

Granular Instrumental Variables: Estimation and Inference

Jinyong Hahn*	Niu He [†]	Zhipeng Liao [‡]	Wenyu Zhou [§]
UCLA	UCLA	UCLA	Zhejiang University

June 11, 2026

Abstract

We develop an estimation and inference framework for granular instrumental variables (GIVs) in models with latent aggregate shocks. Our key insight is that valid GIVs are characterized by the orthogonal complement of the factor-loading space. This characterization yields a feasible procedure for constructing GIVs when factor loadings are unknown and does not require a large cross-sectional dimension. We provide practical procedures for inference and specification testing, and apply the framework to estimate the aggregate equity market multiplier. Our empirical results reveal substantial heterogeneity in equity demand elasticities across investor sectors and may provide nuanced support for the inelastic-markets hypothesis.

JEL Classification: C13, C26, C51, G12

Keywords: Granular Instrumental Variables; Latent Aggregate Shocks; Identification and Inference; Asset Demand; Market Elasticity.

*Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: hahn@econ.ucla.edu

[†]Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: niuhe@g.ucla.edu

[‡]Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: zhipeng.liao@econ.ucla.edu

[§]International Business School, Zhejiang University, Haining, Zhejiang 314400, China. Email: wenyuzhou@intl.zju.edu.cn.

1 Introduction

Understanding how aggregate outcomes respond to shocks is a central objective in economics and finance. In many applications, researchers observe a large cross section of entities that are simultaneously exposed to a small number of common shocks. Examples include firms responding to aggregate demand conditions, financial institutions adjusting portfolios in response to market forces, and countries reacting to global macroeconomic shocks. A common feature of these environments is that the variables of interest are jointly determined in equilibrium, creating endogeneity problems that complicate identification and estimation of structural parameters.

Recently, Gabaix and Koijen (2024) proposed a novel identification strategy based on granular instrumental variables (GIVs). The approach builds on the insight from the granularity literature that when a small number of firms, industries, countries, investors, or borrowers account for a non-negligible share of aggregate activity, idiosyncratic shocks to these units may survive aggregation and influence aggregate outcomes.¹ Exploiting this feature, GIV extracts the idiosyncratic component of observables after controlling for common factors and aggregates these components using size weights to construct instruments for causal parameters such as elasticities and multipliers. Under suitable restrictions on the covariance structure of the idiosyncratic shocks, the resulting instruments are orthogonal to equilibrium disturbances and can therefore identify structural parameters. Unlike earlier approaches that use idiosyncratic shocks to variables excluded from the estimating equation,² GIV constructs instruments from the idiosyncratic component of the variables entering the estimating equation itself. As a result, the methodology does not rely on traditional excluded instruments, which are often difficult to justify or unavailable in practice.

Several recent papers extend the baseline GIV framework. Banafti and Lee (2022) study

¹See, among others, Gabaix (2011); Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012); di Giovanni, Levchenko, and Mejean (2014); Baqaee and Farhi (2019); Gaubert and Itskhoki (2021).

²See, for instance, Leary and Roberts (2014); Amiti and Weinstein (2018); Amiti, Itskhoki, and Konings (2019).

inference in large- n , large- T settings with unknown factors and loadings. Baumeister and Hamilton (2023) develop a likelihood-based approach, while Qian (2023) allows for heterogeneous spillovers. GIV has also become an important identification device in empirical macroeconomics and finance, including applications to stock-market demand, exchange rates, bank lending, and asset pricing.³

Despite its growing importance, several econometric questions remain unresolved. The central challenge is that valid GIV construction requires knowledge of the factor-loading space associated with latent aggregate shocks, which is rarely observed in practice. Existing implementations therefore estimate latent factors and construct GIVs in a second step, a strategy that typically relies on a large cross-sectional dimension and strong normalization assumptions. Moreover, little is known about the consequences of estimating the factor-loading space for identification, estimation, and inference.

This paper develops an estimation and inference framework for structural models identified by GIVs when the factor-loading space is unknown. Rather than estimating latent factors and constructing instruments in a second step, we show that the relevant GIV space can be recovered directly from the covariance structure of the observables. Specifically, the admissible GIVs are generated by the orthogonal complement of the factor-loading space, which can be identified from the eigenspace associated with the smallest eigenvalues of the covariance matrix. This insight transforms the construction of GIVs into a covariance-based problem and yields a feasible procedure for constructing instruments directly from the data. The resulting estimator remains valid even when the number of entities is fixed and therefore does not require the cross-sectional dimension to diverge with the sample size.

The characterization also provides a transparent identification strategy. We show that all admissible GIVs are generated by the orthogonal complement of the column space spanned by the factor loadings and the vector of ones. When the factor-loading space is unknown, the

³Recent applications include Galaasen, Jamilov, Juelsrud, and Rey (2020); Camanho, Hau, and Rey (2022); Ma, Xiao, and Zeng (2022); Dong, Kang, and Peress (2025); many others are discussed in Gabaix and Koijen (2024).

relevant orthogonal complement can be consistently recovered from the eigenspace associated with the smallest eigenvalues of the covariance matrix of the observables. This result yields a feasible GIV estimator and forms the basis for inference with estimated GIVs.

Building on this characterization, we establish consistency and asymptotic normality of the feasible GIV estimator and develop practical procedures for inference and specification testing. In particular, we derive feasible standard errors, establish the asymptotic validity of an over-identification J -test when the GIVs are estimated, and propose a BIC-type criterion for determining the dimension of the factor-loading space. Monte Carlo simulations show that the feasible estimator performs similarly to an oracle estimator that knows the true factor-loading space and that the proposed inference procedures perform well in finite samples.

Beyond estimation and inference, the paper clarifies several identification issues that arise when factor loadings are unknown. We show that certain restrictions commonly interpreted as normalizations on the factor loadings instead impose substantive restrictions on the latent factors. We also examine the identification strategy in Gabaix and Koijen (2024) and show that the moment conditions in their Proposition 7 may fail to identify the structural parameters when factor loadings are unknown. These findings highlight the challenges of identification in the presence of latent aggregate shocks and motivate the alternative characterization developed in this paper. The proposed framework avoids these restrictions and extends naturally to settings with additional exogenous regressors, unbalanced panels, and heterogeneous demand elasticities.

Finally, we return to the estimation of the aggregate equity market multiplier in demand-based asset pricing (Gabaix and Koijen, 2023). This application is particularly relevant because the factor-loading space is unknown and only twelve investor sectors are available in the data. As a result, the large-cross-section justification underlying existing GIV procedures is difficult to invoke directly, making the setting a natural environment in which to assess the practical importance of estimating GIVs when factor loadings are unobserved. Applying our framework, we obtain estimates and conduct inference for the aggregate multiplier together

with a formal specification test of the underlying GIV moment conditions. The empirical results provide evidence consistent with highly inelastic aggregate equity demand.

The remainder of the paper is organized as follows. Section 2 introduces the main ideas in a simplified framework. Section 3 develops the general model, establishes identification, and presents estimation and inference procedures based on estimated GIVs. Section 4 studies identification when factor loadings are unknown, demonstrates a failure of identification in the moment conditions proposed by Gabaix and Koijen (2024), and develops extensions to models with exogenous regressors, unbalanced panels, and heterogeneous demand elasticities. Section 5 reports Monte Carlo evidence, Section 6 presents an empirical application to the aggregate equity market multiplier, and Section 7 concludes. Proofs and additional technical and empirical results are collected in the Online Appendix.

Notation. We use K to denote a generic strictly positive constant that may vary from place to place but does not depend on the sample size T . We write $a \equiv b$ to indicate that a is defined as b . For any positive integer k , let $\mathbf{I}_k$, $\mathbf{1}_k$, and $\mathbf{0}_k$ denote the $k \times k$ identity matrix, the $k \times 1$ vector of ones, and the $k \times 1$ vector of zeros, respectively. For any vector $x_t \in \mathbb{R}^n$ (possibly indexed by t) and any weight vector $\omega \in \mathbb{R}^n$ satisfying $\omega^\top \mathbf{1}_n = 1$, define the weighted average $x_{\omega,t} \equiv \omega^\top x_t$. For any matrix A , let $\text{col}(A)$ and $\text{rank}(A)$ denote its column space and rank, respectively. We use $\|A\|$ and $\|A\|_o$ to denote the Frobenius norm and operator norm of A , respectively, and M_A to denote the orthogonal projection matrix onto the orthogonal complement of $\text{col}(A)$. For any square matrix A , let $\rho_{\min}(A)$ and $\rho_{\max}(A)$ denote its smallest and largest eigenvalues, respectively. For any two matrices A and B , let $\text{diag}(A, B)$ denote the block-diagonal matrix with A and B on its main diagonal, and let $A \otimes B$ denote their Kronecker product. Finally, for any square matrix A , let $\text{vech}(A)$ denote the half-vectorization of A , obtained by stacking the elements of its lower triangular part (including the diagonal) column by column.

2 A Simplified Framework

We first illustrate the intuition of GIV using the following simplified model:

$$y_{i,t} = \phi p_t + \eta_t + u_{i,t}, \quad (1)$$

$$p_t = \psi y_{S,t} + \varepsilon_t, \quad (2)$$

where $y_{i,t}$ denotes the log demand of entity $i \in \{1, \dots, n\}$ at time t , and p_t is the log price common to all entities. The latent variables η_t and $u_{i,t}$ with $\mathbb{E}[u_{i,t}] = 0$ represent aggregate and idiosyncratic demand shocks, respectively. The parameter ϕ measures the demand elasticity. Equation (2) describes the supply side, where $y_{S,t} \equiv S^\top y_t$ denotes aggregate demand, $S \equiv (s_i)_{i \leq n}$, and $y_t \equiv (y_{i,t})_{i \leq n}$, with s_i denoting the market share of entity i . The term ε_t is the supply shock, and ψ denotes the supply elasticity. Although stylized, this model is widely used in the macroeconomics and finance literature (Gabaix and Koijen, 2023; Camanho, Hau, and Rey, 2022).

Let $u_t \equiv (u_{i,t})_{i \leq n}$. The following assumptions are maintained throughout this section:

$$\text{Cov}(\varepsilon_t, u_t) = \sigma_{\varepsilon u} \mathbf{1}_n^\top, \quad \text{Cov}(\eta_t, u_t) = \sigma_{\eta u} \mathbf{1}_n^\top, \quad \text{Var}(u_t) = \sigma_u^2 \mathbf{I}_n, \quad (3)$$

where $\sigma_{\varepsilon u}$ and $\sigma_{\eta u}$ are constants that need not be zero. These conditions ensure the validity of the GIVs constructed in the literature and considered in this section.⁴

Following Gabaix and Koijen (2024), we use the equally weighted average $y_{e,t} \equiv e^\top y_t$, where $e \equiv n^{-1} \mathbf{1}_n$, to construct a GIV defined as

$$z_t(e) \equiv y_{S,t} - y_{e,t}. \quad (4)$$

⁴Gabaix and Koijen (2024) impose the stronger restrictions $\sigma_{\varepsilon u} = 0$ and $\sigma_{\eta u} = 0$ for the identification of ϕ and ψ (see the first sentence of the paragraph containing their display (3)). As we show below, these restrictions are not necessary. Moreover, as shown in the next section, identification and estimation of ϕ and ψ do not require the distributions of the aggregate and idiosyncratic shocks to be time-invariant. In particular, their variances and covariances, such as $\text{Var}(u_{i,t})$ and $\text{Cov}(\eta_t, u_t)$, are allowed to vary over time.

From the demand equation (1), this GIV satisfies $z_t(e) = u_t^\top(S - e)$. Together with the first condition in (3), this implies

$$\mathbb{E}[(p_t - \psi y_{S,t})z_t(e)] = \mathbb{E}[\varepsilon_t u_t^\top](S - e) = \sigma_{\varepsilon u} \mathbf{1}_n^\top (S - e) = 0. \quad (5)$$

Similarly,

$$\begin{aligned} \mathbb{E}[(y_{e,t} - \phi p_t)z_t(e)] &= \mathbb{E}[(\eta_t + u_{e,t})u_t^\top](S - e) \\ &= \text{Cov}(\eta_t, u_t)(S - e) + e^\top \text{Var}(u_t)(S - e) \\ &= \sigma_{\eta u} \mathbf{1}_n^\top (S - e) + \sigma_u^2 e^\top (S - e) = 0. \end{aligned} \quad (6)$$

The GIV $z_t(e)$ thus provides moment conditions (5) and (6) for identifying and estimating the elasticities ϕ and ψ .

The above identification strategy can be generalized to construct generic GIVs

$$z_t(a) \equiv y_{S,t} - y_{a,t} = (S - a)^\top u_t, \quad (7)$$

where $a \in \mathbb{R}^n$ satisfies

$$a \neq S \quad \text{and} \quad a^\top \mathbf{1}_n = 1. \quad (8)$$

Under these conditions, moment restrictions analogous to (5) and (6) can be constructed:

$$\mathbb{E}[(p_t - \psi y_{S,t})z_t(a)] = 0, \quad (9)$$

$$\mathbb{E}[(y_{e,t} - \phi p_t)z_t(a)] = 0. \quad (10)$$

The GIV in (4) corresponds to the special case $a = e$.

When $n > 2$, there exist multiple vectors a satisfying (8). Therefore multiple GIVs are available and ψ and ϕ become over-identified. This provides a natural motivation for using multiple GIVs both to improve efficiency and to conduct specification tests of instrument va-

lidity. We next characterize the resulting set of moment conditions and clarify its connection to the GIV-based approach.

The restrictions in (3) imply the following moment conditions:

$$\mathbb{E}[(y_t - \phi p_t \mathbf{1}_n)(y_t - \phi p_t \mathbf{1}_n)^\top] = (\mathbb{E}[\eta_t^2] + 2\sigma_{\eta u}) \mathbf{1}_n \mathbf{1}_n^\top + \sigma_u^2 \mathbf{I}_n, \quad (11)$$

$$\mathbb{E}[(p_t - \psi y_{S,t})(y_t - \phi p_t \mathbf{1}_n)] = (\mathbb{E}[\varepsilon_t \eta_t] + \sigma_{\varepsilon u}) \mathbf{1}_n, \quad (12)$$

which together provide $n(n+3)/2$ moment conditions. To separate the parameters of interest from the nuisance parameters, let $Q \equiv (q_1, \dots, q_n)$ be an $n \times n$ orthonormal matrix with

$$q_1 \equiv n^{-1/2} \mathbf{1}_n \quad \text{and} \quad q_j \equiv (j(j-1))^{-1/2} \left(\sum_{l \leq j-1} \ell_l - (j-1) \ell_j \right) \quad \text{for } j \geq 2, \quad (13)$$

where ℓ_j denotes the j th canonical basis vector of $\mathbb{R}^n$. The following lemma provides a non-redundant representation of these moment conditions.

Lemma 1. *The non-redundant restrictions in (11) can be equivalently written as*

$$\mathbb{E}[(y_{e,t} - \phi p_t) Q_{-1}^\top y_t] = \mathbf{0}_{n-1}, \quad (14)$$

$$\text{vech}(\mathbb{E}[Q_{-1}^\top y_t y_t^\top Q_{-1}] - \sigma_u^2 \mathbf{I}_{n-1}) = \mathbf{0}_{n(n-1)/2}, \quad (15)$$

$$\mathbb{E}[(y_{e,t} - \phi p_t)^2] - (\mathbb{E}[\eta_t^2] + 2\sigma_{\eta u}) - n^{-1} \sigma_u^2 = 0, \quad (16)$$

while the restrictions in (12) can be equivalently written as

$$\mathbb{E}[(p_t - \psi y_{S,t}) Q_{-1}^\top y_t] = \mathbf{0}_{n-1}, \quad (17)$$

$$\mathbb{E}[(p_t - \psi y_{S,t})(y_{e,t} - \phi p_t)] - (\mathbb{E}[\varepsilon_t \eta_t] + \sigma_{\varepsilon u}) = 0, \quad (18)$$

where $Q_{-1} \equiv (q_2, \dots, q_n)$ and $\{q_j\}_{j=1}^n$ is defined in (13).

The moment conditions in (14) and (17) are equivalent to those in (9) and (10), and can be directly used within a GMM framework for estimation and inference of the unknown

elasticities ψ and ϕ .⁵ In contrast, the restrictions in (15)–(16) and (18) involve only nuisance parameters, namely σ_u^2 , $\mathbb{E}[\eta_t^2] + 2\sigma_{\eta u}$, and $\mathbb{E}[\varepsilon_t \eta_t] + \sigma_{\varepsilon u}$.

