Granular Instrumental Variables*

Xavier Gabaix and Ralph S.J. Koijen

July 30, 2019. Comments welcome.

Abstract

In many settings, there is a dearth of instruments, which hampers economists' ability to investigate causal relations. We propose a quite general way to construct instruments: "granular instrumental variables" (GIVs). In the economies we study, a few large firms or countries account for a large share of economic activity. As they are large, their idiosyncratic shocks affect aggregate outcomes. This makes those idiosyncratic shocks valid instruments for aggregate shocks. We provide a methodology to extract idiosyncratic shocks from the data, this way creating GIVs. Those GIVs allow us to then estimate parameters of interest, including causal elasticities.

We first illustrate the idea in a basic supply and demand framework: we achieve a novel identification of supply and demand elasticities, based on idiosyncratic shocks to supply or demand. We then show how the procedure can be adapted to handle many enrichments. We provide initial illustrations of the procedure with two applications. First, we measure how "sovereign yield shocks" spill over to other countries in the Eurozone. Second, we estimate short-term supply and demand elasticities in the oil market. Our estimates match well existing estimates that use much more complex and labor-intensive (e.g., narrative) methods. We sketch how GIVs could be useful to estimate a host of other causal parameters in economics.

*xgabaix@fas.harvard.edu, ralph.koijen@chicagobooth.edu. We thank Antonio Coppola, Rishab Guha, Dong Ryeol Lee, and Ashesh Rambachan for excellent research assistance. We thank Isaiah Andrews, Gary Chamberlain, Mark Gertler, Oleg Itskhoki, Jens Kvaerner, Erik Loualiche, Serena Ng, Jean-Marc Robin, Kelly Shue, Jim Stock, Harald Uhlig, Dacheng Xiu and seminar participants at the Bank of England, Berkeley Haas, Chicago Booth, Columbia, CREST, Harvard, HEC, NY Fed, NYU, Oslo, Stanford GSB, UCL, UCLA, Virginia, Wharton, and the NBER (corporate finance, and conference on "the rise of mega-firms") for comments.

1 Introduction

In many settings, there is a dearth of instruments, which hampers economists' understanding of causal relations (Ramey (2016); Nakamura and Steinsson (2018); Stock and Watson (2016, 2018); Chodorow-Reich (2017)). We propose a general way to construct instruments: "granular instrumental variables" (GIVs). Those instruments in turn allow to tease apart causal relations in a wide variety of economic contexts.

In the economies we study, many decisions are taken by a few large actors, such as firms, industries or countries. As has been observed before, those actors are often large, and their idiosyncratic shocks (e.g., productivity shocks) affect the aggregate ones.[1]

Those idiosyncratic firm- or country-level shocks are valid instruments for the aggregate shocks. Now, we need a methodology to extract those idiosyncratic shocks, and the paper presents such a methodology. This creates GIVs. Those GIVs then allow us to estimate parameters of interest.

We first illustrate the idea in a basic static setup with supply and demand (Section 2). It is a classic case, and we show how GIVs allow for a novel estimation procedure: they yield an instrument that allows us to estimate the elasticities of supply and demand. Indeed, idiosyncratic demand shocks to large firms or countries give a valid instrument for demand change – and thus allow one to estimate the elasticity of supply. They also allow us to estimate the elasticity of demand: the idiosyncratic demand shock of a large firm impacts the price, which changes the demand of other firms. We formalize those ideas, and present a way to "optimally extract" idiosyncratic shocks, this way constructing optimal GIVs.

We will see that some conditions are needed to obtain valid GIVs. Hence, GIVs are not quite a "free lunch" in constructing instruments, but a "very cost-effective lunch" that yields instruments at a modest cost.

Once the ideas are in place, we show in Section 3 how the procedure can be extended to handle many enrichments, such as feedback loops, heterogeneity, and several exogenous factors. We specify the procedure within this general framework.

Empirical illustrations We provide preliminary empirical results for two applications: sovereign yield spillovers and the equilibrium of global crude oil markets.

First, we study sovereign yield spillovers in the Eurozone in 2009-2018. If a country has an increase in its sovereign yield spread (i.e., on the yield on its government debt, minus the comparable yield for Germany), how much does that "spill over" to other Eurozone countries? We present a simple model that allows to think about that, and delivers a theoretically-grounded functional form for the shape of the spillovers (the modeling device we use is partial mutualization of debt, and we argue that other devices are likely to give a similar reduced form). Then, we use GIVs to

[1] Hence, economies are "granular": their shocks are made of incompressible "grains" of economic activity, at the firm, industry, or country level. This theme is laid out in Gabaix (2011), and developed in Acemoglu et al. (2012); di Giovanni and Levchenko (2012); Carvalho and Grassi (2019).

estimate that spillover. Specifically, we run a factor model for yield increases, and trace the impact of an idiosyncratic sovereign yield shock of one country on the other countries' yields. We find a "multiplier" of 1.5 (1 signifying the absence of spillovers). One implication is that an idiosyncratic increase of 200bps in the Italian yield spread leads to an increase of 20bps in the other countries' yield spreads, implying a "pass-through" of 0.1.

Second, we use GIVs to estimate the short-term demand and supply elasticities in the global crude oil market. Since the seminal work by Kilian (2009), who uses an ordered VAR to identify the shocks in a structural VAR, an active literature explores sign restrictions, informative priors, and narrative methods to estimate these elasticities (see for instance Kilian and Murphy (2014), Baumeister and Hamilton (2019), and Caldara et al. (2018)). We use country-level oil supply growth to construct the GIV, after removing common factors using principal components and OPEC membership to construct an OPEC factor. We find that the granularly identified elasticities are in the range documented in the literature. Moreover, given the apparent importance of demand shocks in crude oil markets during the last 15 years, future work can use disaggregated data on (net) imports, inventories, and oil consumption to sharpen the estimates and to estimate more general models to understand price fluctuations in oil markets.

Uses of GIVs GIVs allow to "democratize" and "automatize" instruments, especially in macro-finance where they have been rare. In standard practice, finding an instrument is a heroic and very ingenious affair. For instance, the "China shock" (entry of China in the World Trade Organization, Autor et al. (2013)) depends on detailed historical knowledge, and applies only to a specific time period. With GIVs, we can have a more systematic way to obtain instruments, that can apply more generally and over many time periods.

Once one thinks about causality and GIV procedures, the answers to many interesting questions feel suddenly within reach. We sketch a few here, hoping that they will inspire other researchers to investigate those and related topics with the help of GIVs.

Doom loops are the notion that when banks do badly, this will hamper the financial health of the state (as the state may need to bailout banks), and hence will increase the yield on the sovereign debt. This in turn will create a fall in bank returns, in a "doom loop" (Farhi and Tirole (2017)). How important are those doom loops quantitatively? Using the idiosyncratic returns of large banks, GIVs allow to estimate both channels, from banks to state and from state to bank. We plan to pursue this application.

If the Turkish Lira (to take a concrete example) appreciates, how does that affect Turkey's exports and borrowing? One could handle that via idiosyncratic demand shocks by large investment funds for the Turkish Lira (Koijen and Yogo (2019) provide a methodology for demand systems).[2]

If there is an export boom, what's the impact on the exchange rate, and the rest of the economy? Idiosyncratic shocks to large exporters will be useful for that, as recent research has shown them

[2] See Caballero and Simsek (2018) and Gabaix and Maggiori (2015) for models along those lines.

to be very large (di Giovanni et al. (2014); Gaubert and Itskhoki (2018); Kramarz et al. (2016)).

