
\end{align}
$$

From (2.99) to (2.100) we make use of the fact that $\widehat{r}$ is a consistent estimator for r, i.e. $\mathbb{P}(\widehat{r} = r) \rightarrow 1$. From (2.100) to (2.101) we make use of the fact that conditional on $\widehat{r} = r$, $\widehat{\phi}^j_{\widehat{r}} = \widehat{\phi}^j_r$. Therefore,

$$
\left|\mathbb{P}\left(\sqrt{T}(\widehat{\phi}^j_{\widehat{r}} - \phi^j) \leq x\right) - \mathbb{P}\left(\sqrt{T}(\widehat{\phi}^j_{r} - \phi^j) \leq x\right)\right| \rightarrow 0. \tag{2.102}
$$

2.12.4 Estimation of the High-Dimensional Precision Matrix via ℓ_1-Penalized Bregman Divergence

An alternative to using the POET like procedure of [37] to estimate a high dimensional precision matrix is to use graphical Lasso methods, as in [71]. One thought would be to directly estimate the precision matrix using graphical models, say by applying the graphical Lasso procedure to the composite error, $v_{it} = \boldsymbol{\lambda}_i'\boldsymbol{\eta}_t + u_{it}$, or by using a local (nodewise) graphical method as in [61] and applying to it v_{it}.

However, these approaches rule out the presence of an approximate factor structure, as they assume unconditional sparsity of the composite error term v_{it}. It is clear that sparsity of v_{it} fails given our (pervasive) factor structure, as pointed out by [72], [73] and [74].

An alternative, hybrid, approach to estimation of high dimensional covariance matrices is to adopt an approximate factor structure, thereby decomposing the process into a low rank part (common component), plus a sparse part (idiosyncratic component) like the approach we adopted in the main text of the paper. Except now we will use a graphical model in estimation of the precision matrix, as opposed to the POET estimation procedure. Thus, we can employ a hybrid approach known as the *factor-adjusted graphical lasso* model or simply FGL of [75]. The FGL approach imposes conditional sparsity similar to POET with the exception that sparsity is imposed on the precision matrix, $\mathbf{\Sigma}_u^{-1}$, of the idiosyncratic term rather than the covariance matrix, $\mathbf{\Sigma}_u$. That is, once the low dimensional common factors are conditioned on, $\mathbf{\Sigma}_u^{-1}$ is assumed to be sparse in the sense that many of the off-diagonal elements are zero. Note that with FGL, sparsity is assumed on the precision matrix, $\mathbf{\Sigma}_u^{-1}$, for the idiosyncratic term and not on the precision matrix of the composite error, $\mathbf{\Sigma}_v^{-1}$ as in a traditional graphical method.

Along the POET procedures, [37], [76] and [41] amongst others, estimate the high dimensional covariance matrix via thresholding techniques and then invert the estimate to obtain an estimate of the precision matrix. Whereas, graphical methods directly estimate the precision matrix. The FGL approach is essentially a hybrid of the two approaches. We adopt the FGL approach of [75], although in their paper endogeneity is not a concern. To that end, suppose momentarily that the idiosyncratic error is observed

$$
\begin{aligned}
u_{it} &= y_{it} - \boldsymbol{\lambda}_i' \boldsymbol{\eta}_t - \phi^s p_t, \\
&= y_{it} - \boldsymbol{\psi}_i' \boldsymbol{f}_t, \qquad (2.103)
\end{aligned}
$$

where $\boldsymbol{\psi}_i := \begin{pmatrix} \boldsymbol{\lambda}_i' & \phi^s \end{pmatrix}'$ and $\boldsymbol{f}_t$ is defined as in the main text. We apply the graphical Lasso procedure of [71] to (2.103), to obtain $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ as the solution to the ℓ_1-penalized Bregman Divergence. Bregman divergence is simply a measure of distance between two objects defined in terms of a strictly convex function, say $f(\cdot)$. Introduce $\mathcal{S}_{++}$ as the set of symmetric positive definite matrices, then, for $\boldsymbol{A}_1$, $\boldsymbol{A}_2 \in \mathcal{S}_{++}$, the Bregman Divergence in this context, is defined as

$$
d_f(\boldsymbol{A}_1, \boldsymbol{A}_2) := f(\boldsymbol{A}_1) - f(\boldsymbol{A}_2) - \langle \nabla f(\boldsymbol{A}_2), \boldsymbol{A}_1 - \boldsymbol{A}_2 \rangle \qquad (2.104)
$$

where $f(\cdot)$ is strictly convex and continuously differentiable. The Bregman Divergence, d_f, can be viewed as the difference of $f(\boldsymbol{A}_1)$ from the first-order approximation of $f(\boldsymbol{A}_1)$ around $\boldsymbol{A}_2$. Moreover, (2.104) nests some important loss functions as special cases for particular choices of $f(\cdot)$, e.g. when $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$, d_f becomes the Mahalanobis distance, which reduces to the squared norm when $\boldsymbol{B} = \boldsymbol{I}$ and when $f(\boldsymbol{x}) = \sum_i x_i \log x_i$, we obtain d_f as the Kullback-Leibler divergence. When one sets $f(\boldsymbol{A}) = -\log \det(\boldsymbol{A})$, then $\nabla f(\boldsymbol{A}) = -\boldsymbol{A}_2^{-1}$,[23] and the Bregman Divergence takes

[23] See Section A.4.1 of [77] for an elegant derivation of this gradient.

the following familiar form for $\boldsymbol{A}_1 = \boldsymbol{\Sigma}_u^{-1}$ and $\boldsymbol{A}_2 = \widehat{\boldsymbol{\Sigma}}_u^{-1}$

$$\begin{aligned} d_f(\boldsymbol{\Sigma}_u^{-1}, \widehat{\boldsymbol{\Sigma}}_u^{-1}) &= -\log\det(\boldsymbol{\Sigma}_u^{-1}) + \langle \widehat{\boldsymbol{\Sigma}}_u, \boldsymbol{\Sigma}_u^{-1} - \widehat{\boldsymbol{\Sigma}}_u \rangle + c_1 \\ &= -\log\det(\boldsymbol{\Sigma}_u^{-1}) + \text{tr}(\widehat{\boldsymbol{\Sigma}}_u \boldsymbol{\Sigma}_u^{-1}) + c_2 \end{aligned} \tag{2.105}$$

where (2.105) can be viewed as the negative Gaussian log-likelihood of the data, partially maximized with respect to the mean parameter. Adding an ℓ_1-penalty on the off-diagonal elements of $\boldsymbol{\Sigma}_u^{-1}$ to (2.105) gives us $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ as the solution to the ℓ_1-penalized Bregman Divergence

$$\widehat{\boldsymbol{\Sigma}}_u^{-1}(\rho) = \underset{\boldsymbol{\Sigma}_u^{-1} \in \mathcal{S}_{++}}{\text{arg min}} \; \{-\log\det(\boldsymbol{\Sigma}_u^{-1}) + \text{tr}(\widehat{\boldsymbol{\Sigma}}_u \boldsymbol{\Sigma}_u^{-1}) + \rho ||\boldsymbol{\Sigma}_u^{-1}||_1\}, \tag{2.106}$$

where ρ is the tuning hyperparameter and only here $||\boldsymbol{\Sigma}_u^{-1}||_1 := \sum_{i \neq j} |\Sigma_{u,ij}^{-1}|$ is defined to not penalize the diagonal elements. The routine can be easily implemented in the `R` package `glassoFast` or `CVglasso`.

However, as noted in [78], there are theoretical and practical benefits to modify (2.106) to the so-called weighted FGL (effectively just adaptive Lasso)

$$\widehat{\boldsymbol{\Sigma}}_u^{-1}(\rho) = \underset{\boldsymbol{\Sigma}_u^{-1} \in \mathcal{S}_{++}}{\text{arg min}} \; \{-\log\det(\boldsymbol{\Sigma}_u^{-1}) + \text{tr}(\widehat{\boldsymbol{\Sigma}}_u \boldsymbol{\Sigma}_u^{-1}) + \rho \sum_{i \neq j} \widehat{W}_{ii} \widehat{W}_{jj} |\Sigma_{u,ij}^{-1}|\}, \tag{2.107}$$

where $\widehat{\boldsymbol{W}}^2 = \text{diag}(\widehat{\boldsymbol{\Sigma}}_u)$. We suggest iterating between estimation of $\boldsymbol{\Sigma}_u^{-1}$ by optimizing (2.107) with the graphical Lasso algorithm and estimation of $\phi^s(\boldsymbol{z}, \boldsymbol{\Sigma}_u^{-1})$ as in (2.15) in Algorithm $3'$ below. For each iteration, we optimally select the penalty hyperparameter, ρ, via cross-validation.

We do not explore the theoretical properties of the sampling error induced by this weighted

FGL estimation procedure, but in some unreported Monte Carlo evidence we find that it performs well. The algorithm below details the overidentified estimation procedure for the case when $k_x = 0$.

Algorithm $3'$ Efficient GMM-FGL for ϕ^s (when $k_x = 0$):

- *Step 1:* Run PCA on (2.9) and obtain $\widehat{z}_t = \boldsymbol{S}'\widehat{\boldsymbol{Q}}\widetilde{\boldsymbol{y}}_{\cdot t}$ as the sample counterpart of (2.10).
- *Step 2:* Initialize $\widehat{\boldsymbol{\Sigma}}_u^{-1} = \boldsymbol{I}_N$.
- *Step 3:* Estimate (2.21) to obtain $\widehat{\varepsilon}$, initialize $\widehat{\boldsymbol{W}}_s = (\widehat{\boldsymbol{Z}}_s'\widehat{\boldsymbol{Z}}_s)^{-1}$ and obtain $\widehat{\boldsymbol{\theta}}^s_{2SLS}(\widehat{\boldsymbol{Z}}_s, \widehat{\boldsymbol{\Sigma}}_u^{-1})$.
- *Step 4:* Obtain $\boldsymbol{y}_{\widehat{E}}(\widehat{\boldsymbol{\Sigma}}_u^{-1})$.
- *Step 5:* Update $\widehat{\boldsymbol{W}}_s = \left(\frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{Z}}_{st}\widehat{\boldsymbol{Z}}_{st}'\widehat{u}^2_{\widehat{E}t}\right)^{-1}$, where $\widehat{u}_{\widehat{E}t} = y_{\widehat{E}t} - \widehat{\boldsymbol{\theta}}^s_{GMM}(\widehat{\boldsymbol{Z}}_s, \widehat{\boldsymbol{\Sigma}}_u^{-1})'\boldsymbol{f}_t$ and construct $\widehat{\boldsymbol{\theta}}^s_{GMM}(\widehat{\boldsymbol{Z}}_s, \widehat{\boldsymbol{\Sigma}}_u^{-1})$ as in (2.22).
- *Step 6:* Construct the sample counterpart of (2.103) to update $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ via (2.107) and update $\boldsymbol{y}_{\widehat{E}}(\widehat{\boldsymbol{\Sigma}}_u^{-1})$.
- *Step 7:* Iterate *Step 4* through *Step 6* until convergence.

Note, to obtain $\widehat{\boldsymbol{\psi}}_i$ in the sample counterpart of (2.103), we have $\widehat{\boldsymbol{\Lambda}} = T^{-1}\boldsymbol{y}_{\cdot\cdot}'\widehat{\boldsymbol{\eta}}$ and $\widehat{\phi}^s_{GMM}$ is an element of $\widehat{\boldsymbol{\theta}}^s_{GMM}$. In view of Algorithm 2, Algorithm $3'$ can be further extended to the case when $k_x > 0$. However, for brevity we omit the details.

Chapter 3

Unit-specific GIVs with Applications to the Automobile and Banking Industries

3.1 Introduction

Economic phenomena are inherently dynamic in nature. A natural justification for intertemporal relationships in economics is simply due to state dependence; which is the notion that the probability of experiencing an event today depends on whether the event was experienced yesterday. Dynamic relationships have allowed researchers to tease out speed of adjustment estimates, e.g., in financial settings: capital structure dynamics towards target leverage, amongst many other structural estimates. Moreover, misspecified models which ignore dynamics can lead to inconsistent estimates. Although the benefits of incorporating dynamics are important, weakly exogenous regressors impose a cost as well.

Dynamic panels with incidental parameters. The inclusion of weakly exogenous regressors generally leads to considerable challenges for estimation. In the panel data framework, this was analytically established in the seminal work of [79] for the case of a lagged dependent variable with individual fixed effects with N large and T fixed. In the fixed-T regime, the problem is more severe in that the structural estimates of interest are inconsistent in the presence of incidental parameters; this negative result is not surprising in light of the seminal work of [80]. Fortunately, when T is also large, the asymptotic bias (inconsistency) can be consistently estimated. This lead to generalizations of the so-called Nickell-bias in more recent work on large N and T panels by [81] who derive an analytical formula for the bias with general weakly exogenous regressors with individual and time effects and also by [82] who derive an analytical formula for the bias with a lagged dependent variable and interactive effects (factor error structure). An alternative approach to consistently estimating the bias based on an analytical expression entails employing jackknife techniques which also perform automatic bias reduction, yet do not require an analytical formula for the bias. Jackknife techniques date back to the seminal work of [83] to overcome time series bias of order $1/T$ and were generalized to the large N and T panel data context by [84]. To that end, [85] develop both analytical and jackknife bias correction techniques for large nonlinear panel data models with individual and time effects.

In the fixed-T case, the common approach to deal with this bias is the use of GMM techniques on suitably transformed equations which rid the problem of incidental parameters, as in [2], [86] and [87] who make use of internal instruments under a fixed T and large N asymptotics. These procedures can suffer from the many/weak instruments problem as T grows. Nevertheless, [88] derive the asymptotics for both T and N large for these approaches.