Specifically, the moment conditions in (15) identify σ_u^2 and also yield additional restrictions that do not depend on unknown parameters. Conditional on ϕ , ψ , and σ_u^2 , the quantities $\mathbb{E}[\eta_t^2] + 2\sigma_{\eta u}$ and $\mathbb{E}[\varepsilon_t \eta_t] + \sigma_{\varepsilon u}$ are just-identified by (16) and (18), respectively. Since σ_u^2 is over-identified by (15), jointly estimating ϕ , ψ , and σ_u^2 using (14), (15), and (17) may yield more efficient estimators of ϕ and ψ than those based only on (14) and (17); see, for example, Akerberg, Chen, Hahn, and Liao (2014).

The analysis in this section relies on a simplified demand specification in which the aggregate shock η_t enters with a known and homogeneous loading across entities. In many applications, however, aggregate shocks may have heterogeneous effects that are not directly observed, giving rise to a more general factor structure. In the next section, we extend the GIV framework to this setting, where the demand equation includes unobserved factors with unknown loadings. This introduces new identification and estimation challenges, as the moment conditions derived above are no longer directly applicable when the factor loadings are unknown.

3 Granular IVs in a General Model

In this section, we study estimation and inference using GIVs in a more general model in which the demand equation (1) incorporates a set of unobserved factors with unknown factor

⁵To establish this equivalence, note first that (14) and (17) are constructed using GIVs of the form $q_j^\top y_t$ for $j \geq 2$. For any $a \in \mathbb{R}^n$ satisfying (8), the corresponding GIV is $(S - a)^\top y_t$, where $(S - a)^\top \mathbf{1}_n = 0$. Since Q_{-1} spans the subspace orthogonal to $\mathbf{1}_n$, it follows that $S - a$ can be written as a linear combination of the columns of Q_{-1} . Hence, the moment conditions in (9) and (10) are implied by those in (14) and (17). Conversely, for each $j \geq 2$, the vector $S - q_j$ satisfies (8), implying that (14) and (17) are implied by (9) and (10).

loadings.⁶ Specifically, we consider

$$y_t = \phi p_t \mathbf{1}_n + \lambda \eta_t + u_t, \quad (19)$$

$$p_t = \psi y_{S,t} + \varepsilon_t. \quad (20)$$

Here η_t denotes an $r \times 1$ vector of unobserved factors, and λ is an $n \times r$ matrix of factor loadings. The vector η_t may also include a constant term, in which case the corresponding loading captures unobserved entity fixed effects. While the supply equation appears identical to (2), we now define

$$y_{S,t} = S_t^\top y_t,$$

where $S_t \equiv (s_{i,t})_{i \leq n}$, and $s_{i,t}$ is nonnegative and predetermined at time t .⁷

In contrast to the simplified model studied in the previous section, we now allow the factor loadings λ to be unknown, which renders the earlier identification and estimation results inapplicable and constitutes the main challenge addressed in this section. One approach to handling the unobserved factors η_t is to estimate them from y_t (after partialling out p_t and entity fixed effects) using principal component analysis (see, e.g., Gabaix and Koijen (2023) and Banafiti and Lee (2022)). However, as noted in the literature (see, e.g., Bai (2003)), the consistency of the estimated factors typically requires the number of entities n to diverge, which stands in sharp contrast to most applications of GIVs, where the number of entities is relatively small.

Another approach, proposed in Gabaix and Koijen (2024), attempts to identify λ jointly with the other unknown parameters in the model under normalization restrictions and conditions similar to (but stronger than) Assumption 1 below. However, as we show in Subsection 4.1, their identification strategy fails to identify the factor loadings and may lead to invalid

⁶The model (19)-(20) can be further extended to include exogenous regressors in both the demand and supply equations without affecting the nature of the estimation and inference procedures proposed in this section; see Subsection 4.2 for details.

⁷Throughout this section, we assume that the number of entities n is fixed over t . The identification strategy, as well as the estimation and inference procedures proposed in this section, also apply to settings in which n varies over time; see Subsection 4.3 for details.

inference.

The method proposed in this section is based on an identification result that applies for any n , whether finite or diverging. Although we follow Gabaix and Koijen (2024) and construct our inference procedures under an asymptotic framework with fixed n , as discussed in Subsection 4.3, the method can be straightforwardly extended to settings in which n is large or even exceeds T .

3.1 Identification

In this subsection, we first establish identification of the demand and supply elasticities ϕ and ψ given λ , thereby extending the results in Lemma 1. We then provide a constructive identification result for the orthogonal complement of $\text{col}((\mathbf{1}_n, \lambda))$, which forms the basis for the estimation and inference procedures developed in the next subsection. We begin by stating the conditions required for identification.

Assumption 1. (i) $\mathbb{E}[u_t] = \mathbf{0}_n$ and $\text{Var}(u_t) = \sigma_{u,t}^2 \mathbf{I}_n$; (ii) $\text{Cov}(\eta_t, u_t) = \Gamma_{\eta u, t} \mathbf{1}_n^\top$, where $\Gamma_{\eta u, t}$ is an $r \times 1$ vector; (iii) $\text{Cov}(\varepsilon_t, u_t) = \sigma_{\varepsilon u, t} \mathbf{1}_n^\top$, where $\sigma_{\varepsilon u, t}$ is a finite scalar; (iv) $T^{-1} \sum_{t \leq T} \mathbb{E}[\eta_t \eta_t^\top]$ is nonsingular and $n > \bar{r}$, where $\bar{r} \equiv \text{rank}((\mathbf{1}_n, \lambda))$.

Assumption 1(i)–(iii) generalize the conditions in (3) by allowing the joint distribution of the demand shocks u_t , the supply shocks ε_t , and the factors η_t to vary over time. Under these conditions, we obtain

$$\mathbb{E}[(y_t - \phi p_t \mathbf{1}_n)(y_t - \phi p_t \mathbf{1}_n)^\top] = \lambda \mathbb{E}[\eta_t \eta_t^\top] \lambda^\top + \lambda \Gamma_{\eta u, t} \mathbf{1}_n^\top + \mathbf{1}_n \Gamma_{\eta u, t}^\top \lambda^\top + \sigma_{u, t}^2 \mathbf{I}_n, \quad (21)$$

$$\mathbb{E}[(p_t - \psi y_{S, t})(y_t - \phi p_t \mathbf{1}_n)] = \lambda \mathbb{E}[\varepsilon_t \eta_t] + \sigma_{\varepsilon u, t} \mathbf{1}_n, \quad (22)$$

which provide a total of $n(n+3)/2$ moment conditions for the unknown parameters ϕ , ψ , $\sigma_{u, t}^2$, $\mathbb{E}[\eta_t \eta_t^\top]$, $\Gamma_{\eta u, t}$, $\mathbb{E}[\varepsilon_t \eta_t]$, and $\sigma_{\varepsilon u, t}$. Assumption 1(iv) is primarily imposed to ensure identification of $\text{col}((\mathbf{1}_n, \lambda))$, whose orthogonal complement will be used to construct the GIVs.

We now reorganize the moment conditions (21)-(22) according to their roles in identifying the different parameters.

Lemma 2. *Let $\bar{\lambda} \equiv (n^{-1/2}\mathbf{1}_n, \bar{\lambda}_{-1})$ be an $n \times \bar{r}$ orthonormal matrix spanning $\text{col}((\mathbf{1}_n, \lambda))$, and let $\bar{\lambda}_\perp$ denote its orthonormal complement. Then the non-redundant restrictions in (21) can be equivalently expressed as*

$$\mathbb{E}\left[(y_{e,t} - \phi p_t)\bar{\lambda}_\perp^\top y_t\right] = \mathbf{0}_{n-\bar{r}}, \quad (23)$$

$$\mathbb{E}\left[\bar{\lambda}_{-1}^\top y_t y_t^\top \bar{\lambda}_\perp\right] = \mathbf{0}_{(\bar{r}-1) \times (n-\bar{r})}, \quad (24)$$

$$\text{vech}\left(\mathbb{E}\left[\bar{\lambda}_\perp^\top y_t y_t^\top \bar{\lambda}_\perp\right] - \sigma_{u,t}^2 \mathbf{I}_{n-\bar{r}}\right) = \mathbf{0}_{(n-\bar{r}+1)(n-\bar{r})/2}, \quad (25)$$

and

$$\begin{aligned} & \text{vech}\left(\mathbb{E}\left[\bar{\lambda}^\top (y_t - \phi p_t \mathbf{1}_n)(y_t - \phi p_t \mathbf{1}_n)^\top \bar{\lambda}\right]\right) \\ &= \text{vech}\left(\bar{\lambda}^\top \left(\lambda \mathbb{E}[\eta_t \eta_t^\top] \lambda^\top + \lambda \Gamma_{\eta u, t} \mathbf{1}_n^\top + \mathbf{1}_n \Gamma_{\eta u, t}^\top \lambda^\top + \sigma_{u,t}^2 \mathbf{I}_n\right) \bar{\lambda}\right), \end{aligned} \quad (26)$$

while the restrictions in (22) can be equivalently written as

$$\mathbb{E}\left[(p_t - \psi y_{S,t})\bar{\lambda}_\perp^\top y_t\right] = \mathbf{0}_{n-\bar{r}}, \quad (27)$$

$$\mathbb{E}\left[\bar{\lambda}^\top (p_t - \psi y_{S,t})(y_t - \phi p_t \mathbf{1}_n)\right] - (\bar{\lambda}^\top \lambda \mathbb{E}[\varepsilon_t \eta_t] + \sigma_{\varepsilon u, t} \bar{\lambda}^\top \mathbf{1}_n) = \mathbf{0}_{\bar{r}}. \quad (28)$$

Lemma 2 shows that the key moment conditions for identifying ϕ and ψ are given by (23) and (27), which are constructed from the generalized GIVs $\bar{\lambda}_\perp^\top y_t$. The moment conditions associated with the diagonal elements of

$$\mathbb{E}\left[\bar{\lambda}_\perp^\top y_t y_t^\top \bar{\lambda}_\perp\right] - \sigma_{u,t}^2 \mathbf{I}_{n-\bar{r}} = \mathbf{0}_{(n-\bar{r}) \times (n-\bar{r})}$$

provide identifying restrictions for $T^{-1} \sum_{t \leq T} \sigma_{u,t}^2$. In contrast, the moment conditions in (24), as well as the off-diagonal elements of the matrix above, do not involve any unknown param-

eters and are therefore redundant from an identification standpoint. Nevertheless, they may be useful for improving the efficiency of the GMM estimator and for testing specification assumptions, such as Assumption 1(i). Finally, (26) and (28) impose restrictions on the nuisance parameters $T^{-1} \sum_{t \leq T} \mathbb{E}[\eta_t \eta_t^\top]$, $T^{-1} \sum_{t \leq T} \Gamma_{\eta u, t}$, $T^{-1} \sum_{t \leq T} \mathbb{E}[\varepsilon_t \eta_t]$, and $T^{-1} \sum_{t \leq T} \sigma_{\varepsilon u, t}$, conditional on the identification of ϕ , ψ , and $T^{-1} \sum_{t \leq T} \sigma_{u, t}^2$.

The moment conditions in (23) and (27) for the identification of ϕ and ψ rely on the generalized GIVs $\bar{\lambda}_\perp^\top y_t$. These instruments are, however, infeasible in practice when the factor loading matrix λ is unknown. We therefore next show how to identify the column space of $\bar{\lambda}$, which in turn determines the space spanned by $\bar{\lambda}_\perp$.

To this end, let $M_{\mathbf{1}_n} \equiv \mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^\top$ and define the demeaned variables

$$\tilde{y}_t \equiv M_{\mathbf{1}_n} y_t, \quad \tilde{\lambda} \equiv M_{\mathbf{1}_n} \lambda, \quad \tilde{u}_t \equiv M_{\mathbf{1}_n} u_t. \quad (29)$$

Applying $M_{\mathbf{1}_n}$ to both sides of (19) yields

$$\tilde{y}_t = \tilde{\lambda} \eta_t + \tilde{u}_t. \quad (30)$$

Combining (30) with Assumption 1, we obtain

$$\mathbb{E}[\tilde{y}_t \tilde{y}_t^\top] = \tilde{\lambda} \mathbb{E}[\eta_t \eta_t^\top] \tilde{\lambda}^\top + \sigma_{u, t}^2 M_{\mathbf{1}_n}. \quad (31)$$

Averaging (31) over t then yields

$$\bar{\Sigma}_{\tilde{y}} = \tilde{\lambda} \left(T^{-1} \sum_{t \leq T} \mathbb{E}[\eta_t \eta_t^\top] \right) \tilde{\lambda}^\top + \bar{\sigma}_u^2 M_{\mathbf{1}_n} \quad (32)$$

where

$$\bar{\Sigma}_{\tilde{y}} \equiv T^{-1} \sum_{t \leq T} \mathbb{E}[\tilde{y}_t \tilde{y}_t^\top] \quad \text{and} \quad \bar{\sigma}_u^2 \equiv T^{-1} \sum_{t \leq T} \sigma_{u, t}^2.$$

Here the matrix $\bar{\Sigma}_{\tilde{y}}$ is identified and can be consistently estimated. The following lemma

shows that given the identification of $\bar{\Sigma}_{\tilde{y}}$, (32) is sufficient to identify both $\bar{\sigma}_u^2$ and the subspace orthogonal to the column space of $\bar{\lambda}$.

Lemma 3. *Under Assumption 1(iv),*

$$\bar{\sigma}_u^2 = \min_{a \in \mathcal{B}_{\mathbf{1}_n}} a^\top \bar{\Sigma}_{\tilde{y}} a, \quad (33)$$

where

$$\mathcal{B}_{\mathbf{1}_n} \equiv \{a \in \mathbb{R}^n : \mathbf{1}_n^\top a = 0, \|a\| = 1\}.$$

Moreover, the set of minimizers of (33) spans $\text{col}(\bar{\lambda}_\perp)$.

Lemma 3 provides a constructive characterization of $\bar{\lambda}_\perp$, which is central to the construction of generalized GIVs. In particular, $\bar{\lambda}_\perp$ can be recovered as the eigenspace associated with the smallest eigenvalue of $\bar{\Sigma}_{\tilde{y}}$ restricted to $\mathcal{B}_{\mathbf{1}_n}$. Intuitively, this corresponds to extracting directions of cross-sectional variation in y_t that are orthogonal to both the common factor structure and the aggregate component spanned by $\mathbf{1}_n$.

The minimization problem in (33), however, is defined over the constrained set $\mathcal{B}_{\mathbf{1}_n}$, which is not directly convenient for implementation. To facilitate computation, we next provide an equivalent representation that transforms this constrained problem into an unconstrained eigenvalue problem in $\mathbb{R}^{n-1}$.

Lemma 4. *Suppose Assumption 1(iv) holds. Consider the minimization problem:*

$$\min_{\tilde{a} \in B_{n-1}} \tilde{a}^\top Q_{-1}^\top \bar{\Sigma}_y Q_{-1} \tilde{a}, \quad (34)$$

where

$$\bar{\Sigma}_y \equiv T^{-1} \sum_{t \leq T} \mathbb{E}[y_t y_t^\top], \quad B_{n-1} \equiv \{\tilde{a} \in \mathbb{R}^{n-1} : \|\tilde{a}\| = 1\}.$$

Then a is a minimizer of (33) if and only if there exists a minimizer $\tilde{a}$ of (34) such that $a = Q_{-1} \tilde{a}$.

The solutions to (34) are given by the normalized eigenvectors associated with the smallest eigenvalue of the symmetric matrix $Q_{-1}^\top \bar{\Sigma}_y Q_{-1}$. By Lemma 3 and Lemma 4, these eigenvectors, after left multiplication by Q_{-1} , span the same subspace as $\bar{\lambda}_\perp$, thereby providing a feasible representation of the generalized GIVs. In practice, $Q_{-1}^\top \bar{\Sigma}_y Q_{-1}$ can be consistently estimated by its sample analogue, so the GIVs can be implemented via standard eigenvalue decomposition without requiring knowledge of the factor loadings λ .

