Efficiency of Granular Instrumental Variables and Asymptotic Distributions with Estimated Factor Loadings

PRELIMINARY DRAFT

Gokul Gopalan Ramachandran* Kenichi Nagasawa* Eric Renault*

November 9, 2026

Abstract

We study efficiency and inference for granular instrumental variable estimators with a fixed number of cross-sectional units. Using optimal instrument results for conditional moment models, we show that the efficient instrument equals the share-weighted granular instrument of Gabaix and Koijen (2024) plus an additional term. Using the optimal instrument generally leads to efficiency gains. We also show that a broader class of weights identifies the parameter in the individual-level equations, which can be used to yield further efficiency gains. When factor loadings are unknown, we provide conditions under which the loadings and the idiosyncratic error covariance matrix are identified while allowing for cross-sectional heteroscedasticity. We then propose GMM estimators whose inference accounts for the estimation of the factor structure. In an application to the global oil market, we find evidence of heteroscedastic idiosyncratic shocks. We provide estimates of oil demand and supply elasticities, which are robust to heteroscedasticity.

Keywords: Granular instrumental variables, factor models, optimal instruments.

*Department of Economics, University of Warwick.

1 Introduction

In many economic settings of interest, granularity plays an important role in explaining the variation of aggregate variables (Acemoglu et al., 2012; Gabaix, 2011). For example, Amiti and Weinstein (2018) find that idiosyncratic shocks to large Japanese banks explain 30–40% of the fluctuations in aggregate lending and investment, and di Giovanni et al. (2014) show that in the universe of French firms, firm-specific shocks contribute about as much to aggregate sales volatility as shocks common to all firms within a sector or country. Recently, Gabaix and Koijen (2024, henceforth GK) propose an approach that exploits such granularity to construct an instrumental variable (IV), which they call the granular IV method. Due to the presence of granularity in a variety of economic settings, the granular IV method has quickly become a popular tool in empirical macroeconomics and finance.[1]

Despite its popularity, several theoretical aspects of granular IV estimators remain to be understood. One issue is efficiency of granular IV estimators. GK establish that the share-weighted sum of granular instruments is asymptotically efficient within a class of linear IV estimators, but it remains open whether more efficient estimators exist in a larger class of estimators. Another important practical issue is statistical inference. A key input to the granular IV approach is the factor structure of individual-level variables, and in most empirical applications, the factor loadings need to be estimated in a preliminary step. Many empirical works have estimated factor structures from a small number of cross-sectional units in settings where a handful of economic agents have dominant presence. With a fixed number of cross-sectional units, factor loadings cannot be consistently estimated without additional restrictions, and it is important to understand the scope and limitations of what we can learn from the granular IV approach under reasonable assumptions on data generating processes. Some of the existing studies have used principal component analysis (PCA) to estimate the factor-loading space. With fixed N, spherical idiosyncratic covariance is a standard sufficient restriction for this approach under the factor-model rank and orthogonality conditions; consistency is not generally guaranteed with unequal idiosyncratic variances. In addition, standard errors that ignore estimation of the loading space can invalidate inference even when that space is consistently estimated.

In this paper, we contribute to the granular IV literature by addressing the aforementioned efficiency and inference issues. Specifically, we use optimal instrument results in conditional moment models to characterize the asymptotically efficient granular IV estimator. Our result shows that the efficient instrument is the sum of two terms: the share-weighted granular instrument proposed by GK plus an additional term. The additional term can contain information not captured by GK's scalar instrument, creating scope for efficiency improvement. Coincidence occurs under further restrictions, such as centered elliptical idiosyncratic shocks with spherical covariance. This efficiency result is established under the known factor loadings assumption,

[1] A partial list of empirical papers applying granular IV methods includes Galaasen et al. (2020); Kundu and Vats (2021); Ma et al. (2021); Adrian et al. (2022); Schubert et al. (2022); Aldasoro et al. (2023); Flynn and Sastry (2024); Holm-Hadulla and Thürwächter (2024) in macroeconomics and Chodorow-Reich et al. (2021); Gabaix and Koijen (2021); Camanho et al. (2022); Ma et al. (2022); Bippus et al. (2023); Dong et al. (2025); Ling et al. (2025) in finance.

following the main analysis of GK. We also establish a new class of moment conditions that identify the structural parameter in individual-level equations (denoted by ϕ^d in GK). GK show that to identify the parameter in the individual equations, one should use the inverse of idiosyncratic variances as weights to form an IV moment condition. We show that there are valid weights other than the inverse of idiosyncratic variances and that by choosing appropriate weights, we can improve the efficiency of granular IV estimators.

For inference issues, we first provide a set of sufficient conditions under which unknown factor loadings and covariance matrix of the idiosyncratic shocks are identified. Importantly, we allow for cross-sectional heteroscedasticity. Accommodating deviations from equal variances may be important as we find some evidence for cross-sectional heteroscedasticity in our empirical application of global oil markets. For our identification results, we build on Anderson and Rubin (1956) and we assume sufficient variations in factor loadings across units and the number of cross-sectional units being sufficiently larger than the number of factors. In the appendix, we extend the identification result by assuming a known block diagonal structure of the covariance matrix i.e., allowing for correlations of the errors within blocks. Building on the identification results, we propose a GMM implementation of granular IV estimators and characterize its asymptotic properties. Importantly, statistical inference based on the GMM estimator is asymptotically valid under unknown factor loadings and idiosyncratic-error covariance matrix.

Using the GMM estimators developed in this paper, we estimate supply and demand elasticities in global oil markets. As mentioned above, in this empirical setting, we find evidence for unequal variances of idiosyncratic shocks across oil supplying countries, which warrants granular IV approaches accommodating (cross-sectional) heteroscedastic errors. Our GMM estimators remain valid with unequal idiosyncratic variances, and using our methods, the demand elasticity estimate is 0.081 (standard error 0.039) and the supply elasticity estimate is -0.069 (standard error 0.038), both of them falling within the range of estimates found in the existing literature.

Related literature. Gabaix and Koijen (2024) (GK) have proposed granular IV approaches where a factor model in a panel data setting is employed to formalize the use of individual specific idiosyncratic shocks as instruments. We use their econometric framework for identification and estimation as our starting point. Our work is concurrent to Hahn et al. (2026).[2] They develop inference for granular IVs constructed from an estimated covariance eigenspace. Our results in Section 4 are closely related to asymptotic distributional results established by Hahn et al. (2026). Their idiosyncratic covariance is spherical across units, although its common variance may vary over time. Our GMM inference accommodates unequal diagonal entries, and our appendix extends identification to known covariance blocks. In addition, we establish new results on asymptotic efficiency of granular IV estimators. Qian (2023) extends granular IV frameworks by allowing for heterogeneous elasticity parameters in individual equations. Baumeister and Hamilton (2023) propose a full-information approach that jointly estimates the factor structure and structural parameter using maximum

[2] A preliminary version of our work was first presented at 2024 NBER-NSF Time Series Conference.

likelihood estimation. The above papers all focus on fixed N asymptotic frameworks. Banafti and Lee (2022) and Gopalan Ramachandran (2026) study asymptotic properties of granular IV estimators under large N, T asymptotic frameworks.

For our efficincy results, we apply the existing optimal instrument results in conditional moment models (Chamberlain, 1987; Newey, 1993). Our identification analysis of factor models builds on Anderson and Rubin (1956). Our results do not seem to directly follow from existing results (e.g., Bekker and ten Berge, 1997; Williams, 2020) because the covariance matrix of error terms exhibit correlations due to cross-sectional de-meaning to remove time fixed effects. For our inference results, we build on the extensive literature of two-step M-estimators (e.g., Newey and McFadden, 1994).

Roadmap Section 2 presents the econometric model and reviews GK's construction of granular insrtumental variables. After reviewing granular IV methods, Sections 3 and 4 present our new results. Section 3 derives efficient granular IV estimators under known factor loadings assumption. Section 4 develops estimation and inference when factor loadings are estimated in a preliminary step. Section 5 applies our proposed methods to elasticity estimation of global oil markets.

2 Econometric Model and Construction of Granular IV

In this section, we describe our econometric model and review the construction of granular IV proposed by GK. After reviewing the granular IV method, we present new results in Sections 3 and 4.

2.1 Model

We follow GK and consider a panel data model with a fixed number of cross-sectional units and a growing number of time periods. Specifically,

$$x_t = \psi p_t + \varepsilon_t, \tag{1}$$

$$y_{it} = \phi p_t + \lambda_i^{\intercal} F_t + u_{it}, \qquad i = 1, \ldots, N, \tag{2}$$

$$x_t = \sum_{i=1}^{N} S_i y_{it}, \tag{3}$$

where researchers observe $p_t, x_t, \{S_i, y_{it}\}_{i=1}^N$ for $t = 1, \ldots, T$ and $\sum_{i=1}^N S_i = 1$. Eq. (1)-(2) represent structural equations, and ψ, ϕ are the parameters of interest. A leading application of this model is a demand-supply system, and Eq. (3) is a market clearing condition. To fix ideas, we use a global oil market as a running example. In the oil market example, p_t denotes a percentage deviation of oil price at time t relative to the equilibrium price, x_t is a percentage deviation of the global oil demand, y_{it} is a percentage deviation of oil supply for country i, and S_i denotes the equilibrium market share of country i. In this example, Eq. (1) represents the aggregate demand function for oil, and Eq. (2) is the supply curve of country i. The

demand/supply labels are interchangeable, and it depends on data availability/specific applications whether the aggregate equation corresponds to a demand or supply curve.

A key component of granular IV approaches is (3), the aggregate variable represented as a share-weighted sum of the individual-level variables. The share-weighted form arises if a level is transformed to a percentage deviation from the mean.[3] As we will see below, this share-weighted form, combined with interactive fixed effects structure, enables construction of instruments within the model.

We model the error term of y_{it} as the sum of interactive fixed effects $\lambda_i^\intercal F_t$ and an idiosyncratic shock u_{it}. For some $r \in \mathbb{N}$, we let $\lambda_i, F_t \in \mathbb{R}^r$ where $r \geq 1$ is assumed to be known.[4] In the oil market example, F_t may represent global supply shocks, and λ_i captures how the aggregate shocks affect country i. Since the coefficient vector λ_i varies across i, this model allows for heterogeneous responses to the same global shocks e.g., some countries are less susceptive to external shocks than others. While the term $\lambda_i^\intercal F_t$ represents an aggregate supply shock, u_{it} is a country i specific shock. This distinction of aggregate versus individual shocks is conceptually important for the granular IV approach as we discuss below.

For simplicity of expositions, we omit exogenous control variables in the above equations and consider the time-invariant share $S_1, \ldots, S_N$. We present analysis of a general model with control variables and time-varying shares in the appendix.

We maintain the following assumptions. Recall that we focus on the fixed-N, large T-asymptotics.

Assumption 1. *Let $y_t = (y_{1t}, \ldots, y_{Nt})^\intercal \in \mathbb{R}^N$ and $u_t = (u_{1t}, \ldots, u_{Nt})^\intercal \in \mathbb{R}^N$.*

(i) $\psi \neq \phi$. $\Lambda \in \mathbb{R}^{N \times r}$ and $S = (S_1, \ldots, S_N)^\intercal \in \mathbb{R}^N$ are non-random. Λ is of full column rank and $\sum_{i=1}^N S_i = 1$.

(ii) The time series $\{p_t, u_t, \varepsilon_t, F_t\}_{t \geq 1}$ is a weakly stationary process with finite second moments. Researchers observe $\{y_t, x_t, p_t\}_{t=1}^T$ and S.

Assumption 2. *(i) $\mathbb{E}[\varepsilon_t | u_t] = 0$ with probability one.*

(ii) $\mathbb{E}[F_t | u_t] = 0$ with probability one.

(iii) Let $\Omega = \mathbb{E}[u_t u_t^\intercal]$ and $M_\Lambda = I - \Lambda(\Lambda^\intercal \Lambda)^{-1}\Lambda^\intercal$ where $I \in \mathbb{R}^{N \times N}$ is the identity matrix. Ω is non-singular and $M_\Lambda \Omega S \neq 0$.

(iv) For some scalar $\sigma_\varepsilon > 0$, $\mathbb{E}[\varepsilon_t^2 | u_t] = \sigma_\varepsilon^2$ with probability one.

[3] Let X_t be an aggregate variable of interest and $\{Y_{it}\}_{i=1}^N$ be individual-level variables, and suppose that these variables are related by $X_t = \sum_{i=1}^N Y_{it}$. Let $\bar{X}$ be the "long-run" mean of X_t and $\bar{Y}_i$ be the corresponding value of Y_{it}. Define $x_t = (X_t - \bar{X})/\bar{X}$ and $y_{it} = (Y_{it} - \bar{Y}_i)/\bar{Y}_i$. We have $x_t = \sum_{i=1}^N y_{it} S_i$ where $S_i := \bar{Y}_i / \bar{X}$ denotes the "share" of unit i.

[4] Under suitable identification and regularity conditions, the number of factors can be consistently estimated with fixed N as $T \to \infty$. If $\Pr(\widehat{r} = r) \to 1$, replacing r by $\widehat{r}$ leaves the pointwise first-order asymptotic distributions of the resulting estimators unchanged. We maintain the known-r assumption to simplify the exposition.

Assumption 1 collects regularity conditions. The restriction $\psi \neq \phi$ is needed for a unique solution of (1)-(3). We treat Λ and S as non-stochastic objects as they are time-invariant. In the appendix, we present results accommodating time-varying shares. The requirement of full column rank of Λ is without loss of generality, and $\sum_{i=1}^{N} S_i = 1$ is a scale normalization. We assume finiteness of second moments as we need it for key objects to be well-defined. Weak stationarity can be weakened to allow for different marginal distributions across time, but we maintain it to simplify expositions. Below we add more restrictions on the time series for further analysis.

Assumption 2 is a set of key identifying restrictions for granular IV approaches. Conditions (i)-(ii) are a version of instrument exclusion restrictions. In the oil market example, the restriction means that individual-level supply shocks u_{it}'s are orthogonal to the aggregate demand shock ε_t and the global supply shocks F_t. The individual-level equation (2) has the interactive fixed effects structure to make it plausible that the term u_{it} consists of only country-level supply shocks, not containing global shocks.

Condition (iii) is an instrument relevance requirement. To interpret this condition, it is instructive to consider the special case where $\lambda_i = 1$ for $i = 1, \ldots, N$ (i.e., the model only contains time fixed effects) and $\Omega = \sigma_u^2 I$ for some $\sigma_u^2 > 0$. In this special case, the condition becomes $S_i \neq \frac{1}{N}$ for some $i = 1, \ldots, N$. This requirement is related to the granularity structure; it excludes a uniform market share i.e., some suppliers have to be larger than others.

Condition (iv) is not required for the validity of granular IV methods, but GK use this condition for their analysis (see their proof of Proposition 2) and we maintain it here to make our analysis of efficient estimators comparable to the optimality result of GK.

2.2 Construction of Granular IV

Here we review the construction of granular IV proposed by GK. We stack Eq. (2) for $i = 1, \ldots, N$ and express it in matrix form

$$y_t = \phi p_t \iota + \Lambda F_t + u_t$$

where $\iota \in \mathbb{R}^N$ is the vector of ones. Without loss of generality, we assume that the first element of λ_i equals 1 for all $i = 1, \ldots, N$ i.e., the first column of Λ equals ι. This normalization means that the interactive fixed effects include time fixed effects. For exposition, suppose Λ is known or identified in a preliminary step. GK point out that the idiosyncratic shock vector can be partially recovered by

$$M_\Lambda y_t = M_\Lambda u_t, \qquad M_\Lambda := I - \Lambda(\Lambda^\intercal \Lambda)^{-1}\Lambda^\intercal. \tag{4}$$

Then, we may use the vector $M_\Lambda y_t$ as instruments to estimate ψ and ϕ. Below, we treat estimation of ψ and ϕ separately, and for ease of reference, we call Eq. (1) the *aggregate* equation and Eq. (2) the *individual* equations.

2.2.1 Validity of IV in Aggregate Equation

The vector $M_\Lambda y_t$ in (4) is a valid IV for x_t in (1) because the instrument exogeneity follows from Assumption 2(i):

$$\mathbb{E}[M_\Lambda y_t \varepsilon_t] = \mathbb{E}[M_\Lambda u_t \varepsilon_t] = \mathbb{E}[M_\Lambda u_t \mathbb{E}[\varepsilon_t | u_t]] = 0.$$

Also, the instrument relevance condition is plausible because

$$\mathbb{E}[M_\Lambda y_t p_t] = \mathbb{E}\left[M_\Lambda u_t \left(\frac{1}{\phi - \psi}\varepsilon_t + \frac{1}{\psi - \phi} S^\intercal (\Lambda F_t + u_t)\right)\right] = \frac{1}{\psi - \phi} M_\Lambda \mathbb{E}[u_t u_t^\intercal] S$$

where the form of p_t is derived using the equilibrium condition (3) and the second equality uses Assumption 2(i)-(ii). Then, Assumption 2(iii) implies that $\mathbb{E}[M_\Lambda y_t p_t] \neq 0$ holds.

Remark 2.1. For identification of ψ in the aggregate equation, Assumption 2(i) is essential for the instrument exogeneity. On the other hand, Assumption 2(ii) can be dispensed with. In the above argument, Assumption 2(ii) is part of sufficient conditions for the instrument relevance, but $\mathbb{E}[F_t | u_t] = 0$ is not necessary for the instrument relevance. This is in contrast to identification in the individual equations. As seen below, Assumption 2(ii) is essential for identification of ϕ. □

Remark 2.2. A weaker condition for the instrument exogeneity is available: the vector $\mathbb{E}[u_t \varepsilon_t]$ lies in the linear space spanned by the columns of Λ. This weaker version means that the individual-level shocks u_t are allowed to have a limited form of correlation with the disturbance term in the aggregate equation. Hahn et al. (2026) have first pointed out this weaker sufficient condition for the instrument exogeneity. For our efficiency results, we need the conditional moment restriction, and we maintain Assumption 2(i). □

2.2.2 Validity of IV in Individual Equations

Identification of ϕ in (2) is less obvious because $M_\Lambda y_t = M_\Lambda u_t$ is correlated with the error term in (2). Specifically, the error term in (2) is $\lambda_i^\intercal F_t + u_{it}$ and, in general,

$$\mathbb{E}[M_\Lambda y_t (\lambda_i^\intercal F_t + u_{it})] = M_\Lambda \mathbb{E}[u_t (\lambda_i^\intercal F_t + u_{it})] = M_\Lambda \mathbb{E}[u_t u_{it}] \neq 0 \quad i = 1, \ldots, N.$$

GK propose to aggregate the individual equations using appropriate weights which make $M_\Lambda y_t$ and the weighted u_{it}'s uncorrelated. Specifically, assuming Ω is diagonal i.e., $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$, GK propose to use the weight $H = \frac{1}{\sum_{i=1}^N \sigma_i^{-2}} (\sigma_1^{-2} \ldots \sigma_N^{-2})^\intercal$. In the general notation, this corresponds to $H = \frac{1}{\iota^\intercal \Omega^{-1} \iota} \Omega^{-1} \iota$. The weighted error term is $H^\intercal (\Lambda F_t + u_t)$, and we have

$$\mathbb{E}[M_\Lambda y_t H^\intercal (\Lambda F_t + u_t)] = M_\Lambda \mathbb{E}[u_t (\Lambda F_t + u_t)^\intercal] H = M_\Lambda \mathbb{E}[u_t u_t^\intercal] H = M_\Lambda \iota (\iota^\intercal \Omega^{-1} \iota)^{-1} = 0$$

where the second equality uses Assumption 2(ii) and the last equality follows from the first column of Λ being ι. The instrument relevance holds as it turns out to be the same condition for the aggregate equation case.

3 Efficient Granular IV Estimators

In this section, we establish asymptotic efficiency results for granular IV estimators. We follow the main analysis of GK and assume the following conditions.

Assumption 3. *The linear space spanned by the columns of* Λ *is known.*

As explained by GK, Assumption 3 is weaker than it may appear. Researchers may be willing to assume

$$\Lambda = XL$$

where $X \in \mathbb{R}^{N\times k}$ is the matrix whose ith row X_i is a vector of time-invariant characteristics of ith unit and $L \in \mathbb{R}^{k\times r}$ is some unknown non-stochastic matrix. Observing X implies $\mathrm{Im}(\Lambda) \subseteq \mathrm{Im}(X)$. The spaces coincide if $\mathrm{rank}(X) = \mathrm{rank}(XL) = r$; for example, this holds when $k = r$, X has full column rank and L is nonsingular. In this case Assumption 3 holds.

3.1 Aggregate Equation

As explained in Section 2.2, we can identify ψ in (1) via the conditional moment restriction

$$\mathbb{E}[(x_t - \psi p_t)|M_\Lambda y_t] = 0. \tag{5}$$

To derive an efficient estimator of ψ, we briefly review the conditional moment results of Chamberlain (1987); Newey (1993). Let X, Z be random vectors and $m(X,\theta)$ be a moment function known up to the parameter θ. A conditional moment model is a setting in which

$$\mathbb{E}[m(X,\theta)|Z] = 0 \text{ almost surely} \quad \Longleftrightarrow \quad \theta = \theta_0.$$

With serially uncorrelated instrumented errors and suitable regularity conditions, an efficient GMM estimator of θ_0 uses the unconditional moment

$$\mathbb{E}\left[g(Z)^\intercal m(X,\theta)\right], \qquad g(Z) = \mathbb{V}[m(X,\theta_0)|Z]^{-1}\frac{\partial}{\partial\theta^\intercal}\mathbb{E}[m(X,\theta)|Z]\Big|_{\theta=\theta_0}$$

where $g(Z)$ is called the optimal instrument.

Applying the above to our context, and assuming $\psi \neq 0$ for the outcome-predictor expression, we obtain the optimal instrument (up to a multiplicative constant)

$$\mathbb{V}[\varepsilon_t|M_\Lambda y_t]^{-1}\mathbb{E}[p_t|M_\Lambda y_t] = \frac{1}{\sigma_\varepsilon^2\psi}\mathbb{E}[x_t|M_\Lambda y_t] \propto S^\intercal\mathbb{E}[y_t|M_\Lambda y_t]$$

where the equality uses Assumption 2(iv), $p_t = \frac{1}{\psi}(x_t - \varepsilon_t)$, and Assumption 2(i). Then, we can construct an efficient IV estimator

$$\hat{\psi} = \frac{\sum_{t=1}^T S^\intercal\widehat{\mathbb{E}}[y_t|M_\Lambda y_t]x_t}{\sum_{t=1}^T S^\intercal\widehat{\mathbb{E}}[y_t|M_\Lambda y_t]p_t} \tag{6}$$

where $\widehat{\mathbb{E}}[y_t|M_\Lambda y_t]$ is a preliminary estimator of $\mathbb{E}[y_t|M_\Lambda y_t]$. To formally state our efficiency result, we introduce another set of assumptions. Write $Z_t = M_\Lambda y_t$, $h_t = S^\intercal\mathbb{E}[y_t \mid Z_t]$, and $\widehat{h}_t = S^\intercal\widehat{\mathbb{E}}[y_t \mid Z_t]$.

Assumption 4. *$\psi \neq 0$, and $\{(y_t, p_t, \varepsilon_t)\}_{t\geq 1}$ is strictly stationary and ergodic. With $\mathcal{F}_{t-1} = \sigma(y_s, p_s, \varepsilon_s : s < t)$, assume $\mathbb{E}[\varepsilon_t \mid Z_t, \mathcal{F}_{t-1}] = 0$. The preliminary estimator satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{h}_t - h_t)^2 = o_{\mathbb{P}}(1), \qquad \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\widehat{h}_t - h_t)\varepsilon_t = o_{\mathbb{P}}(1).$$

The time-series exclusion makes $g(Z_t)\varepsilon_t$ a martingale difference for every square-integrable instrument $g(Z_t)$; it holds, for example, under independent observations and Assumption 2(i).

Theorem 1. *Suppose that Assumptions 1-4 hold. Then, the IV estimator (6) satisfies*

$$\sqrt{T}(\widehat{\psi} - \psi) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\mathbb{E}[(\mathbb{E}[p_t \mid Z_t])^2]}\right).$$

It is asymptotically efficient among regular GMM estimators using finite vectors of square-integrable functions of Z_t as instruments for (5). In particular, it is asymptotically at least as efficient as the GK estimator using $S^{\mathsf{T}} M_\Lambda y_t$, provided that $\mathbb{E}[S^{\mathsf{T}} M_\Lambda y_t p_t] \neq 0$.

The proof in Appendix C.2 verifies feasibility and the optimal-instrument variance bound (Chamberlain, 1987; Newey, 1993). Our contribution here is to shed a new light onto the granular IV estimator proposed by GK. The comparison follows because $S^{\mathsf{T}} M_\Lambda y_t$ belongs to this instrument class. General serial correlation in the instrumented errors requires a long-run variance analysis and is outside this efficiency claim.

Let $P_\Lambda = \Lambda(\Lambda^{\mathsf{T}}\Lambda)^{-1}\Lambda^{\mathsf{T}}$. By the identity $I = M_\Lambda + P_\Lambda$,

$$S^{\mathsf{T}}\mathbb{E}[y_t | M_\Lambda y_t] = S^{\mathsf{T}} M_\Lambda y_t + S^{\mathsf{T}}\mathbb{E}[P_\Lambda y_t | M_\Lambda y_t].$$

Thus, we can view our optimal instrument as the sum of GK's granular instrument ($S^{\mathsf{T}} M_\Lambda y_t$) and the additional term $S^{\mathsf{T}}\mathbb{E}[P_\Lambda y_t | M_\Lambda y_t]$. To characterize coincidence, recall $Z_t = M_\Lambda y_t = M_\Lambda u_t$. The equilibrium equation and the conditional exclusions give

$$h_t = S^{\mathsf{T}}\mathbb{E}[y_t \mid Z_t] = \frac{\psi}{\psi - \phi} S^{\mathsf{T}}\mathbb{E}[u_t \mid Z_t] = \frac{\psi}{\psi - \phi}\left(S^{\mathsf{T}} Z_t + S^{\mathsf{T}}\mathbb{E}[P_\Lambda u_t \mid Z_t]\right).$$

Thus h_t coincides with GK's instrument up to a nonzero scale if and only if $S^{\mathsf{T}}\mathbb{E}[P_\Lambda u_t \mid Z_t] = c S^{\mathsf{T}} Z_t$ almost surely for a constant $c \neq -1$.

For illustration, suppose u_t is centered and elliptically distributed with finite, positive definite covariance Ω. Its conditional mean is linear in the full vector Z_t (Cambanis et al., 1981):

$$\mathbb{E}[u_t \mid Z_t] = \Omega M_\Lambda (M_\Lambda \Omega M_\Lambda)^{+} Z_t,$$

where A^{+} denotes the Moore–Penrose inverse. This makes the optimal instrument linear in Z_t, but does not generally make it proportional to the particular scalar $S^{\mathsf{T}} Z_t$. A sufficient covariance restriction for coincidence is $P_\Lambda \Omega M_\Lambda = 0$, which gives $\mathbb{E}[P_\Lambda u_t \mid Z_t] = 0$. In particular, this holds when $\Omega = \sigma_u^2 I$.