Constructing (granular) instrumental variables. In this paper, we consider a large dynamic panel data model with factor error structure and endogeneity even after controlling for common factors and propose the use of internal instruments to overcome the endogeneity problem. There are some existing methodologies which seek to eliminate the need to *find* an instrument (i.e., by using internal instruments). A leading example is the [2] framework in the context of estimating the speed of adjustment or state dependence parameters using dynamic panel data models with fixed effects (i.e., no cross-sectional dependence), in which higher order lags of the dependent variable serve as instruments for the included lags of the dependent variable. [1] (hereafter GK) illustrate that when the market under consideration is sufficiently concentrated, then one can use size-weighted idiosyncratic shocks to individual micro units, at each time period t, as an instrument for endogenous aggregate variables. This instrument was coined as the Granular Instrumental Variables approach (hereafer GIV) where "granular" is referring to the notion of dominant units having non-negligible affects on aggregate outcomes, see [7]. GK formulate their asymptotics under a fixed N large T regime under some strong assumptions, such as known factor loadings and $i.i.d.$ idiosyncratic errors. These assumptions were subsequently relaxed in [89] who allowed unknown factor loadings and a general covariance matrix for the idiosyncratic errors when both N and T to grow jointly to ∞ jointly. At the same time, to maintain instrumental relevance of the GIV, i.e., size-weighted idiosyncratic shocks, as $N \to \infty$, [89] assume that the size-distribution of the cross-sectional units follows a strictly stationary power-law in tail. Additionally, [89] also overidentified the structural parameters which enables standard overidentification tests to be carried out to asses model validity. [89] labeled their refinement to the GIV methodology as Feasible Granular Instrumental Variables or FGIV for short. However, both GK and [89] consider GIVs for aggregate endogenous variables. In many contexts, indeed the endogenous variables may be unit-specific, which limits the applicability of the GIV/FGIV in their current forms. This paper aims to fill that gap.

Contributions. In this paper, we generalize the set of admissible DGPs in our FGIV framework by allowing weakly exogenous covariates, although we will focus on the special case of a lagged dependent variable, in the presence of a pervasive factor error structure (i.e., incidental parameters). A novel iterative estimation approach is developed to achieve unbiased estimation of the structural parameters because we can no longer simply cross-sectionally demean our structural model to get rid of our endogenous variable to form the instrument using PCA as we did in [89] because in this paper, the endogenous variable is unit-specific, so the aforementioned demeaning procedure does not leave us with a simple factor structure. It is well known that estimation of the factor structure in the presence of covariates, x_{it}, generally leads to biased estimates, see [25] or [82]. However, by overidentifying the system, we find that heteroskedasticity along the cross-sectional dimension no longer leads to an asymptotic bias as it does in [25] and [82]. We propose an iterative GMM-PCA approach with a split-panel jackknife (SPJ) bias-correction technique to correct for the bias of order $1/T$ arising from the lagged dependent variable. We call this iterative estimation algorithm the SPJ GMM-PCA approach.

Notation. Throughout, let X_{wt} denote the cross-sectionally weighted average of a (random) variable X_{it}, $X_{wt} = \sum_{i=1}^{N} w_i X_{it}$. Weights used frequently throughout the paper are the share weights, or relative size weights, but which we simply refer to as size weights, $\underset{N\times 1}{\boldsymbol{S}} := (S_1, S_2, \ldots, S_N)$. Unless otherwise specified, we denote the L^2-norm as $||\cdot||$ or sometimes just $||\cdot||_2$; if another norm is used, it will be explicitly noted. Given a square matrix A, $\gamma_{max}(A)$ denotes the maximum eigenvalue of A. Joint convergence of N and T will be denoted as $(N,T) \xrightarrow{j} \infty$ while restrictions on relative rates of convergence will be explicitly noted. The operation $\xrightarrow{p}$ denotes convergence in probability while the operation $\overset{d}{\rightarrow}$ denotes convergence in distribution. The equation $y_n = \mathcal{O}_p(x_n)$ states that the vector of random variables y_n is at most of order x_n in probability; $a_n = \mathcal{O}(b_n)$ states that the deterministic sequence $\{a_n\}$ is at most of order b_n.

3.2 Model

In this paper we accommodate a unit specific endogenous covariate, where the endogeneity can arise due to an omitted variable problem, simultaneity (although we focus on estimation of a single equation) or spillover effects. We will overcome this endogeneity problem using a unit-specific GIV approah. The unit specific approach is in contrast to [1] and [89], where the endogenous object was an aggregate variate (e.g., size-weighted yield spreads, r_{St}, or global crude oil price, p_t). The model we consider is given by the following panel autoregressive distributed lag (hereafter ARDL) model with multifactor error structure

$$y_{it} = \delta x_{it} + \gamma y_{it-1} + \boldsymbol{\psi}' \boldsymbol{w}_{it} + v_{it}, \tag{3.1}$$

$$v_{it} = \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{it}, \tag{3.2}$$

where the scalar x_{it} is our endogenous variate. We abstract away from the presence of strictly exogenous covariates (whether they be aggregate, cross-section specific, or general), by setting $\boldsymbol{\psi} = 0$ in our exposition, as they introduce unnecessarily cumbersome notations. We also set the lag order of the dependent variable to 1 and the lag order of the covariates, $x_{it}, \boldsymbol{w}_{it}$ to 0 without loss of generality. As such, we can write the model compactly as

$$y_{it} = \boldsymbol{\beta}'\boldsymbol{X}_{it} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{it}, \tag{3.3}$$

where $\underset{2\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \delta & \gamma \end{pmatrix}'$ and $\underset{2\times 1}{\boldsymbol{X}_{it}} = \begin{pmatrix} x_{it} & y_{it-1} \end{pmatrix}'$. At first glance, (3.3) resembles the model of [90] and [25] if $\boldsymbol{X}_{it}$ is strictly exogenous or [91] and [82] if $\boldsymbol{X}_{it}$ is weakly exogenous.[1] Strict exogeneity is formally expressed as $\mathbb{E}(u_{it}|\boldsymbol{X}_{i\cdot}, \boldsymbol{\lambda}_i, \boldsymbol{f}) = 0 \; \forall \, i, t$, where $\underset{T\times 2}{\boldsymbol{X}_{i\cdot}} = \begin{pmatrix} \boldsymbol{x}_{i\cdot} & \boldsymbol{y}_{i\cdot,-1} \end{pmatrix}'$, $\boldsymbol{y}_{i\cdot,-1} = \begin{pmatrix} y_{i0} & \dots & y_{T-1} \end{pmatrix}'$ and $\underset{T\times r}{\boldsymbol{f}} = \begin{pmatrix} \boldsymbol{f}_1 & \dots & \boldsymbol{f}_T \end{pmatrix}'$, which of course would imply that $\mathbb{E}(u_{it}\boldsymbol{X}_{i\cdot}|\boldsymbol{\lambda}_i, \boldsymbol{f}) = 0 \; \forall \, i, t$, i.e., the idiosyncratic errors, u_{it}, are uncorrelated with all leads and lags of $\boldsymbol{X}_{it}$, once we condition on the factors and loadings. This condition simply cannot hold for weakly exogenous regressors because of feedback, which we will explicitly see shortly. Moreover, we allow the first column of $\boldsymbol{X}_{it}$, namely x_{it}, to be endogenous, that is $\mathbb{E}(u_{it}\boldsymbol{x}_{i\cdot}|\boldsymbol{y}_{i\cdot,-1}, \boldsymbol{\lambda}_i, \boldsymbol{f}) \neq 0$. Indeed, controlling for common factors in the composite error term can take care of a substantial portion of the endogeneity in u_{it}. Nevertheless, there are ample circumstances in economics where this assumption is too strong, e.g., [1] and [89] to name a few.

The dynamic nature of the panel renders the second column of $\boldsymbol{X}_{it}$, namely $\boldsymbol{y}_{it-1}$, as weakly exogenous, i.e., there is feedback from previous period's idiosyncratic shocks into future values of the dependent variable. To see why y_{it-1} is weakly exogenous, one can recursively substitute into (3.1) to arrive at

$$y_{it} = \gamma^t y_{i0} + \delta \sum_{j=0}^{t-1} \gamma^j x_{it-j} + \boldsymbol{\lambda}_i' \sum_{j=0}^{t-1} \gamma^j \boldsymbol{f}_{t-j} + \sum_{j=0}^{t-1} \gamma^j u_{it-j}, \tag{3.4}$$

[1] The CCE frameworks of [90] and [91] allow slope heterogeneity, weak factors and the number of factors need not be known but at the cost of imposing restrictions on the DGP of the covariates.

which makes it explicitly clear that the regressor y_{it-1} exhibits $\mathbb{E}(y_{it-1}u_{is}|\boldsymbol{x}_{i\cdot}, \boldsymbol{\lambda}_i, \boldsymbol{f}) \neq 0 \; \forall \, s < t$. Note that we assume that the process started long ago in the past to abstract away from issues pertaining to initial conditions. That is,

$$y_{it} = \delta \sum_{j=0}^{\infty} \gamma^j x_{it-j} + \boldsymbol{\lambda}_i' \sum_{j=0}^{\infty} \gamma^j \boldsymbol{f}_{t-j} + \sum_{j=0}^{\infty} \gamma^j u_{it-j}. \tag{3.5}$$

As mentioned above, we relax the strict exogeneity assumption on x_{it}, by allowing for endogeneity of x_{it} to persist even after controlling for common factors. This is our primary point of departure from the aforementioned papers. However, various types of this endogeneity problem have been investigated in (large) linear dynamic panels with multifactor error structure (aka interactive effects). [92] allow for endogenous covariates by assuming the existence of external instruments and develop an estimation procedure that relies on specifying a linear DGP for the covariates and then proxying the common factors as in [90]. On the other hand, [93] examine the case where y_{it-1} is measured with error, thus rendering it an endogenous variate and develop a minimum distance least squares (MD-LS) iterative approach. We don't consider measurement error in this paper, thus our endogeneity problem is more similar to that of [92] except that we do not impose any restrictions on the generating process of x_{it} and we propose the use of *internal* instruments to overcome our endogeneity problem using a suitably modified GIV/FGIV method as opposed to assuming the existence of external instruments, which we detail in Section 3.3.

3.3 Estimation

OLS-PCA The typical estimation approach for (large) linear dynamic panels with multifactor error structure entails an iterative principal components and least-squares approach (hereafter

OLS-PCA), [25], [82]. This stems from the observation that if $\boldsymbol{\beta}$ were known, then we are left with the problem of estimating an approximate factor model and if the interactive effects, $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$, were known we are left with a linear regression problem. That is,

$$y_{it}^*(\boldsymbol{\beta}) := y_{it} - \boldsymbol{\beta}'\boldsymbol{X}_{it} = \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{it}, \tag{3.6}$$

$$y_{it}^*(C_{it}) := y_{it} - C_{it} = \boldsymbol{\beta}'\boldsymbol{X}_{it} + u_{it}, \tag{3.7}$$

where we let $C_{it} := \boldsymbol{\lambda}_i'\boldsymbol{f}_t$ represent the interactive effects, which is also freqeuntly referred to as the common component in the literature. This gives rise to the OLS-PCA iterative procedure

$$\widehat{\boldsymbol{\beta}}_{ols} = \left(\sum_{i=1}^{N} \boldsymbol{X}_{i\cdot}'\boldsymbol{X}_{i\cdot}\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_{i\cdot}'\boldsymbol{y}_{i\cdot}^*(\widehat{C}_{ols}), \tag{3.8}$$

$$\left[\frac{1}{NT}\sum_{i=1}^{N} \boldsymbol{y}_{i\cdot}^*(\widehat{\boldsymbol{\beta}}_{ols})\boldsymbol{y}_{i\cdot}^{*\prime}(\widehat{\boldsymbol{\beta}}_{ols})\right]\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{f}}\,\boldsymbol{V}_{NT}, \tag{3.9}$$

$$\widehat{\boldsymbol{\Lambda}}' = \frac{1}{T}\widehat{\boldsymbol{f}}'\boldsymbol{y}_{i\cdot}^*(\widehat{\boldsymbol{\beta}}_{ols}), \tag{3.10}$$

where $\boldsymbol{V}_{NT}$ is the $r \times r$ diagonal matrix of eigenvalues ordered from greatest to smallest.[2] The properties of the OLS-PCA estimator are well understood, e.g., [25] and [82]. An alternative approach is to use the CCE estimators of [90] (in the static case) or [91] (in the dynamic case), which proxy the factor structure using cross-sectional averages of the dependent and independent variables under suitable restrictions.[3]

[2] The scaling $(NT)^{-1}$ in (3.15) is to ensure a proper limit for $\boldsymbol{V}_{NT}$ and does not affect $\widehat{\boldsymbol{f}}$, see Proposition A.1 in [25].

[3] More specifically in the static case without endogeneity wrt u_{it}, when N is large $\bar{u}_t \to 0$, which implies that $\bar{\boldsymbol{\lambda}}'\boldsymbol{f}_t = \bar{y}_t - \boldsymbol{\beta}'\bar{\boldsymbol{X}}_t$. When $\boldsymbol{X}_{it} = \boldsymbol{\Gamma}_i\boldsymbol{f}_t + \boldsymbol{e}_{it}$, $\boldsymbol{f}_t$ is sufficiently approximated by *linear* combinations of the dependent variable and importantly the independent variables as well when $\dim(\boldsymbol{X}_{it}) + 1 \geq r$ or $\mathbb{cov}(\boldsymbol{\Gamma}_i, \boldsymbol{\lambda}_i) = 0$, then [90] established regressions augmented with aforementioned cross-sectional averages behave *as if* the factors are controlled for, i.e., the unobserved cross-sectional dependence is accounted for. The CCE estimators of Pesaran enjoy excellent finite sample properties when the restrictions on $\boldsymbol{X}_{it}$ are imposed and satisfied.