3.2 Estimation and inference with GIVs

Building on Lemmas 2–4 in the previous subsection, the moment conditions (23) and (27) for the identification and estimation of $\theta \equiv (\phi, \psi)^\top$ can now be written as

$$\mathbb{E}[\bar{g}_T(\theta; A)] = \mathbf{0}_{2(n-\bar{r})}, \quad (35)$$

where

$$\bar{g}_T(\theta; A) \equiv T^{-1} \sum_{t \leq T} \begin{pmatrix} A^\top y_t (y_{e,t} - \phi p_t) \\ A^\top y_t (p_t - \psi y_{S,t}) \end{pmatrix}. \quad (36)$$

Here $A \equiv Q_{-1} A_0$, where Q_{-1} is an $n \times (n-1)$ matrix defined in Lemma 1, and A_0 is an $(n-1) \times (n-\bar{r})$ matrix collecting the eigenvectors corresponding to the smallest $n-\bar{r}$ eigenvalues of

$$S_y \equiv Q_{-1}^\top \bar{\Sigma}_y Q_{-1}.$$

Since A_0 depends on the unknown population covariance matrix, the moment function $\bar{g}_T(\theta; A)$ is not directly feasible in practice.

To construct feasible moment conditions, we replace A in (36) with $\hat{A} \equiv Q_{-1} \hat{A}_0$, where $\hat{A}_0$ collects the eigenvectors corresponding to the smallest $n-\bar{r}$ eigenvalues of

$$\hat{S}_y \equiv Q_{-1}^\top \hat{\Sigma}_y Q_{-1}, \quad \text{where } \hat{\Sigma}_y \equiv T^{-1} \sum_{t \leq T} y_t y_t^\top.$$

The GIV estimator is then defined as

$$\hat{\theta}(\hat{A}) \equiv \arg \min_{\theta \in \Theta} \bar{g}_T(\theta; \hat{A})^\top W_{0,T}(\hat{A}) \bar{g}_T(\theta; \hat{A}), \quad (37)$$

where

$$W_{0,T}(\hat{A}) \equiv ((\mathbf{I}_2 \otimes \hat{A}^\top) W_{0,T}(\mathbf{I}_2 \otimes \hat{A}))^{-1}, \quad (38)$$

and $W_{0,T}$ is a user-specified symmetric positive definite $2n \times 2n$ matrix.

Since $\bar{g}_T(\theta; \hat{A})$ is linear in θ , the GIV estimator admits the closed-form representation

$$\hat{\theta}(\hat{A}) = (D_{1,T}(\hat{A})^\top W_{0,T}(\hat{A}) D_{1,T}(\hat{A}))^{-1} (D_{1,T}(\hat{A})^\top W_{0,T}(\hat{A}) D_{2,T}(\hat{A})), \quad (39)$$

where $D_{j,T}(\hat{A}) \equiv (\mathbf{I}_2 \otimes \hat{A}^\top) D_{j,T}$ for $j = 1, 2$, and

$$D_{1,T} \equiv T^{-1} \sum_{t \leq T} \text{diag}(y_t p_t, y_t y_{S,t}), \quad D_{2,T} \equiv T^{-1} \sum_{t \leq T} \begin{pmatrix} y_t y_{e,t} \\ y_t p_t \end{pmatrix}. \quad (40)$$

We next present sufficient conditions for establishing the asymptotic properties of $\hat{\theta}(\hat{A})$ within the same framework as Gabaix and Koijen (2024), where the number of entities n is fixed and the number of observations T (indexed by t) tends to infinity.⁸ Let $\{\mu_j\}_{j \leq n-1}$ denote the eigenvalues of S_y arranged in increasing order, and let $A_{0,\perp}$ denote the matrix collecting the eigenvectors associated with $\{\mu_j\}_{n-\bar{r}+1 \leq j \leq n-1}$.

Assumption 2. (i) $|\phi\psi - 1| \geq K^{-1}$ and $|\phi| + |\psi| + \|\lambda\| \leq K$; (ii) for $a, b \in \{u, \eta, \varepsilon\}$,

$$T^{-1/2} \sum_{t \leq T} (a_t b_t^\top - \mathbb{E}[a_t b_t^\top]) = O_p(1);$$

$$(iii) \mu_{n-\bar{r}+1} - \bar{\sigma}_u^2 > K^{-1} \text{ and } \bar{\sigma}_u^2 > K^{-1}; (iv) \max_{t \leq T} \mathbb{E}[u_t^\top u_t + \varepsilon_t^2 + \eta_t^\top \eta_t] \leq K.$$

⁸The asymptotic properties of $\hat{\theta}(\hat{A})$, as well as inference for the unknown parameter θ , can be extended to the case where both n and T diverge by applying techniques from the many-moments literature; see, for example, Han and Phillips (2006) and Newey and Windmeijer (2009).

Assumption 2(i) ensures that the demand and supply system admits a well-defined reduced form and, consequently, a unique equilibrium. Assumption 2(ii) guarantees that the population second moments of $(u_t, \varepsilon_t, \eta_t)$ are approximated by their sample counterparts at the rate $T^{-1/2}$. From Lemma 3, we have $\mu_j = T^{-1} \sum_{t \leq T} \sigma_{u,t}^2$ for all $j \leq n - \bar{r}$. Therefore, Assumption 2(iii) imposes an eigenvalue gap condition on S_y , which is essential for consistent estimation of the eigenspace $\bar{\lambda}_\perp$. This condition can be verified under a lower bound condition on $\rho_{\min}((\mathbf{1}_n, \lambda)^\top (\mathbf{1}_n, \lambda))$, or on $\rho_{\min}(\lambda^\top \lambda)$ when $\mathbf{1}_n \in \text{col}(\lambda)$; see Lemma C.1 in Online Appendix C for details. Assumption 2(iii) also requires that the variance of the idiosyncratic demand shock be bounded away from zero, which is important for maintaining sufficient identification strength of the GIVs. Finally, Assumption 2(iv), together with 2(i), ensures that the second moments of y_t and p_t are well defined.

Assumption 3. *The sequence of market shares $\{S_t\}$ satisfies: (i)*

$$T^{-1/2} \sum_{t \leq T} (a_t b_t^\top - \mathbb{E}[a_t b_t^\top]) = O_p(1)$$

for $a_t, b_t \in \{S_t^\top u_t, \eta_t \otimes S_t, \varepsilon_t\}$, or $a_t \in \{u_t, \eta_t\}$ and $b_t \in \{S_t^\top u_t, \eta_t \otimes S_t\}$; (ii) $\mathbf{1}_n^\top S_t = 1$ for all t .

Assumption 3(i) imposes a set of high-level moment conditions ensuring that sample averages involving the weighted aggregates, such as $S_t^\top u_t$, satisfy a standard $T^{-1/2}$ law of large numbers. In particular, it requires that interactions between market-share weights and the structural shocks u_t , η_t and ε_t exhibit sufficiently weak temporal dependence and possess finite second moments. This condition is analogous to Assumption 2(ii), and is implied by it when S_t is time-invariant. More generally, both conditions can be verified under standard mixing or martingale difference assumptions. Assumption 3(ii) is a normalization condition requiring that the elements of S_t sum to one.

Let $v_t \equiv y_{e,t} - \phi p_t$. For $b \in \{v, \varepsilon\}$, define

$$\xi_{b,t} \equiv y_t b_t - \mathbb{E}[y_t b_t] + (y_t y_t^\top - \mathbb{E}[y_t y_t^\top]) \Upsilon \left(T^{-1} \sum_{t \leq T} \mathbb{E}[y_t b_t] \right), \quad (41)$$

where

$$\Upsilon \equiv Q_{-1} A_{0,\perp} (\bar{\sigma}_u^2 \mathbf{I}_{\bar{r}-1} - \Lambda_\perp)^{-1} A_{0,\perp}^\top Q_{-1}^\top, \quad \Lambda_\perp \equiv \text{diag}((\mu_j)_{n-\bar{r}+1 \leq j \leq n-1}).$$

The random vectors $\xi_{b,t}$, for $b \in \{v, \varepsilon\}$, represent the estimation errors in the moment conditions used to estimate ϕ and ψ , respectively. The first component, $y_t b_t - \mathbb{E}[y_t b_t]$, captures the sampling variation that would arise even if the factor loading matrix λ were known. The second component reflects the additional estimation error induced by replacing $\bar{\lambda}_\perp$ with its estimator, and hence accounts for the impact of estimating the loading matrix on the moment conditions.

Assumption 4. (i) $V^{-1/2} T^{-1/2} \sum_{t \leq T} \xi_t \rightarrow_d N(0, \mathbf{I}_{2n})$ where $\xi_t \equiv (\xi_{v,t}^\top, \xi_{\varepsilon,t}^\top)^\top$ with $\rho_{\min}(V) \geq K^{-1}$; (ii) $W_{0,T} = W_0 + o_p(1)$, where W_0 is a nonrandom symmetric matrix with $K^{-1} \leq \rho_{\min}(W_0) \leq \rho_{\max}(W_0) \leq K$; (iii) $\|T^{-1} \sum_{t \leq T} A^\top \mathbb{E}[y_t p_t]\| \geq K^{-1}$ and $\|T^{-1} \sum_{t \leq T} A^\top \mathbb{E}[y_t y_{S,t}]\| \geq K^{-1}$; (iv) there exists a matrix $\hat{V}$ such that $\hat{V} = V + o_p(1)$.

Assumption 4(i) concerns the asymptotic distribution of $V^{-1/2} T^{-1/2} \sum_{t \leq T} \xi_t$, which can be established via a central limit theorem. Here V denotes the variance matrix of $T^{-1/2} \sum_{t \leq T} \xi_t$. Assumption 4(ii) ensures consistent estimation of the weighting matrix, while Assumption 4(iii) guarantees that the GIVs provide sufficient identification strength for ϕ and ψ to be $T^{1/2}$ -estimable. The latter condition essentially requires that the weighted mean of the entities' market shares, $S_u \equiv T^{-1} \sum_{t \leq T} \sigma_{u,t}^2 \mathbb{E}[S_t]$ does not lie in $\text{col}((\mathbf{1}_n, \lambda))$. It can be verified under suitable primitive conditions; see Lemma C.2 in Online Appendix C for details. Finally, Assumption 4(iv) requires the existence of a consistent estimator of V .⁹

⁹A consistent estimator of V can be constructed using the estimated shocks $\hat{v}_t$ and $\hat{\varepsilon}_t$ obtained from the

To simplify the notation for the asymptotic variance of the GIV estimator, define

$$\Gamma(D_1, W_0, A) \equiv (D_1(A)^\top W_0(A) D_1(A))^{-1} D_1(A)^\top W_0(A),$$

where $D_1(A) \equiv (\mathbf{I}_2 \otimes A^\top) D_1$,

$$D_1 \equiv T^{-1} \sum_{t \leq T} \mathbb{E}[\text{diag}(y_t p_t, y_t y_{S,t})] \quad \text{and} \quad W_0(A) \equiv ((\mathbf{I}_2 \otimes A^\top) W_0 (\mathbf{I}_2 \otimes A))^{-1}.$$

The following theorem establishes the asymptotic distribution of the GIV estimator.

Theorem 1. *Under Assumptions 1, 2, 3, and 4(i)–(iii),*

$$(\Gamma(D_1, W_0, A) V(A) \Gamma(D_1, W_0, A)^\top)^{-1/2} T^{1/2} (\hat{\theta}(\hat{A}) - \theta) \rightarrow_d N(0, \mathbf{I}_2), \quad (42)$$

where

$$V(A) \equiv (\mathbf{I}_2 \otimes A^\top) V (\mathbf{I}_2 \otimes A).$$

Moreover, if Assumption 4(iv) also holds, then

$$\Gamma(D_{1,T}, W_{0,T}, \hat{A}) \hat{V}(\hat{A}) \Gamma(D_{1,T}, W_{0,T}, \hat{A})^\top = \Gamma(D_1, W_0, A) V(A) \Gamma(D_1, W_0, A)^\top + o_p(1), \quad (43)$$

where

$$\hat{V}(\hat{A}) \equiv (\mathbf{I}_2 \otimes \hat{A}^\top) \hat{V} (\mathbf{I}_2 \otimes \hat{A}),$$

and $\Gamma(D_{1,T}, W_{0,T}, \hat{A})$ is defined analogously to $\Gamma(D_1, W_0, A)$ with D_1 , W_0 , and A replaced by $D_{1,T}$, $W_{0,T}$, and $\hat{A}$, respectively.

Theorem 1 shows that the asymptotic variance of the GIV estimator is minimized when $W_0 = V$. Accordingly, the optimal weighting matrix can be obtained by setting $W_{0,T} = \hat{V}$

GIV estimator $\hat{\theta}(\hat{A})$ with identity weighting matrix $W_{0,T} = \mathbf{I}_{2n}$; see, for example, (A.2) in the implementation algorithm in Online Appendix A. The consistency of this variance estimator is established in Theorem D.1 in the Online Appendix.

in (38), which yields

$$W_{*,T}(\hat{A}) \equiv \hat{V}(\hat{A})^{-1}.$$

Let $\hat{\theta}^*(\hat{A})$ denote the corresponding optimally weighted GIV estimator. It then follows that

$$(D_1(A)^\top V(A)^{-1} D_1(A))^{1/2} T^{1/2} (\hat{\theta}^*(\hat{A}) - \theta) \rightarrow_d N(0, \mathbf{I}_2). \quad (44)$$

Standard errors for $\hat{\theta}^*(\hat{A})$ can be constructed from the square roots of the diagonal elements of

$$(TD_{1,T}(\hat{A})^\top W_{*,T}(\hat{A}) D_{1,T}(\hat{A}))^{-1}, \quad (45)$$

whose validity follows from (43).

Since there are $2(n - \bar{r})$ moment conditions for the identification and estimation of ϕ and ψ , these parameters are over-identified whenever $n > \bar{r} + 1$. In this case, the validity of the GIVs can be assessed using an over-identification test.

Theorem 2. *Under Assumptions 1, 2, 3, and 4,*

$$T\bar{g}_T(\hat{\theta}^*(\hat{A}); \hat{A})^\top W_{*,T}(\hat{A}) \bar{g}_T(\hat{\theta}^*(\hat{A}); \hat{A}) \rightarrow_d \chi^2(2(n - \bar{r} - 1)).$$

Theorem 2 establishes the asymptotic distribution of the J-test statistic under the null hypothesis that the moment conditions (35) are valid. When the GIVs are invalid, the power of the J-test follows from standard GMM arguments and is therefore omitted for brevity.

Remark 1. *Theorems 1 and 2 establish estimation and inference procedures for the demand and supply elasticities based on the full set of moment conditions. In some applications, however, interest may center on a single structural parameter, making it natural to estimate the demand and supply elasticities separately using the corresponding subsets of moment conditions.*

For example, when the demand elasticity is the primary parameter of interest, estimation

may be based on the moment functions

$$\bar{g}_{\phi,T}(\phi; A) \equiv T^{-1} \sum_{t \leq T} A^\top y_t (y_{e,t} - \phi p_t). \quad (46)$$

Using arguments analogous to those in the proof of Theorem 1, the resulting GIV estimator can be shown to be $T^{1/2}$ -consistent and asymptotically normal. Its asymptotic variance and standard error can be constructed in the same manner as those of the joint GMM estimator $\hat{\theta}^*(\hat{A})$. In particular, a formula analogous to (45) applies after removing the components associated with the supply-side moment conditions.

Remark 2. The preceding discussion also extends naturally to specification testing. In particular, the J -test constructed from the moment conditions in (46) and the corresponding GIV estimator provides a test of the validity of the demand-side moment restrictions. Under correct specification, the resulting J -statistic converges in distribution to $\chi^2(n - \bar{r} - 1)$. Analogous estimation, inference, and specification-testing results hold when the analysis is based solely on the supply-side moment conditions.

3.3 Estimating the number of GIVs

Construction of the GIVs requires knowledge of the rank $\bar{r}$ of the matrix $(\mathbf{1}_n, \lambda)$, which may not be feasible in practice. In this subsection, we propose a Bayesian information criterion (BIC) for consistent estimation of $\bar{r}$. The construction is motivated by Lemma 3 and Lemma 4.

Specifically, let $\{\hat{\mu}_j\}_{j \leq n-1}$ denote the eigenvalues of $\hat{S}_y$ arranged in increasing order. Define the information criterion

$$\text{BIC}_T(j) \equiv \frac{T}{n-j} \sum_{s=1}^{n-j} \frac{(\hat{\mu}_s - \hat{\mu}_1)^2}{2\hat{\mu}_s^2} + j \log(T), \quad (47)$$

for $j \in \mathcal{J}$, where $\mathcal{J} \equiv \{1, \dots, n-1\}$. The estimator $\hat{r}$ of $\bar{r}$ is then given by

$$\hat{r} = \arg \min_{j \in \mathcal{J}} \text{BIC}_T(j). \quad (48)$$

The consistency of $\hat{r}$ is established in Theorem 3.