Outside these cases, the optimal instrument may also be nonlinear in Z_t.[5] When h_t is not proportional to $S^{\mathsf{T}} Z_t$, the variance comparison in Theorem 1 is strict.

[5] For example, let $N \geq 2$, $\Lambda = \iota$ and $u_{it} = \sigma(\epsilon_{it} - 1)$ with $\sigma > 0$ and ϵ_{it} i.i.d. Exp(1) across i and t, independent of the

3.2 Individual Equations

As noted in Section 2.2, identification of ϕ in the individual equations requires appropriate weights H such that

$$\mathbb{E}[M_\Lambda y_t H^\intercal (y_t - \phi p_t \iota)] = 0.$$

By Assumption 2(ii), the above equality is equivalent to

$$\begin{aligned} \mathbb{E}[M_\Lambda y_t H^\intercal u_t] = 0 \quad &\iff \quad M_\Lambda \mathbb{E}[u_t u_t^\intercal] H = 0 \\ &\iff \quad M_\Lambda \Omega H = 0 \\ &\iff \quad \Omega H \in \mathrm{Im}(\Lambda) \end{aligned}$$

where $\mathrm{Im}(\Lambda)$ is the linear space spanned by the columns of Λ. GK propose to use $H \propto \Omega^{-1}\iota$ where ι is the first column of Λ, but the above discussion indicates that there is a larger class of weight vectors H that is valid for the identification of ϕ. Because this is a new insight in the literature, we document it as a formal result. To construct H, we need the second-moment matrix Ω. It equals the covariance matrix of u_t when $\mathbb{E}[u_t] = 0$. For this section, we assume that it is known up to a constant.

Assumption 5. *The second-moment matrix $\Omega = \mathbb{E}[u_t u_t^\intercal]$ is known up to a scale.*

A leading case for Assumption 5 is when the shocks have zero means, equal variances, and zero cross-sectional correlations, so that $\mathbb{E}[u_t u_t^\intercal] = \sigma_u^2 I$ for some $\sigma_u^2 > 0$. However, in many applications, Assumption 5 may not hold. With additional restrictions on the data generating process, the identification result of ϕ extends to settings in which Assumption 5 does not hold; see details in Section 4.

For nonsingular Ω, let $\mathcal{C} = \{\gamma \in \mathbb{R}^r : \iota^\intercal \Omega^{-1} \Lambda \gamma \neq 0\}$ and for $\gamma \in \mathcal{C}$, define

$$H_\gamma = \frac{1}{\iota^\intercal \Omega^{-1} \Lambda \gamma} \Omega^{-1} \Lambda \gamma$$

where the normalization by $(\iota^\intercal \Omega^{-1} \Lambda \gamma)^{-1}$ is for convenience. Using H_γ as weights, we identify ϕ under the following conditions.

Theorem 2. *Suppose Assumptions 1, 3, and 5 hold, Ω is nonsingular, and*

$$\mathbb{E}[F_t u_t^\intercal] = 0, \qquad \mathbb{E}[M_\Lambda y_t p_t] \neq 0.$$

Then, ϕ is identified via

$$\phi \mathbb{E}[M_\Lambda y_t p_t] = \mathbb{E}[M_\Lambda y_t H_\gamma^\intercal y_t]$$

for $\gamma \in \mathcal{C} = \{\gamma \in \mathbb{R}^r : \iota^\intercal \Omega^{-1} \Lambda \gamma \neq 0\}$.

centered (ε_t, F_t), and let $S \neq \iota/N$. Conditional on $Z_t = z$, the cross-sectional mean $N^{-1}\iota^\intercal u_t$ has an exponential distribution shifted by $-\sigma - \min_i z_i$, with rate N/σ. Hence

$$S^\intercal \mathbb{E}[y_t \mid Z_t] = \frac{\psi}{\psi - \phi} \left[S^\intercal Z_t - \min_{1 \le i \le N} Z_{it} - \sigma \left(1 - \frac{1}{N}\right) \right].$$

The proof is in Appendix C.3. The relevance condition follows from Assumption 2(i) and (iii), together with $\mathbb{E}[F_t u_t^\intercal] = 0$, by the calculation in Section 2.2.1. Since there are many choices of γ that enables identification of ϕ, a natural question is whether there exists some optimal choice of γ. We answer this question below using the asymptotic variance of IV estimators as the criterion. Before addressing the choice of γ, we turn to the question of optimal instruments.

3.2.1 Conditional Moment Restriction

Now, in analogy to the analysis of the aggregate equation, we characterize the optimal instrument for ϕ for a fixed choice of weights H. For this purpose, we strengthen the unconditional moment of Theorem 2 to the conditional moment:

$$\mathbb{E}[H^\intercal(y_t - \phi p_t \iota)|M_\Lambda y_t] = \mathbb{E}[H^\intercal(\Lambda F_t + u_t)|M_\Lambda u_t] = \mathbb{E}[H^\intercal u_t|M_\Lambda u_t] = 0 \tag{7}$$

where the second equality uses Assumption 2(ii) and $H = H_\gamma$ as defined above for some fixed $\gamma \in \mathcal{C}$. Furthermore, we assume conditional homoscedasticity of the composite error:

$$\mathbb{V}[H^\intercal(\Lambda F_t + u_t) \mid M_\Lambda u_t] = \sigma_H^2 \tag{8}$$

for some $\sigma_H^2 > 0$. Note that GK also impose an analogous conditional homoscedasticity restriction in the derivation of the asymptotic variance of their IV estimator (Proposition 2). The restrictions (7)–(8) hold, for example, if $(u_t^\intercal, F_t^\intercal)^\intercal$ is jointly Gaussian with mean zero. Under the maintained factor exclusion and $\Omega H \in \mathrm{Im}(\Lambda)$, the composite error is uncorrelated with $M_\Lambda u_t$; joint Gaussianity then implies independence and hence both restrictions.

Let $e_{H,t} = H^\intercal(\Lambda F_t + u_t)$ and $Z_t = M_\Lambda y_t$. For the time-series efficiency comparison, the sequential exclusion in Assumption 4 must also hold for this composite error, as specified below.

Theorem 3. *Fix $H = H_\gamma$ for $\gamma \in \mathcal{C}$. Suppose that Assumptions 1–5 and (7)–(8) hold. In addition, assume $\mathbb{E}[e_{H,t} \mid Z_t, \mathcal{F}_{t-1}] = 0$, with $\mathcal{F}_{t-1}$ as in Assumption 4. Among regular GMM estimators using finite vectors of square-integrable functions of Z_t as instruments for $\mathbb{E}[H^\intercal(y_t - \phi p_t \iota) \mid Z_t] = 0$, the optimal instrument is, up to a nonzero constant,*

$$h_H^*(Z_t) = \frac{\mathbb{E}[p_t \mid Z_t]}{\sigma_H^2} \propto \mathbb{E}[p_t \mid Z_t] \propto S^\intercal \mathbb{E}[y_t \mid Z_t].$$

The corresponding asymptotic variance bound is $\sigma_H^2 / \mathbb{E}[(\mathbb{E}[p_t \mid Z_t])^2]$.

The proof is in Appendix C.4. The additional sequential exclusion holds under independent observations and (7). The result concerns the population instrument; feasible estimation additionally requires the preliminary-estimation conditions of Assumption 4 with $e_{H,t}$ replacing ε_t. This theorem implies that as long as (7)-(8) (and other assumptions) hold, the form of the optimal instrument does not depend on the weight vector H. (It allows for the variance σ_H^2 to depend on H in (8).) In particular, the optimal instrument $S^\intercal \mathbb{E}[y_t|M_\Lambda y_t]$ is the same as the optimal instrument used to estimate ψ, albeit under different assumptions.

The fact that the form of valid instruments coincide in the aggregate and individual equations is an interesting phenomenon, which Banafti and Lee (2022, p.6) call "miraculous." Our analysis shows that this interesting coincidence directly follows from the computation of optimal instruments.

3.2.2 Choice of Weights

Under our assumptions, the form of the optimal instrument is invariant with respect to H. However, the asymptotic variance of the IV estimator depends on H. Let $z_t \in \mathbb{R}$ be a valid instrument, not necessarily the optimal instrument e.g., $z_t = S^\intercal \mathbb{E}[y_t | M_\Lambda y_t]$ or $z_t = S^\intercal M_\Lambda y_t$. We consider the unconditional moment

$$\mathbb{E}[z_t H^\intercal (y_t - \phi p_t \iota)] = 0.$$

Under appropriate regularity conditions, the asymptotic variance of the IV estimator is

$$\frac{\mathbb{E}[z_t^2 H^\intercal (y_t - \phi p_t \iota)(y_t - \phi p_t \iota)^\intercal H]}{\mathbb{E}[z_t p_t]^2}$$

and given the instrument z_t, the smallest asymptotic variance can be derived by solving

$$\min_{\gamma \in \mathbb{R}^r:\ \iota^\intercal \Omega^{-1} \Lambda \gamma = 1} \gamma^\intercal \Lambda^\intercal \Omega^{-1} \mathbb{E}\left[z_t^2 (y_t - \phi p_t \iota)(y_t - \phi p_t \iota)^\intercal\right] \Omega^{-1} \Lambda \gamma$$

where we use $H \propto \Omega^{-1}\Lambda\gamma$ from Theorem 2. Since the objective is a quadratic form in γ and the constraint is linear, the minimizer has the closed form

$$\gamma^* = \frac{Q^{-1}\Lambda^\intercal \Omega^{-1} \iota}{\iota^\intercal \Omega^{-1} \Lambda Q^{-1} \Lambda^\intercal \Omega^{-1} \iota}, \qquad Q = \Lambda^\intercal \Omega^{-1} \mathbb{E}\left[z_t^2 (y_t - \phi p_t \iota)(y_t - \phi p_t \iota)^\intercal\right] \Omega^{-1} \Lambda,$$

provided Q is positive definite, and estimating ϕ with $H^* = \Omega^{-1}\Lambda\gamma^*$ achieves the smallest asymptotic variance among the class of IV estimators using z_t.

4 Estimation and Inference with Unknown Factor Loadings

In this section, we develop granular IV estimators in settings where the factor loadings are unknown. Specifically, we consider the model (1)-(3), where Assumption 3 does not hold. We rewrite the individual equations

$$y_t = \phi p_t \iota + \Lambda F_t + u_t = \phi p_t \iota + F_{1t} \iota + \Psi \tilde{F}_t + u_t$$

where we partition $F_t = (F_{1t}\ \tilde{F}_t^\intercal)^\intercal$ and $\Lambda = [\iota\ \Psi]$. Write $\underline{r} = r - 1$, so that $\tilde{F}_t \in \mathbb{R}^{\underline{r}}$ and $\Psi \in \mathbb{R}^{N \times \underline{r}}$. We assume without loss of generality that the first element of the factors is time fixed effects, that $r \geq 2$, and that

$$\iota^\intercal \Psi = 0. \tag{9}$$

Here inclusion of time fixed effects is without loss of generality because we need to difference out the ϕp_t term for Ψ to be estimated. The normalization (9) is also without loss of generality. Let $M_\iota \in \mathbb{R}^{N \times N}$ be the "within-transformation" matrix

$$M_\iota = I - \frac{1}{N} \iota \iota^\intercal.$$

Starting from an arbitrary loading matrix $L \in \mathbb{R}^{N \times r}$, decompose $L = M_\iota L + \frac{1}{N}\iota\iota^\intercal L$ so that

$$L\tilde{F}_t = (M_\iota L)\tilde{F}_t + \iota\Big(\tfrac{1}{N}\iota^\intercal L\tilde{F}_t\Big),$$

and the second term is absorbed into the time fixed effect $F_{1t}\iota$. Since $[\iota\ L]$ and $[\iota\ M_\iota L]$ span the same linear space, replacing L by $\Psi = M_\iota L$ leaves the model unchanged.

We consider multi-step estimation procedures where Ψ is estimated in a preliminary step (up to some linear transformation; details below). Following GK, we consider the asymptotic framework in which the number of cross-sectional units is fixed. As well-understood in the literature, consistent estimation of factor loadings in the fixed-N framework requires stronger restrictions than in the large N, T framework. One of our contributions in this section is to provide a set of sufficient conditions under which Ψ is consistently estimable.

We should emphasize that the efficiency results developed in Section 3 do not apply to the settings when the factor loadings are unknown because the knowledge of factor loadings (Assumption 3) is crucial for Theorems 1 and 3. In this section, our motivation is to establish estimation and inference results for the granular IV estimators used in the empirical literature. We leave the issue of efficiency in the case of unknown Λ for future research.

4.1 Estimation of Factor Loadings

Left-multiplying the stacked version of (2) by M_ι,

$$M_\iota y_t = \Psi\tilde{F}_t + M_\iota u_t$$

where the endogenous variable p_t is differenced out and the within-transformation leaves the loadings unchanged, $M_\iota\Psi = \Psi$, by the normalization (9). Under Assumption 1, Λ is of full column rank, and thus, Ψ is of full column rank. We impose the following standard conditions.

Assumption 6. *Assumption 1(ii) holds,* $\mathbb{E}[u_t] = 0$, $\mathbb{C}[\tilde{F}_t, u_t] = 0$, *and* $\mathbb{V}[\tilde{F}_t] =: \Sigma_{\tilde{F}}$ *is non-singular.*

Since N is fixed, we need to impose some restrictions on the covariance matrix of the individual-level error term u_t to identify Ψ. For the main result, we assume that shocks are uncorrelated across cross-sectional units.

Assumption 7. $\Omega = \mathbb{E}[u_t u_t^\intercal]$ *is a diagonal matrix. We write* $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$.

We allow for variances of individual shocks to vary across units. This assumption accommodates, for instance, countries with larger market shares to have a different magnitude of volatility than other countries. In the empirical application in Section 5, we found evidence for substantial differences in variance across cross-sectional units. Since mis-specification generally leads to inconsistency, accommodating heteroscedasticity may be important. Hahn et al. (2026) impose spherical idiosyncratic covariance across units (allowing the

common variance to vary over time), and GK discuss an extension of their method allowing for unknown Ω but maintains the assumption of known Λ. We accommodate settings where neither of Assumptions 3 and 5 holds. In the appendix, we generalize Assumption 7 and consider a setting in which error terms are allowed to correlate within known blocks of cross-sectional units.

Relaxing the spherical error assumption comes with a price, and we need stronger conditions on the factor loading matrix Λ. Specifically, we impose the following.

Assumption 8. *For $\mathcal{I} \subset \{1,\dots,N\}$ with $|\mathcal{I}| = r$, let $\Lambda_{\mathcal{I}} = [\lambda_i^{\intercal}]_{i\in\mathcal{I}}$ denote the submatrix of Λ whose rows consist of $\lambda_i^{\intercal}$, $i \in \mathcal{I}$. For every $\ell \in \{1,\dots,N\}$ there exist disjoint index sets $\mathcal{I}_\ell, \mathcal{I}'_\ell \subset \{1,\dots,N\} \setminus \{\ell\}$ with $|\mathcal{I}_\ell| = |\mathcal{I}'_\ell| = r$ such that $\Lambda_{\mathcal{I}_\ell}$ and $\Lambda_{\mathcal{I}'_\ell}$ are nonsingular.*

We borrow this condition from Anderson and Rubin (1956, Theorem 5.1). It is a strengthening of full column rank condition on Λ. The condition requires that there is variation in λ_i across i's. For example, the condition fails if all the countries have the same factor loadings i.e., they are affected by global supply shocks in the exactly same manner.

Under Assumptions 6-8, we have identification of Ψ up to rotation. For our purpose, rotational indeterminancy of Ψ is not an issue as we are interested in the projection matrix $\Psi(\Psi^{\intercal}\Psi)^{-1}\Psi^{\intercal}$. For convenience, we use the following normalization, a modification of IC4 considered in Bai and Li (2012):

$$\Psi \in \mathcal{L} \qquad \text{and} \qquad \mathbb{V}[\tilde{F}_t] = \Sigma_{\tilde{F}} \text{ is diagonal} \tag{10}$$

where, with $\iota_1 \in \mathbb{R}^{\underline{r}}, \iota_2 \in \mathbb{R}^{N-r}$ denoting the vectors of ones,

$$\mathcal{L} = \left\{ \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \\ -(\iota_1^T\Lambda_1 + \iota_2^T\Lambda_2) \end{bmatrix} \in \mathbb{R}^{N\times\underline{r}} : \Lambda_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \lambda_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{\underline{r}1} & \lambda_{\underline{r}2} & \dots & 1 \end{bmatrix} \in \mathbb{R}^{\underline{r}\times\underline{r}}, \quad \Lambda_2 \in \mathbb{R}^{(N-r)\times\underline{r}} \right\}.$$

Elements of $\mathcal{L}$ have columns orthogonal to $\iota \in \mathbb{R}^N$. This modification of IC4 respects the normalization (9).

Lemma 1. *Suppose Assumptions 6-8 hold. Then, $\Psi\Sigma_{\tilde{F}}\Psi^{\intercal}$ and Ω are uniquely determined by $\mathbb{V}[M_\iota y_t]$. Under the normalization (10), we obtain a unique Ψ.*

Given the identification result, we propose to estimate Ψ via GMM with the moment condition

$$m_\Lambda(y,\Psi,\Sigma,\Omega;\mu) = \operatorname{vech}\Big(\tilde{D}\big(M_\iota(y-\mu)(y-\mu)^{\intercal}M_\iota - \Psi\Sigma\Psi^{\intercal} - M_\iota\Omega M_\iota\big)\tilde{D}^{\intercal}\Big)$$

where $\tilde{D} = [\ I_{N-1} \quad -\iota_{N-1}\ ]$ with the $(N-1)\times(N-1)$ identity matrix I_{N-1} and the $(N-1)$-dimensional vector of ones ι_{N-1}, $\operatorname{vech}(\cdot)$ is the half vectorization operator, and μ will be replaced by an estimate of $\mathbb{E}[M_\iota y_t]$. The transformation by $\tilde{D}$ removes redundant moments. Without such transformation, the covariance matrix of the moment vector becomes singular. The choice of $\tilde{D}$ does not affect the efficient GMM estimator: any $(N-1)\times N$ matrix D of full row rank with $D\iota = 0$ can be written as $D = Q\tilde{D}$ for a nonsingular $Q \in$

$\mathbb{R}^{(N-1)\times(N-1)}$, so replacing $\tilde{D}$ with D transforms m_Λ by a nonsingular linear map. Letting $\widehat{\mu} = \frac{1}{T}\sum_{t=1}^{T} M_\iota y_t$, our estimator for Ψ is defined by

$$(\widehat{\Psi}, \widehat{\Sigma}, \widehat{\Omega}) = \underset{\Psi\in\mathcal{L},\Sigma\in\Delta_{\underline{r}},\Omega\in\Delta_N}{\operatorname{argmin}} \widehat{m}_\Lambda(\Psi,\Sigma,\Omega;\widehat{\mu})^{\mathsf{T}}\widehat{W}\widehat{m}_\Lambda(\Psi,\Sigma,\Omega;\widehat{\mu}), \qquad \widehat{m}_\Lambda(\Psi,\Sigma,\Omega;\mu) = \frac{1}{T}\sum_{t=1}^{T} m_\Lambda(y_t,\Psi,\Sigma,\Omega;\mu)$$

where $\Delta_n = \{A \in \mathbb{R}^{n\times n} : A \text{ is diagonal}\}$ and $\widehat{W}$ is a (possibly random) $N(N-1)/2 \times N(N-1)/2$ positive definite matrix that converges in probability to a positive definite matrix.

The estimator $\widehat{\Psi} \in \mathbb{R}^{N\times\underline{r}}$ lies in an affine subspace of $\mathbb{R}^{N\times\underline{r}}$, as we impose $\Psi \in \mathcal{L}$ in (10). In fact, we see that there are only $\underline{r}(N-1-r/2)$ free parameters in Ψ, so this affine subspace has dimension $d_\theta := \underline{r}(N-1-r/2)$, which is smaller than $N\underline{r}$. We work with a linear mapping Θ defined on $\mathbb{R}^{N\times\underline{r}}$ whose restriction to $\mathcal{L}$ is a bijection onto $\mathbb{R}^{d_\theta}$. The explicit definition of Θ is found in Appendix A. Using this bijection, the estimator defined above is equivalently described by $\widehat{\theta} = \Theta(\widehat{\Psi})$, or $\widehat{\Psi} = \Theta^{-1}(\widehat{\theta})$.

4.2 Granular IV Estimator for Aggregate Equation

Given the identification result of factor loadings in Lemma 1, we consider estimation of the structural parameter ψ with factor loadings estimated in the first step. In this subsection and the next, to distinguish the true parameter values from other elements of the parameter space, we subscript the true parameter by 0 e.g., ψ_0, Ψ_0. In particular, $\theta_0 = \Theta(\Psi_0)$ denotes the coordinates of the true factor loading matrix under the parameterization discussed at the end of Section 4.1.

We consider the granular instrument proposed by GK i.e., $S^{\mathsf{T}}M_\Lambda y_t = S^{\mathsf{T}}M_\Psi M_\iota y_t$. We focus on this instrument because the vast majority of existing empirical studies, if not all, use this instrument. In addition, for the unknown factor loadings case, we do not know the form of optimal instruments.

Estimator. Let $w_t = (y_t^{\mathsf{T}}, x_t, p_t)^{\mathsf{T}}$. We base estimation of the parameter ψ on the moment conditions

$$\mathbb{E}\left[m_\psi(w_t,\psi,\Psi)\right] = 0, \qquad m_\psi(w,\psi,\Psi) = S^{\mathsf{T}}M_\Psi M_\iota y\,(x-\psi p) \tag{11}$$

$$\mathbb{E}\left[m_\Lambda(y_t,\Psi,\Sigma,\Omega;\mu)\right] = 0 \tag{12}$$

where with μ evaluated at $\mathbb{E}[M_\iota y_t]$, the equalities hold if and only if $(\psi,\Psi,\Sigma,\Omega) = (\psi_0,\Psi_0,\Sigma_{\bar{F},0},\Omega_0)$.

Since the moment condition (11) is just-identified for ψ, we have a closed form estimate of ψ after plugging in the estimator $M_{\widehat{\Psi}} = I - \widehat{\Psi}(\widehat{\Psi}^{\mathsf{T}}\widehat{\Psi})^{-1}\widehat{\Psi}^{\mathsf{T}}$ where $\widehat{\Psi}$ is the GMM estimator of Ψ based on (12), defined in Section 4.1. That is, our estimator $\widehat{\psi}$ is defined as the solution of the equation

$$\frac{1}{T}\sum_{t=1}^{T} S^{\mathsf{T}}M_{\widehat{\Psi}}M_\iota y_t\left(x_t - \widehat{\psi}p_t\right) = 0, \quad \text{i.e.,} \quad \widehat{\psi} = \frac{\sum_{t=1}^{T} S^{\mathsf{T}}M_{\widehat{\Psi}}M_\iota y_t\, x_t}{\sum_{t=1}^{T} S^{\mathsf{T}}M_{\widehat{\Psi}}M_\iota y_t\, p_t}.$$

Effect of first-step estimation. Below, we will show that under appropriate assumptions, the GMM estimator $\widehat{\theta} = \Theta(\widehat{\Psi})$ is asymptotically linear i.e., there exists a vector-valued function $\vartheta(y_t)$ such that

$\sqrt{T}(\widehat{\theta}-\theta_0)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\vartheta(y_t)+o_{\mathbb{P}}(1)$. Then, the asymptotic distribution of $\widehat{\psi}$ can be characterized from the asymptotic expansion around the true unknown values (Ψ_0,ψ_0)

$$\sqrt{T}\left(\widehat{\psi}-\psi_0\right)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathcal{J}_p^{-1}\left[m_\psi(w_t,\psi_0,\Psi_0)+\mathcal{J}_\theta^{\psi}\ \vartheta(y_t)\right]+o_{\mathbb{P}}(1) \tag{13}$$

where

$$\mathcal{J}_p=-\frac{\partial}{\partial\tilde{\psi}}\mathbb{E}[m_\psi(w_t,\tilde{\psi},\Psi_0)]\big|_{\tilde{\psi}=\psi_0},$$
$$\mathcal{J}_\theta^{\psi}=\frac{\partial}{\partial\theta^{\intercal}}\mathbb{E}[m_\psi(w_t,\psi_0,\Theta^{-1}(\theta))]\big|_{\theta=\theta_0}.$$

The explicit formulas as well as derivations can be found in Appendix A. Note that the preliminary estimation of μ has no first-order effect on (13) because the derivative of $\mathbb{E}[m_\Lambda(y_t,\Psi,\Sigma,\Omega;\mu)]$ with respect to μ vanishes at $\mu=\mathbb{E}[M_\iota y_t]$.

Remark 4.1. A sufficient condition for loading estimation to have no first-order effect is $\mathcal{J}_\theta^{\psi}=0$. In this case, (13) reduces to

$$\sqrt{T}\left(\widehat{\psi}-\psi_0\right)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathcal{J}_p^{-1}m_\psi(w_t,\psi_0,\Psi_0)+o_{\mathbb{P}}(1). \tag{14}$$

This is the expansion of the corresponding infeasible estimator using Ψ_0. Necessity of $\mathcal{J}_\theta^{\psi}=0$ for this equivalence would require additional nondegeneracy conditions on the first-stage limit. The Jacobian formula in Appendix A shows that

$$\mathbb{E}[M_\iota y_t\varepsilon_t]=0$$

is sufficient for $\mathcal{J}_\theta^{\psi}=0$. For example, this holds when $\mathbb{E}[y_t\varepsilon_t]=a\iota$ for any scalar a, since demeaning removes the common direction. In general, the correction must be retained, and its covariance with the known-loading score also enters the asymptotic variance. □

Remark 4.2. In the case of known factor loadings, $S^{\intercal}M_\Lambda y_t$ is the optimal instrument under some conditions. In such instances, ignoring the adjustment term for the asymptotic variance may seem a right thing to do since, in standard cases (see e.g., Newey, 1993), consistent estimation of optimal instruments has no impact on the asymptotic variance of GMM estimators. Our result that the asymptotic variance of the optimal GMM estimator of ψ is impacted by estimation of optimal instruments may appear to contradict the existing results. However, the reason for this difference lies in the form of the conditional moment. For the purpose of GMM estimation of ψ, we only have at our disposal the conditional moment restriction (5)

$$\mathbb{E}[x_t-\psi_0 p_t\,|M_\Lambda y_t]=0$$

and it is not in the form

$$\mathbb{E}[x_t-\psi_0 p_t\,|\mathcal{F}_t]=0$$

where $\mathcal{F}_t$ is a fixed σ-field. While the latter case is the common case for asymptotic theory of optimal instruments, the former case that we consider in this paper is more complicated since the conditioning information set depends on the value of the unknown parameters. □

Asymptotic distribution. To establish asymptotic properties of the two-step granular IV estimator, we impose the following conditions.

Assumption 9. *For a matrix A, let $\|A\|^2 = \mathrm{tr}(A^\intercal A)$. Let Δ_j^+ be the set of $j \times j$ diagonal matrices with positive diagonal elements, and let $\mathcal{P}$ be a compact subset of $\mathcal{L} \times \Delta_{\underline{r}}^+ \times \Delta_N^+$. For this result, the first-stage GMM criterion is minimized over $\mathcal{P}$.*

(i) Assumption 1 and Assumptions 6–8 hold, and $\mathbb{E}[u_t \varepsilon_t] = 0$.