GMM-PCA However, the OLS-PCA and CCE estimators are inconsistent in this framework. In what follows, we construct internal instruments and develop a novel iterative GMM-PCA approach.

To that end, let $\boldsymbol{W}$ be an $\ell \times \ell$ arbitrary positive definite weight matrix, but is optimally set as $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{\Omega}}^{-1}$, where $\widehat{\boldsymbol{\Omega}} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=\max\{L_f, L_u\}}^{T} \widehat{\boldsymbol{Z}}_{it} \widehat{\boldsymbol{Z}}_{it}' \widehat{u}_{it}^2$, with the unit specific instruments given by

$$\underset{\ell \times 1}{\boldsymbol{Z}_{it}} := \begin{pmatrix} \boldsymbol{y}_{it-1} & \boldsymbol{z}_{GIV,it} & \boldsymbol{z}_{GIV,it-1} & \dots & \boldsymbol{z}_{GIV,it-L_u} & \boldsymbol{f}_{t-1} & \dots & \boldsymbol{f}_{t-L_f} \end{pmatrix}', \tag{3.11}$$

where the number of instruments is given by $\ell := (\dim(\boldsymbol{X}_{it}) + L_u + rL_f)$ (where the lagged dependent variable self-instruments), $\dim(\boldsymbol{X}_{it}) = 2$ in our exposition and $\widehat{\boldsymbol{Z}}_{it}$ places a hat on all but the first element of $\boldsymbol{Z}_{it}$. Now, the unit specific instrument matrix stacks the elements of (3.11) across t and is given by

$$\underset{T^* \times \ell}{\boldsymbol{Z}_{i\cdot}} := \begin{pmatrix} \boldsymbol{y}_{i\cdot,-1} & \boldsymbol{z}_{GIV,i\cdot} & \boldsymbol{z}_{GIV,i\cdot,-1} & \dots & \boldsymbol{z}_{GIV,i\cdot,-L_u} & \boldsymbol{f}_{-1} & \dots & \boldsymbol{f}_{-L_f} \end{pmatrix}, \tag{3.12}$$

with $T^* := T - \max\{L_f, L_u\}$, and L_u and L_f determine the number of lagged GIVs, $z_{GIV,it}$, and lagged factors to include as instruments, respectively. Note, for a generic variate x_{it}, we define $\boldsymbol{x}_{i\cdot,-L} := \begin{pmatrix} x_{i1} & \dots & x_{iT-L} \end{pmatrix}'$ to be lagged L periods and stacked over t. Turning now to the unit-specific GIV, let $\boldsymbol{Q} := (\boldsymbol{I}_N - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}')$ and consider a generalization[4] of the unit-specific GIV

[4] The leave-one-out peers' GIV we formulate in (3.13) is a generalization of that introduced in [1] in the sense that our estimator incorporates dynamics and allows for unknown loadings.

introduced in [1]

$$\begin{aligned} z_{GIV,it} &= \boldsymbol{S}'_{(-i)}\boldsymbol{Q}_{(-i)}\boldsymbol{y}^*_{(-i),\cdot t}(\boldsymbol{\beta}) \qquad (3.13)\\ &= \sum_{j,k:,j,k\neq i} S_j Q_{jk} y^*_{kt}(\boldsymbol{\beta}), \end{aligned}$$

which is a leave-one-out GIV that collects peers' idiosyncratic shocks, where

$$\begin{aligned} \boldsymbol{S}_{(-i)} &:= \begin{pmatrix} S_1 & \dots & S_{i-1} & S_{i+1} & \dots & S_N \end{pmatrix}' \\ \boldsymbol{y}^*_{(-i),\cdot t}(\boldsymbol{\beta}) &:= \begin{pmatrix} y^*_{1t}(\boldsymbol{\beta}) & \dots & y^*_{i-1t}(\boldsymbol{\beta}) & y^*_{i+1t}(\boldsymbol{\beta}) & \dots & y^*_{Nt}(\boldsymbol{\beta}) \end{pmatrix}' \\ \boldsymbol{Q}_{(-i)} &:= (Q_{jk})_{j,k\neq i}, \end{aligned}$$

are the corresponding $(N-1)\times 1$ vectors and $(N-1)\times(N-1)$ matrix which has removed the ith datum. It is apparent that the degree of overidentification is given by $\ell-$ (No. of endog.) $= \ell - 1 = 1 + L_u + rL_f$. One could treat the triple (r, L_u, L_f) as hyperparameters and conduct a grid search to see which triple yields the smallest value of the loss function. In this paper, we assume that L_u and L_f are independent of (N,T) and we set $L_u = L_f = 1$ from here on out for ease of exposition.[5]

The estimator is made feasible much like the OLS-PCA estimator. The factor structure is estimated iteratively along with the regression coefficients, $\boldsymbol{\beta}$. The iterative GMM-PCA procedure

[5] One could let the number of moments grow, as in [2], but we do not pursue the moment proliferation extension in this paper.

is then concisely given by iterating the following equations

$$\widehat{\boldsymbol{\beta}}_{gmm} = \left(\sum_{i=1}^{N} \boldsymbol{X}_{i\cdot}' \widehat{\boldsymbol{Z}}_{i\cdot} \widehat{\boldsymbol{\Omega}}^{-1} \widehat{\boldsymbol{Z}}_{i\cdot}' \boldsymbol{X}_{i\cdot}\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_{i\cdot}' \widehat{\boldsymbol{Z}}_{i\cdot} \widehat{\boldsymbol{\Omega}}^{-1} \widehat{\boldsymbol{Z}}_{i\cdot}' \boldsymbol{y}_{i\cdot}^{*}(\widehat{\boldsymbol{C}}_{i\cdot}), \tag{3.14}$$

$$\left[\frac{1}{NT}\sum_{i=1}^{N} \boldsymbol{y}_{i\cdot}^{*}(\widehat{\boldsymbol{\beta}}_{gmm}) \boldsymbol{y}_{i\cdot}^{*}(\widehat{\boldsymbol{\beta}}_{gmm})'\right] \widehat{\boldsymbol{f}} = \widehat{\boldsymbol{f}}\, \boldsymbol{V}_{NT}, \tag{3.15}$$

$$\widehat{\boldsymbol{\Lambda}}' = \frac{1}{T}\widehat{\boldsymbol{f}}' \boldsymbol{y}_{\cdot\cdot}^{*}(\widehat{\boldsymbol{\beta}}_{gmm}), \tag{3.16}$$

where the estimators for the factors, loadings (and therefore the common component) and the $r \times r$ matrix of ordered eigenvalues, V_{NT}, are all functions of $\widehat{\boldsymbol{\beta}}_{gmm}$. Whereas, in the OLS-PCA procedure the estimators for the factors, loadings and V_{NT} are all functions of $\widehat{\boldsymbol{\beta}}_{ols}$, but in both cases we supress this dependence for notational ease. The system governed by (3.14), (3.15) and (3.16) is generalizing the OLS-PCA procedure in the sense that endogeneity is allowed to persist even after controlling for interactive effects. We now list the detailed steps of this iterative procedure below in Algorithm 5.

Interpretation of the instruments. We begin by interpreting the leave-one-out GIV, $z_{GIV,it}$. This instrument is composed of peers' idiosyncratic shocks, i.e., $z_{GIV,it} = z_{GIV,it}(u_{kt}) \;\forall\, k \neq i$. Thus, it is clear that we must further assume now that u_{it} and u_{jt} are uncorrelated for all $i \neq j$, which gives instrumental exogeneity, $\mathbb{E}(z_{GIV,it} u_{it}) = 0 \;\forall\, i, t$.[6] We argue that exogeneity is plausible because having controlled for observables, dynamics and latent common factors, $z_{GIV,it}$ is then

[6] We require a diagonal covariance matrix for the idiosyncratic errors in this framework, which is in contrast to [89], who allowed for a non-diagonal covariance matrix for the idiosyncratic errors, i.e., $\mathbb{E}(\boldsymbol{u}_{\cdot t}\boldsymbol{u}_{\cdot t}') = \boldsymbol{\Sigma}_u$, such that the rate at which the off-diagonals are growing in N is suitably restricted, e.g., using assumptions as in [37]. The reason a non-diagonal covariance matrix is possible in [89] is because the endogenous object was an aggregate variate and thus the moment used for estimation entailed aggregating the panel using precision weights; this aggregation remarkably rendered the aggregated regression error used in estimation as orthogonal to the instrument regardless of the covariance structure in u_{it}, see [89] for more details. Here, however, we do not aggregate the panel such that a non-diagonal covariance matrix is allowed and hence the assumption that $\mathbb{E}(\boldsymbol{u}_{\cdot t}\boldsymbol{u}_{\cdot t}') = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ is necessary for instrumental exogeneity and can not be relaxed.

Algorithm 5 FGIV GMM-PCA

```
OLS initialization: ignore the factor structure and initialize $\widehat{\boldsymbol{\beta}}_{gmm}$ as in (3.8), with $\widehat{C}_{it} = \boldsymbol{\lambda}_i'\boldsymbol{f}_t = 0 \ \forall \ i, t$.
while true do
    Update $y^*_{it}(\widehat{\boldsymbol{\beta}}_{gmm}) = y_{it} - \widehat{\boldsymbol{\beta}}'_{gmm}\boldsymbol{X}_{it}$.
    if $T < N$ then
        PCA: take eigendecomposition of the $T \times T$ matrix $(NT)^{-1}\sum_{i=1}^{N}\boldsymbol{y}^*_{i\cdot}(\widehat{\boldsymbol{\beta}}_{gmm})\boldsymbol{y}^*_{i\cdot}(\widehat{\boldsymbol{\beta}}_{gmm})'$ and obtain $\widehat{\boldsymbol{f}} = \sqrt{T} \times$ ($r$ eigenvectors corresponding to the $r$ largest eigenvalues) and obtain $\widehat{\boldsymbol{\Lambda}}' = T^{-1}\widehat{\boldsymbol{f}}'\boldsymbol{y}^*_{\cdot\cdot}(\widehat{\boldsymbol{\beta}}_{gmm})$.
    else
        PCA: take eigendecomposition of the $N \times N$ matrix $(NT)^{-1}\sum_{t=1}^{T}\boldsymbol{y}^*_{\cdot t}(\widehat{\boldsymbol{\beta}}_{gmm})\boldsymbol{y}^*_{\cdot t}(\widehat{\boldsymbol{\beta}}_{gmm})'$ and obtain $\widehat{\boldsymbol{\Lambda}} = \sqrt{N} \times$ ($r$ eigenvectors corresponding to the $r$ largest eigenvalues) and obtain $\widehat{\boldsymbol{f}} = N^{-1}\boldsymbol{y}^*_{\cdot\cdot}(\widehat{\boldsymbol{\beta}}_{gmm})\widehat{\boldsymbol{\Lambda}}$.
    end if
    Update $C_{it} = \boldsymbol{\lambda}_i'\boldsymbol{f}_t$ and $y^*_{it}(\widehat{C}_{it}) = y_{it} - \widehat{C}_{it}$.
    Construct instruments: obtain $\widehat{\boldsymbol{Q}} = \boldsymbol{I}_N - \widehat{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Lambda}})^{-1}\widehat{\boldsymbol{\Lambda}}'$ and form the unit specific GIV as the sample counterpart of (3.13). For given $r$, $L_u$ and $L_f$, form the full unit specific instrument matrix as the sample counterpart of (3.12).
    Initialize $\widehat{\boldsymbol{\Omega}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=\max\{L_f,L_u\}}^{T}\widehat{\boldsymbol{Z}}_{it}\widehat{\boldsymbol{Z}}'_{it}$ and obtain $\widehat{\boldsymbol{\beta}}_{gmm}$ as in (3.14).
    GMM: obtain $\widehat{\boldsymbol{\Omega}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=\max\{L_f,L_u\}}^{T}\widehat{\boldsymbol{Z}}_{it}\widehat{\boldsymbol{Z}}'_{it}(y_{it} - \widehat{\boldsymbol{\beta}}'_{gmm}\boldsymbol{X}_{it} - \widehat{C}_{it})^2$ and update $\widehat{\boldsymbol{\beta}}_{gmm}$ as in (3.14).
end while
end
```

hopefully providing a meaningful source unit-specific exogenous variation. Moreover, regarding instrumental relevance, we take $\mathbb{E}(z_{GIV,it}x_{it}) \neq 0 \;\forall\, i,t$, as given and we believe it is a rather mild assumption since idiosyncratic shocks have been shown to be important drivers of firm-level dynamics, e.g., [94], [95], [96] and [97] to name a few. Furthermore, the literature on granularity, e.g., [7], [12], [13] and [98], have further demonstrated that idiosyncratic shocks are important drivers of aggregate dynamics, let alone firm-level dynamics.

Turning now to interpretation of the lagged common factors as instruments. This is similar in spirit to [89], [5] and [6], who also use common factors as instruments. In our case, we use *lagged* factors as instruments; the use of lagged model objects as instruments is in some sense similar to [2], where higher order lagged dependent variables are used as instruments. However, in this setting, lagged dependent variables do not satisfy exogeneity, whereas lagged factors do, i.e., $\mathbb{E}(\boldsymbol{f}_{t-s}u_{it}) = 0 \; \forall s \geq 1$. Lagged factors are also relevant instruments under relatively mild conditions, (1) either the endogenous covariate contains autoregressive terms in its DGP (which we otherwise do not restrict) and/or (2) the factors are serially correlated. Both of these aforementioned conditions for relevance of the lagged factors as instruments are likely to be the rule rather than the exception in many economic applications. Either one of these conditions indeed leed to relevance, $\mathbb{E}(\boldsymbol{f}_{t-s}x_{it}) \neq 0 \;\forall\, i,t, \;\forall s \geq 1$.