Theorem 3. *Under Assumptions 1, 2 and 3, we have $\hat{r} = \bar{r}$ with probability approaching 1.*

We conclude this subsection by providing intuition for the construction of (47). The criterion $\text{BIC}_T(j)$ consists of two components. The first term, $T(n-j)^{-1} \sum_{s=1}^{n-j} (\hat{\mu}_s - \hat{\mu}_1)^2 / (2\hat{\mu}_s^2)$, is decreasing in j and captures an over-fitting effect analogous to that in classical regression settings, where j reflects the dimension or complexity of the model. The second term, $j \log(T)$, is strictly increasing in j and serves as a penalty on model complexity. The estimator $\hat{r}$ in (48) therefore balances the trade-off between goodness-of-fit and model complexity.

The eigenvalues $\{\hat{\mu}_j\}_{j \leq n-1}$ are $T^{1/2}$ -consistent estimators of $\{\mu_j\}_{j \leq n-1}$ under Assumptions 2(i, ii). Since $\mu_s = \bar{\sigma}_u^2$ for $s \in \{1, \dots, n - \bar{r}\}$, it follows that for $j \geq \bar{r}$,

$$\frac{T}{n-j} \sum_{s=1}^{n-j} \frac{(\hat{\mu}_s - \hat{\mu}_1)^2}{2\hat{\mu}_s^2} = O_p(1).$$

Consequently, the penalty term $j \log(T)$ dominates the first term in (47). Since $j \log(T)$ is strictly increasing in j , $\text{BIC}_T(j)$ is asymptotically minimized at $\bar{r}$ over $j \in \{\bar{r}, \dots, n-1\}$. On the other hand, for $j < \bar{r}$, we have $n-j \geq n - \bar{r} + 1$, so Assumption 2(iii) implies that $\mu_{n-j} - \mu_1$ is bounded away from zero. Combined with the $T^{-1/2}$ consistency of $\{\hat{\mu}_j\}_{j \leq n-1}$, this yields

$$\frac{T}{n-j} \sum_{s=1}^{n-j} \frac{(\hat{\mu}_s - \hat{\mu}_1)^2}{2\hat{\mu}_s^2} \geq \frac{T}{n-j} \frac{(\hat{\mu}_{n-j} - \hat{\mu}_1)^2}{2\hat{\mu}_{n-j}^2} \geq K^{-1}T((\mu_{n-j} - \mu_1)^2 - O_p(T^{-1/2})),$$

which diverges at rate T . This term therefore dominates the penalty term of order $\log(T)$, implying that $\text{BIC}_T(j) > \text{BIC}_T(\bar{r})$ wpa1 for all $j < \bar{r}$. Combining the two cases, $\text{BIC}_T(j)$ is asymptotically minimized at $j = \bar{r}$ over $j \in \mathcal{J}$, which ensures the consistency of $\hat{r}$.

4 Extensions and Discussion

This section provides further discussion and extensions of the main results established in the previous section. First, we show that the identification strategy in Gabaix and Koijen (2024) may fail when the factor loadings are unknown, potentially leading to inconsistent estimation and invalid inference under standard GMM procedures. Second, we demonstrate that our estimation and inference procedures can be straightforwardly extended to settings with exogenous regressors in the demand and supply equations and to data with unbalanced features. The latter extension shows that our methods remain applicable even when the number of entities n is large and may exceed T .

4.1 Identification failure of factor loadings in Gabaix and Koijen (2024)

In this subsection, we show that the identification strategy in Gabaix and Koijen (2024), in particular their Proposition 7, may fail when λ is unknown. Gabaix and Koijen (2024) impose a normalization on the factor loadings λ by setting the loadings of the first factor $\eta_{1,t}$ to be $\mathbf{1}_n$, and the loadings of the remaining factors $\eta_{2,t}$, denoted by λ_{-1} , to satisfy¹⁰

$$\mathbf{1}_n^\top \lambda_{-1} = \mathbf{0}_{r-1}^\top, \quad n^{-1} \lambda_{-1}^\top \lambda_{-1} = \mathbf{I}_{r-1}. \quad (49)$$

Moreover, they assume¹¹

$$\text{Var}(u_t) = \sigma_u^2 \mathbf{I}_n, \quad \text{Cov}(\eta_t, u_t) = \mathbf{0}_{r \times n}, \quad \mathbb{E}[\eta_t] = \mathbf{0}_r, \quad \text{Cov}(u_t, \varepsilon_t) = \mathbf{0}_n. \quad (50)$$

¹⁰See the second paragraph above Proposition 7 in Gabaix and Koijen (2024). In the same paragraph, they also impose the restriction that $\text{Var}(\lambda_{-1}^\top y_t)$ is diagonal with distinct diagonal entries ($\lambda_{-1}^\top y_t$ here is equal to $n\tilde{\eta}_t$ in their notation). As we show below, this additional restriction should be interpreted as an assumption on the latent factors rather than a normalization on the factor loadings. Moreover, imposing this restriction does not resolve the identification issue in their approach.

¹¹See the second paragraph on page 2279, display (3), and Assumption 3 in Gabaix and Koijen (2024).

For notational simplicity, we abstract from exogenous regressors in both the demand and supply equations. The demand equation (19) can then be written as

$$y_t = \phi p_t \mathbf{1}_n + \eta_{1,t} \mathbf{1}_n + \lambda_{-1} \eta_{2,t} + u_t, \quad (51)$$

while the supply equation remains the same as (20).

Gabaix and Koijen (2024) propose using both $\lambda_{\perp}^{\top} y_t$ and $\lambda_{-1}^{\top} y_t$ as IVs to construct moment conditions for identifying and estimating ϕ , ψ , and λ_{-1} .¹² Because $\mathbf{1}_n^{\top} \lambda_{-1} = \mathbf{0}_{r-1}^{\top}$, one can partial out $\eta_{2,t}$ by premultiplying (51) by $e^{\top}$, yielding

$$y_{e,t} - \phi p_t = \eta_{1,t} + u_{e,t}. \quad (52)$$

Moreover, because $n^{-1} \lambda_{-1}^{\top} \lambda_{-1} = \mathbf{I}_{r-1}$, premultiplying (51) by $\lambda_{-1}^{\top}$, we also obtain

$$\lambda_{-1}^{\top} y_t = n \eta_{2,t} + \lambda_{-1}^{\top} u_t. \quad (53)$$

From Assumption 1(i) and (50), it follows that the product of (53) and the second term on the right of (52) is such that

$$\mathbb{E}[u_{e,t} \lambda_{-1}^{\top} y_t] = n \mathbb{E}[\eta_{2,t} u_{e,t}] + \lambda_{-1}^{\top} \mathbb{E}[u_t u_{e,t}] = \mathbf{0}_{r-1}. \quad (54)$$

However, the product $\mathbb{E}[\eta_{1,t} \lambda_{-1}^{\top} y_t]$ of the first term on the right of (52) and (53) may be nonzero due to possible correlation between $\eta_{1,t}$ and $\eta_{2,t}$. Therefore, $\lambda_{-1}^{\top} y_t$ cannot be directly used together with $y_{e,t} - \phi p_t$ to form valid moment conditions. To address this issue, Gabaix and Koijen (2024) introduce the regression coefficient b_y of $\eta_{1,t}$ on $y_t^{\top} \lambda_{-1}$ such that

$$\mathbb{E}[(\eta_{1,t} - y_t^{\top} \lambda_{-1} b_y) \lambda_{-1}^{\top} y_t] = \mathbf{0}_{r-1}. \quad (55)$$

¹²They correspond to $z_t(m^y)$ and $\tilde{\eta}_t(m^y)$ in Proposition 4 of Gabaix and Koijen (2024).

Combining this with (52) and (54) yields

$$\mathbb{E}[(y_{e,t} - \phi p_t - y_t^\top \lambda_{-1} b_y) \lambda_{-1}^\top y_t] = \mathbf{0}_{r-1}. \quad (56)$$

Since $y_{e,t} - \phi p_t - y_t^\top \lambda_{-1} b_y = u_{e,t} - u_t^\top \lambda_{-1} b_y + \eta_{1,t} - n\eta_{2,t}^\top b_y$, it follows from Assumption 1(i) and (50) that

$$\mathbb{E}[(y_{e,t} - \phi p_t - y_t^\top \lambda_{-1} b_y) \lambda_{-1}^\top y_t] = \mathbb{E}[(u_{e,t} - u_t^\top \lambda_{-1} b_y + \eta_{1,t} - n\eta_{2,t}^\top b_y) \lambda_{-1}^\top u_t] = \mathbf{0}_{n-r}.$$

Together with (56), this yields the moment conditions from the demand equation:

$$\mathbb{E} \left[(y_{e,t} - \phi p_t - y_t^\top \lambda_{-1} b_y) \begin{pmatrix} \lambda_{-1}^\top y_t \\ \lambda_{-1}^\top y_t \end{pmatrix} \right] = \mathbf{0}_{n-1}. \quad (57)$$

Similarly, $\lambda_{-1}^\top y_t$ cannot be directly used together with $p_t - \psi y_{S,t}$ to identify ψ , because $\lambda_{-1}^\top y_t$ contains $\eta_{2,t}$, which may be correlated with ε_t . Gabaix and Koijen (2024) therefore propose

$$\mathbb{E}[(p_t - \psi y_{S,t} - y_t^\top \lambda_{-1} b_p) \lambda_{-1}^\top y_t] = \mathbf{0}_{r-1},$$

where b_p is defined by

$$\mathbb{E}[(\varepsilon_t - y_t^\top \lambda_{-1} b_p) \lambda_{-1}^\top y_t] = \mathbf{0}_{r-1}.$$

Moreover, by Assumption 1(i) and (50),

$$\mathbb{E}[(p_t - \psi y_{S,t} - y_t^\top \lambda_{-1} b_p) \lambda_{-1}^\top y_t] = \mathbb{E}[(\varepsilon_t - n\eta_{2,t}^\top b_p - u_t^\top \lambda_{-1} b_p) \lambda_{-1}^\top u_t] = \mathbf{0}_{n-r},$$

which provides additional moment conditions. Therefore, the moment conditions from the

supply equation are

$$\mathbb{E} \left[(p_t - \psi y_{S,t} - y_t^\top \lambda_{-1} b_p) \begin{pmatrix} \lambda_{-1}^\top y_t \\ \lambda_{-1}^\top y_t \end{pmatrix} \right] = \mathbf{0}_{n-1}. \quad (58)$$

The moment conditions in (57) and (58) coincide with those in Proposition 4 of Gabaix and Koijen (2024) when additional exogenous variables are excluded. Gabaix and Koijen (2024) argue in their Proposition 4 that these conditions identify ϕ , ψ , b_y , and b_p when λ_{-1} is known. When λ_{-1} is unknown, they propose the additional moment conditions

$$\mathbb{E}[(M_{\mathbf{1}_n} - \lambda_{-1}(\lambda_{-1}^\top \lambda_{-1})^{-1} \lambda_{-1}^\top) y_t y_t^\top \lambda_{-1}] = \mathbf{0}_{n \times (r-1)}, \quad (59)$$

which correspond to equation (50) in Proposition 7 of Gabaix and Koijen (2024). To verify (59), note that $M_{\mathbf{1}_n} \lambda_{-1} = \lambda_{-1}$ by (49). Using (51),

$$(M_{\mathbf{1}_n} - \lambda_{-1}(\lambda_{-1}^\top \lambda_{-1})^{-1} \lambda_{-1}^\top) y_t = (M_{\mathbf{1}_n} - \lambda_{-1}(\lambda_{-1}^\top \lambda_{-1})^{-1} \lambda_{-1}^\top) u_t.$$

Hence, (59) follows from Assumption 1(i).

Since (59) provides $n(r-1)$ moment conditions for $n(r-1)$ unknown entries in λ_{-1} , it may seem to deliver exact identification.¹³ However, as shown in the lemma below, this is not the case: the restrictions in (59) fail to uniquely identify λ_{-1} , even up to rotation.

Lemma 5. *Suppose that $\mathbb{E}[y_t y_t^\top]$ is finite and nonsingular, and that the conditions in (50) hold. Let $\{d_j\}_{j=1}^{n-1}$ be an orthonormal basis of eigenvectors associated with the nonzero eigenvalues of $\mathbb{E}[\tilde{y}_t \tilde{y}_t^\top]$. For any subset $J \subset \{1, \dots, n-1\}$ with $|J| = r-1$, let D_J collect the columns d_j , $j \in J$. Then $n^{1/2} D_J$ satisfies (49) and (59).*

Lemma 5 shows that the moment condition (59), together with the normalization in (49), fails to identify $\text{col}(\lambda_{-1})$. Indeed, (49) and (59) admit $\binom{n-1}{r-1}$ different choices of D_J , whose

¹³Indeed, Gabaix and Koijen (2024) state at the top of page 2294 that “The new moment (50) identifies $\check{\lambda}$.” In our notation, their moment (50) corresponds to (59), while their $\check{\lambda}$ is denoted here by λ_{-1} .

column spaces are generally distinct. Moreover,

$$\text{col}((d_1, \dots, d_{n-1})) = \text{col}(M_{\mathbf{1}_n}) = \text{col}((\lambda_{-1}, \lambda_{\perp})).$$

Therefore, if $\text{col}(D_J) \neq \text{col}(\lambda_{-1})$, then the orthogonal complement of $(\mathbf{1}_n, D_J)$, denoted by $D_{J,\perp}$, need not be orthogonal to the true factor space $\text{col}(\lambda_{-1})$. In particular,

$$D_{J,\perp}^\top \lambda_{-1} \neq \mathbf{0}_{(n-r) \times (r-1)}$$

may hold. Consequently, the candidate GIVs $D_{J,\perp}^\top y_t$ may still contain components of the latent factors η_t , since

$$D_{J,\perp}^\top y_t = D_{J,\perp}^\top \lambda_{-1} \eta_{2,t} + D_{J,\perp}^\top u_t.$$

As a result, moment conditions constructed from $D_{J,\perp}^\top y_t$ may fail to eliminate the latent factor component and therefore need not provide valid identifying restrictions for the structural parameters.

To make the identification problem more explicit, consider the special case $r = n - 1$. In this case, $\lambda_{\perp}$ is one-dimensional, and the moment conditions in (57) and (58) provide only exact identification for the unknown parameters ϕ , ψ , b_y , and b_p , given the true factor loading matrix λ_{-1} and its orthogonal complement $\lambda_{\perp}$. Therefore, identification of these parameters ultimately relies on identification of $\text{col}(\lambda_{-1})$.

However, as discussed above, (57) and (58) admit $n - 1$ different choices of D_J , and hence $n - 1$ corresponding choices of $D_{J,\perp}$. Since $\lambda_{\perp}$ is one-dimensional, at least $n - 2$ of these choices satisfy

$$D_{J,\perp}^\top \lambda_{-1} \neq \mathbf{0}_{1 \times (r-1)}.$$

For such choices, the corresponding candidate GIVs $D_{J,\perp}^\top y_t$ retain latent factor components and therefore generally fail to identify the true elasticities ϕ and ψ . Consequently, GMM estimation based on these invalid instruments would generally converge to pseudo-true val-

ues rather than the true structural parameters. Moreover, since different choices of $D_{J,\perp}$ generally lead to different pseudo-true values, the resulting limits need not even be uniquely determined.

Lemma 5 establishes the non-identification of the factor loadings under the normalization in (49). In addition to (49), Gabaix and Koijen (2024) also assume that $\text{Var}(\lambda_{-1}^\top y_t)$ is a diagonal matrix with distinct diagonal entries.¹⁴ Under (49) and (50), however,

$$\text{Var}(\lambda_{-1}^\top y_t) = \lambda_{-1}^\top \mathbb{E}[y_t y_t^\top] \lambda_{-1} = n^2 \mathbb{E}[\eta_{2,t} \eta_{2,t}^\top] + n \sigma_u^2 \mathbf{I}_{r-1}.$$

Hence, the diagonal structure imposed on $\text{Var}(\lambda_{-1}^\top y_t)$ amounts to additional restrictions on the latent factors, requiring that the components of $\eta_{2,t}$ are uncorrelated and have distinct variances. Since $\mathbb{E}[y_t y_t^\top]$ is unknown, this condition should be viewed as an assumption on the latent factors rather than a restriction on the factor loadings.¹⁵

Even after imposing this restriction together with (49), the moment condition (59) still fails to identify $\text{col}(\lambda_{-1})$. Indeed, by (49), (50), and (51),

$$\mathbb{E}[\tilde{y}_t \tilde{y}_t^\top] = M_{1_n} \mathbb{E}[y_t y_t^\top] M_{1_n} = \lambda_{-1} \mathbb{E}[\eta_{2,t} \eta_{2,t}^\top] \lambda_{-1}^\top + \sigma_u^2 M_{1_n}.$$

Under the additional assumption that $\mathbb{E}[\eta_{2,t} \eta_{2,t}^\top]$ is diagonal with distinct diagonal entries, the matrix $\mathbb{E}[\tilde{y}_t \tilde{y}_t^\top]$ has $r - 1$ distinct eigenvalues larger than σ_u^2 , while σ_u^2 itself is an eigenvalue with multiplicity $n - r$. The eigenspace associated with σ_u^2 is spanned by $\lambda_\perp$. Consequently, there are at least $(r - 1)(n - r) + 1$ such choices of D_J with distinct column spaces.¹⁶ Since

¹⁴See the second paragraph above Proposition 7 in Gabaix and Koijen (2024).