Do firm-specific hiring, investment and innovations spill over to peer firms operating in the same product market (i.e., what is the sign and magnitude of strategic complementaries)? Idiosyncratic innovation shocks to some large players will help construct the GIV.

How much do constraints of financial intermediaries (e.g., broker dealers) matter for asset prices? The GIV will rely on idiosyncratic shocks to intermediary wealth, which may be related to shocks to other parts of the banks.

How much do international shock propagate (e.g., how does a boom in Germany transmit to the rest of Europe)? Using idiosyncratic shocks (differentiating between productivity and demand shocks) to countries will help us answer that question.

Likewise, how do regional "micro" shocks propagate into macro shocks? GIVs allow to measure that, and also estimate a micro-to-macro multiplier.[3]

Related literature We relate to a number of literatures. We offer some brief pointers here, while offering a longer discussion in Section 6.2.

Instruments for macro. An active literature discusses identification strategies in macro (Ramey (2016); Nakamura and Steinsson (2018); Stock and Watson (2018); Chodorow-Reich (2017)). We add to it, by proposing to use GIVs, which are quite ubiquitous. There are lots of idiosyncratic shocks, and GIVs allow us to construct them systematically.

Origins of aggregate fluctuations. A growing literature finds that a sizable amount of volatility is "granular" in nature – coming from idiosyncratic shocks to firms or industries (see Long and Plosser (1983); Gabaix (2011), Acemoglu et al. (2012); di Giovanni and Levchenko (2012); di Giovanni et al. (2014); Baqaee and Farhi (2018a); Pasten et al. (2017); Carvalho and Grassi (2019)). We provide tools that can tease that apart in the presence of loops. Datasets used in this literature can be revisited forming GIVs which allow us to investigate causal relations. Gabaix (2011) introduces the notion of "granular residual" – a weighted sum of proxies for idiosyncratic shocks, and shows how idiosyncratic shocks to firms appear to explain about one third of GDP fluctuations. But that paper does not take the crucial step to use this kind of concept as an instrument to measure causal relations, e.g. in a demand and supply setting.

The idea that we propose is, in retrospect, so natural that we suspected it may have been already proposed in the literature, perhaps in a forgotten paper in the 1940s. However, after searching the literature and consulting with many experts, we could not find it. We are quite sure that this idea has not been systematically implemented in mainstream economic applications. The idea to use idiosyncratic shocks as instruments to estimate spillover effects has been explored in several creative papers, as we discuss in more detail in Section 6.2, such as Leary and Roberts (2014b), Amiti and Weinstein (2018), and Amiti et al. (2019). However, the typical approach is to use idiosyncratic shocks to other variables that are excluded from the main estimating equation

[3] We are pursuing this last question in ongoing work.

to instrument for the actions of competing firms. We, instead, use the idiosyncratic shocks in the estimating equation directly. In addition, we allow for more flexible exposures to unobserved common shocks in extracting idiosyncratic shocks.

Plan The reader is encouraged to read Section 2, Section 4 and Section 5 at first, and then go to Section 3, which contains the general procedure, in a second reading. Section 6 presents a number of extensions and robustness checks, including an application to the estimation of differentiated product demand systems, and discusses more extensively the link with the rest of the literature. Section 7 presents simulations. Section 8 concludes. Long proofs are in Section 9 and the online appendix.

2 The basics

2.1 Notations

We will throughout use the following notations. For a vector $X = (X_i)_{i=1\ldots N}$ and a series of relative weights S_i with $\sum_i S_i = 1$, we let

$$X_E \coloneqq \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{1}$$

$$X_S \coloneqq \sum_{i=1}^{N} S_i X_i, \tag{2}$$

$$X_\Gamma \coloneqq X_S - X_E, \tag{3}$$

so that X_E is the equal weighted average of the vector's elements, X_S is the size weighted average, and X_Γ is their difference.

We will also have shocks u_i that are uncorrelated and with variance $\sigma^2_{u_i}$. Then, we will define the "pseudo-equal weights" or "inverse variance weights"

$$\tilde{E}_i \coloneqq \frac{1/\sigma^2_{u_i}}{\sum_j 1/\sigma^2_{u_j}}, \tag{4}$$

which satisfy $\sum_i \tilde{E}_i = 1$, and are equal to $\tilde{E}_i = \frac{1}{N}$ when all the $\sigma^2_{u_i}$ are equal. Then $X_{\tilde{E}} \coloneqq \sum_{i=1}^{N} \tilde{E}_i X_i$. We'll also define

$$\tilde{\Gamma}_i \coloneqq S_i - \tilde{E}_i. \tag{5}$$

Then, $X_{\tilde{\Gamma}}$ will be the "granular residual" in a number of settings. It is the size-weighted sample average of X minus the "inverse-variance" weighted sample average of X. It will be an optimal proxy for idiosyncratic shocks.

We use the notation β^e for the estimator of a parameter β; $\mathbb{E}_T$ for the sample temporal mean, $\mathbb{E}_T[Y_t] := \frac{1}{T}\sum_{t=1}^{T} Y_t$; C_t for a vector of controls; ι for a vector of 1's, I for the identity matrix, of the appropriate dimension given the context; $V^Y = \mathbb{E}[Y_t Y_t']$ for a variance-covariance matrix; $X \perp Y$ to say that random variables X and Y are uncorrelated.

2.2 A very simple example with no feedback loop

We introduce GIVs by considering a very simple example with supply and demand.

2.2.1 Basic model

For clarity, we lay out a concrete economic model of the equilibrium in, for instance, the oil market. Demand by country i at date t is $D_{it} = \bar{Q} S_i (1 + y_{it})$, where $\bar{Q}$ is the average total world production, y_{it} is a demand disturbance term, and S_i is country $i's$ share of demand, normalized to follow:

$$\sum_{i=1}^{N} S_i = 1. \tag{6}$$

The demand disturbance is assumed to be the sum of a common shock η_t, and an idiosyncratic shock u_{it}:

$$y_{it} = \lambda_i \eta_t + u_{it}. \tag{7}$$

For now we consider the case with uniform loadings,

$$\lambda_i = 1, \tag{8}$$

but we will relax that soon.

All shocks are i.i.d. across dates. Then, total world demand is $D_t = \sum_i D_{it} = \bar{Q}(1 + y_{St})$, where $y_{St} := \sum_i S_i y_{it}$ is the size-weighted average demand disturbance. We suppose that supply is $Q_t = \bar{Q}\left(1 + \frac{p_t - \varepsilon_t}{\alpha}\right)$, where $p_t = \frac{P_t - \bar{P}}{\bar{P}}$ is the proportional deviation from $\bar{P}$, which is thus the average price of oil. Then, to equilibrate supply and demand ($D_t = Q_t(p_t)$), we must satisfy: $\bar{Q}(1 + y_{St}) = \bar{Q}\left(1 + \frac{p_t - \varepsilon_t}{\alpha}\right)$. That is, the deviation of the price from the average satisfies:

$$p_t = \alpha y_{St} + \varepsilon_t. \tag{9}$$

It depends on the size-weighted average demand shock, $y_{St} = \sum_i S_i y_{it}$.

The classic problem is that we cannot estimate α by OLS. Indeed, a direct regression of p_t on y_{St} (that is, a regression of the form $p_t = \alpha y_{St} + \varepsilon_t$) would be biased, as ε_t and η_t (hence ε_t and y_{St}) can be correlated.