The omitted variable interpretation and the hidden factor. To illustrate the omitted variable interpretation in this setting we follow a two-stage least squares formulation and set $L_f = L_u = 0$ (this parameterization yields the just identified case). Consider the linear projection of the endogenous covariate, x_{it}, onto everything that is exogenous with respect to u_{it}

$$x_{it} = \alpha_0 + \alpha_1 y_{it-1} + \alpha_2 \boldsymbol{\lambda_i'} \boldsymbol{f}_t + \pi z_{it} + \zeta_{it}, \tag{3.17}$$

$$= \boldsymbol{\alpha}' \boldsymbol{g}_{it} + \pi z_{it} + \zeta_{it}, \tag{3.18}$$

where $\boldsymbol{\alpha} := \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 \end{pmatrix}'$ and $\boldsymbol{g}_{it} := \begin{pmatrix} 1 & y_{it-1} & \boldsymbol{\lambda}_i' \boldsymbol{f}_t \end{pmatrix}'$ and z_{it} is the part of x_{it} that is uncorrelated with u_{it}: $\mathbb{E}(u_{it}\boldsymbol{g}_{it}|\boldsymbol{\lambda}_i, \boldsymbol{f}, \boldsymbol{y}_{i\cdot,-1}) = \mathbb{E}(u_{it}z_{it}|\boldsymbol{\lambda}_i, \boldsymbol{f}, \boldsymbol{y}_{i\cdot,-1}) = 0$. Whereas, ζ_{it} is the part of x_{it} that is correlated with u_{it}: $\mathbb{E}(u_{it}\zeta_{it}|\boldsymbol{\lambda}_i, \boldsymbol{f}, \boldsymbol{y}_{i\cdot,-1}) \neq 0$. This follows simply from the properties of linear projections and the fact that we allow endogeneity: $\mathbb{E}(u_{it}x_{it}|\boldsymbol{\lambda}_i, \boldsymbol{f}, \boldsymbol{y}_{i\cdot,-1}) \neq 0$. Finally, z_{it} is the instrumental variable that satisfies exogeneity with respect to u_{it} and ζ_{it}. We take $z_{it} = \boldsymbol{S}'_{(-i)} \boldsymbol{u}_{(-i),\cdot t}$ to be peers' size-weighted idiosyncratic shocks, with the ith datum removed, which we proxy as in (3.13). In principle, one could construct z_{it} using idiosyncratic shocks from other external sources, e.g., as in [96], so long as it satisfies instrumental relevancy. One could interpret the size-weighted idiosyncratic shocks themselves, $u_{St} := f_{t,hidden}$ as a hidden factor, from which we construct a unit-specific exogenous source of variation. That is,

$$x_{it} = \alpha_0 + \alpha_1 y_{it-1} + \alpha_2 \boldsymbol{\lambda_i'} \boldsymbol{f}_t + \pi u_{St} + \zeta^*_{it}, \tag{3.19}$$

$$= \alpha_0 + \alpha_1 y_{it-1} + \alpha_2 \boldsymbol{\lambda_i'} \boldsymbol{f}_t + \pi \underbrace{u_{(-i),St}}_{:= z_{it}} + \underbrace{\zeta^*_{it} + S_i u_{it},}_{:= \zeta_{it}}. \tag{3.20}$$

(3.20) makes it clear that the projection error, ζ_{it}, is composed of an exogenous component, ζ^*_{it}, and an endogenous component, $S_i u_{it}$. (3.20) also shows that ζ_{it} is very likely to be heteroskedastic.

3.4 Asymptotic Theory

OLS-PCA. From (3.8), we have

$$\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} = \boldsymbol{A}^{-1}_{\widehat{M}X}\left(\frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{X}'_{i\cdot}\boldsymbol{M}_{\widehat{f}}\boldsymbol{f}\boldsymbol{\lambda}_i + \frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{X}'_{i\cdot}\boldsymbol{M}_{\widehat{f}}\boldsymbol{u}_{i\cdot}\right), \tag{3.21}$$

where $\boldsymbol{M}_{\widehat{f}} := (\boldsymbol{I}_T - \widehat{\boldsymbol{f}}\widehat{\boldsymbol{f}}')$ and $\boldsymbol{A}_{\widehat{M}X} := \frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{X}'_{i\cdot}\boldsymbol{M}_{\widehat{f}}\boldsymbol{X}_{i\cdot}$. Considerable manipulation of (3.21) leads to the following representation

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) &= \boldsymbol{A}^{-1}_{QXM}\left[\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_{it}\boldsymbol{u}_{it} + \frac{\boldsymbol{B}_1}{T} + \frac{\boldsymbol{B}_2}{N} + \frac{\boldsymbol{B}_3}{T}\right] + o_p\left(\frac{1}{\sqrt{NT}}\right). &(3.22)\\ &= \boldsymbol{A}^{-1}_{QXM}\left[\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_{it}\boldsymbol{u}_{it} + \mathcal{O}_p(T^{-1}) + \mathcal{O}_p(N^{-1}) + \mathcal{O}_p(T^{-1})\right] + o_p\left(\frac{1}{\sqrt{NT}}\right) &(3.23)\end{aligned}$$

where the biases, $\boldsymbol{B}_j = \mathcal{O}_p(1)$ are due to [25] (his Proposition A.2 and Lemma A.8) and [82] (their Theorem 4.1 and Corollary 4.2). Let $\underset{N\times T}{\tilde{\boldsymbol{X}}_{c,\cdot\cdot}} := \boldsymbol{Q}_\Lambda\boldsymbol{X}'_{c,\cdot\cdot}\boldsymbol{M}_f$ for each covariate $c = 1, \ldots, k$, where $\boldsymbol{Q}_\Lambda$ is defined in (3.13) and $\underset{T\times N}{\boldsymbol{X}_{c,\cdot\cdot}}$ is covariate specific but stacked over i and t. Then define $\underset{k\times 1}{\tilde{\boldsymbol{X}}_{it}}$ for a given (i,t) by $\tilde{\boldsymbol{X}}_{it} := \begin{pmatrix}\tilde{\boldsymbol{X}}_{1,it} & \ldots & \tilde{\boldsymbol{X}}_{k,it}\end{pmatrix}'$. Finally, $\underset{k\times k}{\boldsymbol{A}_{QXM}} := \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_{it}\tilde{\boldsymbol{X}}'_{it}$. Thus, multiplying (3.23) by $\sqrt{NT}$ we have

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) = \boldsymbol{A}^{-1}_{QXM}\left[\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_{it}\boldsymbol{u}_{it} + \mathcal{O}_p\left(\sqrt{\frac{N}{T}}\right) + \mathcal{O}_p\left(\sqrt{\frac{T}{N}}\right) + \mathcal{O}_p\left(\sqrt{\frac{N}{T}}\right)\right] + o_p(1). \tag{3.24}$$

(3.24) makes it clear that sequences of (N, T) such that $N/T \to \kappa^2 > 0$ are required. Unfortunately, this gives rise to a limiting distribution of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})$ that is generally not centered at zero due to the three $\mathcal{O}_p(1)$ bias terms that will not vanish in (3.24). The asymptotic distribution is given by

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(\left(\kappa \boldsymbol{b}_1 + \kappa^{-1}\boldsymbol{b}_2 + \kappa \boldsymbol{b}_3\right), \mathbb{V}(\widehat{\boldsymbol{\beta}}_{ols})\right), \tag{3.25}$$

where $\kappa \boldsymbol{b}_j = \underset{(N,T)\to\infty}{\text{plim}} \boldsymbol{A}_{MXM}^{-1} \kappa \boldsymbol{B}_j$ for $j = 1, 3$ and $\kappa^{-1}\boldsymbol{b}_2 = \underset{(N,T)\to\infty}{\text{plim}} \boldsymbol{A}_{MXM}^{-1} \kappa^{-1}\boldsymbol{B}_2$.

(3.25) may seem like a negative result, however it is an improvement compared to the fixed T case. In particular, the biases arise due to the incidental parameter problem of [80]. See [99] for a survey of the problem. When T is large, the general inconsistency problem associated with incidental parameters is transformed to an asymptotic bias problem which can be taken care of analytically or via split-panel jackknife.

Some interesting special cases exist, in which the biases are exactly zero: (1) when the panel model is static (no weakly exogenous regressors) then $\boldsymbol{b}_1 = 0$; (2) when u_{it} are homoskedastic and uncorrelated across i, then $\boldsymbol{b}_2 = 0$; (3) when u_{it} are homoskedastic and serially uncorrelated over t, then $\boldsymbol{b}_3 = 0$. The second and third cases are rather intuitive, when u_{it} is homoskedastic or uncorrelated over i or t, there are no incidental parameters along the cross-sectional or time dimensions, respectively.

Bias-corrections. As $(N, T) \xrightarrow{j} \infty$, such that $N/T \to \kappa^2 > 0$ we have

$$(\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) = \frac{\boldsymbol{b}_1}{T} + \frac{\boldsymbol{b}_2}{N} + \frac{\boldsymbol{b}_3}{T} + o_p\left(\frac{1}{NT}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right) \tag{3.26}$$

[25] developed an analytical bias corrected estimator by constructing consistent estimates of $\boldsymbol{b}_2$ and $\boldsymbol{b}_3$. [82] extended Bai's static panel to a dynamic panel, which gives rise to $\boldsymbol{b}_1$ and they subsequently developed an analytical bias corrected estimator for this term. It is well known that an alternative approach to analytical bias correction is to use a nonparametric estimator for the resulting biases by using the so-called split-panel jackknife (SPJ hereafter) approach of [84], based on the seminal work of [83], in a time series context. The SPJ is also employed by [85] and [81] to name a few.

In this context, the SPJ estimator reduces biases of $\mathcal{O}_p(N^{-1}) + \mathcal{O}_p(T^{-1})$ to $o(C_{NT}^{-2})$ where $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. This has been shown to provide substanial benefits in numerous econometric models, [84]. In contrast to analytical formulae for the biases, the SPJ recenters the limiting distribution at zero by using a linear combination of subpanels to nonparameterically estimate the biases. The SPJ recenters the distribution obtained in (3.25) around zero, such that the SPJ estimator satisfies

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_{ols}^{spj} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(\underset{k\times 1}{\mathbf{0}}, \mathbb{V}(\widehat{\boldsymbol{\beta}}_{ols})\right), \tag{3.27}$$

where $\widehat{\boldsymbol{\beta}}_{ols}^{spj} := 3\widehat{\boldsymbol{\beta}}_{ols} - \bar{\boldsymbol{\beta}}_{N,T/2} - \bar{\boldsymbol{\beta}}_{N/2,T}$ with $\bar{\boldsymbol{\beta}}_{N,T/2} := \frac{1}{2}\left(\boldsymbol{\beta}_{N,:T/2} + \boldsymbol{\beta}_{N,T/2:}\right)$ where $\boldsymbol{\beta}_{N,:T/2}$ denotes the least squares estimator using all cross-sectional units and only the first half of the panel's time dimension and similarly $\boldsymbol{\beta}_{N,T/2:}$ is using all cross-sectional units and only the second half of the panel's time demnsion; $\bar{\boldsymbol{\beta}}_{N/2,T}$ is defined similarly with respect to the cross-seciton. To see why

SPJ works, note that the SPJ can be rewritten as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{ols}^{spj} &= 3\widehat{\boldsymbol{\beta}}_{ols} - \bar{\boldsymbol{\beta}}_{N,T/2} - \bar{\boldsymbol{\beta}}_{N/2,T} \\
(\widehat{\boldsymbol{\beta}}_{ols}^{spj} - \boldsymbol{\beta}) &= (\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) - (\bar{\boldsymbol{\beta}}_{N,T/2} - \widehat{\boldsymbol{\beta}}_{ols}) - (\bar{\boldsymbol{\beta}}_{N/2,T} - \widehat{\boldsymbol{\beta}}_{ols}) \\
&= \left[\frac{\boldsymbol{b}_1}{T} + \frac{\boldsymbol{b}_2}{N} + \frac{\boldsymbol{b}_3}{T} + o_p\left(\frac{1}{NT}\right)\right] - \left[\frac{\boldsymbol{b}_1}{T} + \frac{\boldsymbol{b}_3}{T} + o_p\left(\frac{1}{NT}\right)\right] - \left[\frac{\boldsymbol{b}_2}{N} + o_p\left(\frac{1}{NT}\right)\right] \\
&= o_p\left(\frac{1}{NT}\right)
\end{aligned}$$

Unfortunately, in practice, partitioning the cross-section to respect the cross-sectional dependence and/or clusters is not trivial since no natural ordering exists in the cross-section, as it does in the time dimension, i.e. if $|i - j|$ is large in magnitude, it need not imply that units i and j are uncorrelated. [85] suggest clustering when possible and taking the average of P random partitions to remove the effects of taking only a single arbitrary cross-sectional split. This approach, albeit valid, can be computationally cumbersome.