(ii) The true parameter $(\Psi_0, \Sigma_{\tilde{F},0}, \Omega_0)$ lies in the interior of $\mathcal{P}$ in the coordinates (θ, η, ω) defined below.

(iii) $\{M_\iota y_t, M_\iota y_t y_t^\intercal, M_\iota y_t \varepsilon_t, M_\iota y_t p_t\}_{t \geq 1}$ is weakly stationary and for each $i = 1, \ldots, N$,

$$\limsup_{T\to\infty} \left\| \mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^T M_\iota y_t y_{it}\right] \right\| < \infty, \qquad \limsup_{T\to\infty} \left\| \mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^T M_\iota y_t \varepsilon_t\right] \right\| < \infty,$$

$$\limsup_{T\to\infty} \left\| \mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^T M_\iota y_t\right] \right\| < \infty, \qquad \limsup_{T\to\infty} \left\| \mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^T M_\iota y_t p_t\right] \right\| < \infty.$$

(iv) $\mathcal{J}_p = \mathbb{E}[S^\intercal M_{\Psi_0} M_\iota y_t p_t] \neq 0$.

(v) For $\eta \in \mathbb{R}^{\underline{r}}$ and $\omega \in \mathbb{R}^N$, let $\mathrm{diag}(\eta)$ and $\mathrm{diag}(\omega)$ be the diagonal matrices with diagonal elements η and ω, and let η_0 and ω_0 be the vectors of diagonal elements of $\Sigma_{\tilde{F},0}$ and Ω_0, respectively. With $\mu_0 = \mathbb{E}[M_\iota y_t]$, define

$$\begin{aligned}
\mathcal{J}_\theta^\Lambda &= \frac{\partial}{\partial \theta^\intercal} \mathbb{E}\big[m_\Lambda\big(y_t, \Theta^{-1}(\theta), \Sigma_{\tilde{F},0}, \Omega_0; \mu_0\big)\big]\big|_{\theta=\theta_0}, \\
\mathcal{J}_\Sigma^\Lambda &= \frac{\partial}{\partial \eta^\intercal} \mathbb{E}\big[m_\Lambda\big(y_t, \Psi_0, \mathrm{diag}(\eta), \Omega_0; \mu_0\big)\big]\big|_{\eta=\eta_0}, \\
\mathcal{J}_\omega^\Lambda &= \frac{\partial}{\partial \omega^\intercal} \mathbb{E}\big[m_\Lambda\big(y_t, \Psi_0, \Sigma_{\tilde{F},0}, \mathrm{diag}(\omega); \mu_0\big)\big]\big|_{\omega=\omega_0}, \\
\mathcal{J}^\Lambda &= \begin{bmatrix} \mathcal{J}_\theta^\Lambda & \mathcal{J}_\Sigma^\Lambda & \mathcal{J}_\omega^\Lambda \end{bmatrix}.
\end{aligned}$$

The explicit formulas are given in Appendix A. $\widehat{W}$ converges in probability to W, which is positive definite. The matrices $(\mathcal{J}^\Lambda)^\intercal W \mathcal{J}^\Lambda$ and $(\mathcal{J}_\theta^\Lambda)^\intercal W \mathcal{J}_\theta^\Lambda$ are invertible.

The first condition ensures validity of the factor and aggregate moments. First-stage identification follows from Lemma 1 and the normalization (10). The second condition is standard in non-linear GMM models. The compactness assumption facilitates establishing consistency of the estimator. The third condition imposes weak stationarity and restricts the degree of serial correlation. Although we have focused on martingale difference sequences in the previous sections, the formulation of condition (iii) allows for serially correlated time series observations. The fourth condition is the instrument relevance condition for the scalar instrument $S^\intercal M_{\Psi_0} M_\iota y_t$. Under the exclusions in Assumption 2(i)–(ii), the formula in Appendix A gives $\mathcal{J}_p = (\psi_0 - \phi_0)^{-1} S^\intercal M_{\Psi_0} M_\iota \Omega_0 S$. Thus, under these exclusions, the fourth condition is equivalent to $S^\intercal M_{\Psi_0} M_\iota \Omega_0 S \neq 0$; it ensures that (11) identifies ψ_0 and that $\mathcal{J}_p^{-1}$ in (13) is well-defined. Under $\Omega_0 = \sigma_u^2 I$, we have

$S^\intercal M_{\Psi_0} M_\iota \Omega_0 S = \sigma_u^2 \|M_\Lambda S\|^2$, and the condition is equivalent to Assumption 2(iii). Under heteroscedasticity, however, the condition is stronger than Assumption 2(iii), which is the relevance condition for the vector instrument $M_\Lambda y_t$. Since $M_\Lambda \Omega_0$ is not symmetric in general, $S^\intercal M_\Lambda \Omega_0 S$ can be zero even if $M_\Lambda \Omega_0 S \neq 0$. The fifth condition imposes invertibility of some matrices in order to establish an asymptotic linear representation of the estimator of factor loadings.

Theorem 4. *Under Assumption 9, there exists a vector-valued function $\vartheta(y_t)$ such that $\mathbb{E}[\vartheta(y_t)] = 0$ and $\sqrt{T}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^T \vartheta(y_t) + o_{\mathbb{P}}(1)$. In addition, the asymptotic linear representation (13) holds.*

The functional form of $\vartheta(y)$ is given in Appendix A.3, and the proof is in Appendix C.5. Given the linear representation of the estimator, one can establish asymptotic normality of the estimator using an appropriate central limit theorem. We deliberately state the asymptotic result in this way to accommodate different sets of assumptions for asymptotic normality.

It is straightforward to form a plug-in estimator of the influence function $\mathcal{J}_p^{-1}[m_\psi(w_t, \psi_0, \Psi_0) + \mathcal{J}_\theta^\psi \, \vartheta(y_t)]$. See the next subsection and Appendix A.4 for plug-in estimators. Given an estimate of the influence function, we can estimate the asymptotic variance by placing appropriate restrictions on serial dependence of the observations. For instance, suppose that the time series observations of the influence function are serially uncorrelated (e.g., a martingale difference sequence). Then, one can estimate the asymptotic variance by computing

$$\frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{J}}_p^{-2} \left[m_\psi(w_t, \widehat{\psi}, \widehat{\Psi}) + \widehat{\mathcal{J}}_\theta^\psi \, \widehat{\vartheta}(y_t) \right]^2$$

where $\widehat{\ }$ indicates estimated objects. If researchers want to allow for serial correlation, then they should use a heteroscedasticity and autocorrelation consistent (HAC) estimator (e.g., Newey and West, 1987; Andrews, 1991).

4.3 Granular IV Estimator for Individual Equations

By Theorem 2, ϕ_0 is identified using weights $H \propto \Omega_0^{-1} \Lambda_0 \gamma$ for any $\gamma \in \mathbb{R}^r$ with $\iota^\intercal \Omega_0^{-1} \Lambda_0 \gamma \neq 0$, where $\Lambda_0 = [\ \iota \quad \Psi_0\ ]$ is the factor loading matrix, and Section 3.2.2 shows that γ can be chosen to minimize the asymptotic variance of the resulting IV estimator. The estimator we propose is the feasible counterpart of this recipe. We replace (Ψ_0, Ω_0) with the GMM estimates of Section 4.1, and we choose γ by the same formula as in Section 3.2.2, except that the matrix Q is replaced with a matrix Q_Υ that accounts for the first-step estimation of (Ψ_0, Ω_0).

Estimators. We base estimation of the parameter ϕ on the moment conditions (12) and

$$\mathbb{E}\left[m_\phi(w_t, \phi, \Psi_0, \Omega_0; \gamma)\right] = 0, \qquad m_\phi(w, \phi, \Psi, \Omega; \gamma) = S^\intercal M_\Psi M_\iota y \big(y - \phi p \iota\big)^\intercal \Omega^{-1} \Lambda \gamma, \tag{15}$$

where $\Lambda = [\ \iota \quad \Psi\ ]$. Under the normalization $\iota^\intercal \Omega_0^{-1} \Lambda_0 \gamma = 1$ of Theorem 2, we have $(y_t - \phi p_t \iota)^\intercal \Omega_0^{-1} \Lambda_0 \gamma = y_t^\intercal \Omega_0^{-1} \Lambda_0 \gamma - \phi p_t$, and (15) holds if and only if $\phi = \phi_0$.

Let $(\widehat{\Psi}, \widehat{\Omega})$ be the first-step estimates from Section 4.1 and $\widehat{\Lambda} = [\ \iota \quad \widehat{\Psi}\ ]$. For $\gamma \in \mathbb{R}^r$, let $\widehat{\phi}(\gamma)$ be the solution of

$$\frac{1}{T}\sum_{t=1}^{T} m_\phi\big(w_t, \widehat{\phi}(\gamma), \widehat{\Psi}, \widehat{\Omega}; \gamma\big) = 0, \quad \text{i.e.,} \quad \widehat{\phi}(\gamma) = \frac{\sum_{t=1}^{T} S^\intercal M_{\widehat{\Psi}} M_\iota y_t\, y_t^\intercal \widehat{\Omega}^{-1}\widehat{\Lambda}\gamma}{\big(\iota^\intercal \widehat{\Omega}^{-1}\widehat{\Lambda}\gamma\big)\sum_{t=1}^{T} S^\intercal M_{\widehat{\Psi}} M_\iota y_t\, p_t}.$$

Since $\widehat{\phi}(c\gamma) = \widehat{\phi}(\gamma)$ for any $c \neq 0$, the normalization of γ does not affect the estimates. Our granular IV estimator is computed in three steps.

1. *Initial estimator.* Let $\widehat{\phi}^{\text{ini}} = \widehat{\phi}(\gamma^{\text{ini}})$ for a fixed user-chosen $\gamma^{\text{ini}} \in \mathbb{R}^r$ satisfying $\iota^\intercal \Omega_0^{-1}\Lambda_0\gamma^{\text{ini}} \neq 0$. For example, $\gamma^{\text{ini}} = e_{1,r} = (1, 0, \ldots, 0)^\intercal \in \mathbb{R}^r$ always satisfies this condition and yields the weights $H \propto \widehat{\Omega}^{-1}\iota$ and $\widehat{\phi}^{\text{ini}}$ is the feasible version of the estimator of GK.

2. *Optimal weights.* Using $\widehat{\phi}^{\text{ini}}$, compute $\widehat{\Upsilon}_t \in \mathbb{R}^r$, $t = 1, \ldots, T$, defined in (19) below, $\widehat{Q}_\Upsilon = \frac{1}{T}\sum_{t=1}^{T} \widehat{\Upsilon}_t\widehat{\Upsilon}_t^\intercal$, and
$$\widehat{\gamma} = \frac{\widehat{Q}_\Upsilon^{-1}\widehat{\Lambda}^\intercal\widehat{\Omega}^{-1}\iota}{\iota^\intercal\widehat{\Omega}^{-1}\widehat{\Lambda}\widehat{Q}_\Upsilon^{-1}\widehat{\Lambda}^\intercal\widehat{\Omega}^{-1}\iota},$$
which minimizes $\widehat{V}(\gamma) = \widehat{\mathcal{J}}_p^{-2}\gamma^\intercal\widehat{Q}_\Upsilon\gamma$ over $\{\gamma \in \mathbb{R}^r : \iota^\intercal\widehat{\Omega}^{-1}\widehat{\Lambda}\gamma = 1\}$.

3. *Final estimator.* Let $\widehat{\phi} = \widehat{\phi}(\widehat{\gamma})$.

The formula for $\widehat{\gamma}$ is the formula for γ^* in Section 3.2.2 with (Λ, Ω, Q) replaced by $(\widehat{\Lambda}, \widehat{\Omega}, \widehat{Q}_\Upsilon)$.

Choice of $\widehat{\gamma}$. To motivate the second step, fix $\gamma \in \mathbb{R}^r$ normalized so that $\iota^\intercal\Omega_0^{-1}\Lambda_0\gamma = 1$. Theorem 5 below shows that $\widehat{\phi}(\gamma)$ admits the asymptotic linear representation

$$\sqrt{T}\big(\widehat{\phi}(\gamma) - \phi_0\big) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \mathcal{J}_p^{-1}\big[m_\phi(w_t, \phi_0, \Psi_0, \Omega_0; \gamma) + \mathcal{J}_\theta^\phi(\gamma)\vartheta(y_t) + \mathcal{J}_\omega^\phi(\gamma)\varsigma(y_t)\big] + o_{\mathbb{P}}(1). \tag{16}$$

Here, $\vartheta(y_t)$ is the influence function of $\widehat{\theta}$ from Theorem 4, $\varsigma(y_t)$ is the influence function of $\widehat{\omega}$, the vector of diagonal elements of $\widehat{\Omega}$ (see Appendix A.3), and

$$\mathcal{J}_\theta^\phi(\gamma) = \frac{\partial}{\partial\theta^\intercal}\mathbb{E}\big[m_\phi\big(w_t, \phi_0, \Theta^{-1}(\theta), \Omega_0; \gamma\big)\big]\big|_{\theta=\theta_0}, \qquad \mathcal{J}_\omega^\phi(\gamma) = \frac{\partial}{\partial\omega^\intercal}\mathbb{E}\big[m_\phi\big(w_t, \phi_0, \Psi_0, \operatorname{diag}(\omega); \gamma\big)\big]\big|_{\omega=\omega_0},$$

whose explicit formulas are given in Appendix A. The last two terms in (16) account for the first-step estimation of Ψ_0 and Ω_0. The derivative with respect to ϕ is $-\frac{\partial}{\partial\tilde{\phi}}\mathbb{E}[m_\phi(w_t, \tilde{\phi}, \Psi_0, \Omega_0; \gamma)]|_{\tilde{\phi}=\phi_0} = \mathbb{E}[S^\intercal M_{\Psi_0} M_\iota y_t p_t] = \mathcal{J}_p$, the same Jacobian as in the aggregate equation, because the normalization $\iota^\intercal\Omega_0^{-1}\Lambda_0\gamma = 1$ removes the dependence on γ.

The key observation is that the influence function in (16) is linear in γ: m_ϕ depends on γ only through $\Lambda_0\gamma$, and $\mathcal{J}_\theta^\phi(\gamma)$ and $\mathcal{J}_\omega^\phi(\gamma)$ are linear in γ. Hence, there exists a random vector $\Upsilon_t \in \mathbb{R}^r$, which does not depend on γ, such that

$$m_\phi(w_t, \phi_0, \Psi_0, \Omega_0; \gamma) + \mathcal{J}_\theta^\phi(\gamma)\vartheta(y_t) + \mathcal{J}_\omega^\phi(\gamma)\varsigma(y_t) = \gamma^\intercal\Upsilon_t \qquad \text{for all } \gamma \in \mathbb{R}^r. \tag{17}$$

The explicit form of Υ_t is given in Appendix A. If $\{\Upsilon_t\}$ is serially uncorrelated (e.g., a martingale difference sequence), the asymptotic variance of $\sqrt{T}(\widehat{\phi}(\gamma)-\phi_0)$ is $V(\gamma)=\mathcal{J}_p^{-2}\gamma^{\mathsf{T}}Q_\Upsilon\gamma$ with $Q_\Upsilon=\mathbb{E}[\Upsilon_t\Upsilon_t^{\mathsf{T}}]$. When Q_Υ is positive definite, the minimizer of $V(\gamma)$ subject to $\iota^{\mathsf{T}}\Omega_0^{-1}\Lambda_0\gamma=1$ has the closed form

$$\gamma^*=\frac{Q_\Upsilon^{-1}\Lambda_0^{\mathsf{T}}\Omega_0^{-1}\iota}{\iota^{\mathsf{T}}\Omega_0^{-1}\Lambda_0Q_\Upsilon^{-1}\Lambda_0^{\mathsf{T}}\Omega_0^{-1}\iota},\qquad V(\gamma^*)=\frac{\mathcal{J}_p^{-2}}{\iota^{\mathsf{T}}\Omega_0^{-1}\Lambda_0Q_\Upsilon^{-1}\Lambda_0^{\mathsf{T}}\Omega_0^{-1}\iota}. \tag{18}$$

The constraint set is never empty because the first element of $\Lambda_0^{\mathsf{T}}\Omega_0^{-1}\iota$ is $\iota^{\mathsf{T}}\Omega_0^{-1}\iota>0$. This is the minimization problem of Section 3.2.2 with $z_t=S^{\mathsf{T}}M_{\Lambda_0}y_t$, except that Q_Υ includes the contribution of the first-step estimation error. The vector $\widehat{\Upsilon}_t$ used in the second step is the plug-in estimate of Υ_t defined by

$$m_\phi\big(w_t,\widehat{\phi}^{\text{ini}},\widehat{\Psi},\widehat{\Omega};\gamma\big)+\widehat{\mathcal{J}}_\theta^\phi(\gamma)\widehat{\vartheta}(y_t)+\widehat{\mathcal{J}}_\omega^\phi(\gamma)\widehat{\varsigma}(y_t)=\gamma^{\mathsf{T}}\widehat{\Upsilon}_t\qquad\text{for all }\gamma\in\mathbb{R}^r \tag{19}$$

where $\widehat{\mathcal{J}}_\theta^\phi,\widehat{\mathcal{J}}_\omega^\phi,\widehat{\vartheta},\widehat{\varsigma}$ are plug-in estimates of $\mathcal{J}_\theta^\phi,\mathcal{J}_\omega^\phi,\vartheta,\varsigma$ (the explicit formulas are in Appendix A.4). The criterion $\widehat{V}(\gamma)=\widehat{\mathcal{J}}_p^{-2}\gamma^{\mathsf{T}}\widehat{Q}_\Upsilon\gamma$ estimates $V(\gamma)$, and $\widehat{\gamma}$ estimates γ^*. Since $\widehat{\mathcal{J}}_p$ does not depend on γ, its inclusion does not change the minimizing weights.

Remark 4.3. The choice γ^* in (18) minimizes the asymptotic variance of $\widehat{\phi}(\gamma)$ only when $\{\Upsilon_t\}$ is serially uncorrelated. Under serial correlation, $\widehat{\phi}$ remains consistent and Theorem 5 below continues to hold with γ^* defined in (18), but γ^* is no longer the minimizer of the asymptotic variance, and $\widehat{\phi}$ may be inefficient. In this case, the asymptotic variance of $\sqrt{T}(\widehat{\phi}(\gamma)-\phi_0)$ is $\mathcal{J}_p^{-2}\gamma^{\mathsf{T}}Q_\Upsilon^{\text{LR}}\gamma$, where

$$Q_\Upsilon^{\text{LR}}=\lim_{T\to\infty}\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Upsilon_t\right]=\sum_{h=-\infty}^{\infty}\mathbb{E}\big[\Upsilon_t\Upsilon_{t-h}^{\mathsf{T}}\big]$$

is the long-run variance of Υ_t. The objective remains a quadratic form in γ, so the efficient choice of γ is given by (18) with Q_Υ replaced by Q_Υ^{LR}. Accordingly, researchers who wish to target efficiency under serial correlation can replace $\widehat{Q}_\Upsilon$ in the second step with a HAC estimator of Q_Υ^{LR} (e.g., Newey and West, 1987; Andrews, 1991) computed from $\{\widehat{\Upsilon}_t\}_{t=1}^T$. Regardless of the choice of γ, inference under serial correlation should be based on a HAC estimator of the asymptotic variance, as discussed after Theorem 4. □

Asymptotic distribution. To formally state asymptotic properties of $\widehat{\phi}$, we impose the following conditions.

Assumption 10. *(i)* $\mathbb{E}[u_tF_t^{\mathsf{T}}]=0$.

(ii) The matrix $(\mathcal{J}_\omega^\Lambda)^{\mathsf{T}}W\mathcal{J}_\omega^\Lambda$ *is invertible.*

(iii) For Υ_t *defined in* (17), $\{\Upsilon_t\}_{t\geq1}$ *is weakly stationary,* $\frac{1}{T}\sum_{t=1}^T\Upsilon_t\Upsilon_t^{\mathsf{T}}\to_{\mathbb{P}}Q_\Upsilon$, *and* $Q_\Upsilon=\mathbb{E}[\Upsilon_t\Upsilon_t^{\mathsf{T}}]$ *is positive definite.*

The first condition ensures validity of the individual-equation moment. Assumption 6 already gives $\mathbb{E}[u_t\tilde{F}_t^{\mathsf{T}}]=0$, so the additional exclusion concerns F_{1t}. The second condition is used to establish an asymptotic linear representation of $\widehat{\omega}$. The third condition makes $V(\gamma)=\mathcal{J}_p^{-2}\gamma^{\mathsf{T}}Q_\Upsilon\gamma$ a positive definite quadratic form in γ, so that $V(\gamma)$ has the unique minimizer (18) over the affine set $\{\gamma\in\mathbb{R}^r:\iota^{\mathsf{T}}\Omega_0^{-1}\Lambda_0\gamma=1\}$.

Theorem 5. *Under Assumptions 9 and 10, there exists a vector-valued function* $\varsigma(y_t)$ *such that* $\mathbb{E}[\varsigma(y_t)] = 0$ *and* $\sqrt{T}(\widehat{\omega} - \omega_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varsigma(y_t) + o_{\mathbb{P}}(1)$. *In addition:*

(i) For any fixed $\gamma \in \mathbb{R}^r$ *with* $\iota^{\intercal} \Omega_0^{-1} \Lambda_0 \gamma = 1$, *the asymptotic linear representation (16) holds.*

(ii) Let γ^* *be the minimizer of* $V(\gamma)$ *over* γ *satisfying* $\iota^{\intercal} \Omega_0^{-1} \Lambda_0 \gamma = 1$, *given in (18). We have* $\widehat{\gamma} \to_{\mathbb{P}} \gamma^*$ *and*

$$\sqrt{T}\big(\widehat{\phi} - \phi_0\big) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathcal{J}_p^{-1} \big[m_\phi(w_t, \phi_0, \Psi_0, \Omega_0; \gamma^*) + \mathcal{J}_\theta^\phi(\gamma^*) \vartheta(y_t) + \mathcal{J}_\omega^\phi(\gamma^*) \varsigma(y_t) \big] + o_{\mathbb{P}}(1).$$

The proof is in Appendix C.6. Part (ii) shows that the influence function of $\widehat{\phi}$ is the same as that of the infeasible estimator $\widehat{\phi}(\gamma^*)$. In other words, the estimation error $\widehat{\gamma} - \gamma^*$ is of second order and does not contribute to the asymptotic variance. This insensitivity with respect to γ results from the derivative (with respect to γ) of the population moment being zero when evaluated at the true parameter values (e.g., Wooldridge, 2010, Ch.12), which can be seen from the discussion leading to Theorem 2.

Similar to Theorem 4, we characterize the influence function of $\widehat{\phi}$, and one can use an appropriate central limit theorem to deduce asymptotic normality of the GMM estimator. Estimation of the asymptotic variance is also straightforward. Appendix A.4 presents plug-in estimators of unknown parameters that are present in the influence function, and using these estimates, we can construct a plug-in estimator of the influence function. As discussed after Theorem 4, researchers can construct an estimate of the asymptotic variance under different assumptions on serial correlation.

5 Empirical Application

We apply the granular IV estimators of Section 4 to estimate the price elasticities in the global crude oil market. We compare our estimates with those based on the granular IV construction of GK and with the maximum likelihood estimation (MLE) of Baumeister and Hamilton (2023). We find that the estimates of supply and demand elasticities are comparable to those in the literature. The global short-run price elasticity of crude oil supply is 0.081, and the global short-run price elasticity of crude oil consumption is −0.069.

5.1 Data

We use the monthly crude oil production data published by the U.S. Energy Information Administration (EIA) online and in the *Monthly Energy Review*, which tracks output for various countries from 1973 onward. Our analysis uses the six top oil producing countries with consistent crude oil data since 1973 (United States, USSR, Saudi Arabia, China, Iran, and United Arab Emirates) and an aggregate category for the "Rest of the World" ($N = 7$), in that order, covering data up to May 2024. The EIA holds data only for the USSR before 1992. Hence, we aggregate the production of all successor countries of the USSR post-1992 into a single time series and call it USSR. Among the major oil producers, we drop Iraq, which shows high non-price volatility due to wars and conflicts. Following the literature (Baumeister and Hamilton, 2023), we assume that the

global crude oil supply has a factor structure with $r = 3$, i.e., the time fixed effect and two additional factors, so that $\Psi \in \mathbb{R}^{N\times 2}$. Assumption 8 requires, for each ℓ, two disjoint sets of r units other than ℓ, which is possible only if $N \geq 2r + 1$. This necessary condition for the identification of the factor loadings in Lemma 1 holds (with equality) in our application.

Crude oil production has two primary uses: it is either refined into petroleum products or stored as inventory. The EIA provides data on refined petroleum production, which typically exceeds the amount of crude oil input since multiple products are derived from oil. Hence, to estimate total crude oil consumption, we track changes in oil inventories. By calculating the difference between production and changes in inventory each month, we generate a time series of crude oil consumption over our study period.

For inventory data, we follow established methods from the literature (Hamilton, 2009a; Kilian and Murphy, 2014; Baumeister and Hamilton, 2019). The EIA has weekly inventory data for US crude oil stocks starting in 1982, but Baumeister and Hamilton (2019) extend this series back to April 1953 using the *Monthly Energy Review*. We incorporate this extended series, which allows us to push our data back to January 1973. We estimate the total oil stocks in OECD countries indirectly. Starting from January 1980, we have data from the EIA on the combined inventories of crude oil and petroleum products for both the US and the OECD. To estimate the total OECD crude oil stock, we calculate the ratio of combined inventories between the OECD and the US and assume that the same ratio applies to crude oil stocks. Since the data for the OECD's combined inventories only begin in January 1980, we extend this ratio backward, assuming the same value applies before that date. We take this calculated value of OECD crude oil inventory as a proxy for global crude oil inventory. This proxy suffers from measurement errors and excludes China. We are inherently assuming that the inventory response of non-OECD countries to price and other time-varying factors is the same as that of OECD countries.

We use the European Brent spot price as the price of crude oil. The EIA has a time series on Brent prices from May 1987 onwards, and Baumeister et al. (2022) extend it back to January 1973. For our dataset, we use the extension from 1973 to 1987 and the EIA series afterwards. To adjust for inflation, we deflate nominal WTI prices using the U.S. Consumer Price Index (CPIAUCSL, FRED), which tracks prices for all urban consumers, with 1982–84 as the base period.

We do not observe the consumption of crude oil in the data. We observe the total consumption of petroleum and petroleum products derived from crude oil. Crude oil refinement leads to multiple product streams, and the data on consumption of these oil products come from sources different from those on the production of crude oil. For these reasons, we compute the total consumption of crude oil, C_t, indirectly, using the identity

$$C_t = Q_t - \Delta I_t \tag{20}$$

where Q_t is the total production in a time period and ΔI_t is the change in the inventory level in the same period. We observe the production of country i in period t, Q_{it}, and the total production is $Q_t = \sum_{i=1}^{N} Q_{it}$. The identity (20) states that the total production of crude oil in a time period is used in two ways: it is either

consumed, in the sense that it is processed into petroleum and petroleum products for final consumption, or stored as inventory.