However, suppose that we form the GIV:

$$z_t := y_{\Gamma t} = y_{St} - y_{Et} = \sum_{i=1}^{N} S_i y_{it} - \sum_{i=1}^{N} \frac{1}{N} y_{it}. \tag{10}$$

Then, we have, using $u_{St} := \sum_{i=1}^{N} S_i u_{it}$, $u_{Et} := \sum_{i=1}^{N} \frac{1}{N} u_{it}$, that

$$y_{St} = \sum_{i=1}^{N} S_i y_{it} = \eta_t + u_{St}, \qquad y_{Et} = \sum_{i=1}^{N} \frac{1}{N} y_{it} = \eta_t + u_{Et},$$

so $z_t := y_{St} - y_{Et} = (\eta_t + u_{S_t}) - (\eta_t + u_{Et})$ satisfies

$$z_t = u_{St} - u_{Et} =: u_{\Gamma t}. \tag{11}$$

Note that $z_t := y_{St} - y_{Et}$ is just constructed from observables. It is the difference between the size-weighted demand and the equal-weighted demand. Intuitively, it captures the "idiosyncratic demand" by large units, as shown by $z_t = u_{\Gamma t}$.

We assume that the shocks u_{it} are idiosyncratic, in the sense that:

$$\mathbb{E}\left[u_{it}\varepsilon_t\right] = 0 \text{ for all } i, t. \tag{12}$$

This "exogeneity" or "exclusion" assumption needs to be discussed in each economic application — as we will below. More minor, for simplicity, the u_{it} are here i.i.d. over time, but the u_{it}, u_{jt} could be correlated.[4]

Then, we have

$$\mathbb{E}\left[z_t\varepsilon_t\right] = 0: \text{ Exogeneity}, \tag{13}$$

and

$$\mathbb{E}\left[z_t y_{St}\right] \neq 0: \text{ Relevance}. \tag{14}$$

Hence, $z_t = u_{\Gamma t}$ is a valid instrument (and as Proposition 3 will show, an optimal one). We call it a "granular instrumental variable" (GIV).

Given that $p_t - \alpha y_{St} = \varepsilon_t$, we have

$$\mathbb{E}\left[\left(p_t - \alpha y_{St}\right) z_t\right] = 0, \tag{15}$$

that is, $\mathbb{E}\left[p_t z_t\right] - \alpha \mathbb{E}\left[y_{St} z_t\right] = 0$, which gives the supply elasticity α, by

$$\alpha = \frac{\mathbb{E}\left[p_t z_t\right]}{\mathbb{E}\left[y_{St} z_t\right]}. \tag{16}$$

[4]If we have disaggregated data for both supply and demand, we can relax that condition (12), see Section 12.5.

Indeed, in practice, we might estimate α using sample means:[5]

$$\alpha_T^e := \frac{\frac{1}{T}\sum_t p_t z_t}{\frac{1}{T}\sum_t y_{St} z_t}. \tag{17}$$

We now state a formal proposition.[6]

Proposition 1 (Consistency of the GIV estimator in this example). *Suppose that $\mathbb{E}[u_{it}\varepsilon_t] = 0$ (but the u_{it} can have an arbitrary distribution, with mean 0), and we have a succession of i.i.d. dates t. Form the GIV estimator $z_t := y_{\Gamma t}$. Then, z_t identifies the price elasticity, by $\alpha = \frac{\mathbb{E}[p_t z_t]}{\mathbb{E}[y_{St} z_t]}$. In other terms, the GIV estimator for the price elasticity α, $\alpha_T^e := \frac{\frac{1}{T}\sum_t p_t z_t}{\frac{1}{T}\sum_t y_{St} z_t}$, is a consistent estimator.*

Precision of the GIV estimator We define the excess Herfindahl as $h = \sqrt{-\frac{1}{N} + \sum_{i=1}^N S_i^2}$. In the context of industries, for example, a higher $h \in \left[0, \sqrt{1 - \frac{1}{N}}\right]$ means that the industry is more concentrated: an industry where all the firms have the same size features $h = 0$.

The quantity h will prove to be analytically useful, since if $(u_i)_{i=1\ldots N}$ is a series of uncorrelated random variables with mean 0 and common variance σ_u^2, then the volatility of the GIV $z_t = u_{\Gamma t}$ is:

$$\sigma_{u_\Gamma} = h\sigma_u. \tag{18}$$

The next proposition states the conditions under which we have a precise estimator (its proof is in Section 9).

Proposition 2 (Precision of the GIV estimator in this example). *The above estimator based on the granular instrument variable (GIV) $y_{\Gamma t}$ achieves identification of the elasticity parameter α, at the following rate, for $T \to \infty$:*

$$\sqrt{T}\left(\alpha_T^e - \alpha\right) \sim \mathcal{N}\left(0, \sigma_\alpha^2\right),$$

where $\sigma_\alpha = \frac{\sigma_\varepsilon}{\sigma_{u_\Gamma}}$. If we assume further than the u_{it} are i.i.d. with variance σ_u^2,

$$\sigma_\alpha = \frac{\sigma_\varepsilon}{h\sigma_u}, \tag{19}$$

where h is the excess Herfindahl:

$$h := \sqrt{-\frac{1}{N} + \sum_{i=1}^N S_i^2}. \tag{20}$$

[5] One can also use z_t to estimate α by OLS, as in the regression $p_t = \alpha z_t + \varepsilon_t^0$, or even $p_t = \alpha u_{\Gamma t} + \beta \eta_t^e + \varepsilon_t^p$. Section 13 develops this.

[6] It holds under mild regularity conditions on the joint distribution of $u_{it}, \eta_t, \varepsilon_t$ given that the data are i.i.d. across dates.

So in order to have a precise estimate (low σ_α), we need: some large units (in order to have a large excess Herfindahl h), and that idiosyncratic shocks are large compared to aggregate shocks (large $\sigma_u/\sigma_\varepsilon$).

This simple example illustrates the basic idea. The reader might at this point have in mind a number of questions and objections: What if the factor structure is non-trivial (for instance, we don't have $\lambda_i = 1$ in (7))? What if the demand is sensitive to price? Is the GIV that we constructed the best instrument we can find? What happens if there are more feedback loops?

The next subsections are devoted to answering them in turn.

2.2.2 Time-varying size weights

Suppose that we have a time-varying size $S_{i,t-1}$, so that the demand increase is $\sum_i S_{i,t-1} y_{it}$, with $(1, S_{i,t-1})\, u_{it} \perp (\eta_t, \varepsilon_t)$. Then everything goes through without problems, replacing S_i by $S_{i,t-1}$ throughout. The basic GIV becomes: $z_t = y_{S_{t-1},t} - y_{Et} = \sum_i \left(S_{i,t-1} - \frac{1}{N}\right) y_{it}$.

2.2.3 Model with an enriched factor structure

The model might have a richer factor structure, with r factors, i.e. instead of (7) we have:

$$y_{it} = \sum_{f=1}^{r} \lambda_i^f \eta_t^f + u_{it}, \tag{21}$$

or, in vector form:

$$y_t = \Lambda \eta_t + u_t, \tag{22}$$

where Λ is a $N \times r$ matrix, and $u_t \perp (\eta_t, \varepsilon_t)$.