Bai and Liao (2017) An alternative approach entails using large covariance matrices to remove the $\boldsymbol{b}_2$ term from the limiting distribution using the so-called efficient principal components approach (hereafter EPC) of [41]. Their approach is motivated by the special case where $\boldsymbol{b}_2$ is $\mathbf{0}$ when u_{it} are homoskedastic and uncorrelated across i. As such, [41] altered the OLS-PCA objective function in a similar spirit to that of generalized least squares (hereafter GLS)[7]

$$\widehat{\boldsymbol{\beta}}_{epc}(\boldsymbol{\Sigma}_u^{-1}) = \arg\min_{\boldsymbol{\beta}} \min_{\boldsymbol{\Lambda},\{\boldsymbol{f}_t\}} \sum_{t=1}^{T} (\boldsymbol{y}_{\cdot t} - \boldsymbol{X}_{\cdot t} - \boldsymbol{\Lambda}\boldsymbol{f}_t)'\boldsymbol{\Sigma}_u^{-1}(\boldsymbol{y}_{\cdot t} - \boldsymbol{X}_{\cdot t} - \boldsymbol{\Lambda}\boldsymbol{f}_t), \tag{3.28}$$

where the objective function for OLS-PCA did not take the cross-sectional correlation structure in

[7] [41] refrain from referring to (3.28) as a GLS estimator since they allow u_{it} to be serially correlated over t. If u_{it} were $i.i.d.$ over t then (3.28) is indeed a GLS estimator.

u_{it} into account. That is, $\widehat{\boldsymbol{\beta}}_{ols} = \widehat{\boldsymbol{\beta}}_{epc}(\boldsymbol{I}_N)$. For a real symmetric positive definite matrix $\boldsymbol{\Sigma}_u^{-1}$, there exists a unique matrix $\boldsymbol{C} := \boldsymbol{\Sigma}_u^{-1/2}$ such that $\boldsymbol{CC} = \boldsymbol{\Sigma}_u^{-1}$. Then (3.28) is optimizing over *transformed* sum of squared errors $\sum_{t=1}^{T} \boldsymbol{u}_{\cdot t}^{*\prime} \boldsymbol{u}_{\cdot t}^{*}$ where $\boldsymbol{u}_{\cdot t}^{*} := \boldsymbol{C}\boldsymbol{u}_{\cdot t}$ and most importantly we have $\mathbb{E}(\boldsymbol{u}_{\cdot t}^{*}\boldsymbol{u}_{\cdot t}^{*\prime}) = \boldsymbol{C}\boldsymbol{\Sigma}_u\boldsymbol{C} = \boldsymbol{I}_N$, by construction, which renders $\boldsymbol{b}_2 = \boldsymbol{0}$. Note that the EPC is generally not a GLS estimator unless u_{it} are serially uncorrelated and homoskedastic over t. [41] proceed to formulate an analytical bias-corrected estimator to correct for the $\boldsymbol{b}_3$ term and they do not consider weakly exogenous regressors, therefore a $\boldsymbol{b}_1$ term does not arise in their setting.

GMM-PCA. In light of [41], we see that weighting the observations inversely proportional to their cross-sectional covariances eliminated the $\boldsymbol{b}_2$ term from the limiting distribution. We argue that since our GMM-PCA procedure, defined in (3.14), is weighting the observations by the variance of the moment condition used in estimation, $\boldsymbol{\Omega}^{-1}$, that the corresponding $\boldsymbol{b}_2$ term will also be zero here. Indeed, our simulations lend credence to this argument. Intuitively, we note that the weight matrix in GMM can be thought of as a GLS estimator in this transformed regression which is potentially heteroskedastic over i

$$\underset{\ell\times 1}{\boldsymbol{Z}_{i\cdot}'\boldsymbol{y}_{i\cdot}} = \boldsymbol{Z}_{i\cdot}'\boldsymbol{X}_{i\cdot}\boldsymbol{\beta} + \boldsymbol{Z}_{i\cdot}'\boldsymbol{f}\boldsymbol{\lambda}_i + \boldsymbol{Z}_{i\cdot}'\boldsymbol{u}_{i\cdot}, \tag{3.29}$$

since $\mathbb{V}(\boldsymbol{Z}_{i\cdot}'\boldsymbol{u}_{i\cdot}|\boldsymbol{Z}_{i\cdot}) = \boldsymbol{\Omega}$, thus $\boldsymbol{\Omega}^{-1}\mathbb{V}(\boldsymbol{Z}_{i\cdot}'\boldsymbol{u}_{i\cdot}|\boldsymbol{Z}_{i\cdot}) = \boldsymbol{I}_\ell$, by construction. This suggests that $\boldsymbol{b}_2$ should theoretically be zero and this is what we numerically confirm via simulations.

As in the OLS-PCA case, let $\underset{N\times T}{\tilde{\boldsymbol{X}}_{c,\cdot\cdot}} := \boldsymbol{Q}_\Lambda \boldsymbol{X}_{c,\cdot\cdot}'\boldsymbol{M}_f$ for each covariate $c = 1, \ldots, k$, where $\underset{T\times N}{\boldsymbol{X}_{c,\cdot\cdot}}$ is covariate specific but stacked over i and t, then define $\underset{k\times 1}{\tilde{\boldsymbol{X}}_{it}}$ for a given (i, t) by $\tilde{\boldsymbol{X}}_{it} := \begin{pmatrix} \tilde{\boldsymbol{X}}_{1,it} & \ldots & \tilde{\boldsymbol{X}}_{k,it} \end{pmatrix}'$.

In light of [41] and [89], it can be shown that the estimator defined in (3.14) has the following representation

$$\widehat{\boldsymbol{\beta}}_{gmm} - \boldsymbol{\beta} = \left(\frac{1}{NT} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_{i\cdot}' \boldsymbol{Z}_{i\cdot} \boldsymbol{\Omega}^{-1} \boldsymbol{Z}_{i\cdot}' \tilde{\boldsymbol{X}}_{i\cdot} \right)^{-1} \left(\frac{1}{NT} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_{i\cdot}' \boldsymbol{Z}_{i\cdot} \boldsymbol{\Omega}^{-1} \boldsymbol{Z}_{i\cdot}' \boldsymbol{u}_{i\cdot} + \frac{\boldsymbol{B}_1}{T} \right) + o_p\left(\frac{1}{\sqrt{NT}} \right),$$

thus,

$$\begin{aligned} \sqrt{NT}(\widehat{\boldsymbol{\beta}}_{gmm} - \boldsymbol{\beta}) &= \left(\frac{1}{NT} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_{i\cdot}' \boldsymbol{Z}_{i\cdot} \boldsymbol{\Omega}^{-1} \boldsymbol{Z}_{i\cdot}' \tilde{\boldsymbol{X}}_{i\cdot} \right)^{-1} \left(\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_{i\cdot}' \boldsymbol{Z}_{i\cdot} \boldsymbol{\Omega}^{-1} \boldsymbol{Z}_{i\cdot}' \boldsymbol{u}_{i\cdot} + \kappa \boldsymbol{B}_1 \right) + o_p(1), \\ &\xrightarrow{d} \mathcal{N}\left(\kappa \boldsymbol{b}_1, \mathbb{V}(\widehat{\boldsymbol{\beta}}_{gmm}) \right), \end{aligned} \tag{3.30}$$

where $\kappa \boldsymbol{b}_1 = \text{plim}\, \boldsymbol{A}_{\Omega ZX}^{-1} \kappa\, \boldsymbol{B}_1$ with $\boldsymbol{A}_{\Omega ZX}^{-1} := \left(\frac{1}{NT} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_{i\cdot}' \boldsymbol{Z}_{i\cdot} \boldsymbol{\Omega}^{-1} \boldsymbol{Z}_{i\cdot}' \tilde{\boldsymbol{X}}_{i\cdot} \right)^{-1}$ and the $k \times k$ covariance matrix is given by

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}_{gmm}) = \left(\mathbb{E}(\tilde{\boldsymbol{X}}_{it} \boldsymbol{Z}_{it}')\, \boldsymbol{\Omega}^{-1}\, \mathbb{E}(\boldsymbol{Z}_{it} \tilde{\boldsymbol{X}}_{it}') \right)^{-1}. \tag{3.31}$$

Given (3.30), the SPJ GMM-PCA estimator defined as

$$\widehat{\boldsymbol{\beta}}_{gmm}^{spj} = 2\widehat{\boldsymbol{\beta}}_{gmm} - \bar{\boldsymbol{\beta}}_{N,T/2}, \tag{3.32}$$

satisfies

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_{gmm}^{spj} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(\underset{k\times 1}{\mathbf{0}}, \mathbb{V}(\widehat{\boldsymbol{\beta}}_{gmm}) \right), \tag{3.33}$$

notably, the theoretical variance does not change in the presence of bias-correction and can be estimated with

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}}_{gmm}) = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\tilde{\boldsymbol{X}}}_{it} \widehat{\boldsymbol{Z}}'_{it} \, \widehat{\boldsymbol{\Omega}}^{-1} \, \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\boldsymbol{Z}}_{it} \widehat{\tilde{\boldsymbol{X}}}'_{it} \right)^{-1}, \tag{3.34}$$

where $\widehat{\tilde{\boldsymbol{X}}}_{it}$ is the sample counterpart of $\underset{N\times T}{\tilde{\boldsymbol{X}}_{c,\cdot\cdot}} := \boldsymbol{Q}_{\Lambda} \boldsymbol{X}'_{c,\cdot\cdot} \boldsymbol{M}_f$ and uses the sample values of $\boldsymbol{Q}_{\Lambda}$ and $\boldsymbol{M}_f$, respectively.

3.5 Monte Carlo

In the experiments below, the first 1000 observations are discarded for each simulation to ensure stationary data. The model is given by

$$y_{it} = 2x_{it} + 0.6y_{it-1} + \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + u_{it},$$

where $\boldsymbol{\beta} = \begin{pmatrix} \delta & \gamma \end{pmatrix}' = \begin{pmatrix} 2 & 0.6 \end{pmatrix}'$ and the loadings λ_{1i}, λ_{2i} are generated as $\mathcal{N}(3,1)$. The endogenous regressor and factors are generated according to

$$\begin{aligned} x_{it} &= \gamma_x x_{it-1} + c_1 \lambda_{1i} f_{1t} + c_2 \lambda_{2i} f_{2t} + c_3 u_{St} + e_{it}, \\ &= 0.6 x_{it-1} + 0.50 \lambda_{1i} f_{1t} + .50 \lambda_{2i} f_{2t} + 1.5 u_{St} + e_{it}, \\ f_{jt} &= \varphi_j f_{jt-1} + \varepsilon_{jt} \quad j = 1, 2, \end{aligned}$$

with $\gamma_x = \varphi_j = 0.6$, and ε_{jt} is generated as $\mathcal{N}(0, 1 - \varphi_j^2)$ for $j = 1, 2$.

Design 1 For Design 1, the idiosyncratic errors are generated as

$$\begin{pmatrix} u_{it} \\ e_{it} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_u^2 & \sigma_{ue} \\ \sigma_{ue} & \sigma_e^2 \end{bmatrix}\right), \tag{3.35}$$

with $\sigma_u^2 = 4$, $\sigma_e^2 = \sigma_{ue} = 1$; thus, $\rho_{ue} = 0.5$. Note that σ_{ue} controls the degree of endogeneity.

Design 2 Whereas, for Design 2 we consider heteroskedasticity over i by defining $v_{it} \sim \mathcal{N}(0, \sigma_{vi}^2)$ where $\sigma_{vi}^2 \sim U[0.8, 1.2]$, $e_{it} \sim \mathcal{N}(0,1)$ and $u_{it} = \sigma_{ue}e_{it} + v_{it} = e_{it} + v_{it}$. Then, $\rho_{ue} \approx 0.5$.

The market share weights are given by

$$S_i = \frac{\mathcal{S}_i}{\sum_j \mathcal{S}_j} \tag{3.36}$$

$$\mathcal{S}_i = \left(\frac{i}{N}\right)^{-\frac{1}{\mu}}. \tag{3.37}$$

We use the SPJ GMM-PCA estimator defined in (3.32) to estimate δ and γ. The results are reported in Table's 3.1 and 3.2.

3.6 Empirical Application to the Demand for New Automobiles

Logit demand with factor structure In the traditional logit demand case without random coefficients the problem from moving from individual to aggregate demand is solved analytically and not via simulation. More specifically, let U_{ijt} denote the utility derived by consumer i from consuming product j in the market at time t, ε_{ijt} denotes the unobserved idiosyncratic error in consumer utility, δ_{jt} denotes the mean utility level which consists of the usual vector of observed product characteristics for this application $\boldsymbol{x}_{jt}$ (taken to be strictly exogenous), prices p_{jt} (taken to be endogenous) as well as unobserved product characteristics ξ_{jt}. As δ_{jt} is generally unknown, we follow the literature and assume ε_{ijt} has Weibull (or type 1 extreme value) distribution function, i.e., $e^{-e^{\varepsilon_{ijt}}}$. This assumption yields the traditional logit model for BLP market shares, i.e., $s_{jt} = \frac{\exp(\delta_{jt})}{1+\sum_{k=1}^{J}\exp(\delta_{kt})}$. The analytically tractable case enables us to treat estimation as a standard linear panel regression model with interactive effects because in this case $\delta_{jt} = \log(s_{jt}) - \log(s_{0t})$ and there is no need to solve for the mean utility by numerical methods , see [100] for more details. As this application is primarily for illustration purposes, we proceed with this simplified model even though the implied substitution patterns are generally restrictive and unrealistic. The demand system is then given by

$$U_{ijt} = \delta_{jt} + \varepsilon_{ijt} \tag{3.38}$$

$$\delta_{jt} = \log(s_{jt}) - \log(s_{0t}) = \boldsymbol{x}'_{jt}\boldsymbol{\beta} - \alpha p_{jt} + \xi_{jt} \tag{3.39}$$

$$\xi_{jt} = \alpha_j + \theta_t + \boldsymbol{\lambda}'_j \boldsymbol{f}_t + u_{jt}. \tag{3.40}$$

Following [101] we assume that the unobserved product characteristics in the mean utility follow a factor model but we explicitly allow for two-way fixed effects, where α_j denote the product fixed effects and θ_t denote the time effects. To accommodate the two-way fixed effects, before applying our iterative GMM-PCA procedure, for a generic a_{jt}, we transform all observables to $\dot{a}_{jt} := a_{jt} - \bar{a}_t - \bar{a}_j + \bar{a}$, where $\bar{a}_j = \frac{1}{T}\sum_{t=1}^{T} a_{jt}$, $\bar{a}_t = \frac{1}{N}\sum_{j=1}^{N} a_{jt}$ and $\bar{a} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{j=1}^{N} a_{jt}$.