¹⁵If one instead imposes the corresponding restriction on the sample second moment $T^{-1} \sum_{t \leq T} y_t y_t^\top$, then for a broad class of data-generating processes, the orthonormalized eigenvectors associated with any collection of $r - 1$ eigenvalues may satisfy the empirical counterpart of this restriction in finite samples due to estimation error in $T^{-1} \sum_{t \leq T} y_t y_t^\top$.

¹⁶Note that one choice is given by $D_J = n^{-1/2} \lambda_{-1}$, while the remaining $(r - 1)(n - r)$ choices are obtained by selecting $r - 2$ columns from λ_{-1} columns and one column from $\lambda_\perp$.

these columns are orthonormal eigenvectors of $\mathbb{E}[\tilde{y}_t \tilde{y}_t^\top]$, it follows that

$$\text{Var}(n^{1/2} D_J^\top y_t) = n D_J^\top \mathbb{E}[y_t y_t^\top] D_J = n D_J^\top \mathbb{E}[\tilde{y}_t \tilde{y}_t^\top] D_J$$

is diagonal with distinct diagonal entries. Since there exist $(r-1)(n-r)+1$ such choices of D_J with distinct column spaces, this again shows that $\text{col}(\lambda_{-1})$ is not identified.

We conclude this subsection with a lemma establishing the rotational non-uniqueness of the factor loadings identified from (49), (57)–(59) in the general case.

Lemma 6. *Suppose that $\{b_y, b_p, \lambda_{-1}, \lambda_\perp\}$ satisfies (57)–(59) given ϕ and ψ . Let C_1 and C_2 be arbitrary orthogonal matrices of dimensions $(r-1) \times (r-1)$ and $(n-r) \times (n-r)$, respectively. Define*

$$\lambda_{C_1} \equiv \lambda_{-1} C_1, \quad \lambda_{C_2} \equiv \lambda_\perp C_2.$$

Then $\{C_1^\top b_y, C_1^\top b_p, \lambda_{C_1}, \lambda_{C_2}\}$ also satisfies (57)–(59) given ϕ and ψ . Moreover, λ_{C_1} and λ_{C_2} satisfy the same normalization and orthogonality restrictions as λ_{-1} and $\lambda_\perp$.

Lemma 6 shows that the factor loadings λ_{-1} and the regression coefficients b_y and b_p are, at best, identified up to an orthonormal rotation. This raises concerns for estimation and inference based on these moment conditions, since standard GMM procedures require uniqueness of the identified parameters and are therefore not directly applicable in this setting.

The non-identification issue here is more challenging to address than in our approach, because $\{b_y, b_p, \lambda_{-1}, \lambda_\perp\}$ are jointly identified together with the demand and supply elasticities. This joint determination complicates both the computation of the GIV estimator and the analysis of its statistical properties. In contrast, our approach separates the identification of the factor loadings from that of the elasticity parameters. We then exploit the invariance properties of the GIV estimator and the J -test statistic to address the fact that the factor loadings are only identified up to rotation, thereby allowing standard GMM estimation and inference to remain valid.

4.2 Model with exogenous regressors

This subsection extends the model studied in the previous section by allowing for additional exogenous regressors in both the demand and supply equations in (19)–(20). Specifically, we consider

$$y_t = \phi p_t \mathbf{1}_n + x_t \beta + \lambda \eta_t + u_t, \quad (60)$$

$$p_t = \psi y_{S,t} + w_t^\top \gamma + \varepsilon_t, \quad (61)$$

where $x_t \equiv (x_{1,t}, \dots, x_{n,t})^\top$ with $x_{i,t} \in \mathbb{R}^{d_x}$, and $w_t \in \mathbb{R}^{d_w}$ denote observed exogenous variables that have direct effects on demand and supply, respectively. The variables x_t include both sector fixed effects and unit-level demand shifters, while w_t captures aggregate supply shifters.¹⁷ Under suitable exogeneity conditions, the main identification and estimation arguments continue to apply after partialling out x_t from the demand equation (60).¹⁸

Specifically, multiplying $M_{\mathbf{1}_n}$ on both sides of (60) yields

$$\tilde{y}_t = \tilde{x}_t \beta + \tilde{\lambda} \eta_t + \tilde{u}_t,$$

where $\tilde{x}_t \equiv M_{\mathbf{1}_n} x_t$, and $\tilde{y}_t$, $\tilde{\lambda}$, and $\tilde{u}_t$ are defined analogously; see (29). Let $\tilde{y}_t^* \equiv M_{\mathbf{1}_n} y_t^*$, where $y_t^* \equiv y_t - x_t \beta$. Then the above equation can be written as

$$\tilde{y}_t^* = \tilde{\lambda} \eta_t + \tilde{u}_t,$$

which takes a form similar to (30).¹⁹ Therefore, we can apply Lemmas 3 and 4 in Subsection

¹⁷Since $\lambda \eta_t$ can be decomposed as $\lambda(\eta_t - \mathbb{E}[\eta_t]) + \lambda \mathbb{E}[\eta_t]$, and $\lambda \mathbb{E}[\eta_t]$ can be absorbed into the sector fixed effects, we assume without loss of generality that $\mathbb{E}[\eta_t] = \mathbf{0}_r$ throughout this subsection.

¹⁸When x_t includes variables excluded from the supply equation, these may serve as IVs for identifying ψ in (61) if they are uncorrelated with ε_t . Similarly, variables in w_t excluded from the demand equation may identify ϕ in (60) if they are uncorrelated with η_t and u_t . Although standard in the classical simultaneous equations literature, this strategy is not widely used in empirical applications of GIV. We therefore do not assume the existence or exogeneity of such excluded variables.

¹⁹Since w_t is invariant across i , it is automatically partialled out in $\tilde{y}$. Therefore, $\tilde{y}_t^*$ effectively partials out the exogenous regressors in both the demand and supply equations.

3.1, with the second moment matrix $\bar{\Sigma}_{\tilde{y}}$ replaced by

$$\bar{\Sigma}_{\tilde{y}^*} \equiv T^{-1} \sum_{t \leq T} \mathbb{E}[\tilde{y}_t^* \tilde{y}_t^{*\top}],$$

to identify the subspace $\text{col}(\bar{\lambda}_\perp)$, which is orthogonal to $(\mathbf{1}_n, \lambda)$.

Given β , the GIVs and moment conditions can be constructed in the same way as in the previous section, with y_t replaced by y_t^* . To proceed, we need to estimate β , which is required to construct an estimator for y_t^* . The unknown parameter β is estimated by

$$\hat{\beta} \equiv \left(\sum_{t \leq T} \tilde{x}_t^\top \tilde{x}_t \right)^{-1} \left(\sum_{t \leq T} \tilde{x}_t^\top \tilde{y}_t \right).$$

If x_t is exogenous, in the sense that it is uncorrelated with both η_t and u_t , then standard least squares theory implies that $\hat{\beta}$ is a $T^{1/2}$ -consistent estimator of β . Given $\hat{\beta}$, define

$$\hat{y}_t^* \equiv y_t - x_t \hat{\beta}, \quad \hat{y}_{e,t}^* \equiv e^\top \hat{y}_t^*.$$

We then obtain $\hat{A} \equiv Q_{-1} \hat{A}_0$, where $\hat{A}_0$ collects the eigenvectors corresponding to the smallest $n - \bar{r}$ eigenvalues of

$$\hat{S}_{y^*} \equiv Q_{-1}^\top \hat{\Sigma}_{\tilde{y}^*} Q_{-1}, \quad \hat{\Sigma}_{\tilde{y}^*} \equiv T^{-1} \sum_{t \leq T} \hat{y}_t^* \hat{y}_t^{*\top}.$$

The moment conditions used to estimate the unknown parameter $\theta \equiv (\phi, \psi, \gamma^\top)^\top$ are

$$\bar{g}_T(\theta; \hat{A}, \hat{\beta}) \equiv T^{-1} \sum_{t \leq T} \begin{pmatrix} \hat{A}^\top \hat{y}_t^* (\hat{y}_{e,t}^* - \phi p_t) \\ \hat{A}^\top \hat{y}_t^* (p_t - \psi y_{S,t} - w_t^\top \gamma) \\ w_t (p_t - \psi y_{S,t} - w_t^\top \gamma) \end{pmatrix}. \quad (62)$$

The GIV estimator is defined as

$$\hat{\theta}(\hat{A}) \equiv \arg \min_{\theta \in \Theta} \bar{g}_T(\theta; \hat{A}, \hat{\beta})^\top W_{0,T}(\hat{A}) \bar{g}_T(\theta; \hat{A}, \hat{\beta}), \quad (63)$$

where

$$W_{0,T}(\hat{A}) \equiv \left(\text{diag}(\hat{A}^\top, \hat{A}^\top, \mathbf{I}_{d_w}) W_{0,T} \text{diag}(\hat{A}, \hat{A}, \mathbf{I}_{d_w}) \right)^{-1}, \quad (64)$$

and $W_{0,T}$ is a user-specified symmetric positive definite $(2n + d_w) \times (2n + d_w)$ matrix. Since $\bar{g}_T(\theta; \hat{A}, \hat{\beta})$ is linear in θ , the GIV estimator takes the same form as in (39), with $D_{j,T}(\hat{A})$ redefined as

$$D_{j,T}(\hat{A}) \equiv \text{diag}(\hat{A}^\top, \hat{A}^\top, \mathbf{I}_{d_w}) D_{j,T}, \quad j = 1, 2, \quad (65)$$

where

$$D_{1,T} \equiv T^{-1} \sum_{t \leq T} \begin{pmatrix} \hat{y}_t^* p_t & \mathbf{0}_n & \mathbf{0}_{n \times d_w} \\ \mathbf{0}_n & \hat{y}_t^* y_{S,t} & \hat{y}_t^* w_t^\top \\ \mathbf{0}_{d_w} & w_t y_{S,t} & w_t w_t^\top \end{pmatrix}, \quad D_{2,T} \equiv T^{-1} \sum_{t \leq T} \begin{pmatrix} \hat{y}_t^* \hat{y}_{e,t}^* \\ \hat{y}_t^* p_t \\ w_t p_t \end{pmatrix}. \quad (66)$$

To conduct inference on θ and test the validity of the moment conditions in (62), one must account for the estimation error in $\hat{\beta}$, since it enters $\bar{g}_T(\theta; \hat{A}, \hat{\beta})$ through both $\hat{y}_t^*$ and $\hat{A}$. Lemma E.1 in Online Appendix E shows that the randomness introduced by the estimation error of $\hat{\beta}$ is of higher order. Therefore, the estimation error of $\hat{\beta}$ is asymptotically negligible and can be ignored. Consequently, the standard errors of $\hat{\theta}(\hat{A})$ and the specification tests can be constructed in the same way as in the previous section. See Algorithm 1 in Online Appendix A for details.

4.3 Model with unbalanced data structure

The data used to estimate the demand and supply equations have, thus far, been assumed to follow a balanced structure. For instance, $y_{i,t}$ denotes the demand of entity i in period t , where $i \in \{1, \dots, n\}$ and $t \in \{1, \dots, T\}$. The analysis in the previous section assumes

a balanced data structure, so that each time period is associated with the same number of entities. In practice, however, entry and exit lead to an unbalanced data structure. As we show below, the identification and estimation approach extends naturally to this setting, provided that the entry and exit decisions of entities are independent of their demand.

Specifically, the demand equation in (19) is generalized as

$$y_t = \phi p_t \mathbf{1}_{n_t} + \lambda \eta_t + u_t,$$

where $y_t \equiv (y_{i,t})_{i \leq n_t}$ and n_t denotes the number of entities present in the market at time t . The factor-loading matrix λ is of dimension $n_t \times r$, and the idiosyncratic shock vector u_t is of dimension $n_t \times 1$. The supply equation in (20) remains unchanged, with $y_{S,t} \equiv S_t^\top y_t$, where S_t is an $n_t \times 1$ vector of market shares.

To construct the GIV, consider a subsample of n_0 entities that are observed in all periods. Let y_t^0 denote the corresponding subvector of y_t . Their demand equation is

$$y_t^0 = \phi p_t \mathbf{1}_{n_0} + \lambda_0 \eta_t + u_t^0, \tag{67}$$

where λ_0 and u_t^0 are the associated submatrices of λ and u_t . Let $Q_{0,-1} \equiv (q_{0,2}, \dots, q_{0,n_0}) \in \mathbb{R}^{n_0 \times (n_0-1)}$ be defined analogously to (13) with n replaced by n_0 , and $\lambda_{0,\perp} \in \mathbb{R}^{n_0 \times (n_0-\bar{r}_0)}$ denotes the orthogonal complement of $(\mathbf{1}_{n_0}, \lambda_0)$ where $\bar{r}_0 \equiv \text{rank}((\mathbf{1}_{n_0}, \lambda_0))$. Since $\lambda_{0,\perp}^\top y_t^0 = \lambda_{0,\perp}^\top u_t^0$, Assumption 1 implies

$$\mathbb{E}[\eta_t u_t^{0\top} \lambda_{0,\perp}] = \mathbf{0}_{r \times (n_0-\bar{r}_0)}, \quad \mathbf{1}_{n_t}^\top \mathbb{E}[u_t u_t^{0\top}] \lambda_{0,\perp} = \mathbf{0}_{n_0-\bar{r}_0}, \quad \mathbb{E}[\varepsilon_t u_t^{0\top} \lambda_{0,\perp}] = \mathbf{0}_{n_0-\bar{r}_0}.$$

These imply $\lambda_{0,\perp}^\top y_t^0$ provides valid moment conditions

$$\mathbb{E} \left[\begin{pmatrix} \lambda_{0,\perp}^\top y_t^0 (y_{e,t} - \phi p_t) \\ \lambda_{0,\perp}^\top y_t^0 (p_t - \psi y_{S,t}) \end{pmatrix} \right] = \mathbf{0}_{2(n_0-\bar{r}_0)},$$

which identify ϕ and ψ .

Lemmas 3 and 4 apply to this subsample. In particular, $\lambda_{0,\perp}$ can be consistently estimated (up to an orthonormal rotation) by $\hat{A} \equiv Q_{0,-1}\hat{A}_0$, where $\hat{A}_0$ collects the eigenvectors corresponding to the smallest $n_0 - \bar{r}_0$ eigenvalues of

$$\hat{S}_{y^0} \equiv Q_{0,-1}^\top \hat{\Sigma}_{y^0} Q_{0,-1}, \quad \hat{\Sigma}_{y^0} \equiv T^{-1} \sum_{t \leq T} y_t^0 y_t^{0\top}.$$

The GIV estimator is defined analogously to (37), with moment function

$$\bar{g}_T^0(\theta; \hat{A}) \equiv T^{-1} \sum_{t \leq T} \begin{pmatrix} \hat{A}^\top y_t^0 (y_{e,t} - \phi p_t) \\ \hat{A}^\top y_t^0 (p_t - \psi y_{S,t}) \end{pmatrix},$$

and a user-specified symmetric positive definite $2n_0 \times 2n_0$ weight matrix, and it admits an explicit form

$$\hat{\theta}(\hat{A}) = (D_{1,T}^0(\hat{A})^\top W_{0,T}(\hat{A}) D_{1,T}^0(\hat{A}))^{-1} (D_{1,T}^0(\hat{A})^\top W_{0,T}(\hat{A}) D_{2,T}^0(\hat{A})), \quad (68)$$

where $D_{j,T}^0(\hat{A}) \equiv (\mathbf{I}_2 \otimes \hat{A}^\top) D_{j,T}^0$ for $j = 1, 2$, and

$$D_{1,T}^0 \equiv T^{-1} \sum_{t \leq T} \text{diag}(y_t^0 p_t, y_t^0 y_{S,t}), \quad D_{2,T}^0 \equiv T^{-1} \sum_{t \leq T} \begin{pmatrix} y_t^0 y_{e,t} \\ y_t^0 p_t \end{pmatrix}. \quad (69)$$

The asymptotic normality of the GIV estimator and the asymptotic distribution of the J -test statistic follow from the same arguments as in the proofs of Theorems 1 and 2, with the appropriate modifications to the D_1 and V matrices. To conserve space, the implementation details of the GIV estimation and inference procedure are provided in Algorithm 2 in Online Appendix A.