Then, in order to construct a valid GIV we simply run a factor model – for example, via Principal Component Analysis (PCA) – and, in essence, we extract the residuals u_{it} to form the GIV. Let us see that more precisely. Suppose that we have estimated the λ vector (e.g. via PCA, as we will detail later). Then, let Q be a $N \times N$ matrix projecting vectors onto a space orthogonal to Λ, so that $Q\Lambda = 0$.[7] Then, $Qy_t = Qu_t$. Hence, via factor analysis, we obtain the transformed residuals $\check{u}_t = Qu_t$. Then, the GIV is formed as:

$$z_t := S'\check{u}_t = S'Qy_t = \Gamma' y_t, \qquad \Gamma := Q'S, \tag{24}$$

[7] For instance, we can take $Q = Q^{\Lambda,W}$, where

$$Q^{\Lambda,W} := I - \Lambda \left(\Lambda' W \Lambda\right)^{-1} \Lambda' W, \tag{23}$$

with $W = (V^u)^{-1}$ (optimally) or $W = I$ for simplicity. This choice satisfies $Q^{\Lambda,W}\Lambda = 0$ and has a number of good properties listed in (177).

so that

$$z_t = \Gamma' u_t. \tag{25}$$

Then, z_t is a valid instrument, since it is composed of idiosyncratic shocks. Since $p_t - \alpha y_{St} = \varepsilon_t$ and $\mathbb{E}[u_t \varepsilon_t] = 0$, we have $\mathbb{E}[(p_t - \alpha y_{St}) z_t] = 0$, i.e. (15).[8]

This generalizes our basic example (7). In that example, we had $Q = I - \iota E'$, so that $\check{u}_{it} = u_{it} - u_{Et}$, and the GIV was: $z_t = \check{u}_{St} = u_{St} - u_{Et}$. We therefore had $\Gamma = Q'S = S - E$.

2.3 A simple demand and supply example with feedback loops

A simple model We next enrich the previous example, and consider a simple supply and demand example that features a "loop." Suppose that demand for some commodity (say, oil) is:

$$y_{it} = \phi^d p_t + \eta_t + u_{it}, \tag{26}$$

and supply is

$$s_t = \phi^s p_t + \varepsilon_t, \tag{27}$$

where η_t, ε_t can be correlated. We can expect that the demand and supply elasticities (respectively ϕ^d and ϕ^s) satisfy $\phi^d < 0 < \phi^s$. Again, to be more formal, y_{it}, s_t, and p_t are understood as percent deviations from the average demand of country i, from supply, and from price, respectively.[9]

In equilibrium, supply equals demand, $y_{St} = s_t$, which gives the price

$$p_t = \frac{u_{St} + \eta_t - \varepsilon_t}{\phi^s - \phi^d}. \tag{28}$$

There is a "loop" because the demand shocks η_t and u_{it} feed into the price p_t, which then in turns affects demand. The equilibrium quantity produced is

$$s_t = y_{St} = \frac{\phi^s u_{St} + \phi^s \eta_t - \phi^d \varepsilon_t}{\phi^s - \phi^d}. \tag{29}$$

The classic problem of estimating supply and demand equilibrium quantity s_t and price p_t is that we cannot regress: $s_t = \beta p_t + \varepsilon_t$, and hope to get $\beta = \phi^s$, as ε_t and p_t are correlated.

However, suppose that we form the GIV, as in (10)

$$z_t := y_{St} - y_{Et}. \tag{30}$$

[8] This shows that the GIV is valid and possible as long as $\Gamma := Q'S \neq 0$. Fortunately, this is generically true. If Γ were close to 0, that would be picked up by very large standard errors.

[9] We take the model of Section 2.2, and simply set $y_{it} = \phi^d p_t + \eta_t + u_{it}$, where $p_t = \frac{P_t - P_*}{P_*}$ is the proportional deviation from the average.

Given that

$$y_{St} = \phi^d p_t + \eta_t + u_{St}, \qquad y_{Et} = \phi^d p_t + \eta_t + u_{Et},$$

we have:

$$z_t = u_{St} - u_{Et} =: u_{\Gamma t}. \tag{31}$$

As in the previous example, we assume that the shocks u_{it} are idiosyncratic:

$$\mathbb{E}\left[u_{it}\eta_t\right] = \mathbb{E}\left[u_{it}\varepsilon_t\right] = 0 \text{ for all } i, t. \tag{32}$$

Then, we have again a valid instrument:

$$\mathbb{E}\left[z_t\varepsilon_t\right] = \mathbb{E}\left[z_t\eta_t\right] = 0: \text{ Exogeneity,}$$

$$\mathbb{E}\left[z_t p_t\right] \neq 0: \text{ Relevance.}$$

Estimations of supply and demand elasticities by GIV The supply equation (27) implies

$$\mathbb{E}\left[\left(s_t - \phi^s p_t\right) z_t\right] = 0, \tag{33}$$

which gives the supply elasticity ϕ^s by

$$\phi^s = \frac{\mathbb{E}\left[s_t z_t\right]}{\mathbb{E}\left[p_t z_t\right]}. \tag{34}$$

Indeed, in practice, we form the sample average $\phi_T^{s,e} = \frac{\mathbb{E}_T[s_t z_t]}{\mathbb{E}_T[p_t z_t]}$.

Now, we want to estimate demand. For that, we make a stronger assumption: we assume that the shocks u_{it} are i.i.d. across i's and not just dates (we will relax this later). Then, this implies[10]

$$\mathbb{E}\left[u_{Et}u_{\Gamma t}\right] = 0. \tag{35}$$

So, given this, we have: $y_{Et} - \phi^d p_t = \eta_t + u_{Et}$, and $\mathbb{E}\left[\left(y_{Et} - \phi^d p_t\right) z_t\right] = 0$. This gives an estimate of the demand elasticity ϕ^d,

$$\phi^d = \frac{\mathbb{E}\left[y_{Et} z_t\right]}{\mathbb{E}\left[p_t z_t\right]}, \tag{36}$$

and the estimator is $\phi_T^{d,e} = \frac{\mathbb{E}_T[y_{Et} z_t]}{\mathbb{E}_T[p_t z_t]}$.

[10] Indeed, in the i.i.d. case we have

$$\mathbb{E}\left[u_{Et}u_{\Gamma t}\right] = \mathbb{E}\left[(\sum_i \frac{1}{N} u_{it})(\sum_i \Gamma_i u_{it})\right] = \frac{1}{N}\sum_i \Gamma_i \sigma_u^2 = 0.$$

as $\sum_i \Gamma_i = 0$. Equation (79) generalizes this to the non-i.i.d. case.

Estimation by OLS and interpreting it as a first- and second-stage IV estimator Let us recast our GIV in the language of applied microeconomics, and estimate the parameters by OLS (as we will often do in the general case). Recall that the solutions are:

$$p_t = \frac{1}{\phi^s - \phi^d} u_{St} + \varepsilon_t^p, \qquad s_t = y_{St} = \frac{\phi^s}{\phi^s - \phi^d} u_{St} + \varepsilon_t^s,$$

where the $\varepsilon_t^p, \varepsilon_t^s$ are linear combinations of ε_t, η_t. So, if we run the OLS regression, with $z_t = u_{\Gamma t}$,

$$p_t = b^p z_t + \varepsilon_t^p, \tag{37}$$

we estimate

$$b^p = \frac{1}{\phi^s - \phi^d}, \tag{38}$$

which is the sensitivity of the price to the supply or demand shock. If we run the OLS regression:

$$y_{St} = b^{y_S} z_t + \varepsilon_t^s, \tag{39}$$

we estimate

$$b^{y_S} = \frac{\phi^s}{\phi^s - \phi^d} = M. \tag{40}$$

In the language of applied microeconomics, one can view the "first stage" as a regression of the price on the GIV (37). The "second stage" is running supply on the instrumented change in the price $b^p z_t$:

$$s_t = \phi^s \left(b^p z_t\right) + \varepsilon_t^s, \tag{41}$$

which gives ϕ^s. Alternatively, one can run the "reduced form equation" (39), which estimates M. The supply elasticity is given by:

$$\phi^s = \frac{b^{y_S}}{b^p}. \tag{42}$$

The demand elasticity is similar. In the second stage we run equal-weighted demand on the instrumented change in the price, $b^p z_t$:

$$y_{Et} = \phi^d \left(b^p z_t\right) + \varepsilon_t^y, \tag{43}$$

which gives the demand elasticity ϕ^d.[11] Alternatively, we can run the reduced form equation $y_{Et} = b^{y_E} z_t + \varepsilon_t^y$ which gives $b^{y_E} = \phi^d M$, and the demand elasticity is $\phi^d = \frac{b^{y_E}}{b^p}$.