In terms of the dotted variables, the model is essentially as our theoretical exposition with the cross-sectional index now representing products $j = 1, \ldots, J$. As the method in this paper requires a balanced panel, we follow [102] and [101] and aggregate the index j from individual car make (which exhibits entry and exit) to manufacturer-size level and assume that consumers choose between aggregate composites of cars. More specifically, the manufacturers are GM, Ford, Chrysler and Other (this group includes Toyota, Volkswagon, Datsen/Nissan, Honda, Suburu and 'Rest' taken as all other manufacturers). Within a manufacturer, the makes are aggregated to the segment level. For example, within GM the makes are aggregated to Chevy, Oldsmobile, Pontiac, Buick and Cadillac. Finally, within a segment, either it is partitioned to form another product based on a size class of 'small', 'large' or 'all' where 'all' denotes no partiioning by size because cars in this segment level are primarily uniformly sized. In total, this yields $J = 23$ product aggregates for the years 1973-1988 ($T = 16$), which can be seen in Table 3.3.

GIV shares vs BLP shares Moreover, the BLP market shares corresponding to make-size, are not the appropiate shares in terms of forming the GIV. More specifically, the BLP shares are formed as

$$s_{jt} = \frac{q_{jt}}{\#\texttt{households}_{jt}}, \tag{3.41}$$

where q_{jt} denotes the quantity of product j purchased in the national automobile market (year) t and $\#\texttt{households}_{jt}$ denotes the number of households (approximately on the order of 100 million). These shares are appropiate for the BLP model formulation as they capture the outside option available to households, namely, to not purchase an automobile, i.e., since it is a durable good. However, for the purpose of constructing the GIV we form the normalized shares as

$$S_{jt} = \frac{q_{jt}}{\sum_i q_{it}}, \tag{3.42}$$

and refer to the latter shares as GIV shares (uppercase convention) and the former shares as BLP shares (lowercase convention). We form the GIV as

$$z_{jt} = \sum_{k,i,t:k,i \neq j} S_{i,t-1} Q_{ki} \dot{y}^{*}_{kt}(\boldsymbol{\beta}), \tag{3.43}$$

where $S_{i,t-1}$ is different from our theoretical exposition in Section 3.3 in so far as the shares are now time-varying so we use peers' one-period lagged shares in the formulation of the instrument.

Results. The results are in Table 3.4. The J-statistic has a p-value of 0.096 which indicates that we fail to reject the null hypothesis of a valid model at the 1% and 5% significance levels. Moreover, both the robust- and effective-F statistics are above the benchmark value of 10, suggesting that the strength of the instruments are a major concern. In column (1) we estimate the parameters using OLS, in column (2) we estimate the parameters using SPJ OLS-PCA and in column (3) we estimate the parameters using SPJ GMM-PCA. In column (1), we find that OLS is not getting the magnitudes correct as it estimates a positive demand elasticity and the space and air conditioning product characteristics yield negative marginal utilities (although insignificant). In column (2), we find that the demand elasticity is negative as one would expect, but it is not significant. In column (3), we find that all product characteristics yield positive marginal utility and demand elasticity is more negative and significant at the 1% level. Taken together, we find that the SPJ GMM-PCA delivers promising results in estimating the demand for new automobiles.

3.7 Empirical Application to Determinants of Banks' Capital Adequacy Ratios

In this empirical application, we illustrate the SPJ GMM-PCA procedure in estimation of the determinants of banks' capital adequacy ratios. The panel is from a random sample of 300 U.S. banks, with each bank observed over 56 quarters from 2006 Q1 - 2019 Q4. The data are publicly available for download at the FDIC website; but we obtained it directly from the Stata package due to [103]. Consider the following model

$$y_{it} = \delta_0 x_{it} + \delta_1 x_{it-1} + \gamma y_{it-1} + \boldsymbol{\psi}_0' \boldsymbol{w}_{it} + \boldsymbol{\psi}_1' \boldsymbol{w}_{it-1} + \boldsymbol{\lambda}_i' \boldsymbol{f_t} + u_{it}, \tag{3.44}$$

where $y_{it} = \texttt{CAR}_{it}$, $x_{it} = \texttt{liquidity}_{it}$, $\boldsymbol{w}_{it} := \begin{pmatrix} \texttt{ROA}_{it} & \texttt{size}_{it} \end{pmatrix}'$. The variable $\texttt{CAR}_{it}$ denotes the capital adequacy ratio, which is proxied by the ratio of tier 1 (core) capital over risk-weighted assets. $\texttt{liquidity}_{it}$ is proxied by the loan-to-deposit ratio. Thus, higher values of the variable $\texttt{liquidity}_{it}$ actually represent lower levels of liquidity. $\texttt{ROA}_{it}$ denotes return on assets at time t, defined as annualized net income expressed as a percentage of average total assets, it is a measure of profitability. $\texttt{size}_{it}$ is proxied by the log of bank's total assets at time t.

The coefficients can be interpreted as follows: the coefficient γ reflects state dependence. That is, it reflects the adjustment costs that prevents banks from achieving optimal levels of capital adequacy instantaneously. The coefficient δ_0 measures the effect of liquidity on capital adequacy behavior. If $\delta_0 > 0$, this implies that lower liquidity levels nudge banks to raise their capital reserves, possibly to offset their risk exposure. This would imply that the Basel III implementation is having effects in the right direction.

In (3.44), we posit that liquidity is correlated with u_{it} even after controlling for common factors and we form the unit-specific GIV as

$$z_{GIV,it} = \sum_{k,j,t:k,j\neq i} S_{j,t-1} Q_{kj} y^*_{kt}(\boldsymbol{\beta}), \tag{3.45}$$

which is how we formed the GIV in the previous empirical application as well, to accommodate time varying shares.

Descriptive statistics and structural break. As the SPJ procedure splits the panel along the time series dimension, it is important that we have stationarity over t. However, our sample entails a structural break which can be seen in the summary statistics in Table 3.5. As a result, we work with the Post-Great Recession Sample from 2009 Q4 - 2019 Q4, which amounts to discarding the first 16 quarters of data.

Results. The results are in Table 3.6. The J-statistic has a p-value of 0.249 which indicates that we fail to reject the null hypothesis of a valid model at all conventional significance levels. When it comes to the F-statistic, we have a robust-F of 14.563, which is above the threshold of 10; whereas, the effective-F is only 5.840, which suggests that weak instruments are of concern in this application, see [104] who show that the robust-F may be high even when instruments are weak. Nevertheless, columns (1), (2) and (3) indicate that although all methods estimate a positive effect of liquidity on capital adequacy ratios, that OLS may be underestimating the effect of liquidity. Focusing on column (2), the point estimate for liquidity is 5.429, thus, on average, a 1 standard deviation increase in liquidity leads to approximately a 1/3 standard deviation increase in capital adequacy ratios. This implies that when banks face a liquidity squeeze they tend to partially absorb that by raising equity. This implies that the Basel III implementation is having effects in the right direction.

3.8 Concluding Remarks

In this paper, the GIV methodology introduced by [1] has been further developed to accommodate large dynamic panels with unit specific endogenous variates, which require unit-specific GIVs. We develop a SPJ GMM-PCA iterative procedure to estimate the structural parameters of interest. Overidentification tests can be carried out to test model validity. In the first application to the demand for new automobiles we fail to reject the null hypothesis of a valid model and the procedure estimates a downward sloping demand curve as economic theory would suggest. Moreover, the product characteristics yield positive marginal utilities, again as expected. Whereas, the other estimation techniques were less favorable. In the second application to the determinants of banks' capital adequacy ratios the procedure results in weak instruments; whereas the OLS-PCA method seems to be taking care of the endogeneity by estimating the interactive effects. Nevertheless, across numerous estimation techniques we find evidence which suggests that banks respond to liquidity crunches by raising their capital adequacy ratios. Future research which further develops the SPJ GMM-PCA asymptotics would be of interest along with considerations of weak factors, slope heterogeneity and unbalanced panels with data not missing at random.

Table 3.1: Mean point estimates and RMSE for design 1.

Finite sample properties for Design 1						
	N	T	$\widehat{\delta}^{spj}_{ols}$	$\widehat{\delta}^{spj}_{gmm}$	$\widehat{\gamma}^{spj}_{ols}$	$\widehat{\gamma}^{spj}_{gmm}$
1	300	60	3.7682	1.974	1.0083	0.594
			(0.2375)	(0.342)	(0.0821)	(0.061)
2	300	100	3.7294	1.998	0.9997	0.596
			(0.2492)	(0.266)	(0.0896)	(0.046)

Notes: We estimate the SPJ GMM-PCA using (3.32) with the full sample estimate coming from Algorithm 5 while the half-panel estimate is the simple average of discarding the first half of the time series observations for each i and running Algorithm 5 and discarding the second half of the time series observations for each i and running Algorithm 5. The SPJ OLS-PCA is estimated similarly, except we iterate (3.8), (3.9) until convergence. For SPJ GMM-PCA, we set $L_f = L_u = 1$, yielding a total number of instruments for x_{it} as 4.

Table 3.2: Mean point estimates and RMSE for design 2.

Finite sample properties for Design 2.						
	N	T	$\widehat{\delta}^{spj}_{ols}$	$\widehat{\delta}^{spj}_{gmm}$	$\widehat{\gamma}^{spj}_{ols}$	$\widehat{\gamma}^{spj}_{gmm}$
1	300	60	4.078	1.951	0.971	0.6016
			(0.321)	(0.876)	(0.102)	(0.100)
2	300	100	3.974	1.968	0.995	0.601
			(0.220)	(0.687)	(0.068)	(0.083)

Notes: We estimate the SPJ GMM-PCA using (3.32) with the full sample estimate coming from Algorithm 5 while the half-panel estimate is the simple average of discarding the first half of the time series observations for each i and running Algorithm 5 and discarding the second half of the time series observations for each i and running Algorithm 5. The SPJ OLS-PCA is estimated similarly, except we iterate (3.8), (3.9) until convergence. For SPJ GMM-PCA, we set $L_f = L_u = 1$, yielding a total number of instruments for x_{it} as 4.

Table 3.3: Summary statistics for product-aggregates for automobile demand application

Product	Make	Size	Manuf.	GIV Mkt.	BLP Mkt.	Price	hp/weight	MP$	Size
No.		Class		Share %	Share %				
				(avg)	(avg)	(avg)	(avg)	(avg)	(avg)
Product-aggregate characteristics									
1	CV (Chevrolet)	small	GM	7.454	0.783	7.251	0.364	2.45	1.136
2	CV	large	GM	10.043	1.099	8.829	0.378	1.714	1.472
3	OD (Oldsmobile)	small	GM	0.834	0.087	7.18	0.359	2.136	1.209
4	OD	large	GM	7.897	0.863	10.701	0.376	1.637	1.556
5	PT (Pontiac)	small	GM	2.117	0.226	6.597	0.349	2.364	1.177
6	PT	large	GM	4.985	0.541	8.412	0.363	1.693	1.499
7	BK (Buick)	all	GM	7.861	0.838	9.914	0.373	1.705	1.474
8	CD (Cadillac)	all	GM	2.760	0.292	19.688	0.383	1.536	1.505
9	FD (Ford)	small	Ford	7.625	0.823	6.049	0.347	2.27	1.188
10	FD	large	Ford	7.907	0.859	8.692	0.354	1.641	1.553
11	MC (Mercury)	small	Ford	1.283	0.133	7.039	0.381	2.301	1.203
12	MC	large	Ford	3.461	0.379	8.920	0.349	1.712	1.537
13	LC (Lincoln)	all	Ford	1.440	0.158	19.354	0.376	1.416	1.670
14	PL (Plymouth)	small	Chrys	2.795	0.292	6.407	0.366	2.374	1.16
15	PL	large	Chrys	1.939	0.220	7.839	0.333	1.618	1.554
16	DG (Dodge)	small	Chrys	2.859	0.302	6.984	0.37	2.516	1.153
17	DG	large	Chrys	1.942	0.218	7.854	0.331	1.611	1.552
18	TY (Toyota)	all	Other	5.201	0.5433	8.507	0.400	2.361	1.110
19	VW (Volkswagen)	all	Other	1.521	0.167	8.974	0.365	2.346	1.062
20	DT/NI (Datsen/Nissan)	all	Other	3.960	0.411	8.959	0.418	2.338	1.108
21	HD (Honda)	all	Other	3.901	0.409	7.397	0.370	2.667	1.036
22	SB (Subaru)	all	Other	0.949	0.100	6.255	0.379	2.855	1.034
23	REST	all	Other	6.487	0.691	17.687	0.399	1.992	1.221
Sample characteristics									
N		23							
T		16							
Total observations		368							

Table 3.4: Parameter estimates of demand for new automobiles

Estimators:	OLS	OLS-PCA	GMM-PCA
	(1)	(2)	(3)
Parameter estimates and standard errors:			
(x1) Horsepower/weight	2.582**	-4.348	0.570
	(1.253)	(5.234)	(1.819)
(x2) Air conditioning indicator	-0.735	-0.326	1.389*
	(0.235)	(0.878)	(0.748)
(x3) Miles per dollar	0.035	0.185	0.549***
	(0.108)	(0.537)	(0.199)
(x4) Space	-0.867***	0.262	1.299*
	(0.317)	(1.092)	(0.752)
(x5) Price	0.089***	-0.043	-0.382***
	(0.017)	(0.061)	(0.114)
Model statistics:			
(N, T)	$(23, 16)$	$(23, 16)$	$(23, 16)$
No. of factors, r	0	5	5
No. of instruments for price	0	0	4
J-statistic p-value			0.096
First stage statistics:			
Robust F-statistic			10.030
Effective F-statistic			14.469
Adjusted R^2			0.455

* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 3.5: Summary statistics for banking application

	Pre/Great Recession			Post-Great Recession			Full Sample		
	2006Q1-2009Q3			2009Q4-2019Q4					
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Banks' characteristics									
Capital adequacy ratio (CAR)	11.283	10.002	4.584	11.162	10.276	3.508	11.197	10.213	3.846
Size	11.930	11.910	1.244	12.300	12.174	1.365	12.194	12.083	1.342
Return on assets (ROA)	0.930	0.916	0.837	0.936	0.946	1.178	0.932	0.924	0.947
Return on equity (ROE)	9.130	8.593	10.160	8.437	8.434	8.047	8.635	8.467	8.708
Liquidity	0.819	0.833	0.255	0.765	0.774	0.216	0.780	0.793	0.229
Sample characteristics									
N	300			300			300		
T	16			40			56		
Total observations	4,800			12,000			16,800		

Table 3.6: Determinants of banks' capital adequacy ratios

Estimators:			
	OLS	OLS-PCA	GMM-PCA
	(1)	(2)	(3)
Parameter estimates and standard errors:			
(x1) Lagged CAR	0.932***	0.844***	0.269***
	(0.014)	(0.193)	(0.067)
(x2) Lagged size	0.360	0.251	0.007
	(0.142)	(0.419)	(0.075)
(x3) Lagged ROA	-0.038	0.155	0.220***
	(0.112)	(0.232)	(0.068)
(x4) Lagged liquidity	-1.530*	-0.036	-2.679
	(0.812)	(2.077)	(2.358)
(x5) Size	-0.383***	-1.558**	-0.384
	(0.141)	(0.709)	(0.466)
(x6) ROA	0.111	0.043	0.182
	(0.149)	(0.280)	(0.160)
(x7) Liquidity	1.717**	5.429**	7.606*
	(0.816)	(2.564)	(4.560)
Model statistics:			
(N, T)	$(300, 40)$	$(300, 40)$	$(300, 40)$
No. of factors, r	0	3	3
No. of instruments for liquidity	0	0	5
J-statistic p-value			0.249
First stage statistics:			
Robust F-statistic			14.563
Effective F-statistic			5.840
Adjusted R^2			0.9164

* significant at 10%, ** significant at 5%, *** significant at 1%.