Covid As the Covid period was structurally different from any other period, we drop observations from March 2020 to January 2022. We also lose some observations due to the construction of year-on-year growth rates and the inclusion of lagged variables. Finally, we have $N = 7$ and $T = 579$.

5.2 Structural Model

We define the year-on-year growth in the production of crude oil for country i as

$$y_{it} = \frac{Q_{it} - Q_{i,t-12}}{.5(Q_{it} + Q_{i,t-12})}.$$

This definition of the growth rate takes into account seasonality in the production of crude oil. We define the average share of country i in total production as

$$S_i = \frac{1}{T}\sum_{t=1}^{T} \frac{Q_{it}}{Q_t}.$$

This is the share vector $S = (S_1, \ldots, S_N)^\intercal$ of the model (1)-(3). To see this, consider the year-on-year growth in the total global production of crude oil:

$$\begin{aligned}\frac{Q_t - Q_{t-12}}{.5(Q_t + Q_{t-12})} &= \frac{\sum_{i=1}^N Q_{it} - \sum_{i=1}^N Q_{i,t-12}}{.5(Q_t + Q_{t-12})} = \sum_{i=1}^N \left[\frac{Q_{it} - Q_{i,t-12}}{.5(Q_{it} + Q_{i,t-12})} \cdot \frac{Q_{it} + Q_{i,t-12}}{Q_t + Q_{t-12}}\right] \\ &\approx \sum_{i=1}^N S_i \frac{Q_{it} - Q_{i,t-12}}{.5(Q_{it} + Q_{i,t-12})} = S^\intercal y_t.\end{aligned}$$

Therefore, the year-on-year growth in the total global production of crude oil is the share-weighted average of the individual country growth rates. For the growth in consumption, recall that we calculate the level of consumption from the production and inventory data using (20). The year-on-year growth in the total global consumption of crude oil is

$$\begin{aligned}x_t := \frac{C_t - C_{t-12}}{.5(C_t + C_{t-12})} &= \frac{Q_t - Q_{t-12}}{.5(C_t + C_{t-12})} - \frac{\Delta I_t - \Delta I_{t-12}}{.5(C_t + C_{t-12})} \\ \approx \frac{Q_t - Q_{t-12}}{.5(Q_t + Q_{t-12})} - \frac{\Delta I_t - \Delta I_{t-12}}{.5(Q_t + Q_{t-12})} &\approx S^\intercal y_t - inv_t\end{aligned}$$

where inv_t is the year-on-year change in the change of inventory levels, relative to the average production. In our application, the market clearing condition (3) is thus replaced by $x_t = S^\intercal y_t - inv_t$.

We consider the model with exogenous controls of Appendix B. The structural equation that governs the production of crude oil in country i is

$$y_{it} = \phi p_t + \delta^\intercal c_t + \lambda_i^\intercal F_t + u_{it}, \qquad i = 1, \ldots, N, \tag{21}$$

where p_t is the year-on-year growth rate of the real oil price (defined as for the production growth) and $c_t = (x_{t-1}, x_{t-2})^\intercal$ contains two lags of the growth in total global consumption. The growth in consumption follows the structural model

$$x_t = \psi p_t + \beta^\intercal c_t + \varepsilon_t. \tag{22}$$

Thus, the individual equations (21) are the supply curves and the aggregate equation (22) is the demand curve, and we have two structural parameters to estimate: the supply elasticity ϕ and the demand elasticity ψ. For this section, we do not impose any structure on the change in inventories.

5.3 Results

We estimate the structural parameters using the granular IV estimators of Sections 4.2 and 4.3, with $N = 7$ and $r = 3$. The global share of crude oil production of each country varies slightly over time, and we use the average over the sample period. Our share vector S, reported in Table 1, indicates market concentration.

USSR	United States	Saudi Arabia	Iran	China	UAE	Rest of the World
0.166	0.120	0.123	0.055	0.046	0.035	0.455

Table 1: Share of individual units in global crude oil production

As discussed in Section 4.1, our GMM estimator of the factor loadings accommodates heteroscedasticity in the idiosyncratic errors, which we indeed detect in our data. We present the estimated variances of the within-transformed idiosyncratic supply shocks $M_\iota u_t$ of the countries in the sample in Table 2.

USSR	United States	Saudi Arabia	Iran	China	UAE	Rest of the World
22.05	46.37	41.94	495.11	23.87	71.77	21.25

Table 2: Estimated variances of the within-transformed idiosyncratic supply shocks

To estimate the supply elasticity ϕ, we need to aggregate the individual equations using an $N \times 1$ weighting vector. As explained in Section 4.3, we use the weighting vector $\widehat{H} = \widehat{\Omega}^{-1}\widehat{\Lambda}\widehat{\gamma}$, where $\widehat{\Lambda} = [\ \iota \quad \widehat{\Psi}\ ]$ and $\widehat{\Omega}$ are the first-step GMM estimates of the factor loadings and the covariance matrix of the idiosyncratic errors, and $\widehat{\gamma}$ is the plug-in estimate of the optimal γ^* in (18), obtained by minimizing $\widehat{V}(\gamma)$. To obtain the standard errors, we use Theorems 4 and 5, which account for the first-step estimation of Ψ_0 and Ω_0.

We summarize the estimates of the structural parameters in Table 3. We estimate the global short-run price elasticity of crude oil supply to be 0.081, closely aligning with prior estimates in the literature. For instance, Baumeister and Hamilton (2023) report a global supply elasticity of 0.054 using a maximum likelihood approach to granular instrumental variables. Similarly, we estimate the global short-run price elasticity of crude oil consumption to be -0.069, consistent with previous findings. Hamilton (2009b) reports short-run demand elasticities ranging from -0.05 to -0.07 based on a meta-analysis of earlier studies.

Baumeister and Hamilton (2023) estimate a global demand elasticity of −0.12 using MLE under a different specification.

Parameter	**GNR (GMM)**
Supply Elasticity	0.081
	(0.039)
Consumption Elasticity	-0.069
	(0.038)

Table 3: Estimation by GNR using GMM

We compare our estimation method with other methods using granular IV. First, we use principal component analysis (PCA) to estimate the factor loadings in place of GMM estimation. PCA is computationally simpler, but with fixed N it need not recover the loading space under unequal idiosyncratic variances. For inference with an estimated eigenspace, see Hahn et al. (2026, Section 3.2), with exogenous controls treated in their Section 4.2 and Appendix E. Their results require spherical idiosyncratic covariance, eigenvalue separation, relevance and time-series regularity conditions; they do not directly justify the present comparison under general cross-sectional heteroscedasticity. We retain PCA as an empirical benchmark without developing a separate PCA theorem. Table 4 presents the granular IV estimates using PCA. The reported standard errors use an eigenvector-perturbation correction for the estimated loadings and contemporaneous squared influence terms, without a HAC adjustment. They are benchmark calculations whose validity for structural inference is not established under unequal idiosyncratic variances or serially correlated influence functions. In this table, r counts the retained principal components after cross-sectional demeaning, in addition to the common direction ι. Column 1 retains three such components. The supply elasticity is substantially larger than our GMM estimate and the estimates typically reported in the literature. The demand elasticity is larger in magnitude than the GMM estimate and at the upper limit of those reported in the literature. With four retained components in column 2, the supply elasticity estimate is still larger than expected, while the demand elasticity estimate now falls within the range reported in the literature, albeit at the upper end.

Parameter	**GNR (PCA)** $r = 3$	**GNR (PCA)** $r = 4$
Supply Elasticity	0.140	0.112
	(0.075)	(0.065)
Consumption Elasticity	-0.090	-0.080
	(0.034)	(0.033)

Table 4: Estimation by GNR using PCA

We also estimate the parameters using the basic granular IV model of GK to illustrate the need to fully

account for the factor structure. In this case, we assume that only time fixed effects are present, i.e., $\Lambda = \iota$, so that $M_\Lambda = M_\iota$ and the granular instrument $S^\intercal M_\iota y_t$ is the difference between the share-weighted and equal-weighted supply growth. Following GK, we use the simplest weighting vector for the supply side, $H = \iota/N$, which corresponds to GK's weight $H \propto \Omega^{-1}\iota$ under the spherical error distribution $\Omega = \sigma_u^2 I$. Table 5 summarizes the results using this GK estimator. The supply elasticity is positive and significant, but notably larger than the values typically reported in the literature. The consumption (demand) elasticity is unexpectedly positive; however, due to the large standard error, this estimate is inconclusive.

Parameter	**GK**
Supply Elasticity	0.560
	(0.159)
Consumption Elasticity	0.019
	(0.023)

Table 5: Estimation by GK

We finally compare our estimates with the MLE of Baumeister and Hamilton (2023).

5.3.1 Comparison with Baumeister and Hamilton (2023)

In this subsection, we apply the MLE method of Baumeister and Hamilton (2023) to our empirical setting. We consider the same structural model as above, after partialling out the lagged controls c_t. The MLE requires a third structural equation to close the system. Following Baumeister and Hamilton (2023), we assume that the observed change in inventory is measured with error and has the structural form

$$inv_t = \kappa p_t + \zeta_t$$

where κ is the price elasticity of inventory change and ζ_t is a combination of an inventory shock and measurement error. Recalling $w_t = (y_t^\intercal, x_t, p_t)^\intercal$ and $inv_t = S^\intercal y_t - x_t$, we can express this system of three structural equations as

$$\mathbf{A}w_t = \boldsymbol{\nu}_t, \qquad \mathbf{A} = \begin{bmatrix} I_N & 0_{N,1} & -\phi\iota \\ 0_{1,N} & 1 & -\psi \\ S^\intercal & -1 & -\kappa \end{bmatrix}, \qquad \boldsymbol{\nu}_t = \begin{bmatrix} \Lambda F_t + u_t \\ \varepsilon_t \\ \zeta_t \end{bmatrix}$$

where $\mathbf{A}$ collects the structural parameters and $\boldsymbol{\nu}_t$ collects all the unobserved shocks.

Following Baumeister and Hamilton (2023), we assume that $\boldsymbol{\nu}_t \sim \mathcal{N}(0, \Sigma_\nu)$, where $\Sigma_\nu = \mathbb{E}[\boldsymbol{\nu}_t \boldsymbol{\nu}_t^\intercal]$. The MLE maximizes the log-likelihood function

$$\ell_T(\rho) = -\frac{T(N+2)}{2}\log 2\pi + \frac{T}{2}\log(|\mathbf{A}|^2) - \frac{T}{2}\log|\Sigma_\nu| - \frac{T}{2}\mathrm{tr}\left[\mathbf{A}^\intercal \Sigma_\nu^{-1}\mathbf{A}\widehat{\Sigma}_w\right] \tag{23}$$

where $\widehat{\Sigma}_w = \frac{1}{T}\sum_{t=1}^T w_t w_t^\intercal$ and ρ collects the unknown parameters specified below.

We give structure to Σ_ν based on the form of the errors. For this, we follow Baumeister and Hamilton (2023) and assume the following structure for the supply and demand shocks:

$$\Lambda F_t + u_t = b_1 f_{1t} + b_2 f_{2t} + u_t,$$
$$\varepsilon_t = h f_{1t} + \xi_t,$$

where $b_1, b_2 \in \mathbb{R}^N$ are factor loadings, h is a scalar, and $f_{1t}, f_{2t}, u_t, \xi_t, \zeta_t$ are mutually uncorrelated. In effect, we impose two additional assumptions on the form of the shocks. The first is that there are two factors and no time fixed effect. The second is that supply and demand shocks are correlated only through the common factor f_{1t}. Contrast this with our approach, in which we do not assume any parametric form for the correlation between ε_t and F_t. Under these assumptions,

$$\Sigma_\nu = \begin{bmatrix} b_1 b_1^\intercal + b_2 b_2^\intercal + \Omega & h b_1 & 0_{N,1} \\ h b_1^\intercal & h^2 + \sigma_\xi^2 & 0 \\ 0_{1,N} & 0 & \sigma_\zeta^2 \end{bmatrix} \tag{24}$$

where Ω is diagonal as in Assumption 7, and we use the normalizations $\mathbb{E}[f_{1t}^2] = \mathbb{E}[f_{2t}^2] = 1$.

Without loss of generality, we further impose that the two loading vectors are orthogonal, $b_1^\intercal b_2 = 0$. To implement this restriction, we define

$$G = \left[(b_1^\intercal b_1) I_N - b_1 b_1^\intercal\right] \begin{bmatrix} I_{N-1} \\ 0_{1,N-1} \end{bmatrix}.$$

We have $b_1^\intercal G = 0_{1,N-1}$, and we set $b_2 = Ga$ for a vector $a \in \mathbb{R}^{N-1}$. The parameters to estimate are the three structural parameters ϕ, ψ, and κ in $\mathbf{A}$, the N elements of b_1, the $N-1$ elements of a, the N diagonal elements of Ω, and the three scalars h, σ_ξ^2, and σ_ζ^2. That is, ρ has $3N+5$ elements, which equals 26 in our empirical setting with $N = 7$.

We summarize the results of this estimation in Table 6. The MLE point estimate of the consumption elasticity closely aligns with ours, but the MLE standard error is larger. The MLE supply elasticity is substantially larger, with a very large standard error. There are several possible reasons for the discrepancy. First, the MLE relies on the assumption that the unobserved shocks are normally distributed, and crude oil shocks may not be normally distributed. Second and more crucially, the MLE requires correct specification. We show in Appendix E that the MLE can severely underperform when the model is misspecified, leading to biased estimates and large standard errors, and the model may well be misspecified in our application.

Parameter	**MLE**
Supply Elasticity	0.185
	(0.396)
Consumption Elasticity	-0.070
	(0.058)

Table 6: Estimation by MLE

The comparisons illustrate the sensitivity of the estimates to the treatment of common factors and idiosyncratic covariance. Our GMM approach accommodates both heterogeneous factor loadings and unequal idiosyncratic variances under the stated identification conditions. Differences between the empirical estimates alone do not establish the bias or relative reliability of the competing methods.

Our method is also a robust alternative to the MLE method developed by Baumeister and Hamilton (2023). We explore the robustness of the MLE and GMM estimation in the granular IV context in Appendix E through Monte Carlo experiments. We find that the MLE is highly sensitive to misspecification of the covariance matrix, whereas our GMM method does not rely on this additional structure and continues to perform well.

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012): "The Network Origins of Aggregate Fluctuations," *Econometrica*, 80, 1977–2016.

Adrian, T., F. Grinberg, N. Liang, S. Malik, and J. Yu (2022): "The Term Structure of Growth-at-Risk," *American Economic Journal: Macroeconomics*, 14, 283–323.

Aldasoro, I., P. Beltrán, F. Grinberg, and T. Mancini-Griffoli (2023): "The macro-financial effects of international bank lending on emerging markets," *Journal of International Economics*, 142, 103733.

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, 525–587.

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

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Bai, J. and K. Li (2012): "Statistical analysis of factor models of high dimension," *The Annals of Statistics*, 40, 436 – 465.

Banafti, S. and T.-H. Lee (2022): "Inferential Theory for Granular Instrumental Variables in High Dimensions," *arXiv preprint arXiv:2201.06605*.

Baumeister, C. and J. D. Hamilton (2019): "Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Demand Shocks," *American Economic Review*, 109, 1873–1910.

——— (2023): "A Full-Information Approach to Granular Instrumental Variables," Tech. rep., Working Paper, UCSD.

Baumeister, C., D. Korobilis, and T. K. Lee (2022): "Energy markets and global economic conditions," *Review of Economics and Statistics*, 104, 828–844.

Bekker, P. A. and J. M. F. ten Berge (1997): "Generic Global Identification in Factor Analysis," *Linear Algebra and its Applications*, 264, 255–263.

Bippus, B., S. Lloyd, and D. Ostry (2023): "Granular Banking Flows and Exchange-Rate Dynamics," Staff Working Paper 1043, Bank of England.

Camanho, N., H. Hau, and H. Rey (2022): "Global Portfolio Rebalancing and Exchange Rates," *Review of Financial Studies*, 35, 5228–5274.

Cambanis, S., S. Huang, and G. Simons (1981): "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, 11, 368–385.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Chodorow-Reich, G., A. Ghent, and V. Haddad (2021): "Asset insulators," *The Review of Financial Studies*, 34, 1509–1539.

di Giovanni, J., A. A. Levchenko, and I. Méjean (2014): "Firms, Destinations, and Aggregate Fluctuations," *Econometrica*, 82, 1303–1340.

Dong, X., N. Kang, and J. Peress (2025): "Fast and Slow Arbitrage: The Predictive Power of (Persistent) Capital Flows for Factor Returns," *Review of Financial Studies*, 38, 2936–2987.

Flynn, J. P. and K. A. Sastry (2024): "The Macroeconomics of Narratives," Working Paper 32602, National Bureau of Economic Research.

Gabaix, X. (2011): "The granular origins of aggregate fluctuations," *Econometrica*, 79, 733–772.

Gabaix, X. and R. S. Koijen (2021): "In search of the origins of financial fluctuations: The inelastic markets hypothesis," Report, National Bureau of Economic Research.

Gabaix, X. and R. S. J. Koijen (2024): "Granular Instrumental Variables," *Journal of Political Economy*, 132, 2274–2303.

Galaasen, S., R. Jamilov, R. Juelsrud, and H. Rey (2020): "Granular Credit Risk," Tech. Rep. w27994, National Bureau of Economic Research, Cambridge, MA.

Gopalan Ramachandran, G. (2026): "Granular Instrumental Variables in Large Panels: Identification and Inference Across Strong, Nearly Weak, and Weak GIV," ArXiv:2607.02095.

Hahn, J., N. He, Z. Liao, and W. Zhou (2026): "Granular Instrumental Variables: Estimation and Inference," ArXiv:2606.14057.

Hamilton, J. D. (2009a): "Causes and Consequences of the Oil Shock of 2007-08," Tech. rep., National Bureau of Economic Research.

——— (2009b): "Understanding crude oil prices," *The energy journal*, 30, 179–206.

Holm-Hadulla, F. and C. Thürwächter (2024): "Granular shocks to corporate leverage and the macroeconomic transmission of monetary policy," Working Paper 2891, European Central Bank.

Kilian, L. and D. P. Murphy (2014): "The role of inventories and speculative trading in the global market for crude oil," *Journal of Applied Econometrics*, 29, 454–478.

Kundu, S. and N. Vats (2021): "Banking Networks and Economic Growth: From Idiosyncratic Shocks to Aggregate Fluctuations," Working Paper 128, European Systemic Risk Board.

Ling, D. C., C. Wang, and T. Zhou (2025): "Granular risks and stock returns: Evidence from commercial real estate," *Real Estate Economics*, 53, 1069–1104.

Ma, Y., T. Paligorova, and J.-L. Peydró (2021): "Expectations and Bank Lending," Working paper.

Ma, Y., K. Xiao, and Y. Zeng (2022): "Mutual Fund Liquidity Transformation and Reverse Flight to Liquidity," *Review of Financial Studies*, 35, 4674–4711.

Newey, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1993): "16 Efficient estimation of models with conditional moment restrictions," in *Econometrics*, Elsevier, vol. 11 of *Handbook of Statistics*, 419–454.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Elsevier, vol. 4, 2111–2245.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Qian, E. (2023): "Heterogeneity-robust granular instruments," *arXiv preprint arXiv:2304.01273*.

Schubert, G., A. Stansbury, and B. Taska (2022): "Employer Concentration and Outside Options," Working paper.

Williams, B. (2020): "Identification of the Linear Factor Model," *Econometric Reviews*, 39, 92–109.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, The MIT Press.

A Formulas Omitted in the Main Text

Here we collect the formulas of Jacobian matrices and the bijection Θ between $\mathcal{L}$ and $\mathbb{R}^{d_\theta}$ discussed in Section 4, the influence functions of the factor model estimator, and plug-in estimators of these objects.

A.1 Parametrization of Factor Loading Matrix

First we define additional notations. For $1 \leq i \leq l$, $l \in \mathbb{N}$, let $e_{i,l} \in \mathbb{R}^l$ be the ith element of the standard basis of $\mathbb{R}^l$. Let $0_{n,m}$ be the $n \times m$ zero matrix, $\mathcal{E}_\ell = \begin{bmatrix} 0_{(N-1-\ell),\ell} & I_{N-1-\ell} & 0_{(N-1-\ell),1} \end{bmatrix} \in \mathbb{R}^{(N-1-\ell)\times N}$, and

$$\mathcal{E} = \begin{bmatrix} \mathcal{E}_1 & \dots & 0_{N-2,N} \\ \vdots & \ddots & \vdots \\ 0_{N-r,N} & \dots & \mathcal{E}_{\underline{r}} \end{bmatrix} \in \mathbb{R}^{d_\theta \times N\underline{r}}$$

where the off-diagonal blocks are all zero matrices.

Now we can explicitly define the bijection Θ between $\mathcal{L}$ and $\mathbb{R}^{d_\theta}$.

$$\Theta(L) = \mathcal{E}\,\mathrm{vec}(L), \qquad L \in \mathbb{R}^{N\times \underline{r}}. \tag{25}$$

This map is linear on $\mathbb{R}^{N\times\underline{r}}$, and its restriction to $\mathcal{L}$ is a bijection onto $\mathbb{R}^{d_\theta}$, since an element of $\mathcal{L}$ is determined by its free entries.

Let $\Psi \in \mathcal{L}$, and $\theta = \Theta(\Psi)$. Differentiating $\mathrm{vec}(\Psi) = \mathrm{vec}(\Theta^{-1}(\theta))$ with respect to θ gives

$$\frac{\partial \mathrm{vec}(\Psi)}{\partial \theta^\intercal} = (\mathcal{E} + \mathcal{B})^\intercal, \qquad \mathcal{B} = \begin{bmatrix} -\mathcal{B}_1 & \dots & 0_{N-2,N} \\ \vdots & \ddots & \\ 0_{N-r,N} & & -\mathcal{B}_{\underline{r}} \end{bmatrix} \tag{26}$$

where $\mathcal{B}_\ell = \begin{bmatrix} 0_{(N-1-\ell),N-1} & \iota_{N-1-\ell} \end{bmatrix} \in \mathbb{R}^{(N-1-\ell)\times N}$ for $\ell = 1, \dots, \underline{r}$. The term $\mathcal{B}$ arises because the last row of an element in $\mathcal{L}$ is determined by the other rows.

A.2 Jacobian matrices

The Jacobians in the asymptotic linear representation (13) of $\widehat{\psi}$ are

$$\begin{aligned} \mathcal{J}_p &= \mathbb{E}[S^\intercal M_{\Psi_0} M_\iota y_t p_t], \\ \mathcal{J}_\theta^\psi &= -\mathbb{E}\left[\left(\varepsilon_t y_t^\intercal M_\iota \otimes S^\intercal\right)\left(I_{N^2} + K_{N,N}\right)\left(\Psi_0(\Psi_0^\intercal\Psi_0)^{-1} \otimes M_{\Psi_0}\right)\right](\mathcal{E}+\mathcal{B})^\intercal, \end{aligned}$$

where $K_{n,m}$ is the square matrix of size nm defined by the equalities $K_{n,m}\mathrm{vec}\,(A) = \mathrm{vec}\,(A^\intercal)$ for all $A \in \mathbb{R}^{n\times m}$. Under Assumption 9, the additional exclusion $\mathbb{E}[u_t F_{1t}] = 0$ gives the simplification

$$\mathcal{J}_p = \frac{1}{\psi_0 - \phi_0} S^\intercal M_{\Psi_0} M_\iota \Omega_0 S.$$

For the Jacobians of the factor model moment in Assumption 9, let $\mathcal{H} \in \mathbb{R}^{N(N-1)/2\times(N-1)^2}$ be the matrix such that $\mathcal{H}\text{vec}(A) = \text{vech}(A)$ for any $A \in \mathbb{R}^{(N-1)\times(N-1)}$, and let $\mathcal{G} = \mathcal{H}(\tilde{D} \otimes \tilde{D}) \in \mathbb{R}^{N(N-1)/2\times N^2}$. Then,

$$\begin{aligned}
\mathcal{J}_\theta^\Lambda &= -\mathcal{G}\big(I_{N^2} + K_{N,N}\big)\big\{(\Psi_0\Sigma_{\tilde{F},0}) \otimes I_N\big\}\,(\mathcal{E}+\mathcal{B})^\intercal\,,\\
\mathcal{J}_\Sigma^\Lambda &= -\mathcal{G}\big(\Psi_0 \otimes \Psi_0\big)\left[\begin{array}{ccc} \text{vec}(e_{1,\underline{r}}e_{1,\underline{r}}^\intercal) & \dots & \text{vec}(e_{\underline{r},\underline{r}}e_{\underline{r},\underline{r}}^\intercal)\end{array}\right],\\
\mathcal{J}_\omega^\Lambda &= -\mathcal{G}\left[\begin{array}{ccc} \text{vec}(e_{1,N}e_{1,N}^\intercal) & \dots & \text{vec}(e_{N,N}e_{N,N}^\intercal)\end{array}\right].
\end{aligned}$$

For the Jacobians of the moment (15) in Section 4.3, let $H_0(\gamma) = \Omega_0^{-1}\Lambda_0\gamma$ and let $\gamma_{-1} \in \mathbb{R}^{\underline{r}}$ denote the subvector of γ removing the first element. Then,

$$\begin{aligned}
\mathcal{J}_\theta^\phi(\gamma) &= -\Big(\mathbb{E}\big[H_0(\gamma)^\intercal(y_t-\phi_0 p_t\iota)y_t^\intercal M_\iota\big]\otimes S^\intercal\Big)\big(I_{N^2}+K_{N,N}\big)\Big(\Psi_0\big(\Psi_0^\intercal\Psi_0\big)^{-1}\otimes M_{\Psi_0}\Big)\,(\mathcal{E}+\mathcal{B})^\intercal\\
&\quad + S^\intercal M_{\Psi_0}\mathbb{E}[M_\iota y_t(y_t-\phi_0 p_t\iota)^\intercal]\Omega_0^{-1}(\gamma_{-1}^\intercal\otimes I_N)(\mathcal{E}+\mathcal{B})^\intercal,\\
\mathcal{J}_\omega^\phi(\gamma) &= -\mathbb{E}\big[S^\intercal M_{\Psi_0}M_\iota y_t\big((\Lambda_0\gamma)^\intercal\otimes(y_t-\phi_0 p_t\iota)^\intercal\big)\big](\Omega_0^{-1}\otimes\Omega_0^{-1})\left[\begin{array}{ccc} \text{vec}(e_{1,N}e_{1,N}^\intercal) & \dots & \text{vec}(e_{N,N}e_{N,N}^\intercal)\end{array}\right].
\end{aligned}$$

Using $(c^\intercal\otimes S^\intercal)\text{vec}(X) = S^\intercal Xc$ and $(\gamma_{-1}^\intercal\otimes I_N)\text{vec}(Z) = Z\gamma_{-1}$, each term of the influence function is linear in γ, and the vector Υ_t in (17) is

$$\begin{aligned}
\Upsilon_t = \Big\{&S^\intercal M_{\Psi_0}M_\iota y_t(y_t-\phi_0 p_t\iota)^\intercal\Omega_0^{-1}\Lambda_0 - S^\intercal\text{vec}^{-1}\big(\mathcal{A}_1\vartheta(y_t)\big)\mathbb{E}[M_\iota y_t(y_t-\phi_0 p_t\iota)^\intercal\Omega_0^{-1}\Lambda_0]\\
&+ S^\intercal M_{\Psi_0}\mathbb{E}[M_\iota y_t(y_t-\phi_0 p_t\iota)^\intercal]\Omega_0^{-1}\text{vec}^{-1}\big((\mathcal{E}+\mathcal{B})^\intercal\vartheta(y_t)\big)\left[\begin{array}{cc} 0_{\underline{r},1} & I_{\underline{r}}\end{array}\right]\\
&- \mathbb{E}\big[S^\intercal M_{\Psi_0}M_\iota y_t(y_t-\phi_0 p_t\iota)^\intercal\big]\text{vec}^{-1}\big(\mathcal{A}_2\varsigma(y_t)\big)\Lambda_0\Big\}^\intercal
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{A}_1 &= \big(I_{N^2}+K_{N,N}\big)\Big(\Psi_0\big(\Psi_0^\intercal\Psi_0\big)^{-1}\otimes M_{\Psi_0}\Big)\,(\mathcal{E}+\mathcal{B})^\intercal,\\
\mathcal{A}_2 &= (\Omega_0^{-1}\otimes\Omega_0^{-1})\left[\begin{array}{ccc} \text{vec}(e_{1,N}e_{1,N}^\intercal) & \dots & \text{vec}(e_{N,N}e_{N,N}^\intercal)\end{array}\right].
\end{aligned}$$

The four terms of $\Upsilon_t^\intercal\gamma$ correspond, in order, to $m_\phi(w_t,\phi_0,\Psi_0,\Omega_0;\gamma)$, the two terms of $\mathcal{J}_\theta^\phi(\gamma)\vartheta(y_t)$, and $\mathcal{J}_\omega^\phi(\gamma)\varsigma(y_t)$.