In practice, we will add controls to those regressions, including estimates of η_t recovered from factor analysis.

From M and b^p, we can recover the elasticities ϕ^s and ϕ^d. This is exactly the same estimate as

[11] Here we used (35), which makes the OLS valid.

the IV estimator, derived earlier in (34), (36).[12]

Standard errors: When "weak instruments" are or are not a problem When estimating via OLS (e.g. b^p and M), the standard errors are reliably estimated by the usual OLS method, even in small samples. When a ratio is implicitly performed (e.g. to estimate ϕ^d, ϕ^s), the 2SLS procedure as in (41) will also give correct standard errors when the instrument is strong enough. A good rule of thumb for the strength of the instrument is that the F statistics (which is the squared t-statistic on b^p) on the first stage (37) should be greater than 10.[13]

2.4 Optimality of the GIV

We come back to the simplest case of Section 2.2.1, for ease of exposition.[14] Above, we have shown that $z_t = y_{\Gamma t}$ allows for identification, for a specific $\Gamma = S - E$. It is easily verified that GIV with weights Γ such that $\sum_i \Gamma_i = 0$ would work. Hence, we can ask for an optimal Γ. The Γ we proposed initially was actually optimal, as we formalize below.

Proposition 3 (Optimal weights Γ for the GIV $y_{\Gamma t}$). *Consider the GIV $z_t = y_{\Gamma t} = \sum_i \Gamma_i y_{it}$, with some weights Γ_i with $\sum_i \Gamma_i = 0$. The idiosyncratic shocks u_{it}'s are assumed to be i.i.d. across time, and have variance-covariance matrix V^u. Then, in the basic supply and demand model of Section 2.2, the asymptotic variance of the estimator α_T^e in 17 (which is $\sigma_\alpha^2 = \lim_{T\to\infty} T var\left(\alpha_T^e - \alpha\right)$) satisfies $\sigma_\alpha^2\left(\Gamma\right) = \frac{\sigma_\varepsilon^2 \mathbb{E}\left[y_{\Gamma t}^2\right]}{\mathbb{E}\left[y_{St} y_{\Gamma t}\right]^2}$. The value*

$$\tilde{\Gamma} = S - \tilde{E}, \qquad \tilde{E} := \frac{\left(V^u\right)^{-1}\iota}{\iota'\left(V^u\right)^{-1}\iota}, \tag{44}$$

gives the optimal GIV estimator, in the sense that for any other Γ that is not collinear to $\tilde{\Gamma}$, the asymptotic variance $\sigma_\alpha^2\left(\Gamma\right)$ is larger. When the shocks are i.i.d., this implies $\tilde{E}_i = \frac{1}{N}$, and when they are uncorrelated, this implies $\tilde{E}_i := \frac{1/\sigma_{u_i}^2}{\sum_j 1/\sigma_{u_j}^2}$, so that $\tilde{E}$ may be called the "precision-weighted quasi-equal" weights.

Hence, the "essence" of the GIV is not to be "size weighted minus value weighted" idiosyncratic shocks, but rather "size weighted minus precision weighted" (i.e. inverse-variance weighted when shocks are uncorrelated) idiosyncratic shocks.

There are two more ways in which the GIV is optimal. First, it is the optimally-weighted GMM estimator.[15] This implies that other combinations of idiosyncratic shocks (besides weighing by

[12] Indeed, the OLS estimators are $M_T^e = \frac{\mathbb{E}_T[y_{St} z_t]}{\mathbb{E}_T[z_t^2]}$ and $\psi_T^e = \frac{\mathbb{E}_T[p_t z_t]}{\mathbb{E}_T[z_t^2]}$. We have $\phi_T^{s,e} = \frac{M_T^e}{\psi_T^e} = \frac{\mathbb{E}_T[y_{St} z_t]}{\mathbb{E}_T[p_t z_t]}$, which is the same as (34), as $s_t = y_{St}$ in equilibrium.

[13] See the literature on weak instruments, e.g. Stock and Yogo (2005); Andrews et al. (2018).

[14] But it will be clear to the reader that the results in the present subsection hold much more generally.

[15] Any moment $\mathbb{E}_T\left[\left(p_t - \alpha y_{St}\right)\left(u_{it} - u_{Et}\right)\right] = 0$ is a valid GMM moment to identify α. It is easy to check that the optimal GMM weighted estimator is our GIV, $\mathbb{E}_T\left[\left(p_t - \alpha y_{St}\right)\left(u_{St} - u_{Et}\right)\right] = 0$

Γ) would not help the precision of the estimator. Second, one can show that it is the maximum likelihood estimator, if we assume that all shocks are Gaussian (see Section 14). Still, the optimality formulation of Proposition 3 is the simplest to use in other contexts.

2.5 Interpreting and diagnosing idiosyncratic shocks

What is an idiosyncratic shock? Mathematically, an idiosyncratic shock is plainly a random variable u_{it} such that $\mathbb{E}_{t-1}[\tilde{\eta}_t u_{it}] = 0$, where $\tilde{\eta}_t = (\eta_t, \varepsilon_t)$ includes all the common shocks. But it may be useful to discuss different types of economic settings that map into that definition.

In some cases it is quite clear – e.g. a random productivity shock, or demand shock. But there are more subtle types of idiosyncratic shocks. One is an "unexpected change in the loading on a common shock". For instance, suppose that OPEC decided to cut down production, which in the language of our example is an aggregate η_t shock. If Saudi Arabia cuts down production by more than anticipated, that is an idiosyncratic shock. Formally, if supply is $y_{it} = \phi^s p_t + \left(\lambda_i + \check{\lambda}_{it}\right)\eta_t + v_{it}$, with $\mathbb{E}_{t-1}\left[(1, \eta_t \tilde{\eta}_t)\,\check{\lambda}_{it}\right] = 0$, then $u_{it} = \check{\lambda}_{it}\eta_t + v_{it}$ is a bona fide idiosyncratic shock. To take another example, suppose that we hear about a change in real estate prices in the economy, η_t, but that a bank i was more exposed to it than anticipated: the market thought the bank's equity would move by $\lambda_i \eta_t$, but it moved by $r_{it} = \left(\lambda_i + \check{\lambda}_{it}\right)\eta_t$ for an expectational surprise $\check{\lambda}_{it}$ with $\mathbb{E}_{t-1}\left[(1, \eta_t \tilde{\eta}_t)\,\check{\lambda}_{it}\right] = 0$. Then, the bank will have an idiosyncratic shock $u_{it} = \check{\lambda}_{it}\eta_t$ as part of its total return r_{it}.

Likewise suppose that the news is that a bank failed a stress test (while it was anticipated it would pass the test). This is an idiosyncratic shock. However, the bank could have failed the test because of some development in the macroeconomy η_t. Then, provided that the factor model allows for a rich enough structure in η_t, the latter will be controlled for.