3.9 Supplementary Appendix

3.9.1 Figures

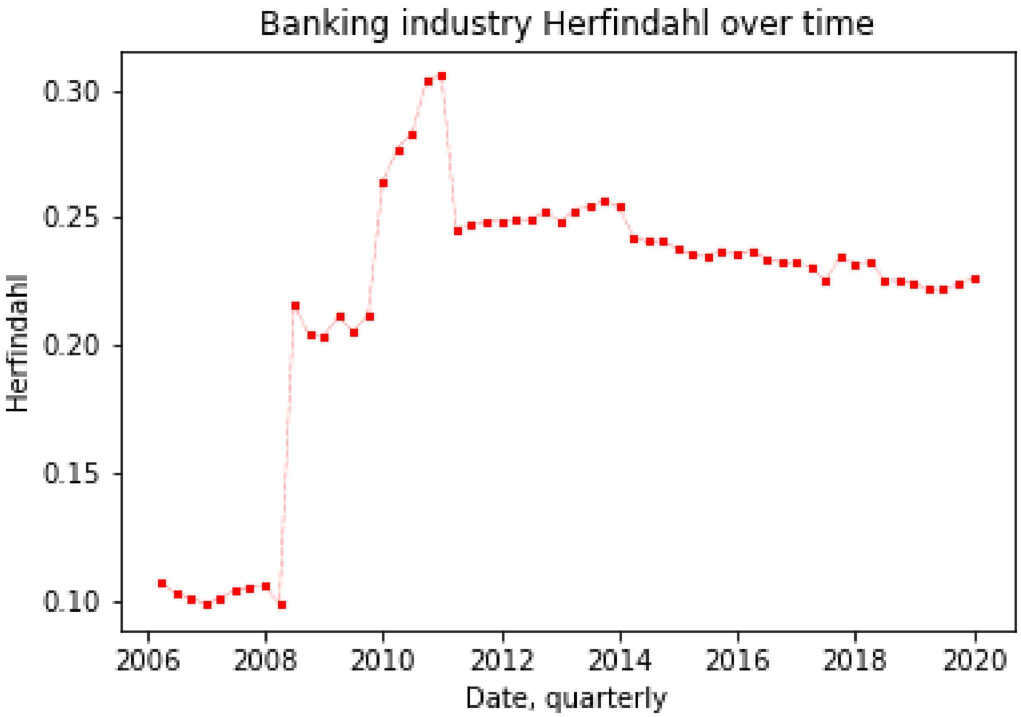

Figure 3.1: Banking industry Herfindahl over time

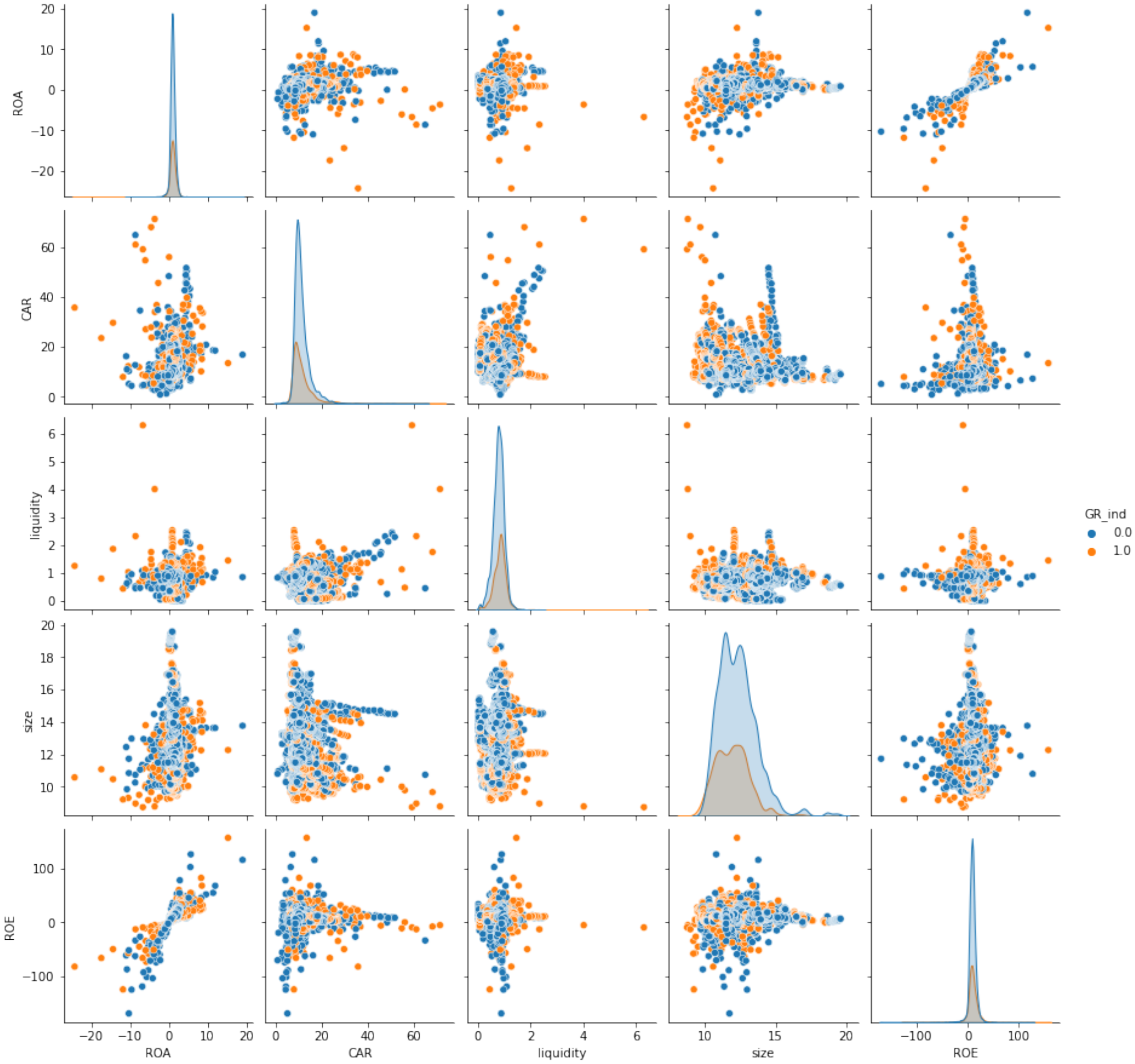

Figure 3.2: Pairwise plots of financial variables for Pre/Great Recession vs. Post-Great Recession

Chapter 4

Conclusions

This dissertation has developed econometric theory along with numerous applications for large panel data models that aim to quantify causal relationships in the absence of a randomized control trial. More concretely, investigation of causal relationships is conducted through estimation and inference using the granular instrumental variables methodology that has been extended along numerous dimensions in each chapter.

In Chapter 2, we further developed the GIV methodology by allowing latent loadings which are treated as unknown parameters to be estimated before constructing the Feasible GIV instrument. We further demonstrate that the sampling error arising from estimating the instrument, factors and a high dimensional precision matrix does not affect the limiting distribution for the structural parameters of interest. We also overidentify the structural parameters, which leads to new and improved results in the global crude oil markets application and demonstrate that the J-test is well sized with simulation evidence. Our Monte Carlo study illustrates that our estimators and algorithms exhibit desirable performance with the finite sample distributions being well approximated by the asymptotic distributions.

In Chapter 3, the GIV methodology has been further developed to accommodate large dynamic panels with unit specific endogenous variates, which require unit-specific GIVs. We develop a SPJ GMM-PCA iterative procedure to estimate the structural parameters of interest. Overidentification tests can be carried out to test model validity. In the first application to the demand for new automobiles we fail to reject the null hypothesis of a valid model and the procedure estimates a downward sloping demand curve as economic theory would suggest. Moreover, the product characteristics yield positive marginal utilities, again as expected. Whereas, the other estimation techniques were less favorable. In the second application to the determinants of banks' capital adequacy ratios the procedure results in weak instruments; whereas the SPJ OLS-PCA method seems to be taking care of the endogeneity by estimating the interactive effects. Nevertheless, across numerous estimation techniques we find evidence which suggests that banks respond to liquidity crunches by raising their capital adequacy ratios.

More fruitful areas of future research would be additional empirical applications of the theoretical results derived in this dissertation. Interesting theoretical extensions would be to allow for random slope coefficients with correlated heterogeneity, the presence of weak factors and unbalanced panels with data not missing at random.

Bibliography

[1] X. Gabaix and R. S. Koijen, “Granular instrumental variables,” *Available at SSRN 3368612*, 2021.

[2] M. Arellano and S. Bond, “Some tests of specification for panel data: Monte carlo evidence and an application to employment equations,” *The Review of Economic Studies*, vol. 58, no. 2, pp. 277–297, 1991.

[3] T. J. Bartik, “Who benefits from state and local economic development policies?,” *WE Upjohn Institute for Employment Research*, 1991.

[4] R. Rigobon, “Identification through heteroskedasticity,” *Review of Economics and Statistics*, vol. 85, no. 4, pp. 777–792, 2003.

[5] J. Bai and S. Ng, “Instrumental variable estimation in a data rich environment,” *Econometric Theory*, vol. 26, no. 6, pp. 1577–1606, 2010.

[6] G. Kapetanios and M. Marcellino, “Factor-gmm estimation with large sets of possibly weak instruments,” *Computational Statistics & Data Analysis*, vol. 54, no. 11, pp. 2655–2675, 2010.

[7] X. Gabaix, “The granular origins of aggregate fluctuations,” *Econometrica*, vol. 79, no. 3, pp. 733–772, 2011.

[8] M. H. Pesaran and C. F. Yang, “Econometric analysis of production networks with dominant units,” *Journal of Econometrics*, vol. 219, no. 2, pp. 507–541, 2020.

[9] J. B. Long and C. I. Plosser, “Real business cycles,” *Journal of Political Economy*, vol. 91, no. 1, pp. 39–69, 1983.

[10] M. Horvath, “Sectoral shocks and aggregate fluctuations,” *Journal of Monetary Economics*, vol. 45, no. 1, pp. 69–106, 2000.

[11] B. Dupor, “Aggregation and irrelevance in multi-sector models,” *Journal of Monetary Economics*, vol. 43, no. 2, pp. 391–409, 1999.

[12] D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi, “The network origins of aggregate fluctuations,” *Econometrica*, vol. 80, no. 5, pp. 1977–2016, 2012.

[13] D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi, “Microeconomic origins of macroeconomic tail risks,” *American Economic Review*, vol. 107, no. 1, pp. 54–108, 2017.

[14] D. D. Gatti, C. Di Guilmi, E. Gaffeo, G. Giulioni, M. Gallegati, and A. Palestrini, “A new approach to business fluctuations: heterogeneous interacting agents, scaling laws and financial fragility,” *Journal of Economic Behavior & Organization*, vol. 56, no. 4, pp. 489–512, 2005.

[15] C. Canals, X. Gabaix, J. Vilarrubia, and D. Weinstein, “Trade shocks, trade balances, and idiosyncratic shocks,” *Banco de España, Documento de Trabajo*, no. 721, 2007.

[16] M. Koren and S. Tenreyro, “Volatility and development,” *The Quarterly Journal of Economics*, vol. 122, no. 1, pp. 243–287, 2007.

[17] S. Blank, C. M. Buch, and K. Neugebauer, “Shocks at large banks and banking sector distress: The banking granular residual,” *Journal of Financial Stability*, vol. 5, no. 4, pp. 353–373, 2009.

[18] Y. Malevergne, P. Santa-Clara, and D. Sornette, “Professor zipf goes to wall street,” tech. rep., National Bureau of Economic Research, 2009.

[19] W. Yan, “Role of diversification risk in financial bubbles,” *Swiss Finance Institute Research Paper*, no. 11-26, 2011.