A.3 Influence Functions of the Factor Model Estimator

Here we give the influence functions ϑ and ς of $\widehat{\theta}$ and $\widehat{\omega}$ in Theorems 4 and 5. With η and ω the vectors of diagonal elements of Σ and Ω as in Assumption 9, we collect the parameters of the factor model in $\pi = (\theta^\intercal,\eta^\intercal,\omega^\intercal)^\intercal \in \mathbb{R}^{d_\theta+\underline{r}+N}$. Let

$$\mathcal{K} = \big((\mathcal{J}^\Lambda)^\intercal W\mathcal{J}^\Lambda\big)^{-1}(\mathcal{J}^\Lambda)^\intercal W \in \mathbb{R}^{(d_\theta+\underline{r}+N)\times N(N-1)/2},$$

and let $\Pi_\theta = \left[\begin{array}{cc} I_{d_\theta} & 0_{d_\theta,\underline{r}+N}\end{array}\right]$ and $\Pi_\omega = \left[\begin{array}{cc} 0_{N,d_\theta+\underline{r}} & I_N\end{array}\right]$ be the matrices that select θ and ω from π. The GMM estimator $\widehat{\pi}$ has the influence function $-\mathcal{K}\,m_\Lambda(y,\Psi_0,\Sigma_{\tilde{F},0},\Omega_0;\mu_0)$, and its θ- and ω-blocks are

$$\vartheta(y) = -\Pi_\theta\mathcal{K}\,m_\Lambda\big(y,\Psi_0,\Sigma_{\tilde{F},0},\Omega_0;\mu_0\big), \qquad \varsigma(y) = -\Pi_\omega\mathcal{K}\,m_\Lambda\big(y,\Psi_0,\Sigma_{\tilde{F},0},\Omega_0;\mu_0\big). \tag{27}$$

As noted after (13), the preliminary estimation of μ does not contribute to these influence functions.

A.4 Plug-in Estimators

We now collect plug-in estimators of the Jacobians and influence functions that appear in Section 4. Each replaces population moments by sample averages and the true parameter values by their estimates. Throughout, $(\widehat{\Psi}, \widehat{\Sigma}, \widehat{\Omega})$ is the GMM estimator of Section 4.1, $\widehat{\omega}$ is the vector of diagonal elements of $\widehat{\Omega}$, $\widehat{\mu} = \frac{1}{T}\sum_{t=1}^{T} M_\iota y_t$, $\widehat{\Lambda} = [\ \iota \quad \widehat{\Psi}\ ]$, and

$$\widehat{\varepsilon}_t = x_t - \widehat{\psi} p_t, \qquad \widehat{v}_t = y_t - \widehat{\phi}^{\text{ini}} p_t \iota, \qquad \widehat{\Xi} = \frac{1}{T}\sum_{t=1}^{T} M_\iota y_t \widehat{v}_t^{\intercal},$$

where $\widehat{\Xi}$ estimates $\mathbb{E}[M_\iota y_t (y_t - \phi_0 p_t \iota)^{\intercal}]$.

Jacobians. The Jacobian $\mathcal{J}_p$, common to the aggregate and individual equations, is estimated by

$$\widehat{\mathcal{J}}_p = \frac{1}{T}\sum_{t=1}^{T} S^{\intercal} M_{\widehat{\Psi}} M_\iota y_t p_t.$$

For the aggregate equation, let

$$\widehat{\mathcal{A}}_1 = \big(I_{N^2} + K_{N,N}\big)\Big(\widehat{\Psi}\big(\widehat{\Psi}^{\intercal}\widehat{\Psi}\big)^{-1} \otimes M_{\widehat{\Psi}}\Big)\,(\mathcal{E} + \mathcal{B})^{\intercal}, \qquad \widehat{\mathcal{J}}_\theta^{\psi} = -\bigg(\Big(\frac{1}{T}\sum_{t=1}^{T} \widehat{\varepsilon}_t y_t^{\intercal} M_\iota\Big) \otimes S^{\intercal}\bigg)\widehat{\mathcal{A}}_1.$$

The Jacobians of the factor model moment involve no expectations of the data, so we evaluate their formulas at the estimates:

$$\begin{aligned}
\widehat{\mathcal{J}}_\theta^{\Lambda} &= -\mathcal{G}\big(I_{N^2} + K_{N,N}\big)\big\{(\widehat{\Psi}\widehat{\Sigma}) \otimes I_N\big\}\,(\mathcal{E} + \mathcal{B})^{\intercal}, \\
\widehat{\mathcal{J}}_\Sigma^{\Lambda} &= -\mathcal{G}\big(\widehat{\Psi} \otimes \widehat{\Psi}\big)\begin{bmatrix} \operatorname{vec}(e_{1,\underline{r}} e_{1,\underline{r}}^{\intercal}) & \dots & \operatorname{vec}(e_{\underline{r},\underline{r}} e_{\underline{r},\underline{r}}^{\intercal}) \end{bmatrix},
\end{aligned}$$

and $\widehat{\mathcal{J}}^{\Lambda} = [\ \widehat{\mathcal{J}}_\theta^{\Lambda} \quad \widehat{\mathcal{J}}_\Sigma^{\Lambda} \quad \mathcal{J}_\omega^{\Lambda}\ ]$, where $\mathcal{J}_\omega^{\Lambda}$ is known. For the individual equations, let $\widehat{H}(\gamma) = \widehat{\Omega}^{-1}\widehat{\Lambda}\gamma$ and

$$\widehat{\mathcal{A}}_2 = \big(\widehat{\Omega}^{-1} \otimes \widehat{\Omega}^{-1}\big)\begin{bmatrix} \operatorname{vec}(e_{1,N} e_{1,N}^{\intercal}) & \dots & \operatorname{vec}(e_{N,N} e_{N,N}^{\intercal}) \end{bmatrix}.$$

Then,

$$\begin{aligned}
\widehat{\mathcal{J}}_\theta^{\phi}(\gamma) &= -\Big(\widehat{H}(\gamma)^{\intercal}\widehat{\Xi}^{\intercal} \otimes S^{\intercal}\Big)\widehat{\mathcal{A}}_1 + S^{\intercal} M_{\widehat{\Psi}} \widehat{\Xi}\,\widehat{\Omega}^{-1}(\gamma_{-1}^{\intercal} \otimes I_N)(\mathcal{E} + \mathcal{B})^{\intercal}, \\
\widehat{\mathcal{J}}_\omega^{\phi}(\gamma) &= -S^{\intercal} M_{\widehat{\Psi}}\Big((\widehat{\Lambda}\gamma)^{\intercal} \otimes \widehat{\Xi}\Big)\widehat{\mathcal{A}}_2,
\end{aligned}$$

where the second line uses $M_\iota y_t\big((\Lambda\gamma)^{\intercal} \otimes v^{\intercal}\big) = (\Lambda\gamma)^{\intercal} \otimes (M_\iota y_t v^{\intercal})$ for $v \in \mathbb{R}^N$. Under Assumptions 9 and 10(i), the sample averages of $M_\iota y_t y_t^{\intercal}$, $M_\iota y_t \varepsilon_t$, and $M_\iota y_t p_t$ converge in probability to their expectations, and thus, the above Jacobian estimators are consistent by the continuous mapping theorem.

Influence functions. Let $\widehat{\mathcal{K}} = \big((\widehat{\mathcal{J}}^{\Lambda})^{\intercal}\widehat{W}\widehat{\mathcal{J}}^{\Lambda}\big)^{-1}(\widehat{\mathcal{J}}^{\Lambda})^{\intercal}\widehat{W}$. The plug-in estimates of the influence functions (27) are

$$\widehat{\vartheta}(y_t) = -\Pi_\theta \widehat{\mathcal{K}}\, m_\Lambda\big(y_t, \widehat{\Psi}, \widehat{\Sigma}, \widehat{\Omega}; \widehat{\mu}\big), \qquad \widehat{\varsigma}(y_t) = -\Pi_\omega \widehat{\mathcal{K}}\, m_\Lambda\big(y_t, \widehat{\Psi}, \widehat{\Sigma}, \widehat{\Omega}; \widehat{\mu}\big).$$

The plug-in estimate of the influence function of $\widehat{\psi}$ in (13) is

$$\widehat{\mathcal{J}}_p^{-1}\big[m_\psi(w_t,\widehat{\psi},\widehat{\Psi})+\widehat{\mathcal{J}}_\theta^{\psi}\,\widehat{\vartheta}(y_t)\big],\qquad m_\psi(w_t,\widehat{\psi},\widehat{\Psi})=S^{\intercal}M_{\widehat{\Psi}}M_\iota y_t\widehat{\varepsilon}_t,$$

which yields the variance estimator given after Theorem 4. For the individual equations, substituting the above estimates into (19) gives the sample analogue of Υ_t:

$$\begin{aligned}\widehat{\Upsilon}_t=\Big\{&S^{\intercal}M_{\widehat{\Psi}}M_\iota y_t\widehat{v}_t^{\intercal}\widehat{\Omega}^{-1}\widehat{\Lambda}-S^{\intercal}\mathrm{vec}^{-1}\big(\widehat{\mathcal{A}}_1\widehat{\vartheta}(y_t)\big)\widehat{\Xi}\,\widehat{\Omega}^{-1}\widehat{\Lambda}\\&+S^{\intercal}M_{\widehat{\Psi}}\widehat{\Xi}\,\widehat{\Omega}^{-1}\mathrm{vec}^{-1}\big((\mathcal{E}+\mathcal{B})^{\intercal}\widehat{\vartheta}(y_t)\big)\begin{bmatrix}0_{\underline{r},1} & I_{\underline{r}}\end{bmatrix}-S^{\intercal}M_{\widehat{\Psi}}\widehat{\Xi}\,\mathrm{vec}^{-1}\big(\widehat{\mathcal{A}}_2\widehat{\varsigma}(y_t)\big)\widehat{\Lambda}\Big\}^{\intercal}.\end{aligned}$$

With $\widehat{Q}_\Upsilon=\frac{1}{T}\sum_{t=1}^T\widehat{\Upsilon}_t\widehat{\Upsilon}_t^{\intercal}$ and $\widehat{\gamma}$ as in Section 4.3, the plug-in estimate of the influence function of $\widehat{\phi}$ in Theorem 5 is $\widehat{\mathcal{J}}_p^{-1}\widehat{\gamma}^{\intercal}\widehat{\Upsilon}_t$ by (17). If the influence function is serially uncorrelated, we estimate $V(\gamma^*)$ in (18) by

$$\widehat{V}(\widehat{\gamma})=\widehat{\mathcal{J}}_p^{-2}\,\frac{1}{T}\sum_{t=1}^T\big(\widehat{\gamma}^{\intercal}\widehat{\Upsilon}_t\big)^2=\widehat{\mathcal{J}}_p^{-2}\,\widehat{\gamma}^{\intercal}\widehat{Q}_\Upsilon\widehat{\gamma}=\frac{\widehat{\mathcal{J}}_p^{-2}}{\iota^{\intercal}\widehat{\Omega}^{-1}\widehat{\Lambda}\widehat{Q}_\Upsilon^{-1}\widehat{\Lambda}^{\intercal}\widehat{\Omega}^{-1}\iota},$$

and otherwise, we apply a HAC estimator to $\{\widehat{\mathcal{J}}_p^{-1}\widehat{\gamma}^{\intercal}\widehat{\Upsilon}_t\}_{t=1}^T$. One may also replace $\widehat{\phi}^{\mathrm{ini}}$ by $\widehat{\phi}$ in $\widehat{v}_t$ when estimating the variance; the first equality still applies, but the last one no longer does because $\widehat{\gamma}$ was computed from $\widehat{Q}_\Upsilon$ based on $\widehat{\phi}^{\mathrm{ini}}$.

Remark A.1. The plug-in estimators admit two properties that are useful for computation. First, each Jacobian estimator coincides with the derivative of the corresponding sample moment function evaluated at the estimates, e.g.,

$$\widehat{\mathcal{J}}_\theta^{\psi}=\frac{\partial}{\partial\theta^{\intercal}}\frac{1}{T}\sum_{t=1}^T m_\psi\big(w_t,\widehat{\psi},\Theta^{-1}(\theta)\big)\Big|_{\theta=\widehat{\theta}},\qquad \widehat{\mathcal{J}}_\theta^{\phi}(\gamma)=\frac{\partial}{\partial\theta^{\intercal}}\frac{1}{T}\sum_{t=1}^T m_\phi\big(w_t,\widehat{\phi}^{\mathrm{ini}},\Theta^{-1}(\theta),\widehat{\Omega};\gamma\big)\Big|_{\theta=\widehat{\theta}},$$

so they can alternatively be computed, or checked, by numerical differentiation. Second, since the derivative of m_Λ with respect to π does not depend on the data, $\widehat{\mathcal{J}}^{\Lambda}$ is the Jacobian of the sample moment $\widehat{m}_\Lambda$ at $\widehat{\pi}$, and the first-order condition of the GMM problem in Section 4.1 reads $(\widehat{\mathcal{J}}^{\Lambda})^{\intercal}\widehat{W}\widehat{m}_\Lambda(\widehat{\Psi},\widehat{\Sigma},\widehat{\Omega};\widehat{\mu})=0$. Hence, $\frac{1}{T}\sum_{t=1}^T\widehat{\vartheta}(y_t)=0$ and $\frac{1}{T}\sum_{t=1}^T\widehat{\varsigma}(y_t)=0$ hold exactly. □

B Model with Exogenous Controls and Time-Varying Shares

We extend the model (1)-(3) considered in the main body of the paper. Let $c_t,\{v_{it}\}_{i=1}^N$ be control variables and we consider the model

$$x_t=\psi p_t+\beta^{\intercal}c_t+\varepsilon_t \tag{28}$$

$$y_{it}=\phi p_t+\delta^{\intercal}v_{it}+\lambda_i^{\intercal}F_t+u_{it},\qquad i=1,\dots,N, \tag{29}$$

$$x_t=\sum_{i=1}^N S_{it}y_{it} \tag{30}$$

where we allow for time-varying shares $S_t=(S_{1t}\dots S_{Nt})^{\intercal}$.

First we state a set of sufficient conditions for identification of ψ,ϕ.

Assumption 1′. *Let d_v be the dimension of v_{it} and $v_t = [v_{1t} \dots v_{Nt}]^\intercal$ be the $N \times d_v$ matrix.*

(i) $\psi \neq \phi$. $\Lambda \in \mathbb{R}^{N \times r}$ *is non-random and of full column rank. For* $t \geq 1$, $\sum_{i=1}^{N} S_{it} = 1$.

(ii) The time series $\{p_t, c_t, v_t, u_t, \varepsilon_t, F_t, S_t\}_{t \geq 1}$ *is a weakly stationary process with finite second moments. Researchers observe* $\{y_t, x_t, p_t, c_t, v_t, S_t\}_{t=1}^{T}$.

Assumption 2′. *(i)* $\mathbb{E}[\varepsilon_t | u_t, c_t, S_t] = 0$ *with probability one.*

(ii) $\mathbb{E}[F_t | u_t, v_t] = 0$ *with probability one.*

(iii) With probability one, $\mathbb{E}[u_t | v_t] = 0$ *and* $\mathbb{E}[u_t u_t^\intercal | S_t] = \mathbb{E}[u_t u_t^\intercal]$. Ω *is non-singular and* $M_\Lambda \Omega S_t \neq 0$ *with probability one.*

(iv) For some scalar $\sigma_\varepsilon > 0$, $\mathbb{E}[\varepsilon_t^2 | u_t, c_t, S_t] = \sigma_\varepsilon^2$ *with probability one.*

B.1 Construction with Preliminary Residualization

We construct a granular instrument for the aggregate equation (28) by first removing the common direction from the individual outcomes and then residualizing the transformed outcomes on the observed controls. Both steps can be implemented before estimating the factor loadings.

Premultiplying the individual equations by $M_\iota = I - \iota(\iota^\intercal \iota)^{-1} \iota^\intercal$ removes the price term:

$$M_\iota y_t = M_\iota v_t \delta + M_\iota \Lambda F_t + M_\iota u_t.$$

Assumption 2′(ii) implies $\mathbb{E}[F_t \mid v_t] = 0$ by iterated expectations. Together with $\mathbb{E}[u_t \mid v_t] = 0$ from Assumption 2′(iii), this gives $\mathbb{E}[M_\iota y_t \mid v_t] = M_\iota v_t \delta$ and

$$r_t := M_\iota y_t - \mathbb{E}[M_\iota y_t \mid v_t] = M_\iota \Lambda F_t + M_\iota u_t. \tag{31}$$

For $\mathbb{E}[M_\iota y_t \mid v_t]$, each component of $M_\iota y_t$ is conditioned on the same full control matrix v_t, and its conditional mean is linear in the entries of that matrix.

Since ι belongs to the column space of Λ, $M_\Lambda M_\iota = M_\Lambda$, and the granular instrument is

$$z_t := M_\Lambda r_t = M_\Lambda u_t. \tag{32}$$

For known Λ, the same instrument can be obtained by first projecting out the loading space and then residualizing on controls:

$$M_\Lambda r_t = M_\Lambda y_t - \mathbb{E}[M_\Lambda y_t \mid v_t].$$

Neither construction requires separate identification of δ; the conditional fitted values suffice even if the corresponding regression coefficients are not unique. In particular, common controls with common coefficients satisfy $v_t = \iota c_t^\intercal$ and are already annihilated by M_ι; in that case, $\mathbb{E}[M_\iota y_t \mid v_t] = 0$ and $z_t = M_\Lambda y_t$.

Estimation of the loading space. When Λ is unknown, the transformed residuals retain a factor structure that can be used to identify its column space. For this argument, maintain weak stationarity of the residualized factor model. Write $\Lambda = [\ \iota \quad \Psi\ ]$ with $M_\iota \Psi = \Psi$ as in Section 4. Then (31) gives

$$r_t = \Psi \tilde{F}_t + M_\iota u_t. \tag{33}$$

Assumption 2′(ii) also implies $\mathbb{C}[\tilde{F}_t, u_t] = 0$. Consequently,

$$\mathbb{V}[r_t] = \Psi \Sigma_{\tilde{F}} \Psi^\intercal + M_\iota \Omega M_\iota, \qquad \Sigma_{\tilde{F}} := \mathbb{V}[\tilde{F}_t]. \tag{34}$$

Thus, if $\Sigma_{\tilde{F}}$ is nonsingular and the diagonal-error and loading-rank restrictions of Assumptions 7-8 hold, Lemma 1 applies to the residualized factor model, after a suitable normalization. It identifies Ω and the column space of Ψ, and hence that of Λ. The factor covariance is unchanged by residualization because $\mathbb{E}[\tilde{F}_t \mid v_t] = 0$ under the maintained assumptions.

B.1.1 Instrument Validity in the Aggregate Equation

By (32), Assumption 2′(i) implies $\mathbb{E}[z_t \varepsilon_t] = 0$. The same assumption gives $\mathbb{E}[c_t \varepsilon_t] = 0$, so

$$\mathbb{E}\left[\begin{pmatrix} z_t \\ c_t \end{pmatrix} \left(x_t - \psi p_t - \beta^\intercal c_t \right) \right] = 0. \tag{35}$$

For identification of (ψ, β) from (35), it is sufficient that

$$\mathbb{E}\left[\begin{pmatrix} z_t \\ c_t \end{pmatrix} \begin{pmatrix} p_t & c_t^\intercal \end{pmatrix} \right] \tag{36}$$

has full column rank. When $Q_c := \mathbb{E}[c_t c_t^\intercal]$ is nonsingular, this condition is equivalent to

$$\mathbb{E}[z_t p_t^\perp] \neq 0, \qquad p_t^\perp := p_t - \mathbb{E}[p_t c_t^\intercal] Q_c^{-1} c_t. \tag{37}$$

Thus, the granular instrument must be correlated with the component of price not explained by the aggregate controls. Under this condition, ψ is uniquely identified.

B.1.2 Instrument Validity in the Individual Equations

Fix a nonrandom $H \in \mathbb{R}^N$ satisfying $H^\intercal \iota = 1$ and $\Omega H \in \mathrm{Im}(\Lambda)$, as in Section 3.2. The weighted equation is

$$H^\intercal y_t = \phi p_t + \delta^\intercal v_t^\intercal H + H^\intercal (\Lambda F_t + u_t).$$

Assumption 2′(ii)–(iii) implies $\mathbb{E}[H^\intercal (\Lambda F_t + u_t) \mid v_t] = 0$, so $v_t^\intercal H$ is a valid instrument. For $z_t = M_\Lambda u_t$, the same assumptions give

$$\mathbb{E}[z_t H^\intercal (\Lambda F_t + u_t)] = M_\Lambda \mathbb{E}[u_t F_t^\intercal] \Lambda^\intercal H + M_\Lambda \Omega H = 0.$$

Thus the instruments satisfy

$$\mathbb{E}\left[\begin{pmatrix} z_t \\ v_t^\intercal H \end{pmatrix} \left(H^\intercal y_t - \phi p_t - \delta^\intercal v_t^\intercal H \right) \right] = 0.$$

Identification of (ϕ, δ) follows from the same rank condition as in the aggregate equation, (36), with c_t replaced by $v_t^\intercal H$. If $\mathbb{E}[v_t^\intercal H H^\intercal v_t]$ is nonsingular, this condition is equivalent to

$$\mathbb{E}[z_t p_{H,t}^\perp] = \mathbb{E}[z_t p_t] \neq 0,$$

where $p_{H,t}^\perp$ is the residual from the population linear projection of p_t on $v_t^\intercal H$. The equality follows from $\mathbb{E}[z_t H^\intercal v_t] = 0$ under Assumption 2′(iii).

The aggregate condition (37) does not admit the simplification $\mathbb{E}[z_t p_t^\perp] = \mathbb{E}[z_t p_t]$ in general, since the assumptions allow $\mathbb{E}[z_t c_t^\intercal] \neq 0$.

B.2 Optimal Instrument

We derive optimal instruments for ψ and ϕ when the coefficients on the controls are nuisance parameters. We treat the loading space and the population residualization used to construct z_t as known; for the individual equations, we also fix the aggregation weights H. The efficiency comparison is relative to the conditional moment specified for each equation below, under independent observations or time-series conditions under which the admissible instrumented errors are serially uncorrelated and the usual central limit theorem applies. General serial dependence requires a long-run variance analysis.

B.2.1 A Review of Optimal Instruments with Nuisance Parameters

Consider the conditional moment model

$$Y_t = \theta_0 p_t + \eta_0^\intercal W_t + e_t, \qquad \mathbb{E}[e_t \mid Z_t] = 0, \tag{38}$$

where θ_0 is scalar and η_0 is a finite-dimensional nuisance parameter. Define

$$a(Z_t) := \mathbb{E}[p_t \mid Z_t], \qquad b(Z_t) := \mathbb{E}[W_t \mid Z_t], \qquad s^2(Z_t) := \mathbb{E}[e_t^2 \mid Z_t].$$

Assume $s^2(Z_t) > 0$ almost surely and the weighted second moments below are finite. Also assume that the matrix

$$Q_b := \mathbb{E}\left[\frac{b(Z_t) b(Z_t)^\intercal}{s^2(Z_t)}\right]$$

is positive definite. The conditional-moment efficiency result of Chamberlain (1987); Newey (1993) gives the joint optimal instrument, up to an invertible constant transformation and sign, as

$$g^*(Z_t) = \frac{1}{s^2(Z_t)} \begin{pmatrix} a(Z_t) \\ b(Z_t) \end{pmatrix}.$$

Applying the conditional-moment efficiency result of Chamberlain (1987) to the full parameter vector and using the standard partitioned-information argument yields the optimal instrument for the scalar parameter of interest; see Newey (1990, p. 107, equation (10)). In the present model, this amounts to projecting the target instrument on the nuisance directions with inverse conditional variance weights. Let

$$\kappa := Q_b^{-1} \mathbb{E}\left[\frac{b(Z_t) a(Z_t)}{s^2(Z_t)}\right], \qquad a_\perp(Z_t) := a(Z_t) - b(Z_t)^\intercal \kappa. \tag{39}$$

The resulting scalar instrument is

$$h^*(Z_t) = \frac{a_\perp(Z_t)}{s^2(Z_t)}. \tag{40}$$

B.2.2 Aggregate Equation

Let $Z_{\psi,t} = (z_t^\intercal, c_t^\intercal, S_t^\intercal)^\intercal$, where $z_t = M_\Lambda u_t$. By iterated expectations, Assumption 2′(i) and (iv) imply

$$\mathbb{E}[x_t - \psi p_t - \beta^\intercal c_t \mid Z_{\psi,t}] = 0, \qquad \mathbb{V}[\varepsilon_t \mid Z_{\psi,t}] = \sigma_\varepsilon^2. \tag{41}$$

Set $a_{\psi,t} = \mathbb{E}[p_t \mid Z_{\psi,t}]$, and suppose that $Q_c = \mathbb{E}[c_t c_t^\intercal]$ is positive definite. Define the residual from the population linear projection of the conditional price predictor on controls:

$$a_{\psi,t}^\perp := a_{\psi,t} - c_t^\intercal \kappa_\psi, \qquad \kappa_\psi := Q_c^{-1}\mathbb{E}[c_t a_{\psi,t}], \qquad \mathbb{E}[(a_{\psi,t}^\perp)^2] > 0. \tag{42}$$

The last inequality is the required relevance after removing controls. The condition $M_\Lambda \Omega S_t \neq 0$ in Assumption 2′(iii) alone does not establish this residual relevance in the model with controls. Applying (40) with $(Y_t, \theta, \eta, W_t) = (x_t, \psi, \beta, c_t)$ yields

$$h_{\psi,t}^* = \frac{a_{\psi,t}^\perp}{\sigma_\varepsilon^2} \propto a_{\psi,t}^\perp. \tag{43}$$

If $\psi \neq 0$, an equivalent expression is

$$h_{\psi,t}^* \propto S_t^\intercal \mathbb{E}[\widetilde{y}_t \mid Z_{\psi,t}], \qquad \widetilde{y}_t := y_t - \iota c_t^\intercal Q_c^{-1}\mathbb{E}[c_t x_t]. \tag{44}$$

This has the same form as the instrument in Section 3, with y_t replaced by $\widetilde{y}_t$. To verify the equivalence, equilibrium and the conditional moment imply

$$m_{\psi,t} := S_t^\intercal \mathbb{E}[y_t \mid Z_{\psi,t}] = \mathbb{E}[x_t \mid Z_{\psi,t}] = \psi a_{\psi,t} + \beta^\intercal c_t. \tag{45}$$

Since (c_t, S_t) is measurable with respect to $Z_{\psi,t}$, $\mathbb{E}[c_t x_t] = \mathbb{E}[c_t m_{\psi,t}]$ and $S_t^\intercal \iota = 1$ give

$$S_t^\intercal \mathbb{E}[\widetilde{y}_t \mid Z_{\psi,t}] = m_{\psi,t} - c_t^\intercal Q_c^{-1}\mathbb{E}[c_t m_{\psi,t}] = \psi a_{\psi,t}^\perp,$$

which establishes (44).