Likewise, the volatility of idiosyncratic shocks can depend on the common shocks. Suppose that $u_{it} = \sigma_t v_{it}$ where σ_t and η_t could be correlated (for instance, σ_t could increase when $|\eta_t|$ is high), but $\mathbb{E}_{t-1}[\sigma_t \tilde{\eta}_t v_{it}] = 0$ (a sufficient condition is that v_{it} independent of $\sigma_t \tilde{\eta}_t$); then, u_{it} is an idiosyncratic shock in the sense that $\mathbb{E}_{t-1}[\tilde{\eta}_t u_{it}] = 0$.

Thresholded and narrative GIVs In applications, it is possible to make further progress by assessing the drivers of the top shocks narratively. One procedure is to simply select the top K shocks by $S_i |\check{u}_{it}|$ (where $\check{u}_{it}$ is the residual from factor analysis, e.g. $\check{u}_{it} = u_{it} - u_{Et}$, and we select across all actors i and dates t), and check in the news what happened on that day (and check that the shocks are idiosyncratic indeed). We did that for some our applications. Formally, that means that we formulate a "thresholded" GIV,

$$z_t^{\tau} = \sum_i \tau\left(S_i \check{u}_{it}\right), \tag{45}$$

using the thresholding function $\tau(x) = x1_{|x|\geq b}$, which only keeps granular shocks bigger than $b > 0$.[16] Then, the GIV procedure works using that "thresholded" GIV (see Section 12.8). This thresholded GIV might also be useful to assess non-linear effects, for instance, in case of demand or supply curves.

After examining those largest shocks by looking at the news, some shocks might be eliminated as not idiosyncratic; we can call $I_t^{\mathcal{N}}$ the set of shocks that are "narratively certified" to be idiosyncratic by this procedure, and form the alternative instrument

$$z_t^{\mathcal{N}} = \sum_{i \in I_t^{\mathcal{N}}} S_i \check{u}_{it}. \tag{46}$$

This is roughly what the "narrative" approach in the literature (e.g. Caldara et al. (2018)) does. Here, in addition, we have a systematic way to select the candidate large shocks (by top values of $S_i |\check{u}_{it}|$), and get the controls η_t^e for the idiosyncratic shocks, when we run the regressions.

Sporadic factors A potential issue is that of a "sporadic factor", i.e. a factor η_t that affects a few actors special ways, but is not recurrent. An example would be a one-off policy announcement by the European Central Bank that they will buy both Italian and Spanish bonds, so that the truth is not that Italy is affecting Spain or vice-versa, but rather the ECB affecting both.

One solution, besides the narrative check that we just detailed, would be to filter out days with a high "sporadicity statistic" $\mathcal{S}_t$ that we now propose. Suppose that for each date we filter out the idiosyncratic shocks $\check{u}_{it}$. For each date and actor i we form $b_{it} = \frac{\check{u}_{it}^2}{\sigma_{u_i,t-1}^2}$, where a high b_{it} is an indicator of extra activity, and $\sigma_{u_i,t-1}^2$ is a predictor of the volatility of u_{it}. We may allow that one entity has a large idiosyncratic shock, but if two (or more) do, this is suspicious, and possibly the sign of a sporadic factor. So, calling $b_{(2)t}$ the activity of the second more active actor, we form $\mathcal{S}_t = b_{(2)t}$.[17] Over the entire sample, we might remove the days with anomalously high sporadicity statistics, e.g. in the top 5% by that metric.

Quasi-experimental instruments and identification by functional form A large literature explores identification by functional form, where consistency of the estimator depends on functional form or distributional assumptions. Classic examples include the Heckman (1978) selection model, identification via heteroskedasticity, as in Rigobon (2003) and Lewbel (2012), and Arellano and Bond (1991) and Blundell and Bond (1998) in the context of dynamic panel data models. The typical concern with these approaches, compared to quasi-experimental instruments that are outside of the model, is that the estimators are inconsistent when the model is misspecified.

In the case of GIVs, we generally start from a structural model that motivates the estimating equation, as for instance in the model of doom loops in Section 17, which prescribes the definition

[16] We adjust b to select a pre-specified expected number K of shocks that survive the thresholding.
[17] We could also sum over the most active K entities, excluding the most active one.

of the size vector S and, in some cases, the characteristics that determine the exposures x_{it}. To extract idiosyncratic shocks, we rely on statistical factor models.[18]

Instead of viewing this last step as a merely statistical exercise that is hard to validate externally, GIVs provide an empirical strategy to understand the economic drivers of the instrument by screening the top shocks narratively. By understanding the nature of the shock based on news coverage (as in the narrative examination we just discussed), for instance, we can ensure that the shocks are truly idiosyncratic and interpretable. For instance, a large negative return associated with a failed stress test of a bank in the context of doom loops, a negative supply shock in Kuwait and Iraq during the First Gulf War, or a positive demand shock in China in the early 2000s in the context crude oil markets, are all valid instruments. While alternative identification methods might rely on functional form assumptions only, GIVs, by being able to screen the shocks economically, provide a systematic way to construct instruments more in the spirit of quasi-experimental instruments.

2.6 An over-identification test with multiple GIVs

Consider the model of demand with a single factor with heterogenous exposures

$$y_{it} = \phi^d p_t + \lambda_i \eta_t + u_{it},$$

and assume we obtained an estimate of the factor, η_t^e, as in Section 2.7. We abstract from estimation error in the factor in this section.

Suppose we construct two different instruments, z_{1t} and z_{2t}. These could include size weighted averages of all $\check{u}_{it}^e$; a subset of the largest realizations of $S_i\,|\check{u}_{it}^e|$; a subset of narratively-checked shocks as in Section 2.5; or z_{1t} might be based on supply shocks and z_{2t} on demand shocks, as in Section 12.5.[19]

One can then estimate separately the parameters of interest (e.g. ϕ^d) based on z_{1t} vs z_{2t}, and see if they are economically different. We can also do a formal test. We form the moment conditions as $\mathbb{E}\left[g_t(\theta)\right] = 0$, where $g_t(\theta) = \left(y_{Et} - \phi^d p_t - \lambda_E \eta_t^e\right)(z_{1t}, z_{2t}, \eta_t^e)'$ and $\theta = (\phi^d, \lambda_E)$. We can simply perform the Sargan-Hansen $J-$test for over-identifying moment conditions. The test statistic is given by

$$J = T g_{Et}' W_T g_{Et} \to^d \chi_1^2, \tag{47}$$

under the null, where $W_T = \left(\frac{1}{T}\sum_{t=1}^T g_t g_t'\right)^{-1}$.[20]

[18]We discuss the robustness of GIVs to various forms of misspecification in Section 2.8.

[19]Yet another GIV procedure is to use characteristics x_{it} measurable at time $t-1$ (e.g., firm size, or GDP per capita, or a bank's credit risk), and form the x-weighted GIV: $z_t^x := \sum_i S_i \check{u}_{it}^e x_{it}$. If the test fails, it's probably the case that x_{it} is economically important and it should have been included as a factor loading in a larger factor model.

[20]If we have K instruments z_{kt}, $k = 1 \dots K$, then the procedure is the same, with $g_t(\theta) = \left(y_{Et} - \phi^d p_t - \lambda_E \eta_t^e\right)(z_{1t}, \dots, z_{Kt}, \eta_t^e)'$, and then $J \to^d \chi_{K-1}^2$.

2.7 A step-by-step user's guide

We can also have a richer factor structure, as in Section 2.2.3, and add observable factors. We can specify demand as

$$y_{it} = \phi^d p_t + \lambda_i \eta_t + mC_{it}^y + u_{it}, \tag{48}$$

where λ_i,η_t could be multidimensional and C_{it}^y is an observable vector of controls. The arguments in this section extend to multiple factors, as illustrated in Section 3.