[20] V. Carvalho and X. Gabaix, “The great diversification and its undoing,” *American Economic Review*, vol. 103, no. 5, pp. 1697–1727, 2013.

[21] S. Schiaffi *et al.*, “The granularity of the stock market: Forecasting aggregate returns using firm-level data,” *Rivista di Politica Economica*, no. 4, pp. 141–169, 2013.

[22] S. Jannati, “Geographic spillover of dominant firms’ shocks,” in *8th Miami Behavioral Finance Conference*, vol. 2019, 2017.

[23] S. C. Lera and D. Sornette, “Quantification of the evolution of firm size distributions due to mergers and acquisitions,” *PloS one*, vol. 12, no. 8, 2017.

[24] J. Bai and S. Ng, “Determining the number of factors in approximate factor models,” *Econometrica*, vol. 70, no. 1, pp. 191–221, 2002.

[25] J. Bai, “Panel data models with interactive fixed effects,” *Econometrica*, vol. 77, no. 4, pp. 1229–1279, 2009.

[26] K. Mohaddes and M. H. Pesaran, “Country-specific oil supply shocks and the global economy: A counterfactual analysis,” *Energy Economics*, vol. 59, pp. 382–399, 2016.

[27] J. Bai, “Inferential theory for factor models of large dimensions,” *Econometrica*, vol. 71, no. 1, pp. 135–171, 2003.

[28] J. Bai, Y. Liao, and J. Yang, “Unbalanced panel data models with interactive effects,” in *The Oxford Handbook of Panel Data*, 2015.

[29] J. Bai and S. Ng, "Matrix completion, counterfactuals, and factor analysis of missing data," *Journal of the American Statistical Association*, vol. 116, no. 536, pp. 1746–1763, 2021.

[30] R. Xiong and M. Pelger, "Large dimensional latent factor modeling with missing observations and applications to causal inference," *arXiv preprint arXiv:1910.08273*, 2019.

[31] J. Bai and S. Ng, "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Econometrica*, vol. 74, no. 4, pp. 1133–1150, 2006.

[32] R. Greenaway-McGrevy, C. Han, and D. Sul, "Asymptotic distribution of factor augmented estimators for panel regression," *Journal of Econometrics*, vol. 169, no. 1, pp. 48–53, 2012.

[33] A. Onatski, "Asymptotics of the principal components estimator of large factor models with weakly influential factors," *Journal of Econometrics*, vol. 168, no. 2, pp. 244–258, 2012.

[34] N. Bailey, G. Kapetanios, and M. H. Pesaran, "Exponent of cross-sectional dependence: Estimation and inference," *Journal of Applied Econometrics*, vol. 31, no. 6, pp. 929–960, 2016.

[35] S. Freyaldenhoven, "Factor models with local factors – determining the number of relevant factors," *Journal of Econometrics*, 2021.

[36] J. Bai and S. Ng, "Approximate factor models with weaker loadings," *arXiv preprint arXiv:2109.03773*, 2021.

[37] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society. Series B, Statistical methodology*, vol. 75, no. 4, 2013.

[38] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 939–967, 2001.

[39] A. Onatski, "Determining the number of factors from empirical distribution of eigenvalues," *The Review of Economics and Statistics*, vol. 92, no. 4, pp. 1004–1016, 2010.

[40] S. C. Ahn and A. R. Horenstein, "Eigenvalue ratio test for the number of factors," *Econometrica*, vol. 81, no. 3, pp. 1203–1227, 2013.

[41] J. Bai and Y. Liao, "Inferences in panel data with interactive effects using large covariance matrices," *Journal of Econometrics*, vol. 200, no. 1, pp. 59–78, 2017.

[42] B. F. Logan, C. Mallows, S. Rice, and L. A. Shepp, "Limit distributions of self-normalized sums," *The Annals of Probability*, vol. 1, no. 5, pp. 788–809, 1973.

[43] A. Pagan, "Econometric issues in the analysis of regressions with generated regressors," *International Economic Review*, vol. 25, no. 1, pp. 221–247, 1984.

[44] G. Chamberlain, "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, vol. 34, no. 3, pp. 305–334, 1987.

[45] H. R. Moon and M. Weidner, “Linear regression for panel with unknown number of factors as interactive fixed effects,” *Econometrica*, vol. 83, no. 4, pp. 1543–1579, 2015.

[46] R. S. Mariano, “Approximations to the distribution functions of the ordinary least-squares and two-stage least-squares estimators in the case of two included endogenous variables,” *Econometrica*, pp. 67–77, 1973.

[47] M. Hatanaka, “On the existence and the approximation formulae for the moments of the k-class estimators,” *The Economic Studies Quarterly (Tokyo. 1950)*, vol. 24, no. 2, pp. 1–15, 1973.

[48] T. Sawa, “Finite-sample properties of the k-class estimators,” *Econometrica*, pp. 653–680, 1972.

[49] K. Takeuchi, “Exact sampling moments of the ordinary least squares, instrumental variable and two-stage least squares estimators,” *International Economic Review*, vol. 11, no. 1, pp. 1–12, 1970.

[50] A. Ullah and A. Nagar, “The exact mean of the two-stage least squares estimator of the structural parameters in an equation having three endogenous variables,” *Econometrica*, pp. 749–758, 1974.

[51] J. Sargan, “On the existence of the moments of 3sls estimators,” *Econometrica*, pp. 1329–1350, 1978.

[52] W. A. Fuller, “Some properties of a modification of the limited information estimator,” *Econometrica*, pp. 939–953, 1977.

[53] G. Hillier and V. Srivastava, “The exact bias and mean square error of the k-class estimators for the coefficient of an endogenous variable in a general structural equation,” *mimeographed, Monash University*, 1981.

[54] T. W. Kinal, “The existence of moments of k-class estimators,” *Econometrica*, vol. 48, no. 1, pp. 241–249, 1980.

[55] D. Staiger and J. H. Stock, “Instrumental variables regression with weak instruments,” *Econometrica*, vol. 65, no. 3, pp. 557–586, 1997.

[56] T. J. Rothenberg, “Approximating the distributions of econometric estimators and test statistics,” *Handbook of Econometrics*, vol. 2, pp. 881–935, 1984.

[57] J. B. Kadane, “Comparison of k-class estimators when the disturbances are small,” *Econometrica*, pp. 723–737, 1971.

[58] L. Kilian, “Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market,” *American Economic Review*, vol. 99, no. 3, pp. 1053–69, 2009.

[59] D. Caldara, M. Cavallo, and M. Iacoviello, “Oil price elasticities and oil price fluctuations,” *Journal of Monetary Economics*, vol. 103, pp. 1–20, 2019.

[60] C. Baumeister and J. D. Hamilton, "Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks," *American Economic Review*, vol. 109, no. 5, pp. 1873–1910, 2019.

[61] L. Callot, M. Caner, A. Ö. Önder, and E. Ulaşan, "A nodewise regression approach to estimating large portfolios," *Journal of Business & Economic Statistics*, vol. 39, no. 2, pp. 520–531, 2021.

[62] D. Darling, "The influence of the maximum term in the addition of independent random variables," *Transactions of the American Mathematical Society*, vol. 73, no. 1, pp. 95–107, 1952.

[63] B. V. Gnedenko and A. N. Kolmogorov, "Limit distributions for sums of independent random variables," 1954.

[64] W. Feller, *An Introduction to Probability Theory and Its Applications Vol II*. John Wiley and Sons, 1971.

[65] R. Durrett, *Probability: Theory and Examples*, vol. 49. Cambridge University Press, 2019.

[66] B. C. Arnold, "Univariate and multivariate pareto models," *Journal of Statistical Distributions and Applications*, vol. 1, no. 1, pp. 1–16, 2014.

[67] A. Jakubowski, "Minimal conditions in p-stable limit theorems," *Stochastic Processes and Their Applications*, vol. 44, no. 2, pp. 291–327, 1993.

[68] A. Jakubowski, "Minimal conditions in p-stable limit theoremsâĂŤii," *Stochastic Processes and Their Applications*, vol. 68, no. 1, pp. 1–20, 1997.

[69] R. A. Davis and T. Hsing, "Point process and partial sum convergence for weakly dependent random variables with infinite variance," *The Annals of Probability*, vol. 23, no. 2, pp. 879–917, 1995.

[70] G. Chamberlain and M. Rothschild, "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica*, vol. 51, no. 5, pp. 1281–1304, 1983.

[71] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[72] M. Barigozzi, C. Brownlees, and G. Lugosi, "Power-law partial correlation network models," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 2905–2929, 2018.

[73] C. Brownlees, E. Nualart, and Y. Sun, "Realized networks," *Journal of Applied Econometrics*, vol. 33, no. 7, pp. 986–1006, 2018.

[74] Y. Koike, "De-biased graphical lasso for high-frequency data," *Entropy*, vol. 22, no. 4, p. 456, 2020.

[75] T.-H. Lee and E. Seregina, "Optimal portfolio using factor graphical lasso," *arXiv preprint arXiv:2011.00435*, 2020.

[76] J. Fan, H. Liu, and W. Wang, "Large covariance estimation through elliptical factor models," *Annals of Statistics*, vol. 46, no. 4, pp. 1383–1414, 2018.

[77] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[78] J. Janková and S. van de Geer, "Inference in high-dimensional graphical models," in *Handbook of Graphical Models*, ch. 14, pp. 325–351, CRC Press, 2018.

[79] S. Nickell, "Biases in dynamic models with fixed effects," *Econometrica*, pp. 1417–1426, 1981.

[80] J. Neyman and E. L. Scott, "Consistent estimates based on partially consistent observations," *Econometrica: Journal of the Econometric Society*, pp. 1–32, 1948.

[81] A. Chudik, M. H. Pesaran, and J.-C. Yang, "Half-panel jackknife fixed-effects estimation of linear panels with weakly exogenous regressors," *Journal of Applied Econometrics*, vol. 33, no. 6, pp. 816–836, 2018.

[82] H. R. Moon and M. Weidner, "Dynamic linear panel regression models with interactive fixed effects," *Econometric Theory*, vol. 33, no. 1, pp. 158–195, 2017.

[83] M. H. Quenouille, "Approximate tests of correlation in time-series 3," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 45, pp. 483–484, Cambridge University Press, 1949.

[84] G. Dhaene and K. Jochmans, "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, vol. 82, no. 3, pp. 991–1030, 2015.

[85] I. Fernández-Val and M. Weidner, "Individual and time effects in nonlinear panel models with large n, t," *Journal of Econometrics*, vol. 192, no. 1, pp. 291–312, 2016.

[86] S. C. Ahn, P. Schmidt, *et al.*, "Efficient estimation of models for dynamic panel data," *Journal of Econometrics*, vol. 68, no. 1, pp. 5–28, 1995.

[87] M. Arellano and O. Bover, "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, vol. 68, no. 1, pp. 29–51, 1995.

[88] J. Alvarez and M. Arellano, "The time series and cross-section asymptotics of dynamic panel data estimators," *Econometrica*, vol. 71, no. 4, pp. 1121–1159, 2003.

[89] S. Banafti and T.-H. Lee, "Inferential theory for granular instrumental variables in high dimensions," *arXiv preprint arXiv:2201.06605*, 2022.

[90] M. H. Pesaran, "Estimation and inference in large heterogeneous panels with a multifactor error structure," *Econometrica*, vol. 74, no. 4, pp. 967–1012, 2006.

[91] A. Chudik and M. H. Pesaran, "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors," *Journal of Econometrics*, vol. 188, no. 2, pp. 393–420, 2015.

[92] M. Harding and C. Lamarche, "Least squares estimation of a panel data model with multifactor error structure and endogenous covariates," *Economics Letters*, vol. 111, no. 3, pp. 197–199, 2011.

[93] N. Lee, H. R. Moon, and M. Weidner, "Analysis of interactive fixed effects dynamic linear panel regression with measurement error," *Economics Letters*, vol. 117, no. 1, pp. 239–242, 2012.

[94] B. Friedrich, L. Laun, C. Meghir, and L. Pistaferri, "Earnings dynamics and firm-level shocks," tech. rep., National Bureau of Economic Research, 2019.

[95] M. Amiti, O. Itskhoki, and J. Konings, "International shocks, variable markups, and domestic prices," *The Review of Economic Studies*, vol. 86, no. 6, pp. 2356–2402, 2019.

[96] M. T. Leary and M. R. Roberts, "Do peer firms affect corporate financial policy?," *The Journal of Finance*, vol. 69, no. 1, pp. 139–178, 2014.

[97] A. Goyal and P. Santa-Clara, "Idiosyncratic risk matters!," *The Journal of Finance*, vol. 58, no. 3, pp. 975–1007, 2003.

[98] C. Yeh, "Revisiting the origins of business cycles with the size-variance relationship," *Federal Reserve Bank of Richmond December*, 2019.

[99] T. Lancaster, "The incidental parameter problem since 1948," *Journal of Econometrics*, vol. 95, no. 2, pp. 391–413, 2000.

[100] S. Berry, J. Levinsohn, and A. Pakes, "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, pp. 841–890, 1995.

[101] H. R. Moon, M. Shum, and M. Weidner, "Estimation of random coefficients logit demand models with interactive fixed effects," *Journal of Econometrics*, vol. 206, no. 2, pp. 613–644, 2018.

[102] S. Esteban and M. Shum, "Durable-goods oligopoly with secondary markets: the case of automobiles," *The RAND Journal of Economics*, vol. 38, no. 2, pp. 332–354, 2007.

[103] S. Kripfganz and V. Sarafidis, "Instrumental-variable estimation of large-t panel-data models with common factors," *The Stata Journal*, vol. 21, no. 3, pp. 659–686, 2021.

[104] J. L. M. Olea and C. Pflueger, "A robust test for weak instruments," *Journal of Business & Economic Statistics*, vol. 31, no. 3, pp. 358–369, 2013.