B.2.3 Individual Equations

As in Section 3.2.1, we strengthen unconditional instrument validity to a conditional moment restriction. Fix a nonrandom H satisfying $H^\intercal \iota = 1$ and $\Omega H \in \mathrm{Im}(\Lambda)$, and let $Z_{\phi,t} = (z_t^\intercal, \mathrm{vec}(v_t)^\intercal)^\intercal$. We impose

$$\mathbb{E}[H^\intercal y_t - \phi p_t - \delta^\intercal v_t^\intercal H \mid Z_{\phi,t}] = \mathbb{E}[H^\intercal u_t \mid Z_{\phi,t}] = 0, \tag{46}$$

where the equality follows from Assumption 2′(ii). Furthermore, writing $e_{H,t} := H^\intercal(\Lambda F_t + u_t)$, we assume conditional homoscedasticity:

$$\mathbb{V}[e_{H,t} \mid Z_{\phi,t}] = \sigma_{\phi,H}^2 > 0. \tag{47}$$

Following Section B.2.2, set $a_{\phi,t} = \mathbb{E}[p_t \mid Z_{\phi,t}]$ and suppose that $Q_H := \mathbb{E}[v_t^\intercal H H^\intercal v_t]$ is positive definite. Define

$$a^\perp_{\phi,H,t} := a_{\phi,t} - H^\intercal v_t \kappa_{\phi,H}, \qquad \kappa_{\phi,H} := Q_H^{-1}\mathbb{E}[v_t^\intercal H a_{\phi,t}], \tag{48}$$

and assume $\mathbb{E}[(a^\perp_{\phi,H,t})^2] > 0$. Since $v_t^\intercal H$ is measurable with respect to $Z_{\phi,t}$, applying (40) with $(Y_t, \theta, \eta, W_t) = (H^\intercal y_t, \phi, \delta, v_t^\intercal H)$ gives

$$h^*_{\phi,H,t} = \frac{a^\perp_{\phi,H,t}}{\sigma^2_{\phi,H}} \propto a^\perp_{\phi,H,t}. \tag{49}$$

Thus, the optimal instrument is the conditional price predictor after projecting out the weighted controls. Under heteroskedasticity, the same argument applies using the inverse conditional variance weights in (39)–(40).

If $\phi \neq 0$, an equivalent outcome-predictor expression is

$$h^*_{\phi,H,t} \propto H^\intercal \mathbb{E}[\widetilde{y}_{H,t} \mid Z_{\phi,t}], \qquad \widetilde{y}_{H,t} := y_t - \iota H^\intercal v_t Q_H^{-1}\mathbb{E}[v_t^\intercal H H^\intercal y_t]. \tag{50}$$

Indeed, (46) gives $H^\intercal \mathbb{E}[y_t \mid Z_{\phi,t}] = \phi a_{\phi,t} + \delta^\intercal v_t^\intercal H$. Projecting both sides on $v_t^\intercal H$ and subtracting the fitted values yields $H^\intercal \mathbb{E}[\widetilde{y}_{H,t} \mid Z_{\phi,t}] = \phi a^\perp_{\phi,H,t}$, just as in Section B.2.2.

In Section 3.2.1, the optimal instrument is invariant to H and coincides with the aggregate instrument. With controls, the projection on $v_t^\intercal H$ generally depends on H, and the two equations can have different conditioning sets and control spaces.

Although the preceding derivation allows time-varying shares, consider first the special case of constant shares and common controls $v_t = \iota c_t^\intercal$. Then $v_t^\intercal H = c_t$, the two conditioning sets contain the same information, and $a^\perp_{\phi,H,t} = a^\perp_{\psi,t}$. Under the respective conditional exclusion and homoscedasticity restrictions, the optimal instruments therefore coincide up to scale; if $\psi \neq 0$, both are proportional to $S^\intercal \mathbb{E}[\widetilde{y}_t \mid Z_{\psi,t}]$ in (44). Without controls and with constant shares, this recovers the form in Section 3.2.1. For time-varying shares, including S_t in $Z_{\phi,t}$ requires imposing the individual conditional exclusion and variance restriction on this larger conditioning set.

B.3 Estimation and Inference with Unknown Factor Loadings

We extend Section 4 to the model with controls, keeping N and the number of factors fixed. The estimators use linear granular instruments; the population efficiency results in Section B.2 do not establish their efficiency when the loadings are estimated. Throughout this subsection, subscript 0 denotes the true parameter value.

B.3.1 Preliminary Residualization and Factor Estimation

Set $b_t = (1, \mathrm{vec}(v_t)^\intercal)^\intercal$, retaining all entries even if some are redundant. This vector is random, with dimension fixed as $T \to \infty$. Let $Q_b := \mathbb{E}[b_t b_t^\intercal]$, which may be singular, and write A^+ for the Moore–Penrose inverse of

a matrix A. Define $B_0 = \mathbb{E}[M_\iota y_t b_t^\intercal] Q_b^+$. By (31),

$$M_\iota y_t = B_0 b_t + r_t, \qquad \mathbb{E}[r_t \mid v_t] = 0, \qquad M_\iota B_0 = B_0.$$

With $\widehat{Q}_b = \frac{1}{T}\sum_{t=1}^T b_t b_t^\intercal$, estimate the conditional mean by multivariate least squares:

$$\widehat{B} = \left(\frac{1}{T}\sum_{t=1}^T M_\iota y_t b_t^\intercal\right) \widehat{Q}_b^+, \qquad \widehat{r}_t = M_\iota y_t - \widehat{B} b_t. \tag{51}$$

In implementation, each component of $M_\iota y_t$ is regressed on the same vector b_t, with separate coefficients. The pseudoinverse selects a coefficient matrix even when coefficients are not unique; the fitted values and residuals are unique. Researchers therefore need not choose a nonredundant basis. For example, if $v_t = \iota c_t$ for a scalar common control, one can retain $b_t = (1, c_t, \ldots, c_t)^\intercal$: the fitted values are identical to those from regressing on $(1, c_t)^\intercal$.

This estimates the fitted values without requiring separate identification of δ_0 from the within-transformed equations. Since b_t includes a constant, $\frac{1}{T}\sum_{t=1}^T \widehat{r}_t = 0$.

Use the parameterization $\pi = (\theta^\intercal, \eta^\intercal, \omega^\intercal)^\intercal$ and the normalized factor model of Section 4.1. Replace its covariance moment by

$$g_t(\pi, B) = \operatorname{vech}\Big(\tilde{D}\left[(M_\iota y_t - B b_t)(M_\iota y_t - B b_t)^\intercal - \Psi\Sigma\Psi^\intercal - M_\iota \Omega M_\iota\right]\tilde{D}^\intercal\Big), \tag{52}$$

and minimize

$$\left(\frac{1}{T}\sum_{t=1}^T g_t(\pi, \widehat{B})\right)^\intercal \widehat{W} \left(\frac{1}{T}\sum_{t=1}^T g_t(\pi, \widehat{B})\right)$$

over the compact parameter set $\mathcal{P}$ in Assumption 9.

The following sufficient conditions cover both structural estimators below. They strengthen the time-series conditions, but do not require the conditional homoscedasticity in Assumption 2′(iv).

Assumption 9′. *(i) Assumption 1′ and Assumption 2′(i)–(iii) hold. The residualized factor model satisfies Assumptions 7– 8, with $\Sigma_{\tilde{F},0}$ positive definite and the normalization (10). The interiority and weighting-matrix/Jacobian conditions in Assumption 9(ii),(v) hold for this factor model.*

(ii) The observed process is strictly stationary and ergodic, $\|S_t\|$ is bounded almost surely, and $\mathbb{E}[\|d_t\|^4] < \infty$, where $d_t = (1, y_t^\intercal, x_t, p_t, c_t^\intercal, b_t^\intercal)^\intercal$.

(iii) With $\mathcal{X}_t = (\operatorname{vec}(d_t d_t^\intercal)^\intercal, [S_t \otimes \operatorname{vec}(d_t d_t^\intercal)]^\intercal)^\intercal$,

$$\limsup_{T\to\infty} \left\| \mathbb{V}\left[T^{-1/2}\sum_{t=1}^T \mathcal{X}_t\right]\right\| < \infty.$$

The vector $\mathcal{X}_t$ collects the products needed for the sample moments and their derivatives. These conditions allow serial dependence; independent observations with the stated moment bounds are a sufficient special case. Relevance conditions are stated separately for each equation.

Allowing singular Q_b does not require a separate rank assumption here. If a is in its null space, then $\mathbb{E}[(a^\intercal b_t)^2] = 0$, so $a^\intercal b_t = 0$ almost surely. Thus the range of $\widehat{Q}_b$ is contained in that of Q_b. Since $\widehat{Q}_b \to_{\mathbb{P}} Q_b$ and the population distribution and dimension are fixed, their ranks agree with probability approaching one, and $\widehat{Q}_b^+ \to_{\mathbb{P}} Q_b^+$. On this event, the least-squares normal equations yield

$$\widehat{B} - B_0 = \left(\frac{1}{T}\sum_{t=1}^{T} r_t b_t^\intercal\right)\widehat{Q}_b^+.$$

This gives the same expansion as with a nonredundant basis, expressed using the selected coefficient matrices.

Write $g_{0t} = g_t(\pi_0, B_0)$ and use $\mathcal{K}$, Π_θ and Π_ω from Appendix A.3. The factor Jacobian is unchanged, since the parameter-dependent covariance in (52) is the same. Define

$$\ell_{B,t} = \operatorname{vec}(r_t b_t^\intercal Q_b^+), \qquad \vartheta_t = -\Pi_\theta \mathcal{K} g_{0t}, \qquad \varsigma_t = -\Pi_\omega \mathcal{K} g_{0t}. \tag{53}$$

Under Assumption 9′,

$$\begin{aligned}\sqrt{T}\operatorname{vec}(\widehat{B} - B_0) &= T^{-1/2}\sum_{t=1}^{T}\ell_{B,t} + o_{\mathbb{P}}(1),\\ \sqrt{T}(\widehat{\theta} - \theta_0) = T^{-1/2}\sum_{t=1}^{T}\vartheta_t + o_{\mathbb{P}}(1), &\qquad \sqrt{T}(\widehat{\omega} - \omega_0) = T^{-1/2}\sum_{t=1}^{T}\varsigma_t + o_{\mathbb{P}}(1).\end{aligned} \tag{54}$$

In particular, estimating B_0 does not add a first-order term to the factor estimator. To see this, the least-squares normal equations give the exact identity

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{r}_t\widehat{r}_t^\intercal = \frac{1}{T}\sum_{t=1}^{T} r_t r_t^\intercal - (\widehat{B} - B_0)\left(\frac{1}{T}\sum_{t=1}^{T} b_t b_t^\intercal\right)(\widehat{B} - B_0)^\intercal.$$

The last term is $O_{\mathbb{P}}(T^{-1})$. Identification follows from the residualized factor model in Section B.1; the consistency and GMM expansion in Appendix C.5 then apply with g_{0t} in place of the original covariance moment.

B.3.2 Aggregate Equation

Write $P(\theta) = M_{[\iota,\Theta^{-1}(\theta)]}$ and $P_0 = P(\theta_0)$. Define

$$h_{\psi,t}(\theta, B) = S_t^\intercal P(\theta)(M_\iota y_t - B b_t), \qquad h_{\psi,t} = h_{\psi,t}(\theta_0, B_0) = S_t^\intercal z_t.$$

The estimator $(\widehat{\psi}, \widehat{\beta})$ solves

$$\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}\widehat{h}_{\psi,t}\\ c_t\end{pmatrix}(x_t - \widehat{\psi} p_t - \widehat{\beta}^\intercal c_t) = 0, \qquad \widehat{h}_{\psi,t} = h_{\psi,t}(\widehat{\theta}, \widehat{B}). \tag{55}$$

Suppose $Q_c = \mathbb{E}[c_t c_t^\intercal]$ is positive definite and let

$$h_{\psi,t}^\perp = h_{\psi,t} - \mathbb{E}[h_{\psi,t} c_t^\intercal] Q_c^{-1} c_t, \qquad \mathcal{J}_{p,\psi} = \mathbb{E}[h_{\psi,t}^\perp p_t] \neq 0. \tag{56}$$

This is the scalar-instrument counterpart of (37); vector relevance alone does not guarantee relevance of this particular aggregation. Equivalently,

$$\widehat{\psi} = \frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{h}^{\perp}_{\psi,t} x_t}{\frac{1}{T}\sum_{t=1}^{T} \widehat{h}^{\perp}_{\psi,t} p_t},$$

where $\widehat{h}^{\perp}_{\psi,t}$ is the sample residual from regressing $\widehat{h}_{\psi,t}$ on c_t.

Let derivatives below be evaluated at (θ_0, B_0), and set

$$\begin{aligned} \mathcal{J}_{\theta,\psi} &= \mathbb{E}\left[\varepsilon_t \frac{\partial h_{\psi,t}(\theta, B)}{\partial \theta^{\intercal}}\right], \\ \mathcal{J}_{B,\psi} &= \mathbb{E}\left[\varepsilon_t \frac{\partial h_{\psi,t}(\theta, B)}{\partial \mathrm{vec}(B)^{\intercal}}\right] = -\mathbb{E}[\varepsilon_t (b_t^{\intercal} \otimes S_t^{\intercal} P_0)]. \end{aligned} \tag{57}$$

Under Assumption 9′ and (56),

$$\sqrt{T}(\widehat{\psi} - \psi_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathcal{J}_{p,\psi}^{-1} \left[h^{\perp}_{\psi,t}\varepsilon_t + \mathcal{J}_{\theta,\psi}\vartheta_t + \mathcal{J}_{B,\psi}\ell_{B,t}\right] + o_{\mathbb{P}}(1). \tag{58}$$

Indeed, eliminate the control equations in (55) and expand its remaining moment at the truth. The derivative in β vanishes because $\mathbb{E}[h^{\perp}_{\psi,t} c_t^{\intercal}] = 0$; estimating the projection of $h_{\psi,t}$ on c_t has no first-order effect because $\mathbb{E}[c_t \varepsilon_t] = 0$. Substitution of (54) gives the result.

Unlike the factor covariance moment, the aggregate moment need not be orthogonal to preliminary residualization. Assumption 2′(i) does not impose exclusion conditional on v_t, so $\mathcal{J}_{B,\psi}$ need not vanish. The stronger restriction $\mathbb{E}[\varepsilon_t \mid v_t, c_t, S_t] = 0$ is sufficient to eliminate this term, but is not required for (58). In particular, the maintained aggregate exclusion already eliminates this term for common controls $v_t = \iota c_t^{\intercal}$.

B.3.3 Individual Equations

First choose a fixed nonrandom aggregation vector $s \in \mathbb{R}^N$ and set

$$h_{\phi,t}(\theta, B) = s^{\intercal} P(\theta)(M_\iota y_t - B b_t), \qquad h_{\phi,t} = s^{\intercal} z_t, \qquad \mathcal{J}_{p,\phi} = \mathbb{E}[h_{\phi,t} p_t] \neq 0.$$

For constant shares, $s = S$ recovers the instrument in Section 4. With time-varying shares, a fixed s preserves validity under the existing exclusions, provided this scalar relevance condition holds. Using S_t instead requires the additional conditions stated below.

As in Section 4, let

$$H(\theta, \omega; \gamma) = \Omega^{-1}[\iota, \Psi]\gamma, \qquad H_0 = H(\theta_0, \omega_0; \gamma), \qquad a_\gamma = \iota^{\intercal} H_0 \neq 0,$$

where γ is fixed. We allow unnormalized weights and retain their normalization in the denominator; H_0/a_γ satisfies the normalization in Section B.1. Let $\widehat{H} = H(\widehat{\theta}, \widehat{\omega}; \gamma)$ and $\widehat{h}_{\phi,t} = h_{\phi,t}(\widehat{\theta}, \widehat{B})$. Define

$$\widehat{\phi}(\gamma) = \frac{\left(\frac{1}{T}\sum_{t=1}^{T} \widehat{h}_{\phi,t} y_t^{\intercal}\right) \widehat{H}}{(\iota^{\intercal}\widehat{H}) \left(\frac{1}{T}\sum_{t=1}^{T} \widehat{h}_{\phi,t} p_t\right)}. \tag{59}$$

The absence of δ from this expression is deliberate. The normal equations in (51) imply $\frac{1}{T}\sum_{t=1}^{T}\widehat{h}_{\phi,t}b_t^{\intercal}=0$, and hence $\frac{1}{T}\sum_{t=1}^{T}\widehat{h}_{\phi,t}v_t=0$. Thus (59) also solves the sample moment with y_t replaced by $y_t-v_t\delta$ for any δ. Neither separate estimation of δ_0 nor nonsingularity of the weighted-control covariance is needed for this estimator of ϕ_0.

Write $e_t=y_t-\phi_0 p_t\iota-v_t\delta_0=\Lambda_0F_t+u_t$. With all derivatives evaluated at the truth, define

$$\begin{aligned}\mathcal{J}_{\theta,\phi}(\gamma)&=\mathbb{E}\left[(e_t^{\intercal}H_0)\frac{\partial h_{\phi,t}}{\partial\theta^{\intercal}}+h_{\phi,t}e_t^{\intercal}\frac{\partial H}{\partial\theta^{\intercal}}\right],\\ \mathcal{J}_{\omega,\phi}(\gamma)&=\mathbb{E}\left[h_{\phi,t}e_t^{\intercal}\frac{\partial H}{\partial\omega^{\intercal}}\right].\end{aligned}\tag{60}$$

Under Assumption 9′ and the stated normalization and relevance conditions,

$$\sqrt{T}\big(\widehat{\phi}(\gamma)-\phi_0\big)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(a_\gamma\mathcal{J}_{p,\phi})^{-1}\left[h_{\phi,t}e_t^{\intercal}H_0+\mathcal{J}_{\theta,\phi}(\gamma)\vartheta_t+\mathcal{J}_{\omega,\phi}(\gamma)\varsigma_t\right]+o_{\mathbb{P}}(1).\tag{61}$$

For the proof, expand

$$\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{h}_{\phi,t}(y_t-\phi_0p_t\iota-v_t\delta_0)^{\intercal}\right)\widehat{H}.$$

Its population value is zero by the individual moment validity in Section B.1. Moreover,

$$\mathcal{J}_{B,\phi}(\gamma)=-\mathbb{E}[(e_t^{\intercal}H_0)(b_t^{\intercal}\otimes s^{\intercal}P_0)]=0,$$

because $\mathbb{E}[e_t\mid v_t]=0$. Divide the expansion by the consistent denominator in (59) and apply (54). This yields (61), including the effect of estimating both the loading space and the weights.

Using time-varying shares. To use $h_{\phi,t}=S_t^{\intercal}z_t$, a sufficient strengthening of Assumption 2′(ii)–(iii) is

$$\mathbb{E}[F_t\mid u_t,v_t,S_t]=0,\qquad\mathbb{E}[u_t\mid v_t,S_t]=0,\tag{62}$$

while retaining $\mathbb{E}[u_tu_t^{\intercal}\mid S_t]=\Omega_0$. These restrictions imply $\mathbb{E}[h_{\phi,t}e_t^{\intercal}H_0]=0$, $\mathbb{E}[h_{\phi,t}v_t]=0$, and $\mathcal{J}_{B,\phi}(\gamma)=0$. However, the sample orthogonality $\frac{1}{T}\sum_{t=1}^{T}\widehat{h}_{\phi,t}v_t=0$ no longer holds automatically. Instead estimate (ϕ,δ) jointly from

$$\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}\widehat{h}_{\phi,t}\\ v_t^{\intercal}\widehat{H}\end{pmatrix}\widehat{H}^{\intercal}(y_t-\widehat{\phi}p_t\iota-v_t\widehat{\delta})=0.\tag{63}$$

If $\mathbb{E}[v_t^{\intercal}H_0H_0^{\intercal}v_t]$ is positive definite and $\mathcal{J}_{p,\phi}\neq0$, the representation (61) continues to hold with s replaced by S_t in the instrument and its derivatives. Estimation of δ_0 has no first-order effect on the granular moment because $\mathbb{E}[h_{\phi,t}v_t^{\intercal}H_0]=0$.

B.3.4 Feasible Inference and Choice of Weights

Let $L_{\psi,t}$ and $L_{\phi,t}(\gamma)$ denote the summands, including their denominators, in (58) and (61). They have mean zero. Suppose additionally that $T^{-1/2}\sum_{t=1}^{T}(\mathcal{X}_t-\mathbb{E}[\mathcal{X}_t])$ converges to a centered normal distribution whose covariance is the limit of its covariance matrices. Then the estimators are jointly asymptotically normal with

covariance equal to the long-run covariance of $(L_{\psi,t}, L_{\phi,t}(\gamma))^\intercal$. The variance bound in Assumption 9′(iii) alone does not assert this central limit theorem.

For implementation, replace expectations and parameters in the influence functions and Jacobians by sample averages and consistent estimates. Use $\widehat{\varepsilon}_t = x_t - \widehat{\psi} p_t - \widehat{\beta}^\intercal c_t$. Even when δ_0 is not separately identified, a feasible residual for e_t is

$$\widehat{e}_t = y_t - \widehat{\phi} p_t \iota - \left(\frac{1}{T} \sum_{s=1}^{T} (y_s - \widehat{\phi} p_s \iota) b_s^\intercal \right) \widehat{Q}_b^{+} b_t. \tag{64}$$

Indeed, $\mathbb{E}[e_t b_t^\intercal] = 0$ and $v_t \delta_0$ lies in the span of b_t. For $L_t = (L_{\psi,t}, L_{\phi,t}(\gamma))^\intercal$, the stated moment conditions give $\frac{1}{T}\sum_{t=1}^{T} \|\widehat{L}_t - L_t\|^2 = o_{\mathbb{P}}(1)$. For serially uncorrelated influence functions, their sample covariance consistently estimates the asymptotic covariance. With serial dependence, use a consistent HAC estimator under the additional dependence and bandwidth conditions required for it (Newey and West, 1987; Andrews, 1991). Divide by T to estimate the covariance of the estimators themselves.

Finally, the numerator of (61) is linear in γ. Write it as $\gamma^\intercal \Upsilon_t^c$ and let $a_0 = \Lambda_0^\intercal \Omega_0^{-1} \iota$. If the long-run covariance Q_Υ^c of Υ_t^c is positive definite, the variance-minimizing weights for this fixed choice of granular instrument are

$$\gamma^* = \frac{(Q_\Upsilon^c)^{-1} a_0}{a_0^\intercal (Q_\Upsilon^c)^{-1} a_0}, \qquad a_0^\intercal \gamma^* = 1. \tag{65}$$

As in Section 4, obtain consistent plug-in weights using an initial estimator, for example $\gamma = e_{1,r}$, for which $a_0^\intercal \gamma = \iota^\intercal \Omega_0^{-1} \iota > 0$. If $\widehat{\gamma} \to_{\mathbb{P}} \gamma^*$, substitution of $\widehat{\gamma}$ adds no first-order term: the population granular moment is zero for every γ, and its corrected sample numerator is linear in γ. Thus (61) holds at γ^* for the final estimator. For the time-varying-share implementation, the weighted-control rank condition must hold at the initial and limiting weights. This variance comparison concerns the choice of H within the specified linear-instrument procedure. Without controls and with constant shares, setting $s = S$ and $a_\gamma = 1$ reduces the two linear representations to those in Theorems 4 and 5.