The GIV procedure is as follows:

1. Estimate a panel regression
$$y_{it} = a_t + mC_{it}^y + e_{it},$$
where $e_{it} = \check{\lambda}_i \eta_t + \check{u}_{it}$.

2. There are two approaches to estimate the factor, η_t, depending on whether one has information about the factor loadings, λ_i.

    (a) If one has information about the factor loadings, $\check{\lambda}_i = bx_i + \zeta_i$, with x_i an observable characteristic with $\mathbb{E}[x_i \zeta_i] = 0$, then we can estimate the factors using cross-sectional regressions of e_{it} on x_i. Denote the estimated factor by $\eta_t^{x,e}$.

    (b) If no information on the factor loadings is available, we can estimate the factors via PCA or factor analysis. Denote the estimated factor by $\eta_t^{PCA,e}$. One can use one of the tests in Bai and Ng (2002); Onatski (2009) to determine the number of factors.

3. We regress $y_{\Gamma t} = \alpha_0 \eta_t^e + \alpha_1 C_{\Gamma t}^e + u_{\Gamma t}^e$, and set $z_t := u_{\Gamma t}^e$.[21] We proceed as above (Section 2.3), e.g. estimate the demand elasticity by the moment condition
$$\mathbb{E}\left[\left(y_{Et} - \phi^d p_t - \lambda_E \eta_t^e - mC_{Et}^y\right)\left(z_t, \eta_t^e, C_{Et}^y\right)\right] = 0.$$

4. We recommend examining the narratively largest shocks (as ranked by $S_i\left|\check{u}_{it}^e\right|$), as in Section 2.5. If one has enough power, one can construct a GIV only based on the narratively-checked top events. If one has several GIVs, one can estimate separately the parameters of interest (e.g. ϕ^d) based on each GIV, and see if they are economically different. One may even do a formal over–identification test (Section 2.6).

2.8 Robustness to misspecification

The GIV procedure is robust to some forms of misspecification, and more fragile to others.

We may keep only the shocks to some actors (in a space I_t), i.e. set $z_t = \sum_{i \in I_t} S_i \check{u}_{it}^e$ (with $\check{u}_{it}^e = u_{it} - u_{Et}$), selecting for example the shocks to the top K entities, the shocks for which we

[21] Here η_t^e can combine the estimates where $\eta_t^{x,e}$ and $\eta_t^{PCA,e}$.

have data, or some subset of the entities based on size. Then again, everything goes through.[22] The estimator is still valid, just not the optimal GIV estimator.

Suppose that we misspecify the vector S of sizes, for example by defining $z_t = \sum_i S_i^\circ \check{u}_{it}^e$ using a wrong vector S°. Then, the IV is still valid, but the OLS can be biased. In our basic example of Section 2.2, we still have $\mathbb{E}\left[(p_t - \alpha y_{St}) z_t\right] = 0$, so that the IV procedure (16) still works. Likewise, in the more complex supply and demand case, the IV relations (34) and (36) still hold. But the OLS relations are slightly biased.[23]

If we assume homogeneous coefficients (e.g. on the elasticities of demand or supply), while in fact they are truly heterogeneous, then again (assuming that η_t was well-estimated in the cross-section) the IV estimates are correct, and so are the OLS estimates.[24] One reason for potential instability of GIV estimates with two or more GIVs (as in Section 2.7) may be that indeed the elasticities are different across actors.[25]

If we misspecify the variance of the u_{it} (but keeping them uncorrelated), things are essentially fine: as $u_E = O_p\left(\frac{1}{\sqrt{N}}\right)$, we do not need $\mathbb{E}\left[u_{\Gamma t} u_{Et}\right] = 0$ to hold exactly, as the term $\mathbb{E}\left[u_{\Gamma t} u_{Et}\right]$ will still be small, of order $O\left(\frac{1}{\sqrt{N}}\right)$, and will vanish for large N.

A threat is that we might not control properly for common factors. Indeed, $z_t = u_{\Gamma t} + \lambda_\Gamma \eta_t - \lambda_\Gamma^e \eta_t^e$, so there is a danger that, even after controlling for η_t^e in the regression we will not completely eliminate the $\lambda_\Gamma \eta_t - \lambda_\Gamma^e \eta_t^e$ error.[26] This danger is greater when $|\lambda_\Gamma|$ is greater, i.e. when loadings are correlated with size. This is a small sample problem (with a large enough T, N we measure η_t, λ accurately). A missing factor may also be detected by the tests of stability of estimates across GIVs of Section 2.7, if different actors have different loadings on the missing factors. As is common practice in the weak factors literature, one can verify the stability of the estimates by adding one or two factors beyond what is recommended by the formal tests of Bai and Ng (2002); Onatski (2009). A second approach in this literature is to make stronger parametric assumptions. In our setting it would be natural to make assumptions about the distribution of the shocks and estimate the model via maximum likelihood (see Section 14). A third approach is to filter out "sporadic factors" as in Section 2.5. A fourth approach is to opt for the narrative GIV of Section 2.5. If one checks the top, say, 10-20 events and they pass the narrative check, one can record that we could not reject the hypothesis of a misspecified factor model. In a more purist way, one could even restrict the GIV to those top events.

[22] For instance, we still have $u_{St} = z_t + \varepsilon_t^{uS}$ with $z_t \perp \varepsilon_t^{uS}$.

[23] Calling $\psi = \frac{\mathbb{E}[z_t u_{St}]}{\mathbb{E}[z_t^2]}$ (which is 1 when $S^\circ = S$), then the OLS above gives (in population) $b^{p,e} = b^p \psi$ and $M^e = M\psi$. For some selection procedures (e.g. selecting the shocks to some pre-specified entities as we discussed), we still have that $\psi = 1$, so that OLS is still valid.

[24] Just, the IV estimates yield ϕ^s, ϕ_E^d, and the OLS coefficients are those corresponding to the interpretation that the elasticity of demand is ϕ_E^d rather that ϕ_S^d (see Section 12.9).

[25] Section 10.1 gives a way to estimate different elasticities ϕ_i.

[26] As we do control for η_t^e in the regression, the bias is due to the residual of $\lambda_\Gamma \eta_t - \lambda_\Gamma^e \eta_t^e$ after controlling for η_t^e.

3 General setup and procedure

The previous section introduced the GIV in a simple context, with no loops or a single loop. We now propose a more general setup with potentially several factors, arbitrary loop structure, and rich heterogeneity. This section can be skipped when reading the paper for the first time and the reader can continue with Section 4.

3.1 Framework

Consider the following model of stationary "actions" y_{it} (e.g. employment, investment, TFP shock, return, etc.) by "actor" i (e.g., a firm or industry i in a closed-economy setting, or a country i in an international setting):

$$y_{it} = \sum_f \lambda_{it}^f F_t^f + u_{it} + k_{y_i} + m C_{it}^y, \tag{49}$$

where each F_t^f is a factor, λ_{it}^f are factor loadings, u_{it} is an idiosyncratic shocks, k_{y_i} is a constant, C_{it}^y is a vector of controls that may include lagged demands and other characteristics. Factor f follows:

$$F_t^f = \alpha^f y_{St} + \eta_t^f + k^f + m^f C_t^F. \tag{50}$$

It depends on an exogenous shock η_t^f, and some on the mean action y_{St}, and on a set of controls C_t^F (different from C_{it}^y). Those controls may include, for instance, lagged values. We assume that the "size" weights have been normalized to add to one, $\sum_i S_i = 1$. We partition the factors into "exogenous factors" $\mathcal{F}^{\text{Exo}}$, where we know $\alpha^f = 0$, and "endogenous" factors $\mathcal{F}^{\text{Endo}}$, where α^f may be non-zero.