Remark 3. *The method developed in this subsection constructs GIVs using data from a subset of n_0 entities. Once the GIVs are obtained, the full data set can be used to construct*

moment conditions for estimating the unknown parameters in the model. Since n_0 may be substantially smaller than both n and T , the proposed approach can be applied in settings where the cross-sectional dimension is large and may even exceed the sample size.

Remark 4. *The flexibility of using only a subset of entities to construct the GIVs also allows the framework to accommodate heterogeneous demand elasticities across entities, provided that a subset of entities is known to share a common elasticity.*

Specifically, suppose that the first n_0 entities share a common demand elasticity $\bar{\phi}$. We may use the demand equations for these entities, i.e., (67), to construct the GIVs $\hat{A}^\top y_t^0$, which can then be employed to form the moment functions

$$\bar{g}_T^{n_0}(\bar{\phi}; \hat{A}) \equiv T^{-1} \sum_{t \leq T} \hat{A}^\top y_t^0 (y_{n_0,t} - \bar{\phi} p_t) \quad \text{and} \quad \bar{g}_T^i(\phi_i; \hat{A}) \equiv T^{-1} \sum_{t \leq T} \hat{A}^\top y_t^0 (y_{i,t} - \phi_i p_t), \quad (70)$$

to estimate the common elasticity $\bar{\phi}$ and the entity-specific elasticities ϕ_i for $i = n_0 + 1, \dots, n$, where $y_{n_0,t} = n_0^{-1} \mathbf{1}_{n_0}^\top y_t^0$. Using arguments analogous to those in the proof of Theorem 1, it can be shown that the resulting GIV estimators of $\bar{\phi}$ and ϕ_i ($i = n_0 + 1, \dots, n$) are $T^{1/2}$ -consistent and asymptotically normal. Together with consistent estimators of their asymptotic variances, these results can be used to conduct inference on heterogeneous demand elasticities and to test hypotheses such as $H_0 : \bar{\phi} = \phi_i$ and $H_0 : \phi_i = \phi_{i'}$ for $i \neq i'$.

5 Simulation Studies

We examine the finite-sample performance of the proposed GIV estimation and inference procedures through Monte Carlo experiments. Subsection 5.1 describes the simulation design, and Subsection 5.2 reports the results.

5.1 Simulation Setting

We consider two simulation designs, a baseline design and an extended design, to investigate the finite-sample performance of the proposed GIV estimator, the associated inference procedures, and the specification test. The baseline design follows the model in (19)–(20), while the extended design augments the demand equation with three exogenous regressors.

Solving the demand-supply system in (60)–(61), with exogenous regressors in the demand equation but no additional exogenous variables in the supply equation, yields the following reduced-form expressions under the extended design:

$$y_t = \left(\mathbf{I}_n + \frac{\phi\psi}{1 - \phi\psi} \mathbf{1}_n S_t^\top \right) (x_t\beta + \lambda\eta_t + u_t) + \frac{\phi}{1 - \phi\psi} \mathbf{1}_n \varepsilon_t, \quad (71)$$

$$p_t = \frac{\psi}{1 - \phi\psi} S_t^\top (x_t\beta + \lambda\eta_t + u_t) + \frac{\varepsilon_t}{1 - \phi\psi}. \quad (72)$$

To generate the simulated data, we first draw the demand and supply shocks η_t , u_t , and ε_t , the exogenous regressors x_t , and the market share vector S_t conditional on the parameter values of ϕ , ψ , β , and λ . These simulated values are then substituted into (71)–(72) to obtain (p_t, y_t) . This procedure generates the simulated observations $\{y_t, p_t, x_t, S_t\}$ for each period t .

The demand and supply shocks are mutually independent and i.i.d. across t , with

$$\eta_t \sim N(\mathbf{0}_{r+1}, \Sigma_\eta), \quad u_t \sim N(\mathbf{0}_n, \mathbf{I}_n), \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad (73)$$

where $\Sigma_\eta \equiv ((0.1)^{|i-j|})_{i,j \leq r+1}$, and σ_ε^2 is set to 0.5. The demand and supply elasticities are set to $\phi = -0.5$ and $\psi = 1.5$, respectively. The exogenous regressors $\{x_{i,t}\}$ are generated i.i.d. from $N(\mathbf{0}_3, \mathbf{I}_3)$ across i and t , independently of $(\eta_t^\top, u_t^\top, \varepsilon_t)^\top$. The market share vector S_t is fixed across t and follows a Pareto rank-size specification: $s_i \propto (i/n)^{-1/\mu_S}$, with tail index $\mu_S = 0.2$, where the shares are normalized to satisfy $\sum_{i=1}^n s_i = 1$. Such a power-law profile is consistent with the size distributions documented for industries, firms, and financial

intermediaries (Gabaix, 2011; Gabaix and Koijen, 2024), and generates the concentrated cross-sectional structure under which granular variation is informative.

The factor loading matrix is specified as $\lambda = (\mathbf{1}_n, \lambda_{-1})$, where λ_{-1} is the $n \times r$ matrix of loadings on the r non-aggregate latent factors. To construct λ_{-1} , we first draw nr independent $N(0, 1)$ random variables to form an $n \times r$ matrix λ_0 . We then project λ_0 onto the orthogonal complement of $(\mathbf{1}_n, S)$ and obtain a preliminary loading matrix $\tilde{\lambda}_0$ through the QR decomposition $M_{(\mathbf{1}_n, S)}\lambda_0 = \tilde{\lambda}_0 R_{\lambda_0}$, where R_{λ_0} is an $r \times r$ upper triangular matrix and the columns of $\tilde{\lambda}_0$ are orthonormal. We then set $\lambda_{-1} = n^{1/2}\tilde{\lambda}_0$. By construction, the resulting loading matrix $\lambda = (\mathbf{1}_n, \lambda_{-1})$ has rank $\bar{r} = r + 1$.

We set $\beta = \mathbf{0}_{d_x}$ in (71)–(72) for both designs. In the baseline design, β is treated as known and the regressors x_t are omitted. In the extended design, however, β is estimated using the procedure described in Subsection 4.2. Because $\beta = \mathbf{0}_{d_x}$, the regressors x_t play no role in the data-generating process. Thus, any difference between the two designs reflects the additional estimation error associated with estimating β and partialling out x_t .

We consider six combinations of the number of entities and the number of non-aggregate latent factors:

$$(n, r) \in \{(5, 1), (5, 2), (8, 3), (8, 5), (10, 5), (10, 7)\}, \quad (74)$$

together with three sample sizes, $T \in \{150, 300, 450\}$. To evaluate the finite-sample performance of the GIV estimator, as well as the size properties of the associated inference and specification tests, we conduct 10,000 Monte Carlo replications for each (n, r, T) cell under each simulation design.

To assess the effect of estimation error arising from the recovery of the subspace orthogonal to $\text{col}((\mathbf{1}_n, \lambda))$ on the performance of the GIV estimator, we consider an oracle GIV estimator constructed under the assumption that $\text{col}((\mathbf{1}_n, \lambda))$ is known. Consequently, the matrix A , whose columns form a basis for the orthogonal complement of $\text{col}((\mathbf{1}_n, \lambda))$ and are used to construct the GIVs, is treated as known.²⁰ The oracle GIV estimator therefore

²⁰In the simulation, A is constructed from the left singular vectors associated with the zero singular values

bypasses both the BIC step for estimating the number of factors and the estimation of the orthogonal complement of $\text{col}((\mathbf{1}_n, \lambda))$. In the extended design, the oracle estimator additionally treats β as known and partials out $x_t\beta$ using the true parameter value. By contrast, the feasible GIV estimator selects $\bar{r}$ using the BIC criterion in (47)–(48), and then constructs the GIVs and the corresponding GIV estimator according to Algorithm 1 of Online Appendix A.

Both the oracle and feasible GIV estimators are evaluated using their finite-sample root mean squared errors (RMSEs), with the results reported in Table 1 of the next subsection. We also investigate the empirical rejection probabilities of the two-sided tests of $H_0 : \phi = -0.5$ and $H_0 : \psi = 1.5$ at the 5% significance level. The results are reported in Table 2 of the next subsection. Inference is conducted using t -tests based on the standard error estimators described in Algorithm 1 of Online Appendix A, together with the asymptotic normality of the GIV estimators.

To evaluate the power of the J -test, we consider a controlled violation of the covariance restrictions underlying GIV validity while keeping the factor-loading matrix $\lambda = (\mathbf{1}_n, \lambda_{-1})$, the structural parameters, and the marginal distributions of u_t , η_t , and ε_t unchanged. Let $b_n = Ad_n$, where $d_n \in \mathbb{R}^{n-r-1}$ is the unit vector obtained by applying the Gram–Schmidt procedure to the first standard basis vector against $A^\top S$. For $\rho \in [0, 1)$, we generate u_t as

$$u_t = u_t^* + \frac{\rho}{\sigma_\varepsilon} \varepsilon_t b_n - \left(1 - (1 - \rho^2)^{1/2}\right) (b_n^\top u_t^*) b_n,$$

where $u_t^* \sim N(\mathbf{0}_n, \mathbf{I}_n)$ is generated together with ε_t and η_t in the same way as u_t in (73).

By construction, $b_n \in \text{col}(A)$ implies $b_n^\top \mathbf{1}_n = 0$ and $b_n^\top \lambda_{-1} = \mathbf{0}_r^\top$, so b_n is orthogonal to $\text{col}(\lambda)$. It is straightforward to verify that for any $\rho \in [0, 1)$, the joint distribution of (u_t, ε_t) remains Gaussian with

$$\text{Var}(u_t) = \mathbf{I}_n \quad \text{and} \quad \text{Cov}(u_t, \varepsilon_t) = \rho \sigma_\varepsilon b_n.$$

in the singular value decomposition of $(\mathbf{1}_n, \lambda)$.

When $\rho = 0$, the design reduces to the correctly specified benchmark design used above to study the finite-sample properties of the GIV estimators and inference procedures. When $\rho > 0$, however, u_t and ε_t become correlated, with heterogeneous correlation patterns determined by b_n . As a result, the orthogonality conditions underlying the GIVs are violated, rendering the GIVs invalid. Moreover, the degree of misspecification increases linearly with ρ .

We use the six (n, r) combinations in (74) together with the three sample sizes T considered above, vary ρ over the grid $\rho_j = 0.02j$ for $j = 0, \dots, 20$, and conduct 10,000 simulation replications for each (n, r, T, ρ) cell under each design. The significance level of the J -test is set at 0.05, and the resulting empirical rejection probabilities are reported in Figure 1 of the next subsection.

5.2 Simulation Results

Table 1 shows that the RMSEs of both the oracle and feasible estimators of ϕ and ψ decline substantially as the sample size T increases. For example, in the baseline design with $(n, r) = (5, 1)$, the RMSE of the oracle estimator of ϕ decreases from 0.129 at $T = 150$, to 0.088 at $T = 300$, and further to 0.072 at $T = 450$. Similar improvements are observed across all configurations and for both structural parameters.

The feasible GIV estimator closely tracks the oracle estimator throughout the simulation designs. In most cases, the difference in RMSE between the oracle and feasible estimators is negligible. For instance, in the baseline design with $(n, r) = (10, 5)$, the RMSEs of the feasible estimator $\hat{\phi}^*(\hat{A})$ are 0.115, 0.080, and 0.064 at $T = 150, 300$, and 450, respectively, compared with the corresponding oracle RMSEs of 0.113, 0.079, and 0.064. Similar qualitative patterns are observed for the GIV estimator of ψ , as well as for both estimators in the extended design with exogenous regressors.

We next examine the size properties of the t -tests for $H_0 : \phi = -0.5$ and $H_0 : \psi = 1.5$ at the 5% significance level. The results, reported in Table 2, show that the empirical rejection probabilities approach the nominal level as the sample size T increases. When the

Table 1: Root Mean Squared Error of the GIV Estimators

(n, r)	T	Baseline ($d_x = 0$)				Extended ($d_x = 3$)			
		Oracle		Feasible		Oracle		Feasible	
		ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ
(5, 1)	150	0.129	0.086	0.130	0.087	0.126	0.085	0.131	0.091
	300	0.088	0.060	0.092	0.065	0.088	0.060	0.089	0.065
	450	0.072	0.049	0.076	0.051	0.071	0.049	0.074	0.050
(5, 2)	150	0.128	0.086	0.150	0.109	0.129	0.086	0.160	0.116
	300	0.089	0.060	0.106	0.078	0.088	0.061	0.103	0.083
	450	0.071	0.049	0.094	0.072	0.071	0.049	0.095	0.071
(8, 3)	150	0.118	0.082	0.119	0.084	0.118	0.082	0.121	0.084
	300	0.081	0.057	0.081	0.058	0.080	0.058	0.081	0.059
	450	0.066	0.047	0.067	0.048	0.066	0.046	0.067	0.052
(8, 5)	150	0.118	0.082	0.142	0.115	0.117	0.082	0.152	0.115
	300	0.080	0.057	0.100	0.070	0.081	0.057	0.111	0.081
	450	0.066	0.046	0.081	0.062	0.066	0.046	0.097	0.080
(10, 5)	150	0.113	0.082	0.115	0.084	0.115	0.081	0.118	0.083
	300	0.079	0.057	0.080	0.057	0.080	0.057	0.081	0.058
	450	0.064	0.046	0.064	0.046	0.064	0.045	0.064	0.046
(10, 7)	150	0.113	0.080	0.145	0.113	0.113	0.081	0.142	0.110
	300	0.079	0.056	0.102	0.083	0.080	0.056	0.107	0.093
	450	0.064	0.045	0.081	0.058	0.064	0.045	0.081	0.059

Notes. This table reports the root mean squared errors of the GIV estimators based on 10,000 Monte Carlo replications, separately for the oracle and feasible estimators, under the baseline design ($d_x = 0$) and the extended design with $d_x = 3$ exogenous regressors. The oracle estimator treats λ (and β when present) as known. The feasible estimator selects $\bar{r}$ using the BIC criterion in (47)–(48) and is computed according to Algorithm 1 in Online Appendix A. Both estimators are evaluated using the same simulated data within each replication. To reduce the influence of extreme draws, the RMSE is computed using the truncated loss function $\min\{5, (\hat{\alpha} - \alpha)^2\}$ for a generic estimator $\hat{\alpha}$ of the parameter α .

sample size is relatively small, i.e., $T = 150$, the tests exhibit modest over-rejection, with the distortion becoming more pronounced in configurations involving a larger number of GIVs. For example, in the baseline design with $(n, r) = (8, 3)$, where there are four GIVs and eight moment conditions, the empirical rejection probabilities of the t -tests based on the feasible GIV estimators for ϕ and ψ are 0.104 and 0.073, respectively, at $T = 150$. By contrast, when $(n, r) = (8, 5)$, where only two GIVs are available, the corresponding rejection probabilities are 0.075 and 0.061, respectively. Similar patterns are observed in the extended design.

The over-rejection observed in these t -tests does not appear to be primarily driven by estimation error in the number of factors or in the null space of $(\mathbf{1}_n, \lambda)$, since the tests based on the oracle GIV estimators, which do not require estimation of these nuisance parameters, display similar finite-sample behavior. Instead, the over-rejection is likely related to the well-known many-moment bias in two-step GMM estimation, of which the GIV estimator is a special case; see, for example, Hansen, Heaton, and Yaron (1996) and Newey and Windmeijer (2009). Several approaches may help mitigate the resulting size distortion in small samples. For example, instead of estimating ϕ and ψ jointly, one may estimate them using two separate GMM procedures. This reduces the number of moment conditions used in each estimation problem, although it may sacrifice some of the efficiency gains from joint GMM estimation. Another possibility is to employ the continuously updated GMM estimator rather than the two-step GMM estimator. While this approach may improve finite-sample inference, it also introduces additional computational burden, since the continuously updated GMM estimator does not admit a closed-form solution.