C Proofs

C.1 Proof of Lemma 1

Notation For $n \in \mathbb{N}_+$, ι_n is the vector of ones in $\mathbb{R}^n$, I_n is the $n \times n$ identity matrix, and $e_{i,n}$ is the ith canonical basis of $\mathbb{R}^n$. □

For $\ell \in \{1, \ldots, N\}$, define the $(N-1) \times N$ matrix $\tilde{D}_\ell$ by inserting $-\iota_{N-1}$ into I_{N-1} between the $\ell - 1$th and ℓth columns. For example,

$$\tilde{D}_2 = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & -1 & 0 & \ldots & 1 \end{bmatrix}, \qquad \tilde{D}_N = \begin{bmatrix} I_{N-1} & -\iota_{N-1} \end{bmatrix}.$$

We have $\tilde{D}_\ell M_\iota = \tilde{D}_\ell$ for $\ell = 1, \ldots, N$, and

$$\tilde{D}_\ell \mathbb{V}[M_\iota y_t] \tilde{D}_\ell^\intercal = \Psi_\ell \Sigma_{\tilde{F}} \Psi_\ell^\intercal + \tilde{D}_\ell \Omega \tilde{D}_\ell^\intercal, \qquad \Psi_\ell = \tilde{D}_\ell \Psi$$

where the ith row of Ψ_ℓ is $(\psi_i - \psi_\ell)^\intercal$ if $i < \ell$ and $(\psi_{i+1} - \psi_\ell)^\intercal$ if $i \geq \ell$ and

$$\tilde{D}_\ell \Omega \tilde{D}_\ell^\intercal = \sum_{i=1}^{N} \sigma_i^2 \tilde{D}_\ell e_{i,N} e_{i,N}^\intercal \tilde{D}_\ell^\intercal.$$

Then, for $\ell = 1, \ldots, N$,

$$\tilde{D}_\ell \Omega \tilde{D}_\ell^\intercal = \sum_{i=1}^{\ell-1} \sigma_i^2 e_{i,N-1} e_{i,N-1}^\intercal + \sum_{i=\ell}^{N-1} \sigma_{i+1}^2 e_{i,N-1} e_{i,N-1}^\intercal + \sigma_\ell^2 \iota_{N-1} \iota_{N-1}^\intercal.$$

Note that off-diagonal elements of $\tilde{D}_\ell \Omega \tilde{D}_\ell^\intercal$ equal σ_ℓ^2. Let $\mathcal{I}_\ell$ and $\mathcal{I}'_\ell$ be subsets of $\{1, \ldots, N\}$ that appear in Assumption 8. Let Σ_ℓ be the $r \times r$ submatrix of $\tilde{D}_\ell \mathbb{V}[M_\iota y_t] \tilde{D}_\ell^\intercal$ whose rows correspond to indices in $\mathcal{I}_\ell$ and columns to indices in $\mathcal{I}'_\ell$. Then, for $s \in \mathbb{R}$,

$$\Sigma_\ell - s\iota_r \iota_r^\intercal = \Psi_{\ell,\mathcal{I}_\ell} \Sigma_{\tilde{F}} \Psi_{\ell,\mathcal{I}'_\ell}^\intercal + (\sigma_\ell^2 - s)\iota_r \iota_r^\intercal = \begin{bmatrix} \Psi_{\ell,\mathcal{I}_\ell} & \iota_r \end{bmatrix} \begin{bmatrix} \Sigma_{\tilde{F}} & 0 \\ 0 & \sigma_\ell^2 - s \end{bmatrix} \begin{bmatrix} \Psi_{\ell,\mathcal{I}'_\ell}^\intercal \\ \iota_r^\intercal \end{bmatrix}$$

and we have

$$\det\left(\Sigma_\ell - s\iota_r \iota_r^\intercal\right) = (\sigma_\ell^2 - s)\det\left(\begin{bmatrix} \Psi_{\ell,\mathcal{I}_\ell} & \iota_r \end{bmatrix}\right) \det(\Sigma_{\tilde{F}}) \det\left(\begin{bmatrix} \Psi_{\ell,\mathcal{I}'_\ell} & \iota_r \end{bmatrix}\right).$$

Since the rows of Ψ_ℓ are $(\psi_i - \psi_\ell)^\intercal$, $\Psi_{\ell,\mathcal{I}_\ell} = \Psi_{\mathcal{I}_\ell} - \iota_r \psi_\ell^\intercal$ and

$$\det\left(\begin{bmatrix} \Psi_{\ell,\mathcal{I}_\ell} & \iota_r \end{bmatrix}\right) = \det\left(\begin{bmatrix} \Psi_{\mathcal{I}_\ell} & \iota_r \end{bmatrix}\right) = (-1)^{r-1} \det\left(\begin{bmatrix} \iota_r & \Psi_{\mathcal{I}_\ell} \end{bmatrix}\right) = (-1)^{r-1} \det(\Lambda_{\mathcal{I}_\ell}).$$

An analogous calculation holds for the determinant involving $\Psi_{\ell,\mathcal{I}'_\ell}$. Thus,

$$\det\left(\Sigma_\ell - s\iota_r \iota_r^\intercal\right) = (\sigma_\ell^2 - s)\det(\Lambda_{\mathcal{I}_\ell}) \det(\Sigma_{\tilde{F}}) \det(\Lambda_{\mathcal{I}'_\ell})$$

and the left-hand side is zero if and only if $\sigma_\ell^2 = s$ because by the hypothesis, $\det(\Lambda_{\mathcal{I}_\ell}) \det(\Sigma_{\tilde{F}}) \det(\Lambda_{\mathcal{I}'_\ell})$ is non-zero.

The above argument establishes the identification of $\Omega = \mathrm{diag}(\sigma_1^2 \ldots \sigma_N^2)$, and thus,

$$\Psi \Sigma_{\tilde{F}} \Psi^\intercal = \mathbb{V}[M_\iota y_t] - M_\iota \Omega M_\iota^\intercal$$

is identified. Under (10), Ψ is uniquely pinned down from $\Psi \Sigma_{\tilde{F}} \Psi^\intercal$ (e.g., Anderson and Rubin, 1956, Lemma 5.1).

C.2 Proof of Theorem 1

Proof. Write $a_t = \mathbb{E}[p_t \mid Z_t]$. Assumption 2(i) and (iv) gives $\mathbb{E}[\varepsilon_t \mid Z_t] = 0$ and $\mathbb{E}[\varepsilon_t^2 \mid Z_t] = \sigma_\varepsilon^2$. Since $x_t = S^\intercal y_t = \psi p_t + \varepsilon_t$, we have $h_t = \psi a_t$. Moreover, the equilibrium calculation in Section 2.2.1 yields

$$\mathbb{E}[Z_t p_t] = \frac{M_\Lambda \Omega S}{\psi - \phi} \neq 0.$$

Consequently, a_t cannot vanish almost surely. Jensen's inequality and finite second moments imply $0 < \mathbb{E}[a_t^2] < \infty$, and hence

$$J := \mathbb{E}[h_t p_t] = \psi \mathbb{E}[a_t^2] \neq 0.$$

Feasible estimator. By stationarity and ergodicity, $T^{-1}\sum_t h_t p_t \to_{\mathbb{P}} J$ and $T^{-1}\sum_t p_t^2 = O_{\mathbb{P}}(1)$. Cauchy–Schwarz and the preliminary-estimation conditions give

$$\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{h}_t - h_t)p_t\right| \le \left\{\frac{1}{T}\sum_{t=1}^{T}(\widehat{h}_t - h_t)^2\right\}^{1/2}\left\{\frac{1}{T}\sum_{t=1}^{T}p_t^2\right\}^{1/2} = o_{\mathbb{P}}(1).$$

Thus the feasible denominator converges to J. The exact IV identity and the other preliminary-estimation condition imply

$$\sqrt{T}(\widehat{\psi} - \psi) = \frac{T^{-1/2}\sum_t \widehat{h}_t\varepsilon_t}{T^{-1}\sum_t \widehat{h}_t p_t} = J^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_t\varepsilon_t + o_{\mathbb{P}}(1).$$

Under Assumption 4, $h_t\varepsilon_t$ is a strictly stationary ergodic martingale difference with variance $\sigma_\varepsilon^2\mathbb{E}[h_t^2] < \infty$. The martingale central limit theorem therefore gives asymptotic normality with variance

$$\frac{\sigma_\varepsilon^2\mathbb{E}[h_t^2]}{J^2} = \frac{\sigma_\varepsilon^2}{\mathbb{E}[a_t^2]}.$$

Efficiency comparison. For any square-integrable instrument $g(Z_t)$ with $\mathbb{E}[g(Z_t)p_t] \neq 0$, the same ergodic and martingale arguments give asymptotic variance

$$V(g) = \frac{\sigma_\varepsilon^2\mathbb{E}[g(Z_t)^2]}{\mathbb{E}[g(Z_t)a_t]^2} \ge \frac{\sigma_\varepsilon^2}{\mathbb{E}[a_t^2]},$$

where the inequality is Cauchy–Schwarz, with equality for $g(Z_t) \propto a_t$. This also bounds regular GMM estimators based on any finite instrument vector $q(Z_t)$. Indeed, writing $d = \mathbb{E}[q(Z_t)p_t]$ and W for the probability limit of the GMM weight matrix, their first-order influence function is $g_q(Z_t)\varepsilon_t$, where

$$g_q(Z_t) = \frac{d^{\mathsf{T}}Wq(Z_t)}{d^{\mathsf{T}}Wd}, \qquad \mathbb{E}[g_q(Z_t)p_t] = 1.$$

The preceding bound applies to g_q. Since $h_t = \psi a_t$, the feasible estimator attains it. Taking $g(Z_t) = S^{\mathsf{T}}Z_t$ proves the comparison with the GK estimator. □

C.3 Proof of Theorem 2

Proof. Because $\Omega = \mathbb{E}[u_t u_t^{\mathsf{T}}]$ is nonsingular, it is positive definite. The first column of Λ is ι, so $e_{1,r} \in \mathcal{C}$ since $\iota^{\mathsf{T}}\Omega^{-1}\iota > 0$; in particular, the admissible class is nonempty. For any $\gamma \in \mathcal{C}$, write $d_\gamma = \iota^{\mathsf{T}}\Omega^{-1}\Lambda\gamma$. Then

$$\iota^{\mathsf{T}}H_\gamma = 1, \qquad \Omega H_\gamma = \Lambda\gamma/d_\gamma.$$

The normalization removes the unknown scale of Ω, and the class of these weights depends only on the known column space of Λ. Using $M_\Lambda y_t = M_\Lambda u_t$ and the factor exclusion gives

$$\begin{aligned}\mathbb{E}\big[M_\Lambda y_t H_\gamma^{\mathsf{T}}(y_t - \phi p_t\iota)\big] &= M_\Lambda\big(\mathbb{E}[u_t F_t^{\mathsf{T}}]\Lambda^{\mathsf{T}} + \Omega\big)H_\gamma \\ &= M_\Lambda\Lambda\gamma/d_\gamma = 0.\end{aligned}$$

Since $\iota^{\mathsf{T}}H_\gamma = 1$, this establishes the stated moment equation. For any candidate value $b \in \mathbb{R}$, it also implies

$$\mathbb{E}\big[M_\Lambda y_t(H_\gamma^{\mathsf{T}}y_t - bp_t)\big] = (\phi - b)\mathbb{E}[M_\Lambda y_t p_t].$$

The vector on the right vanishes only when $b = \phi$, by the relevance condition. Thus the moment uniquely identifies ϕ for every $\gamma \in \mathcal{C}$. □

C.4 Proof of Theorem 3

Proof. Let $a_t = \mathbb{E}[p_t \mid Z_t]$. The relevance argument in Appendix C.2 gives $0 < \mathbb{E}[a_t^2] < \infty$. Since $\iota^\intercal H = 1$, the weighted equation is $H^\intercal y_t = \phi p_t + e_{H,t}$, and the conditional restrictions imply

$$\mathbb{E}[e_{H,t} \mid Z_t] = 0, \qquad \mathbb{E}[e_{H,t}^2 \mid Z_t] = \sigma_H^2.$$

For any square-integrable instrument $g(Z_t)$ with $\mathbb{E}[g(Z_t)p_t] \neq 0$, the sequential exclusion makes $g(Z_t)e_{H,t}$ a martingale difference. Strict stationarity, ergodicity and finite second moments therefore give the IV estimator's asymptotic variance as

$$V_H(g) = \frac{\sigma_H^2 \mathbb{E}[g(Z_t)^2]}{\mathbb{E}[g(Z_t)a_t]^2} \geq \frac{\sigma_H^2}{\mathbb{E}[a_t^2]}.$$

The inequality is Cauchy–Schwarz, with equality for $g(Z_t) \propto a_t$, so $h_H^*(Z_t) = a_t/\sigma_H^2$ attains the bound. The reduction of a regular finite-instrument GMM estimator to an effective scalar instrument in Appendix C.2 applies unchanged with $e_{H,t}$ replacing ε_t, establishing the comparison over the stated GMM class.

Finally, the aggregate conditional exclusion and equilibrium give

$$S^\intercal \mathbb{E}[y_t \mid Z_t] = \mathbb{E}[x_t \mid Z_t] = \psi \mathbb{E}[p_t \mid Z_t] = \psi a_t.$$

Since $\psi \neq 0$ under Assumption 4, the outcome-predictor expression is equivalent up to scale. The conditional price predictor does not depend on H; only the constant σ_H^2 and hence the variance bound depend on the chosen admissible weights. □

C.5 Proof of Theorem 4

Proof. We verify the consistency conditions of Newey and McFadden (1994, Theorem 2.1) and use the GMM first-order expansion in their Section 3.3, equation (3.6). The variance bounds below suffice for an asymptotic linear representation; a central limit theorem is not needed for this result. Write $q_t = M_\iota y_t$, $\mu_0 = \mathbb{E}[q_t]$, and collect the first-stage parameters as $\pi = (\theta^\intercal, \eta^\intercal, \omega^\intercal)^\intercal$, as in Appendix A.3. Define

$$f_t(\pi; \mu) := m_\Lambda\big(y_t, \Theta^{-1}(\theta), \mathrm{diag}(\eta), \mathrm{diag}(\omega); \mu\big), \qquad g_T(\pi; \mu) := T^{-1} \sum_{t=1}^{T} f_t(\pi; \mu).$$

Let $\mathcal{P}_\pi$ be the compact image of $\mathcal{P}$ in these coordinates, and let $g(\pi) = \mathbb{E}[f_t(\pi; \mu_0)]$. Assumptions 6–7 imply $g(\pi_0) = 0$ through the covariance decomposition of q_t.

Centering and consistency. The variance bounds in Assumption 9(iii) give $d_T := \widehat{\mu} - \mu_0 = O_{\mathbb{P}}(T^{-1/2})$ and $\sqrt{T} g_T(\pi_0; \mu_0) = O_{\mathbb{P}}(1)$ by Chebyshev's inequality. Since $\widehat{\mu} = T^{-1} \sum_t q_t$, the sample covariance identity gives, uniformly in π,

$$g_T(\pi; \widehat{\mu}) - g_T(\pi; \mu_0) = -\mathrm{vech}\big(\tilde{D} d_T d_T^\intercal \tilde{D}^\intercal\big) = O_{\mathbb{P}}(T^{-1}).$$

This establishes directly that estimated centering has no first-order effect. Moreover, $f_t(\pi;\mu_0) - \mathbb{E}[f_t(\pi;\mu_0)]$ does not depend on π. Thus, using $g(\pi_0) = 0$,

$$\sup_{\pi\in\mathcal{P}_\pi} \|g_T(\pi;\widehat{\mu}) - g(\pi)\| = O_{\mathbb{P}}(T^{-1/2}).$$

The sample GMM criterion therefore converges uniformly to $g(\pi)^{\intercal} W g(\pi)$. Since $\tilde{D}$ is one-to-one on $\iota^{\perp}$, the reduced moments retain the covariance information used in Lemma 1. That lemma and the normalization (10) identify $(\Psi_0, \Sigma_{\tilde{F},0}, \Omega_0)$ uniquely. Thus, the continuous limiting criterion has the unique minimizer π_0, since $W > 0$. Compactness and Theorem 2.1 of Newey and McFadden (1994) imply $\widehat{\pi} \to_{\mathbb{P}} \pi_0$.

First-stage expansion. The map Θ^{-1} is affine, with derivative (26). Consequently, $G(\pi) := \partial f_t(\pi;\mu)/\partial \pi^{\intercal}$ is continuous and independent of both y_t and μ, and $G(\pi_0) = \mathcal{J}^{\Lambda}$ with the reduced-moment Jacobians in Appendix A. Since π_0 is interior, the first-order condition $G(\widehat{\pi})^{\intercal}\widehat{W} g_T(\widehat{\pi};\widehat{\mu}) = 0$ holds with probability approaching one. Set

$$\overline{G}_T := \int_0^1 G\big(\pi_0 + s(\widehat{\pi} - \pi_0)\big)\, ds \to_{\mathbb{P}} \mathcal{J}^{\Lambda}.$$

Expanding g_T along this segment and using the nonsingular information matrix in Assumption 9(v) yields

$$\begin{aligned}\sqrt{T}(\widehat{\pi} - \pi_0) &= -\big[G(\widehat{\pi})^{\intercal}\widehat{W}\overline{G}_T\big]^{-1} G(\widehat{\pi})^{\intercal}\widehat{W}\sqrt{T} g_T(\pi_0;\widehat{\mu}) \\ &= -\mathcal{K}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t(\pi_0;\mu_0) + o_{\mathbb{P}}(1),\end{aligned}$$

where $\mathcal{K}$ is defined in Appendix A.3. Selecting the θ block gives $\vartheta(y_t) = -\Pi_\theta \mathcal{K} f_t(\pi_0;\mu_0)$ in (27), with $\mathbb{E}[\vartheta(y_t)] = 0$.

Structural expansion. The restriction $\mathbb{E}[u_t\varepsilon_t] = 0$ implies $\mathbb{E}[m_\psi(w_t,\psi_0,\Psi_0)] = S^{\intercal} M_{\Psi_0} M_\iota \mathbb{E}[u_t\varepsilon_t] = 0$. Define

$$b_T(\theta) := T^{-1}\sum_{t=1}^{T} m_\psi\big(w_t,\psi_0,\Theta^{-1}(\theta)\big), \qquad \widehat{\mathcal{J}}_p := T^{-1}\sum_{t=1}^{T} S^{\intercal} M_{\widehat{\Psi}} q_t p_t.$$

The variance bounds and consistency of $\widehat{\Psi}$ imply $\widehat{\mathcal{J}}_p \to_{\mathbb{P}} \mathcal{J}_p \neq 0$ and $\sqrt{T} b_T(\theta_0) = O_{\mathbb{P}}(1)$. Because Ψ_0 has full column rank, the projection matrix is continuously differentiable near θ_0. The derivative of b_T depends on the data only through $T^{-1}\sum_t q_t\varepsilon_t$, which converges to its expectation. Its integral along the segment from θ_0 to $\widehat{\theta}$ consequently converges to $\mathcal{J}_\theta^{\psi}$. The defining IV equation then gives

$$\begin{aligned}\sqrt{T}(\widehat{\psi} - \psi_0) &= \widehat{\mathcal{J}}_p^{-1}\sqrt{T} b_T(\widehat{\theta}) \\ &= \mathcal{J}_p^{-1}\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m_\psi(w_t,\psi_0,\Psi_0) + \mathcal{J}_\theta^{\psi}\sqrt{T}(\widehat{\theta} - \theta_0)\right\} + o_{\mathbb{P}}(1).\end{aligned}$$

Substituting the first-stage expansion proves (13). □

C.6 Proof of Theorem 5

Proof. Use the notation q_t, π, and f_t of Appendix C.5. Selecting the ω block of its first-stage expansion gives

$$\sqrt{T}(\widehat{\omega} - \omega_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varsigma(y_t) + o_{\mathbb{P}}(1), \qquad \varsigma(y_t) = -\Pi_\omega \mathcal{K} f_t(\pi_0; \mu_0),$$

with $\mathbb{E}[\varsigma(y_t)] = 0$, as in (27).

Validity and the fixed-weight expansion. Write $v_t = y_t - \phi_0 p_t \iota = \Lambda_0 F_t + u_t$ and $z_{0t} = S^{\mathsf{T}} M_{\Psi_0} q_t = S^{\mathsf{T}} M_{\Lambda_0} u_t$. The additional exclusion in Assumption 10(i) gives

$$\mathbb{E}[z_{0t} v_t^{\mathsf{T}}] \Omega_0^{-1} \Lambda_0 = S^{\mathsf{T}} M_{\Lambda_0} \big(\mathbb{E}[u_t F_t^{\mathsf{T}}] \Lambda_0^{\mathsf{T}} + \Omega_0 \big) \Omega_0^{-1} \Lambda_0 = 0.$$

Thus the population individual-equation moment is zero for every γ. Let $\Psi(\theta) = \Theta^{-1}(\theta)$, $\Lambda(\theta) = [\ \iota \quad \Psi(\theta)\ ]$, and $\Omega(\omega) = \mathrm{diag}(\omega)$. Define the row vector

$$b_T(\theta, \omega) := T^{-1} \sum_{t=1}^{T} S^{\mathsf{T}} M_{\Psi(\theta)} q_t v_t^{\mathsf{T}} \Omega(\omega)^{-1} \Lambda(\theta).$$

The variance bounds in Assumption 9(iii) imply $\sqrt{T} b_T(\theta_0, \omega_0) = O_{\mathbb{P}}(1)$ and convergence of $T^{-1} \sum_t q_t v_t^{\mathsf{T}}$ to its expectation. The maps defining b_T are continuously differentiable near the truth, since Ψ_0 has full column rank and Ω_0 is positive definite. A componentwise Taylor expansion, with derivatives integrated along the segment between the true and estimated parameters, therefore gives

$$\sqrt{T} b_T(\widehat{\theta}, \widehat{\omega}) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Upsilon_t^{\mathsf{T}} + o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1). \tag{66}$$

Here the first-stage contributions are precisely the two Jacobian terms in (17). Set $a_0 = \Lambda_0^{\mathsf{T}} \Omega_0^{-1} \iota$ and $\widehat{a} = \widehat{\Lambda}^{\mathsf{T}} \widehat{\Omega}^{-1} \iota$. For fixed γ with $a_0^{\mathsf{T}} \gamma = 1$, the exact IV identity is

$$\sqrt{T}\big(\widehat{\phi}(\gamma) - \phi_0\big) = \frac{\sqrt{T} b_T(\widehat{\theta}, \widehat{\omega}) \gamma}{\widehat{\mathcal{J}}_p \, \widehat{a}^{\mathsf{T}} \gamma}.$$

Since $\widehat{\mathcal{J}}_p \to_{\mathbb{P}} \mathcal{J}_p \neq 0$ and $\widehat{a} \to_{\mathbb{P}} a_0$, substituting (66) proves part (i). Scale invariance allows the same argument for any fixed initial vector with $a_0^{\mathsf{T}} \gamma^{\mathrm{ini}} \neq 0$, and hence $\widehat{\phi}^{\mathrm{ini}} - \phi_0 = O_{\mathbb{P}}(T^{-1/2})$.

Consistency of the estimated weights. The first-stage estimates, initial structural estimate, and sample moments entering the plug-in formulas in Appendix A.4 are consistent. To control their effect on the estimated second moment, define the finite-dimensional vector

$$X_t = \big(1, q_t^{\mathsf{T}}, \mathrm{vec}(q_t y_t^{\mathsf{T}})^{\mathsf{T}}, p_t q_t^{\mathsf{T}}\big)^{\mathsf{T}}.$$

The finite second moments in Assumption 9(iii) imply $\mathbb{E}[\|X_t\|^2] < \infty$ and $T^{-1} \sum_t \|X_t\|^2 = O_{\mathbb{P}}(1)$ by Markov's inequality. The formulas for the centered first-stage moments and (19) are linear in the entries of X_t: we

can write $\Upsilon_t = C_0 X_t$ and $\widehat{\Upsilon}_t = \widehat{C} X_t$, where consistency of their coefficients gives $\widehat{C} \to_{\mathbb{P}} C_0$. Consequently,

$$T^{-1} \sum_{t=1}^{T} \|\widehat{\Upsilon}_t - \Upsilon_t\|^2 \leq \|\widehat{C} - C_0\|^2 T^{-1} \sum_{t=1}^{T} \|X_t\|^2 = o_{\mathbb{P}}(1).$$

Cauchy–Schwarz and Assumption 10(iii) then yield

$$\widehat{Q}_\Upsilon - T^{-1} \sum_{t=1}^{T} \Upsilon_t \Upsilon_t^{\intercal} = o_{\mathbb{P}}(1), \qquad \widehat{Q}_\Upsilon \to_{\mathbb{P}} Q_\Upsilon.$$

Since Q_Υ is positive definite and the first element of a_0 is $\iota^{\intercal} \Omega_0^{-1} \iota > 0$, the weight formula is well-defined with probability approaching one. The continuous mapping theorem gives

$$\widehat{\gamma} = \frac{\widehat{Q}_\Upsilon^{-1} \widehat{a}}{\widehat{a}^{\intercal} \widehat{Q}_\Upsilon^{-1} \widehat{a}} \to_{\mathbb{P}} \frac{Q_\Upsilon^{-1} a_0}{a_0^{\intercal} Q_\Upsilon^{-1} a_0} = \gamma^*.$$

Final estimator. The sample normalization $\widehat{a}^{\intercal} \widehat{\gamma} = 1$ gives

$$\begin{aligned} \sqrt{T}(\widehat{\phi} - \phi_0) &= \widehat{\mathcal{J}}_p^{-1} \sqrt{T} b_T(\widehat{\theta}, \widehat{\omega}) \widehat{\gamma} \\ &= \mathcal{J}_p^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Upsilon_t^{\intercal} \gamma^* + o_{\mathbb{P}}(1). \end{aligned}$$

The substitution of γ^* for $\widehat{\gamma}$ costs only $O_{\mathbb{P}}(1) o_{\mathbb{P}}(1)$ by (66); no $\sqrt{T}$ rate for $\widehat{\gamma}$ is needed. Expanding $\Upsilon_t^{\intercal} \gamma^*$ using (17) proves part (ii). □

D Identification of Factor Loadings with Block-Correlated Errors

In this section, we generalize Assumption 7 and allow the idiosyncratic shocks u_{it} to be correlated within known blocks of cross-sectional units. In many applications, mutual uncorrelatedness of the individual shocks is plausible across, but not necessarily within, groups of institutionally linked units. In the oil market example, the supply shocks of OPEC members may comove because of coordinated production decisions, over and above the common variation captured by the factors F_t. We show that the conclusion of Lemma 1 survives under such a block structure, provided that the cross section is sufficiently rich outside every block in a sense made precise below.

Let $\{B_1, \ldots, B_G\}$ be a partition of $\{1, \ldots, N\}$ into G non-empty blocks. For $i \in \{1, \ldots, N\}$, let $b(i)$ denote the index of the block containing unit i, and let $N_g = |B_g|$.

Assumption 11. *The partition $\{B_1, \ldots, B_G\}$ is known to the researcher, and $E[u_{it} u_{jt}] = 0$ for all $i, j \in \{1, \ldots, N\}$ with $b(i) \neq b(j)$.*

Under Assumption 11, after relabeling the units so that blocks are contiguous, $\Omega = E[u_t u_t^{\intercal}]$ is block diagonal, $\Omega = \operatorname{diag}(\Omega_{B_1}, \ldots, \Omega_{B_G})$, where the within-block covariance matrices $\Omega_{B_g} := E[u_{B_g,t} u_{B_g,t}^{\intercal}]$ with $u_{B_g,t} := (u_{it})_{i \in B_g}$ are completely unrestricted. Assumption 7 is the special case in which every block is a singleton.

The identification argument requires that, for each block, the remaining cross-sectional units are sufficiently rich. For $g \in \{1,\dots,G\}$, we say that a pair of index sets $(\mathcal{I},\mathcal{I}')$ is *g-admissible* if

(a) $\mathcal{I},\mathcal{I}' \subset \{1,\dots,N\} \setminus B_g$ with $|\mathcal{I}| = |\mathcal{I}'| = r$; and

(b) no block intersects both $\mathcal{I}$ and $\mathcal{I}'$: for every $g' \in \{1,\dots,G\}$, $B_{g'} \cap \mathcal{I} = \emptyset$ or $B_{g'} \cap \mathcal{I}' = \emptyset$.

Condition (b) implies $\mathcal{I} \cap \mathcal{I}' = \emptyset$ but is stronger. It does not, however, preclude $\mathcal{I}$ (or $\mathcal{I}'$) from containing several units of the same block.

Assumption 12. *For every $g \in \{1,\dots,G\}$, there exists a g-admissible pair $(\mathcal{I}_g,\mathcal{I}'_g)$ such that $\Lambda_{\mathcal{I}_g}$ and $\Lambda_{\mathcal{I}'_g}$ are nonsingular.*

Assumption 12 plays the role of Assumption 8, to which it reduces when all blocks are singletons. Note that a single pair $(\mathcal{I}_g,\mathcal{I}'_g)$ serves the entire block B_g.

Lemma 2. *Suppose that Assumptions 1, 6, 11, and 12 hold. Then, $\Psi\Sigma_{\tilde{F}}\Psi^{\mathsf{T}}$ and Ω are uniquely determined by $\mathbb{V}[M_\iota y_t]$. Under the normalization (10), we obtain a unique Ψ.*

The proof follows the determinant strategy used for Lemma 1, with one new ingredient. Off-diagonal within-block entries $\Omega_{\ell q}$, $\ell \neq q$, cannot be reached by differencing the data with respect to a single base unit; we therefore difference twice, with base unit ℓ on the left of $\mathbb{V}[M_\iota y_t]$ and base unit q on the right. Block-separation of the selected rows and columns then makes the error component of the resulting $r \times r$ moment matrix proportional to $\iota_r\iota_r^{\mathsf{T}}$, with coefficient $\Omega_{\ell q}$, and the entry is recovered as the unique zero of an affine determinant function of the same form as in the proof of Lemma 1.

Proof. We use the notation of Section C.1. In addition, for $\ell \in \{1,\dots,N\}$ and $\mathcal{I} = \{i_1 < \cdots < i_r\} \subset \{1,\dots,N\}\setminus\{\ell\}$, define $D_{\mathcal{I},\ell} \in \mathbb{R}^{r\times N}$ as the matrix whose ath row is $(e_{i_a,N} - e_{\ell,N})^{\mathsf{T}}$, $a = 1,\dots,r$; equivalently, $D_{\mathcal{I},\ell}$ collects the rows of $\tilde{D}_\ell$ corresponding to the indices in $\mathcal{I}$. Every row of $D_{\mathcal{I},\ell}$ is orthogonal to ι_N, so that

$$D_{\mathcal{I},\ell}\,\iota_N = 0 \qquad \text{and} \qquad D_{\mathcal{I},\ell}M_\iota = D_{\mathcal{I},\ell}. \tag{67}$$

Step 1: A moment matrix that isolates one entry of Ω. Fix $g \in \{1,\dots,G\}$ and $\ell,q \in B_g$, where $\ell = q$ is allowed, and let $(\mathcal{I},\mathcal{I}')$ be a g-admissible pair; since $\mathcal{I},\mathcal{I}' \subset \{1,\dots,N\}\setminus B_g$, we have $\ell \notin \mathcal{I}$ and $q \notin \mathcal{I}'$, so $D_{\mathcal{I},\ell}$ and $D_{\mathcal{I}',q}$ are well defined. Define the $r \times r$ matrix

$$\Sigma_{\ell q}(\mathcal{I},\mathcal{I}') := D_{\mathcal{I},\ell}\,\mathbb{V}[M_\iota y_t]\,D_{\mathcal{I}',q}^{\mathsf{T}}.$$

Left-multiplying the stacked individual equations by $D_{\mathcal{I},\ell}$ and using (67), the terms $\phi p_t\iota_N$ and $F_{1t}\iota_N$ drop out and

$$D_{\mathcal{I},\ell}\,y_t = \big(\Psi_{\mathcal{I}} - \iota_r\psi_\ell^{\mathsf{T}}\big)\tilde{F}_t + D_{\mathcal{I},\ell}\,u_t, \qquad \Psi_{\mathcal{I}} := [\psi_i^{\mathsf{T}}]_{i\in\mathcal{I}} \in \mathbb{R}^{r\times(r-1)},$$

and analogously for $D_{\mathcal{I}',q}\,y_t$. By (67), $\Sigma_{\ell q}(\mathcal{I},\mathcal{I}') = \mathrm{Cov}(D_{\mathcal{I},\ell}\,y_t,\, D_{\mathcal{I}',q}\,y_t)$, and combining the previous display with $E[\tilde{F}_t u_t^{\mathsf{T}}] = 0$ (Assumption 6),

$$\Sigma_{\ell q}(\mathcal{I},\mathcal{I}') = \big(\Psi_{\mathcal{I}} - \iota_r\psi_\ell^{\mathsf{T}}\big)\,\Sigma_{\tilde{F}}\,\big(\Psi_{\mathcal{I}'} - \iota_r\psi_q^{\mathsf{T}}\big)^{\mathsf{T}} + D_{\mathcal{I},\ell}\,\Omega\,D_{\mathcal{I}',q}^{\mathsf{T}}. \tag{68}$$

The (a,b) entry of the last term in (68) is, writing $i = i_a \in \mathcal{I}$ and $j = j_b \in \mathcal{I}'$,

$$(e_{i,N} - e_{\ell,N})^{\mathsf{T}}\,\Omega\,(e_{j,N} - e_{q,N}) = \Omega_{ij} - \Omega_{iq} - \Omega_{\ell j} + \Omega_{\ell q} = \Omega_{\ell q},$$

where the second equality holds by Assumption 11: $\Omega_{ij} = 0$ because no block intersects both $\mathcal{I}$ and $\mathcal{I}'$, so $b(i) \neq b(j)$; $\Omega_{iq} = 0$ because $i \notin B_g$ while $q \in B_g$; and $\Omega_{\ell j} = 0$ because $j \notin B_g$ while $\ell \in B_g$. Hence $D_{\mathcal{I},\ell}\,\Omega\,D_{\mathcal{I}',q}^{\mathsf{T}} = \Omega_{\ell q}\,\iota_r\iota_r^{\mathsf{T}}$ and

$$\Sigma_{\ell q}(\mathcal{I},\mathcal{I}') = \left(\Psi_{\mathcal{I}} - \iota_r\psi_\ell^{\mathsf{T}}\right)\Sigma_{\tilde{F}}\left(\Psi_{\mathcal{I}'} - \iota_r\psi_q^{\mathsf{T}}\right)^{\mathsf{T}} + \Omega_{\ell q}\,\iota_r\iota_r^{\mathsf{T}}. \tag{69}$$

Step 2: A determinant identity. For $s \in \mathbb{R}$, (69) factorizes as

$$\Sigma_{\ell q}(\mathcal{I},\mathcal{I}') - s\,\iota_r\iota_r^{\mathsf{T}} = \begin{bmatrix} \Psi_{\mathcal{I}} - \iota_r\psi_\ell^{\mathsf{T}} & \iota_r \end{bmatrix} \begin{pmatrix} \Sigma_{\tilde{F}} & 0 \\ 0 & \Omega_{\ell q} - s \end{pmatrix} \begin{bmatrix} \Psi_{\mathcal{I}'} - \iota_r\psi_q^{\mathsf{T}} & \iota_r \end{bmatrix}^{\mathsf{T}},$$

so that

$$\det\!\big(\Sigma_{\ell q}(\mathcal{I},\mathcal{I}') - s\,\iota_r\iota_r^{\mathsf{T}}\big) = (\Omega_{\ell q} - s)\,\det\!\big(\big[\Psi_{\mathcal{I}} - \iota_r\psi_\ell^{\mathsf{T}}\;\;\iota_r\big]\big)\,\det(\Sigma_{\tilde{F}})\,\det\!\big(\big[\Psi_{\mathcal{I}'} - \iota_r\psi_q^{\mathsf{T}}\;\;\iota_r\big]\big).$$

Adding $\psi_{\ell k}$ times the last column of $[\Psi_{\mathcal{I}} - \iota_r\psi_\ell^{\mathsf{T}} \;\; \iota_r]$ to its kth column, for $k = 1,\dots,r-1$, leaves the determinant unchanged, and

$$\det\!\big(\big[\Psi_{\mathcal{I}} - \iota_r\psi_\ell^{\mathsf{T}}\;\;\iota_r\big]\big) = \det\!\big(\big[\Psi_{\mathcal{I}}\;\;\iota_r\big]\big) = (-1)^{r-1}\det\!\big(\big[\iota_r\;\;\Psi_{\mathcal{I}}\big]\big) = (-1)^{r-1}\det(\Lambda_{\mathcal{I}}),$$

where the last equality holds because the first column of Λ is ι; in particular, the value does not depend on the base unit ℓ. The same computation applies to the determinant involving $\mathcal{I}'$ and q. Since the two sign factors cancel,

$$\det\!\big(\Sigma_{\ell q}(\mathcal{I},\mathcal{I}') - s\,\iota_r\iota_r^{\mathsf{T}}\big) = (\Omega_{\ell q} - s)\,\kappa(\mathcal{I},\mathcal{I}'), \qquad \kappa(\mathcal{I},\mathcal{I}') := \det(\Lambda_{\mathcal{I}})\,\det(\Sigma_{\tilde{F}})\,\det(\Lambda_{\mathcal{I}'}). \tag{70}$$

We stress that (70) holds for any g-admissible pair $(\mathcal{I},\mathcal{I}')$ and uses only Assumption 6 and Assumption 11; in particular, its derivation does not require $\Lambda_{\mathcal{I}}$, $\Lambda_{\mathcal{I}'}$, or $\Sigma_{\tilde{F}}$ to be nonsingular.

Step 3: Identification of Ω. Let $(\Psi^\dagger, \{\tilde{F}_t^\dagger, u_t^\dagger\}_{t\geq 1})$ be any structure that satisfies Assumption 6 and Assumption 11 with the same (known) partition and that generates the same $\mathbb{V}[M_\iota y_t]$, and let $\Omega^\dagger$ and $\kappa^\dagger$ denote the associated objects. Fix g and let $(\mathcal{I}_g,\mathcal{I}_g')$ be the pair of Assumption 12; g-admissibility is a property of the known partition only, so $(\mathcal{I}_g,\mathcal{I}_g')$ is g-admissible for both structures. For any $\ell,q \in B_g$, applying (70) to each structure yields, for all $s \in \mathbb{R}$,

$$(\Omega_{\ell q} - s)\,\kappa(\mathcal{I}_g,\mathcal{I}_g') = \det\!\big(\Sigma_{\ell q}(\mathcal{I}_g,\mathcal{I}_g') - s\,\iota_r\iota_r^{\mathsf{T}}\big) = (\Omega_{\ell q}^\dagger - s)\,\kappa^\dagger(\mathcal{I}_g,\mathcal{I}_g'),$$

where the middle expression is a function of $\mathbb{V}[M_\iota y_t]$ alone. By Assumption 12 and the nonsingularity of $\Sigma_{\tilde{F}}$, $\kappa(\mathcal{I}_g,\mathcal{I}_g') \neq 0$, so all three expressions are non-constant affine functions of s. Matching slopes gives $\kappa^\dagger(\mathcal{I}_g,\mathcal{I}_g') = \kappa(\mathcal{I}_g,\mathcal{I}_g') \neq 0$, and matching the (unique) zeros gives $\Omega_{\ell q}^\dagger = \Omega_{\ell q}$. Since g and $\ell,q \in B_g$ were

arbitrary, all within-block entries of $\Omega^\dagger$ and Ω coincide, while all cross-block entries of both matrices are zero by Assumption 11. Hence $\Omega^\dagger = \Omega$; that is, Ω is uniquely determined by $\mathbb{V}[M_\iota y_t]$, each entry $\Omega_{\ell q}$, $\ell, q \in B_g$, being the unique zero of the non-constant affine map $s \mapsto \det(\Sigma_{\ell q}(\mathcal{I}_g, \mathcal{I}'_g) - s\, \iota_r \iota_r^\intercal)$.

Step 4: Conclusion. Given Ω,

$$\Psi\, \Sigma_{\tilde{F}}\, \Psi^\intercal = \mathbb{V}[M_\iota y_t] - M_\iota \Omega M_\iota^\intercal$$

is uniquely determined as well. Finally, Assumption 12 implies that Λ has a nonsingular $r \times r$ submatrix and hence full column rank, so that $\Psi = M_\iota \Psi$ has full column rank (as noted in Section 4.1). Under (9), Ψ is uniquely pinned down from $\Psi \Sigma_{\tilde{F}} \Psi^\intercal$ (e.g., Anderson and Rubin, 1956, Lemma 5.1). □

Remark D.1 (Block sizes and the number of blocks). Assumption 12 restricts the block structure. First, since condition (b) implies that $\mathcal{I}_g$ and $\mathcal{I}'_g$ are disjoint subsets of $\{1, \dots, N\} \setminus B_g$, each of cardinality r, it requires $N - N_g \geq 2r$ for every g; that is, no block may contain more than $N - 2r$ units. Second, the blocks other than B_g must admit a split into two groups, one containing $\mathcal{I}_g$ and the other containing $\mathcal{I}'_g$; equivalently, some subfamily of $\{B_{g'}\}_{g' \neq g}$ must have total size in the interval $[r,\, N - N_g - r]$. In particular, at least three blocks are required, $G \geq 3$: if $G = 2$, both $\mathcal{I}_g$ and $\mathcal{I}'_g$ would have to be drawn from the single remaining block, contradicting (b). Beyond these counting requirements, nonsingularity of $\Lambda_{\mathcal{I}_g}$ and $\Lambda_{\mathcal{I}'_g}$ requires sufficient variation of the factor loadings across units, exactly as in Assumption 8. □

Remark D.2 (Two blocks: genuine non-identification). The requirement $G \geq 3$ is not an artifact of our proof strategy: with $G = 2$, Ω is generally not identified from $\mathbb{V}[M_\iota y_t]$. Let $a \in \mathbb{R}^N$ have entries $a_i = 1$ for $i \in B_1$ and $a_i = -1$ for $i \in B_2$, and for $c \in \mathbb{R}$ define

$$\Omega(c) := \Omega + c\left(a\, \iota_N^\intercal + \iota_N\, a^\intercal\right),$$

whose (i, j) entry is $\Omega_{ij} + c\,(a_i + a_j)$. Then, (i) $\Omega(c)$ satisfies Assumption 11 because $a_i + a_j = 0$ whenever $b(i) \neq b(j)$; (ii) $\Omega(c)$ is positive definite for $|c|$ sufficiently small whenever Ω is, so it is a legitimate covariance matrix; and (iii) $M_\iota \Omega(c) M_\iota^\intercal = M_\iota \Omega M_\iota^\intercal$ because $M_\iota \iota_N = 0$. Hence, for every small $c \neq 0$, the structures with error covariance Ω and $\Omega(c) \neq \Omega$ (and the same factor component) generate the same $\mathbb{V}[M_\iota y_t]$. When $G \geq 3$, this construction is unavailable: if a symmetric perturbation $a\, \iota_N^\intercal + \iota_N\, a^\intercal$ preserves the block structure, then $a_i + a_j = 0$ for all cross-block pairs, and picking units i, j, k in three distinct blocks yields $a_i = -a_j = a_k = -a_i$, whence $a = 0$. We note that this non-identification is relative to $\mathbb{V}[M_\iota y_t]$, the object on which our sequential procedure is based; other moments, e.g., $\mathrm{Cov}(M_\iota y_t, p_t)$, which involves $M_\iota \Omega S$ through the equilibrium condition, may carry additional information about Ω, but exploiting them would entangle the estimation of the loadings with the structural parameters (ψ, ϕ), and we do not pursue this route. □

Remark D.3 (Block-local factors). Assumption 12 also disciplines the loadings relative to the blocks. Suppose that the loadings of one of the factors vanished for all units outside some block B_g (a "block-local" factor). Then, the corresponding column of $\Psi_{\mathcal{I}}$ would be zero for every $\mathcal{I} \subset \{1, \dots, N\} \setminus B_g$, so $\Lambda_{\mathcal{I}_g}$ would be singular and Assumption 12 would fail for g. This exclusion is not incidental: the contribution of such a factor to

$\mathbb{V}[y_t]$ is supported on $B_g \times B_g$ and is therefore observationally equivalent to additional within-block error covariance, which Assumption 11 leaves unrestricted. Assumption 12 thus formalizes the requirement that every common factor be pervasive across blocks. □

Remark D.4 (Closed form, estimation, and testable implications). Applying the polynomial identity $\det(A - s\,\iota_r\iota_r^{\intercal}) = \det(A) - s\,\iota_r^{\intercal}\operatorname{adj}(A)\,\iota_r$, valid for any $A \in \mathbb{R}^{r\times r}$, to (70) delivers the closed form

$$\Omega_{\ell q} = \frac{\det\big(\Sigma_{\ell q}(\mathcal{I}_g, \mathcal{I}'_g)\big)}{\iota_r^{\intercal}\operatorname{adj}\big(\Sigma_{\ell q}(\mathcal{I}_g, \mathcal{I}'_g)\big)\,\iota_r}, \qquad \ell, q \in B_g,$$

whose denominator equals $\kappa(\mathcal{I}_g, \mathcal{I}'_g) \neq 0$. Replacing $\mathbb{V}[M_\iota y_t]$ with its sample analogue yields a plug-in estimator of Ω, and, under conditions guaranteeing asymptotic normality of the sample second moments, root-T inference follows from the delta method, the denominator being bounded away from zero. When several g-admissible pairs satisfy the rank condition, each delivers the same $\Omega_{\ell q}$, generating overidentifying restrictions that can be incorporated into the GMM estimation of the factor loadings. The block structure itself has testable implications: for $\ell \in B_g$ and $q \in B_{g'}$ with $g \neq g'$, the argument of Steps 1–3 applies verbatim to any pair $(\mathcal{I}, \mathcal{I}')$ with $\mathcal{I}, \mathcal{I}' \subset \{1, \dots, N\} \setminus (B_g \cup B_{g'})$, $|\mathcal{I}| = |\mathcal{I}'| = r$, no block intersecting both $\mathcal{I}$ and $\mathcal{I}'$, and $\Lambda_{\mathcal{I}}$, $\Lambda_{\mathcal{I}'}$ nonsingular; it then identifies $\Omega_{\ell q}$ without imposing $\Omega_{\ell q} = 0$, so that cross-block uncorrelatedness can be tested whenever such pairs exist. More generally, the proof shows that an arbitrary entry $\Omega_{\ell q}$ is identified as soon as there exist $\mathcal{I} \not\ni \ell$ and $\mathcal{I}' \not\ni q$ with $|\mathcal{I}| = |\mathcal{I}'| = r$, $\Lambda_{\mathcal{I}}$ and $\Lambda_{\mathcal{I}'}$ nonsingular, and $\Omega_{ij} = \Omega_{iq} = \Omega_{\ell j} = 0$ known a priori for all $i \in \mathcal{I}$ and $j \in \mathcal{I}'$; block-diagonal Ω is one instance of this scheme, and banded structures arising from spatial dependence are another. □

E Monte Carlo Experiments: Robustness to Misspecification

This section shows through Monte Carlo experiments that our granular IV estimators are robust to misspecification of the covariance matrix of the shocks. Correct specification of this covariance matrix is critical for the MLE of Baumeister and Hamilton (2023), whereas our estimators of Section 4 do not impose any parametric structure on it.

We follow the supply-demand model of the empirical application in Section 5 and proceed in two steps. First, as a baseline, we consider a correctly specified model. Second, we introduce a misspecification that does not affect the validity of the granular instrument: the inventory shock is correlated with the demand shock. We find that the MLE is consistent under correct specification, but it becomes inconsistent and its coverage breaks down under misspecification. Our granular IV estimators perform well in both cases.

As in Section 5, the supply of unit i, the aggregate demand, and the change in inventory follow the

structural equations

$$\begin{aligned} y_{it} &= \phi p_t + \lambda_i^{\intercal} F_t + u_{it}, \qquad i = 1, \ldots, N, \\ x_t &= \psi p_t + \varepsilon_t, \\ inv_t &= \zeta_t, \end{aligned}$$

together with the market clearing condition $x_t = S^{\intercal} y_t - inv_t$. For simplicity, we omit the controls c_t and set the price elasticity of inventory change κ to zero.

E.1 Correct Specification

The shocks have the structure assumed by the MLE in Section 5:

$$\Lambda F_t = b_1 f_{1t} + b_2 f_{2t}, \qquad \varepsilon_t = h f_{1t} + \xi_t,$$

where $b_1, b_2 \in \mathbb{R}^N$, $h \in \mathbb{R}$, and $f_{1t}, f_{2t}, u_t, \xi_t, \zeta_t$ are mutually uncorrelated. Hence, after rescaling b_1, b_2, and h so that the factors have unit variance, the covariance matrix Σ_ν of $\boldsymbol{\nu}_t$ has the structure (24), and the MLE is correctly specified. Since b_1 and b_2 are orthogonal to ι (see below), in the notation of Section 4, $\Lambda = [\ \iota \quad \Psi\ ]$ with $\Psi = [\ b_1 \quad b_2\ ]$ up to the normalization (10), and the time fixed effect F_{1t} is zero.

Data generating process We set $\phi = 0.1$ and $\psi = -0.1$. We fix $N = 10$ and consider the sample sizes $T = 50$, 100, 300, 500, 1000, and 2000. The strength of the granular instrument is governed by the Jacobian (see Appendix A)

$$\mathcal{J}_p = \frac{1}{\psi_0 - \phi_0} S^{\intercal} M_\Lambda \Omega_0 S,$$

which depends on how far the share vector S is from the space spanned by the columns of Λ, including ι. For a realistic instrument strength, we set S close to the share vector of the empirical application (Table 1), extended to 10 units:

$$S = (0.442,\ 0.166,\ 0.124,\ 0.120,\ 0.055,\ 0.046,\ 0.035,\ 0.005,\ 0.005,\ 0.005)^{\intercal}.$$

The elements of b_1 and b_2 are drawn independently from $\mathcal{N}(0, 4)$ and $\mathcal{N}(0, 16)$, respectively, and we then impose that b_1 and b_2 are orthogonal to each other, to ι, and to S. The orthogonality to S prevents an accidentally weak instrument. Since the factor loadings are non-random parameters to be estimated, we draw them once and hold them fixed across the Monte Carlo replications. We draw h from $\mathcal{N}(0, 1)$, and the inventory shock is $\zeta_t \sim \mathcal{N}(0, 0.5)$. The factors are $(f_{1t}, f_{2t})^{\intercal} \sim \mathcal{N}(0, 0.25 I_2)$. The idiosyncratic shocks are $u_t \sim \mathcal{N}(0, \Omega)$ and $\xi_t \sim \mathcal{N}(0, \sigma_\xi^2)$, where $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ with each σ_i^2 drawn from the uniform distribution on $[5, 6]$.

Results We conduct 1000 Monte Carlo replications and report four measures: the median of the estimates, the Monte Carlo bias, the coverage, and the median length of the 95% confidence interval. The coverage is the fraction of replications in which the 95% confidence interval contains the true value.

We report the results for our granular IV estimators in Table 7. As expected, the estimators are consistent. The medians of the estimates are very close to the true values, the bias is very small, and the coverage is close to 95%. The confidence intervals are wider for the smallest sample size $T = 50$, and the precision improves as T increases.

Supply Estimation

N	T	Median	MC Bias	Coverage	CI Length
10	50	0.098	-0.003	0.981	0.182
10	100	0.1	-0.008	0.961	0.093
10	300	0.1	0	0.945	0.048
10	500	0.1	0	0.941	0.037
10	1000	0.1	0	0.938	0.026
10	2000	0.1	0	0.937	0.019

Demand Estimation

N	T	Median	MC Bias	Coverage	CI Length
10	50	-0.099	-0.013	0.936	0.136
10	100	-0.1	-0.021	0.951	0.094
10	300	-0.1	-0.001	0.949	0.055
10	500	-0.099	0	0.949	0.042
10	1000	-0.1	0	0.952	0.03
10	2000	-0.1	0	0.941	0.021

Table 7: Supply and Demand Estimation using GMM under Correct Specification

We report the results for the MLE in Table 8. As expected, the MLE is also consistent. The medians of the estimates are very close to the true values, and for $T \geq 100$, the bias is very small and the coverage is close to 95%. Under correct specification, the MLE should be the most precise, and we find this to be the case: the confidence intervals of the MLE are shorter than those of our granular IV estimators. For the supply elasticity ϕ, the MLE confidence interval is about one third as long as ours for $T = 50$ and about one half as long for $T = 2000$.

Supply Estimation

N	T	Median	MC Bias	Coverage	CI Length
10	50	0.102	-0.041	0.907	0.058
10	100	0.101	-0.007	0.936	0.041
10	300	0.1	0.001	0.947	0.023
10	500	0.1	0	0.931	0.018
10	1000	0.1	0	0.944	0.012
10	2000	0.1	0	0.951	0.009

Demand Estimation

N	T	Median	MC Bias	Coverage	CI Length
10	50	-0.098	0.063	0.891	0.095
10	100	-0.1	0.013	0.936	0.068
10	300	-0.099	0	0.931	0.039
10	500	-0.1	0	0.939	0.03
10	1000	-0.1	0	0.94	0.021
10	2000	-0.1	0	0.932	0.015

Table 8: Supply and Demand Estimation using MLE under Correct Specification

E.2 Misspecification

We now introduce a misspecification. The shocks have the structure

$$\Lambda F_t = b_1 f_{1t} + b_2 f_{2t}, \qquad \varepsilon_t = h f_{2t} + \xi_t, \qquad \zeta_t = h_\zeta \, \xi_t,$$

where $f_{1t}, f_{2t}, u_t, \xi_t$ are mutually uncorrelated. The only difference from the previous subsection is that the inventory shock ζ_t is correlated with the demand shock ε_t through ξ_t. We nonetheless estimate the model by the MLE with the covariance structure (24), which is therefore misspecified. In contrast, Assumption 2(i)-(ii) continues to hold, and the granular instrument remains valid.

Data generating process The data generating process is the same as in the previous subsection, and we draw h_ζ from $\mathcal{N}(0,1)$.

Results We report the results for our granular IV estimators in Table 9. As expected, the estimators remain consistent under this model. The medians of the estimates are very close to the true values, the bias is very small, and the coverage is close to 95%. As under correct specification, the confidence intervals are wider for $T = 50$, and the precision improves as T increases.

Supply Estimation					
N	T	Median	MC Bias	Coverage	CI Length
10	50	0.097	-0.018	0.983	0.192
10	100	0.1	-0.001	0.961	0.094
10	300	0.1	0	0.946	0.049
10	500	0.101	0	0.941	0.037
10	1000	0.1	0	0.938	0.026
10	2000	0.1	0	0.935	0.019
Demand Estimation					
N	T	Median	MC Bias	Coverage	CI Length
10	50	-0.098	0.007	0.924	0.14
10	100	-0.1	-0.01	0.943	0.101
10	300	-0.1	-0.001	0.946	0.058
10	500	-0.1	-0.001	0.956	0.044
10	1000	-0.1	0	0.954	0.031
10	2000	-0.1	0	0.951	0.022

Table 9: Supply and Demand Estimation using GMM under Misspecification

We report the results for the MLE in Table 10. The MLE ceases to be consistent. The median of the estimates of the supply elasticity ϕ, around 0.115, is relatively close to the true value, but it does not approach the true value as T increases. The bias is an order of magnitude larger than under correct specification, and the coverage breaks down: it is far below 95% and decreases as T increases. The median of the estimates of the demand elasticity ψ is far from the true value, the Monte Carlo bias is extremely large, and the coverage is very low.

Supply Estimation					
N	T	Median	MC Bias	Coverage	CI Length
10	50	0.115	-0.161	0.551	0.071
10	100	0.114	-0.231	0.434	0.051
10	300	0.116	-0.102	0.32	0.028
10	500	0.114	-0.136	0.256	0.021
10	1000	0.115	-0.056	0.206	0.015
10	2000	0.114	-0.087	0.14	0.011
Demand Estimation					
N	T	Median	MC Bias	Coverage	CI Length
10	50	-0.033	975.384	0.309	0.094
10	100	-0.037	-28.061	0.281	0.071
10	300	-0.035	206.014	0.209	0.041
10	500	-0.035	175.826	0.18	0.031
10	1000	-0.04	11.578	0.113	0.022
10	2000	-0.042	256.965	0.082	0.016

Table 10: Supply and Demand Estimation using MLE under Misspecification