Finally, we examine the performance of the J -test for assessing the validity of the moment conditions constructed using the GIVs. Table 3 reports the empirical rejection probabilities at the 5% nominal significance level under correct specification. The size behavior is broadly similar to that of the t -tests reported in Table 2: the empirical rejection probabilities approach the nominal level as the sample size T increases, while modest over-rejection is observed in small samples, particularly when the number of moment conditions is relatively large. For example, in the baseline design with $(n, r) = (10, 5)$, the empirical rejection probabilities of the oracle and feasible J -tests are 0.069 and 0.080, respectively, at $T = 150$. In the corresponding extended design, the rejection probabilities are 0.074 and 0.083, respectively. Similar patterns are observed for $(n, r) = (8, 3)$, where the feasible rejection probabilities are 0.082 in both the baseline and extended designs at $T = 150$. As the sample size increases, the empirical rejection probabilities move steadily toward the nominal 5% level across all configurations.

Table 2: Empirical Rejection Probabilities of the Two-sided t -tests

(n, r)	T	Baseline ($d_x = 0$)				Extended ($d_x = 3$)			
		Oracle		Feasible		Oracle		Feasible	
		ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ
(5, 1)	150	0.091	0.063	0.094	0.062	0.089	0.058	0.090	0.061
	300	0.068	0.058	0.067	0.059	0.071	0.058	0.071	0.059
	450	0.063	0.052	0.065	0.053	0.058	0.056	0.059	0.058
(5, 2)	150	0.071	0.058	0.072	0.058	0.073	0.057	0.076	0.059
	300	0.062	0.052	0.064	0.053	0.060	0.057	0.062	0.056
	450	0.056	0.052	0.056	0.052	0.058	0.056	0.057	0.055
(8, 3)	150	0.101	0.069	0.104	0.073	0.104	0.070	0.113	0.073
	300	0.076	0.059	0.078	0.060	0.073	0.062	0.076	0.064
	450	0.068	0.058	0.068	0.059	0.067	0.055	0.071	0.056
(8, 5)	150	0.070	0.058	0.075	0.061	0.069	0.059	0.075	0.061
	300	0.058	0.058	0.058	0.056	0.057	0.054	0.059	0.053
	450	0.056	0.052	0.059	0.054	0.055	0.051	0.057	0.051
(10, 5)	150	0.096	0.072	0.102	0.077	0.101	0.067	0.107	0.073
	300	0.078	0.059	0.079	0.059	0.077	0.061	0.082	0.063
	450	0.068	0.058	0.068	0.057	0.068	0.053	0.066	0.054
(10, 7)	150	0.069	0.057	0.073	0.064	0.068	0.057	0.075	0.062
	300	0.060	0.057	0.061	0.059	0.061	0.051	0.062	0.053
	450	0.057	0.053	0.058	0.056	0.055	0.053	0.057	0.054

Notes. This table reports the empirical rejection rates of the two-sided t -tests of $H_0 : \phi = -0.5$ and $H_0 : \psi = 1.5$ at the 5% nominal significance level, using the asymptotic critical value 1.96, based on 10,000 Monte Carlo replications. Results are reported separately for the oracle and feasible estimators under both the baseline design ($d_x = 0$) and the extended design with $d_x = 3$ exogenous regressors. The oracle estimator treats λ (and β when present) as known. The feasible estimator selects $\bar{r}$ using the BIC criterion in (47)–(48) and is implemented according to Algorithm 1 in Online Appendix A.

As discussed earlier for the t -tests, the observed small-sample over-rejection does not appear to be primarily driven by estimation error in the number of factors or in the null space of $(\mathbf{1}_n, \lambda)$, since the oracle and feasible procedures display very similar finite-sample behavior. Instead, the distortion is likely related to the many-moment nature of the GMM problem.

We next examine the power of the feasible J -test under misspecification. Figure 1 plots the empirical rejection probabilities of the J -test as a function of ρ , which controls the severity

Table 3: Empirical Rejection Probabilities of the J -test

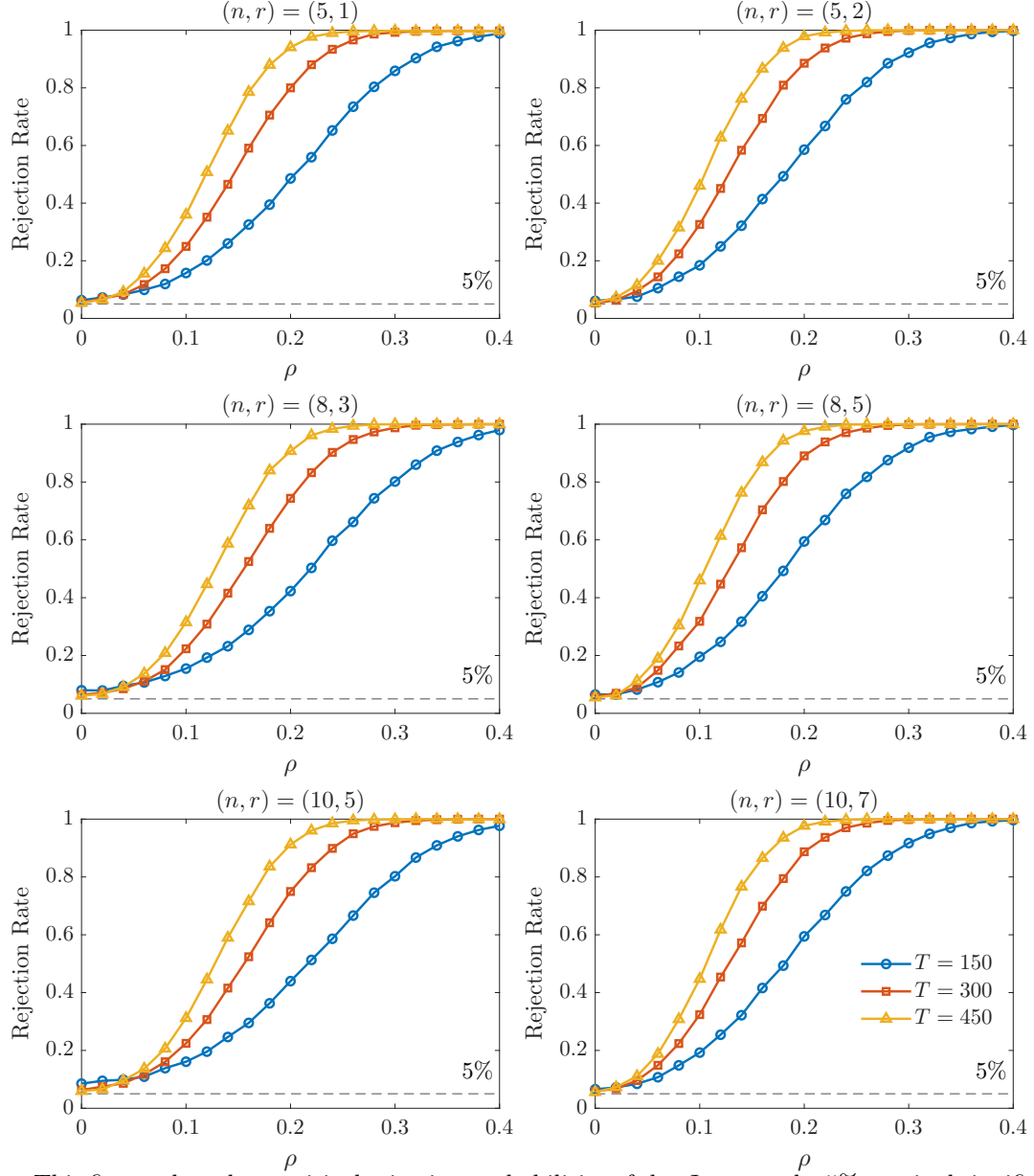
(n, r)	T	Baseline ($d_x = 0$)		Extended ($d_x = 3$)	
		Oracle	Feasible	Oracle	Feasible
(5, 1)	150	0.067	0.071	0.068	0.068
	300	0.060	0.062	0.058	0.059
	450	0.053	0.052	0.054	0.055
(5, 2)	150	0.060	0.061	0.059	0.061
	300	0.056	0.058	0.052	0.055
	450	0.056	0.057	0.058	0.057
(8, 3)	150	0.074	0.082	0.073	0.082
	300	0.064	0.064	0.062	0.067
	450	0.058	0.058	0.060	0.061
(8, 5)	150	0.060	0.064	0.055	0.063
	300	0.057	0.058	0.055	0.058
	450	0.050	0.050	0.053	0.057
(10, 5)	150	0.069	0.080	0.074	0.083
	300	0.055	0.060	0.064	0.069
	450	0.056	0.061	0.053	0.058
(10, 7)	150	0.060	0.065	0.055	0.063
	300	0.054	0.058	0.052	0.057
	450	0.054	0.056	0.053	0.054

Notes. This table reports the empirical rejection rates of the J -test at the 5% nominal significance level based on 10,000 Monte Carlo replications. Results are reported separately for the oracle and feasible estimators under both the baseline design ($d_x = 0$) and the extended design with $d_x = 3$ exogenous regressors. For the oracle estimator, the reference distribution is $\chi^2_{2(n-r-2)}$. For the feasible estimator, the reference distribution is the replication-specific $\chi^2_{2(n-\hat{r}-1)}$. The oracle estimator treats λ (and β when present) as known. The feasible estimator selects $\bar{r}$ using the BIC criterion in (47)–(48) and is implemented according to Algorithm 1 in Online Appendix A.

of the violation of the GIV moment conditions in the baseline design.²¹ By construction, each power curve begins near the nominal 5% level when $\rho = 0$. The rejection probabilities then increase monotonically with ρ , and the power curves become substantially steeper as the sample size T increases. For example, in the baseline design with $(n, r) = (10, 7)$, the test achieves approximately 80% power at $\rho \approx 0.29$ when $T = 150$, $\rho \approx 0.22$ when $T = 300$, and $\rho \approx 0.16$ when $T = 450$. Similar patterns are observed across all configurations. The feasible

²¹The corresponding results for the extended design are reported in Online Appendix F and exhibit similar qualitative patterns.

Figure 1: Empirical Power of the J -test: Baseline Design



Notes. This figure plots the empirical rejection probabilities of the J -test at the 5% nominal significance level in the baseline design ($d_x = 0$), based on 10,000 Monte Carlo replications. Each panel corresponds to one of the six configurations (n, r) listed in (74), and within each panel three curves correspond to the sample sizes $T \in \{150, 300, 450\}$. The horizontal axis is ρ and the vertical axis is the rejection rate. The horizontal dashed line marks the nominal 5% level. Power is reported for the feasible GIV estimator, which selects $\bar{r}$ using the BIC criterion in (47)–(48) and is implemented according to Algorithm 1 in Online Appendix A.

J -test exhibits good power in detecting moderate violations of the moment restrictions at empirically relevant sample sizes.

6 Empirical Application: Aggregate Market Multiplier

A central question in asset pricing is how strongly the aggregate stock market responds to shifts in investor demand for equities. This response is summarized by the aggregate market multiplier, denoted by $\kappa \equiv -\phi^{-1}$, where ϕ is the aggregate demand elasticity. Economically, κ measures the change in aggregate equity value induced by a one-dollar demand shock. The aggregate demand elasticity has become a central object of interest in asset pricing and macro-finance because investor demand, portfolio reallocation, and market segmentation can have important effects on equilibrium asset prices; see, among others, Piazzesi and Schneider (2007), Kojen and Yogo (2019), and Gabaix and Kojen (2023). In frictionless benchmark models, aggregate demand is highly elastic, implying a multiplier close to zero. By contrast, the inelastic-markets hypothesis predicts a substantially larger value of κ .

The size of the aggregate multiplier remains controversial. Standard asset-pricing models imply a macro elasticity of roughly 10 to 20, corresponding to a multiplier of only 0.05 to 0.1.²² Yet empirical estimates of stock-level, factor-level, and aggregate demand elasticities generally point to substantially less elastic demand. For example, Lou (2012) estimates a stock-level multiplier of about 1.2, Pavlova and Sikorskaya (2023) report multipliers between 0.3 and 0.5, and Gabaix and Kojen (2023) report substantially larger multipliers at the aggregate level. These findings are consistent with the economic intuition that aggregate equity demand should be less elastic than demand for individual stocks, since stocks are closer substitutes for one another than for alternative asset classes such as bonds. At the same time, the estimated multipliers are an order of magnitude larger than those implied by standard asset-pricing models, posing a challenge for conventional theories of asset demand. Resolving this discrepancy requires credible identification of the aggregate demand elasticity, a key parameter for quantifying the effects of capital flows, institutional demand shocks, and policy interventions on asset prices.

²²See Appendices F and I of Gabaix and Kojen (2023) for computations of the macro elasticity implied by the model of Lucas (1978), the rare-disaster models of Barro (2006) and Gabaix (2012), and the long-run risks model of Bansal and Yaron (2004).

Obtaining such identification is challenging because prices and quantities are jointly determined in equilibrium. Demand shocks affect prices, while prices simultaneously enter investors' demand equations, rendering simple regressions of demand on prices generally inconsistent. Building on the demand-based asset-pricing framework of Koijen and Yogo (2019), Gabaix and Koijen (2023) address this endogeneity problem by exploiting the latent-factor structure in investor demand and estimate aggregate multipliers ranging from 4.73 to 5.85, with a median estimate close to five. Their approach, however, relies on consistent estimation of the latent demand factors and therefore on asymptotic arguments in which the cross-sectional dimension diverges; see, for example, Bai (2003). In this section, we revisit the aggregate multiplier using the GIV framework developed in Section 3, which permits valid estimation and inference without requiring the number of sectors to grow with the sample size.

Following Gabaix and Koijen (2023), we study the aggregate multiplier through a demand system for U.S. equity holdings. Investors are grouped into n equity-holding sectors. Let $\Delta q_{i,t}$ denote the fractional quarterly change in investor i 's equity holdings at quarter t , with its empirical counterpart defined in (78) below; see Online Appendix G.1 for construction details. The demand equation is given by

$$\Delta q_{i,t} = \phi \Delta p_t + \lambda_i^\top \eta_t + u_{i,t}, \quad (75)$$

where Δp_t denotes the quarterly equity market return, $\eta_t \in \mathbb{R}^r$ represents latent aggregate demand factors, and $u_{i,t}$ is an idiosyncratic demand shock. The parameter ϕ captures the aggregate demand elasticity and is the primary object of interest.²³

Because the aggregate supply of equity is approximately fixed in the short run, market

²³Unlike Gabaix and Koijen (2023), we do not explicitly include aggregate macroeconomic variables, such as GDP growth, in (75). Aggregate variables that enter the demand equation with homogeneous loadings across sectors are eliminated by the orthogonality condition defining the GIVs, since the instruments are constructed to be orthogonal to $\mathbf{1}_n$. Aggregate variables with heterogeneous loadings are absorbed into the latent factor component $\lambda \eta_t$. Under the maintained assumption that the idiosyncratic shocks $u_{i,t}$ are orthogonal to these aggregate components, their omission does not affect identification, estimation, or inference for the demand elasticity ϕ and the aggregate multiplier κ .

clearing implies that size-weighted net demand equals zero:

$$q_{S,t} \equiv S_t^\top \Delta q_t = 0, \quad (76)$$

where $\Delta q_t = (\Delta q_{i,t})_{i \leq n}$ and S_t denotes the vector of predetermined market shares. Combining (75) and (76) yields the equilibrium price equation

$$\Delta p_t = \kappa \left(S_t^\top \lambda \eta_t + u_{S,t} \right), \quad (77)$$

where $u_{S,t} \equiv S_t^\top u_t$. Thus, the aggregate multiplier κ measures the equilibrium price response to aggregate demand shocks, with a less elastic demand (smaller $|\phi|$) corresponding to a larger multiplier.

Our data are drawn from the Financial Accounts of the United States, Table L.224, which reports the equity holdings of major investor sectors.²⁴ Following Gabaix and Koijen (2023), our benchmark analysis uses the sample period 1993Q1–2018Q4 and includes twelve sectors that hold U.S. equities continuously throughout this period. Table G.1 in Online Appendix G lists these sectors together with their average market shares. In the data, the fractional change in sector i 's equity holdings is measured as

$$\Delta q_{i,t} \equiv \frac{w_{i,t}}{w_{i,t-1} R_t} - 1, \quad (78)$$

where $w_{i,t}$ denotes the value of sector i 's equity holdings, R_t is the gross capital-appreciation return on the aggregate stock market, and Δp_t is measured by the quarterly simple return on the CRSP value-weighted index excluding dividends.²⁵ Beyond the benchmark sample, we consider an extended sample spanning 1988Q4–2025Q4, which is the longest period over

²⁴We use the June 2026 vintage of the Financial Accounts, in which corporate equity holdings by sector are reported in Table L.224.

²⁵Online Appendix G.1 details the construction of $\Delta q_{i,t}$ and Δp_t , and Online Appendix G.2 describes the sector classification and market-share weights S_t . Following Gabaix and Koijen (2023), pooled sector-level demand growth is winsorized at the 5th and 95th percentiles.

Table 4: Benchmark Estimates of the Aggregate Equity Market Multiplier

	1993Q1 – 2018Q4		1988Q4 – 2025Q4	
	$n = 12$	$n = 6$	$n = 12$	$n = 6$
OLS	5.42*** (0.43)	10.05*** (0.78)	5.66*** (0.52)	11.42*** (1.11)
FIV	4.42*** (0.87)	−7.13** (2.83)	4.39*** (1.08)	−2.15 (2.76)
GIV	5.05*** (0.30)	8.70*** (0.82)	4.46*** (0.30)	9.42*** (0.93)
J -test (p -value)	< 0.001	0.817	< 0.001	0.594
Estimated $\bar{r}$	1	1	1	1
T	104	104	149	149

Notes. The table reports estimates of the aggregate equity market multiplier, $\hat{\kappa} = -1/\hat{\phi}$, for the full twelve-sector panel ($n = 12$) and the six-sector granular core ($n = 6$). OLS is obtained from a regression of equally weighted demand, $n^{-1} \sum_{i \leq n} \Delta q_{i,t}$, on the market return Δp_t . FIV denotes the factor-residual instrumental-variable estimator of Gabaix and Koijen (2023), implemented using Algorithm 3 in Online Appendix G.3 with two latent factors and four observed factors (GDP growth, size, value, and momentum). GIV denotes the demand-only granular-IV estimator based on the moment conditions in (46). The reported J -test p -value corresponds to the over-identification test, whose asymptotic null distribution is $\chi^2_{n-\bar{r}-1}$, where $\bar{r}$ is selected using the BIC criterion in (47)–(48). Standard errors are reported in parentheses and are computed using the Newey–West estimator. For the OLS and GIV estimators, $\hat{\kappa}$ is obtained from $\hat{\phi}$ through the transformation $\hat{\kappa} = -1/\hat{\phi}$ and the corresponding standard errors are computed using the delta method. Significance levels correspond to two-sided tests of $H_0 : \kappa = 0$. Significance levels: *** 1%, ** 5%, * 10%.

which all twelve sectors are continuously observed.

For comparison, we also report estimates from the factor-residual IV (FIV) estimator of Gabaix and Koijen (2023), together with the OLS estimator. The FIV estimator constructs instruments from estimated idiosyncratic demand shocks obtained after removing observed and latent demand factors.²⁶ The OLS estimator is obtained from a regression of equally weighted demand, $n^{-1} \sum_{i \leq n} \Delta q_{i,t}$, on the market return Δp_t . Table 4 reports the resulting estimates and standard errors.

Across all specifications, the OLS estimate of the aggregate multiplier exceeds its FIV and GIV counterparts. This pattern is consistent with the endogeneity problem discussed earlier. Positive demand shocks increase both holdings and prices, causing an uninstrumented regression to attribute part of the demand shock to the price response. As a result,

²⁶Online Appendix G.3 provides implementation details for the FIV estimator.

the demand elasticity is biased toward zero and the implied multiplier is biased upward. By exploiting granular demand variation that is orthogonal to common demand factors, the GIV estimator corrects this source of bias.

When all twelve sectors are used to construct the GIVs, the estimated aggregate multiplier is 5.05 in the benchmark sample 1993Q1–2018Q4 and 4.46 in the extended sample 1988Q4–2025Q4, close to the corresponding FIV estimates. However, the over-identification test strongly rejects the associated moment restrictions in both samples, with p -values effectively equal to zero. One traditional interpretation of this result is that the instruments are invalid. Alternatively, following a perspective common in the treatment-effects literature, rejection of the over-identification test may reflect heterogeneity in the underlying causal parameters. From this perspective, the evidence suggests that the homogeneous-elasticity specification imposed on all twelve sectors is too restrictive and that demand elasticities may differ substantially across investor sectors.

To investigate this possibility, we restrict attention to the six largest sectors, which together account for more than 97% of total equity holdings in the sample: households, mutual funds and ETFs, the foreign sector, private pension funds, state and local pension funds, and life insurance companies. Relative to the twelve-sector specification, both the OLS and GIV estimates increase substantially. For example, the GIV estimate rises from 5.05 to 8.70 in the benchmark sample and from 4.46 to 9.42 in the extended sample. More importantly, the over-identification test no longer rejects, yielding p -values of 0.817 and 0.594 in the two samples. These findings suggest that the six largest sectors exhibit more homogeneous demand behavior and therefore provide a more credible basis for estimating a common aggregate demand elasticity and the corresponding market multiplier.

The comparison between the FIV and GIV estimators highlights the importance of the fixed- n approach. When all twelve sectors are included, the two estimators deliver qualitatively similar conclusions. The FIV estimates of the aggregate multiplier are 4.42 in the benchmark sample and 4.39 in the extended sample, close to the corresponding GIV estimates

Table 5: Heterogeneous Demand Multipliers by Sector

Sector	1993Q1–2018Q4		1988Q4–2025Q4	
	$\hat{\kappa}$	J -test (p -value)	$\hat{\kappa}$	J -test (p -value)
Granular core ($n = 6$)	8.70*** (0.82)	0.817	9.42*** (0.93)	0.594
Property and casualty insurers	3.57*** (0.31)	0.094	3.96*** (0.38)	0.927
Federal government retirement funds	−18.15 (9.06)	0.010	−17.32 (12.99)	< 0.001
State and local governments	3.70*** (0.45)	0.419	3.65*** (0.33)	0.202
Closed-end funds	6.63*** (2.08)	0.410	6.52*** (1.71)	0.356
Banks	6.68*** (1.70)	0.347	10.68** (4.29)	0.786
Broker-dealers	−40.97 (90.85)	0.064	−251.05 (3058.61)	0.011

Notes. The table reports sector-specific demand multipliers, $\hat{\kappa}_j = -1/\hat{\phi}_j$, estimated using the GIV moment conditions in (70). The GIVs are constructed from the six largest equity-holding sectors (the granular core), which are assumed to share a common demand elasticity. For each sample period, the first row reproduces the corresponding six-sector estimate from Table 4. Standard errors are reported in parentheses and are computed using the Newey–West estimator. The reported p -values correspond to the over-identification J -test. Rejection of the J -test indicates that the moment conditions constructed from the granular core are invalid for the corresponding sector. Standard errors for $\hat{\kappa}_j$ are computed using the delta method, and significance levels correspond to two-sided tests of $H_0 : \kappa_j = 0$. For sectors whose J -test rejects at the 5% level, point estimates and standard errors are reported for completeness only, and statistical significance is not indicated. Significance levels: *** 1%, ** 5%, * 10%.

of 5.05 and 4.46, respectively, and broadly consistent with the estimates reported by Gabaix and Koijen (2023). Both instrumental-variable estimators also yield smaller multipliers than OLS, as expected from the endogeneity bias discussed above.

The contrast becomes much sharper when attention is restricted to the six-sector granular core. In this case, the FIV estimator produces negative multiplier estimates of −7.13 and −2.15, whereas the GIV estimator yields stable and economically meaningful estimates of 8.70 and 9.42. This divergence reflects the different identification strategies underlying the two procedures. The FIV estimator constructs instruments from estimated idiosyncratic

demand shocks obtained after removing latent factors through principal-components analysis and therefore relies on consistent estimation of those factors. With only six sectors, the cross-sectional dimension is too small for this approach to be reliable. By contrast, the fixed- n GIV estimator does not require consistent estimation of latent factors and remains valid when the number of sectors is small. The resulting GIV estimates are therefore more credible for the six-sector specification, which is favored by the over-identification test and appears consistent with the homogeneous-elasticity restriction underlying the aggregate multiplier.

Table 5 sheds light on the source of the rejection of the twelve-sector specification. Using the GIVs constructed from the six-sector core, we estimate sector-specific demand elasticities, ϕ_j , and the corresponding implied market multipliers, $\kappa_j \equiv -\phi_j^{-1}$, for the remaining sectors. Several sectors exhibit substantially smaller multipliers than the six-sector core. Property and casualty insurers and state and local governments have multipliers between 3.5 and 4.0, less than half the core estimate. Closed-end funds have intermediate multipliers of roughly 6.5, while banks exhibit larger and less precisely estimated multipliers. For all of these sectors, the J -test provides little evidence against the validity of the moment conditions constructed from the six-sector core.

The two remaining sectors, federal government retirement funds and broker-dealers, exhibit markedly different behavior. Their estimated multipliers are negative in both samples, implying non-positive estimates of the corresponding demand elasticities. Moreover, these estimates are highly imprecise and economically difficult to interpret. More importantly, the J -test rejects the validity of the moment conditions for federal government retirement funds in both samples. For broker-dealers, the J -test is close to rejection at the 5% level in the shorter sample (p -value = 0.064) and is strongly rejected in the longer sample (p -value = 0.011). Taken together, these findings suggest that the moment conditions constructed from the six-sector core do not provide valid identifying restrictions for these two sectors. Consequently, the corresponding multiplier estimates should not be given an economic interpretation.

Taken together, Tables 4 and 5 provide strong evidence that demand elasticities differ substantially across investor sectors. Among the ten sectors for which the J -test does not reject (treating broker-dealers as invalid given the decisive rejection in the longer sample and the borderline p -value of 0.064 in the shorter sample), the estimated multipliers range from approximately 3.6 to 10.7, with the upper end reflecting the imprecisely estimated bank sector. This heterogeneity explains both the rejection of the twelve-sector specification and the substantially smaller multiplier obtained when all sectors are pooled together. By contrast, the six-sector core appears considerably more homogeneous and yields a stable aggregate multiplier of roughly nine across both sample periods. This estimate is an order of magnitude larger than the frictionless benchmark multiplier of 0.05 to 0.1, providing evidence consistent with the inelastic-markets hypothesis.

7 Conclusion

This paper develops an estimation and inference framework for structural models identified by granular instrumental variables (GIVs). The key insight is that, under suitable cross-sectional restrictions on the idiosyncratic shocks, valid GIVs are characterized by the orthogonal complement of the factor-loading space associated with latent aggregate shocks. This characterization provides a transparent foundation for GIV-based identification and allows demand and supply elasticities to be estimated without conventional excluded instruments or direct observation of the latent factors.

A central contribution of the paper is to show that the relevant orthogonal complement can be identified and consistently estimated directly from the covariance structure of the observables, without first estimating the latent factors themselves. As a result, the proposed framework remains applicable even when the number of entities is fixed and does not require the cross-sectional dimension to diverge with the sample size. Building on this result, we develop feasible procedures for estimation, inference, and specification testing based on esti-

mated GIVs. Monte Carlo evidence shows that the feasible estimator performs similarly to an oracle procedure that knows the true factor-loading space, while the empirical application yields evidence consistent with highly inelastic aggregate equity demand.

The analysis also clarifies several identification issues that arise when factor loadings are unknown. In particular, we show that certain restrictions commonly imposed in existing implementations of the GIV methodology should be interpreted as substantive assumptions on the latent factors rather than innocuous normalizations. We further show that identification can fail for existing moment-condition approaches when the factor-loading space is unknown. The framework developed here avoids these difficulties by focusing directly on the geometry of the factor-loading space and exploiting the eigenspace structure of the covariance matrix of the observables.

More broadly, the results demonstrate that cross-sectional heterogeneity can be used not only to construct granular instruments but also to conduct valid estimation, inference, and specification testing in models with latent aggregate shocks. We hope that the framework developed in this paper will facilitate the use of GIV methods in empirical work and provide a foundation for future research in settings where latent aggregate forces and granular heterogeneity interact.

References

- ACEMOGLU, D., V. M. CARVALHO, A. OZDAGLAR, AND A. TAHBAZ-SALEHI (2012): “The network origins of aggregate fluctuations,” *Econometrica*, 80(5), 1977–2016.
- ACKERBERG, D., X. CHEN, J. HAHN, AND Z. LIAO (2014): “Asymptotic efficiency of semiparametric two-step GMM,” *Review of Economic Studies*, 81(3), 919–943.
- AMITI, M., O. ITSXHOKI, AND J. KONINGS (2019): “International shocks, variable markups, and domestic prices,” *The Review of Economic Studies*, 86(6), 2356–2402.

- AMITI, M., AND D. E. WEINSTEIN (2018): “How much do idiosyncratic bank shocks affect investment? Evidence from matched bank-firm loan data,” *Journal of Political Economy*, 126(2), 525–587.
- BAI, J. (2003): “Inferential theory for factor models of large dimensions,” *Econometrica*, 71(1), 135–171.
- BANAFTI, S., AND T.-H. LEE (2022): “Inferential theory for granular instrumental variables in high dimensions,” *arXiv preprint arXiv:2201.06605*.
- BANSAL, R., AND A. YARON (2004): “Risks for the long run: A potential resolution of asset pricing puzzles,” *The Journal of Finance*, 59(4), 1481–1509.
- BAQAEE, D. R., AND E. FARHI (2019): “The macroeconomic impact of microeconomic shocks: Beyond Hulten’s theorem,” *Econometrica*, 87(4), 1155–1203.
- BARRO, R. J. (2006): “Rare disasters and asset markets in the twentieth century,” *The Quarterly Journal of Economics*, 121(3), 823–866.
- BAUMEISTER, C., AND J. D. HAMILTON (2023): “Uncovering disaggregated oil market dynamics: A full-information approach to granular instrumental variables,” *Working paper, UCSD*.
- CAMANHO, N., H. HAU, AND H. REY (2022): “Global portfolio rebalancing and exchange rates,” *The Review of Financial Studies*, 35(11), 5228–5274.
- DI GIOVANNI, J., A. A. LEVCHENKO, AND I. MEJEAN (2014): “Firms, destinations, and aggregate fluctuations,” *Econometrica*, 82(4), 1303–1340.
- DONG, X., N. KANG, AND J. PERESS (2025): “Fast and slow arbitrage: The predictive power of persistent capital flows for factor returns,” *The Review of Financial Studies*, 38(9), 2936–2987.
- GABAIX, X. (2011): “The granular origins of aggregate fluctuations,” *Econometrica*, 79(3), 733–772.
- (2012): “Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance,” *The Quarterly Journal of Economics*, 127(2), 645–700.

- GABAIX, X., AND R. S. KOIJEN (2023): “In search of the origins of financial fluctuations: The inelastic markets hypothesis,” SSRN Working Paper No. 3686935.
- (2024): “Granular instrumental variables,” *Journal of Political Economy*, 132(7), 2274–2303.
- GALAASEN, S., R. JAMILOV, R. JUELSRUD, AND H. REY (2020): “Granular credit risk,” Discussion paper, National Bureau of Economic Research.
- GAUBERT, C., AND O. ITSKHOKI (2021): “Granular comparative advantage,” *Journal of Political Economy*, 129(3), 871–939.
- HAN, C., AND P. C. PHILLIPS (2006): “GMM with many moment conditions,” *Econometrica*, 74(1), 147–192.
- HANSEN, L. P., J. HEATON, AND A. YARON (1996): “Finite-sample properties of some alternative GMM estimators,” *Journal of Business & Economic Statistics*, 14(3), 262–280.
- KOIJEN, R. S., AND M. YOGO (2019): “A demand system approach to asset pricing,” *Journal of Political Economy*, 127(4), 1475–1515.
- LEARY, M. T., AND M. R. ROBERTS (2014): “Do peer firms affect corporate financial policy?,” *The Journal of Finance*, 69(1), 139–178.
- LOU, D. (2012): “A flow-based explanation for return predictability,” *The Review of Financial Studies*, 25(12), 3457–3489.
- LUCAS, R. E. J. (1978): “Asset prices in an exchange economy,” *Econometrica*, 46(6), 1429–1445.
- MA, Y., K. XIAO, AND Y. ZENG (2022): “Mutual fund liquidity transformation and reverse flight to liquidity,” *The Review of Financial Studies*, 35(10), 4674–4711.
- NEWHEY, W. K., AND F. WINDMEIJER (2009): “Generalized method of moments with many weak moment conditions,” *Econometrica*, 77(3), 687–719.

- PAVLOVA, A., AND T. SIKORSKAYA (2023): “Benchmarking intensity,” *The Review of Financial Studies*, 36(3), 859–903.
- PIAZZESI, M., AND M. SCHNEIDER (2007): “Asset prices and asset quantities,” *Journal of the European Economic Association*, 5(2-3), 380–389.
- QIAN, E. (2023): “Heterogeneity-robust granular instruments,” *arXiv preprint arXiv:2304.01273*.