We model the exposures to factors as either non-parametric (unrestricted λ_i^f) or as parametric:

$$\lambda_{it}^f = \lambda_0^f + \lambda_1^f x_{it}^f, \tag{51}$$

where $x_{it}^f \in \mathbb{R}^\chi$ is an observable (e.g. x_{it} is the de-meaned log size of entity i). For endogenous factors, we treat here the parametric case, and defer the non-parametric case to Section 10.1. We can also treat the semi-parametric case where

$$\lambda_{it}^f = \lambda_0^f + \lambda_1^f x_{it}^f + \zeta_i^f, \tag{52}$$

where ζ_i^f is an extra non-parametric case.

We make the following identifying assumptions, for all f, i, the noise u_{it} are idiosyncratic:

$$u_{it} \perp \left(\eta_t^f, C_t^y, C_t^F, x_{it}^f\right), \tag{53}$$

but the η_t^f may be correlated across f's, and η_t^f may be correlated with the controls, C_t^y and C_t^F.

The u_{it} may have some correlation across i's and can be heteroskedastic, as we discuss later. For expositional simplicity we assume that all dates are i.i.d.

We rewrite model (49) in vector form:

$$y_t = \Lambda_t F_t + u_t + C_t^y m + k_y, \tag{54}$$

with $F_t = \left(F_t^{\text{Endo}\prime}, F_t^{\text{Exo}\prime}\right)'$, $\Lambda_t = \left(\Lambda_t^{\text{Endo}\prime}, \Lambda_t^{\text{Exo}\prime}\right)'$, Λ_t a $N \times r$ matrix, F_t a $r \times 1$ vector, C_t^y an $N \times c$ matrix, m is $c \times 1$, where c is the dimension of the controls, and k_y is a $N \times 1$ vector. The endogenous and exogenous factors, and their loadings, are defined by:[27]

$$F_t^j = \alpha^j y_{St} + \eta_t^j + k_F^j + C_t^F \phi^j,$$

with $j =$ Exo, Endo, with $\alpha^{\text{Exo}} = 0$. We write $\alpha = \left(\alpha^j\right)_{j=\text{Exo, Endo}} = \left(\alpha^f\right)_{f=1\ldots r}$ is an $r \times 1$ vector.

Multipliers Solving for the model gives, $y_{St} = \Lambda_{St} F_t + u_{St} + k_{yS} + C_{St}^y m$, that is,

$$y_{St} = \Lambda_{St} \alpha y_{St} + u_{St} + b_t, \tag{55}$$

where b_t satisfies $b_t \perp u_{St}$.[28] So, we can solve for the aggregate outcome y_{St} as $y_{St} = \frac{u_{St} + b_t}{1 - \Lambda_{St}\alpha}$, that is,

$$y_{St} = M_t \left(u_{St} + b_t\right), \tag{56}$$

where the multiplier $M_t = \frac{dy_{St}}{du_{St}}$ is

$$M_t = \frac{1}{1 - \Lambda_{St}\alpha} = \frac{1}{1 - \sum_{f:\text{Endo}} \Lambda_{St}^f \alpha^f} \tag{57}$$

and measures the total impact of shocks, after going through all feedback loops. Hence, an idiosyncratic shock has an impact that M_t times bigger than its direct effect. Also, the total impact of an idiosyncratic shock on factor f is:

$$F_t^f = M_t \alpha^f u_{St} + b_t^f \tag{58}$$

for another expression $b_t^f \perp u_{St}$. Our regressions will allow to identify M and $M\alpha^f$.

In some cases, we may not observe all endogenous factors, F_t^{Endo}. In this case, we still recover the correct multiplier, M_t, and it should be interpreted as accounting for all feedback loops in the economy, including those operating via the unobservable, endogenous factors. However, we can obviously not estimate α^f for those unobserved factors.

[27] Our initial examples are particular cases of the general procedure, as detailed in Section 12.11.

[28] We have $b_t = k_{yS} + C_{St}^y m + \Lambda_{St}^{\text{Exo}} F_t^{\text{Exo}} + \Lambda_{St}^{\text{Endo}} \left(\eta_t^{\text{Endo}} + k_F^{\text{Endo}} + C_t^F \phi^{\text{Endo}}\right)$, with $k_{yS} := \sum_i S_i k_{y_i}$.

3.2 A step-by-step user's guide

We outline the benchmark procedure, alongside several extensions. We summarize the model as

$$y_t = \Lambda_t F_t + u_t + k_y + C_t^y m,$$
$$F_t^j = \alpha^j y_{St} + \eta_t^j + k_F^j + C_t^F \phi^j,$$

where we focus on the case in which $\Lambda_t^{\text{Endo}} = \lambda_0^{\text{Endo}} \iota$, where λ_0^{Endo} is a scalar.[29] Loadings can be semi-parametric, $\Lambda_t^{\text{Exo}} = \Lambda_0^{\text{Exo}} + \Lambda_1^{\text{Exo}} x_t + \zeta_t$, where $\mathbb{E}\left[\check{\zeta}_t \check{x}_t\right] = 0$, or non-parametric.

We assume that y_t is observed, and we run regressions on the observed factors F_t^f.

1. Define $\check{a}_{it} = a_{it} - a_{Et}$ for a generic variable a_t, and estimate the panel regression

$$\check{y}_t = c + \check{\Lambda}_t^{\text{Exo}} C_t^{F,\text{Exo}} \phi + \check{C}_t^y m + e_t,$$

which removes endogenous factors and estimates the coefficients on the controls. The vector of residuals equals $e_t = \check{\Lambda}_t^{\text{Exo}} \eta_t^{\text{Exo}} + \check{u}_t$.

2. We can estimate the factors, η_t^{Exo}, and denote the estimates by η_t^e, in one of two ways:

   (a) In the case of semi-parametric loadings, $\Lambda_t^{\text{Exo}} = \Lambda_0^{\text{Exo}} + \Lambda_1^{\text{Exo}} x_t + \zeta_t$, we estimate the factors using

$$e_t = \check{x}_t \eta_t^{\text{Exo}} + \epsilon_t,$$

   which is equivalent to a series of cross-sectional regressions of e_t on x_t to estimate η_t^{Exo}. By considering semi-parametric loadings, we do not need to know exactly how the loadings depend on characteristics, x_t, and a noisy signal of exposures suffices to estimate the factors, η_t^{Exo}.[30]

   (b) In the case of non-parametric loadings, we estimate the factors using factor analysis, in case of small N, or principal components analysis (PCA), in case of large N, from e_t. We estimate the number of factors using the methods developed in Bai and Ng (2002).

3. Estimate $\left(M, \alpha^f M\right)$, using OLS, with $Z_t = y_{\Gamma t}$,

$$\begin{aligned} y_{St} &= M Z_t + \gamma_y' \eta_t^e + k_y + C_{St}^y m + e_t^y, \\ F_t^{\text{Endo}} &= \alpha^{\text{Endo}} M Z_t + \gamma_F' \eta_t^e + k_F^{\text{Endo}} + C_t^F \phi^{\text{Endo}} + e_t^{\text{Endo}}. \end{aligned}$$

[29] The procedure in this section extends to the model with multiple endogenous factors, when the factor exposures, λ_i, depend on a vector of characteristics, and to the case where the characteristics may vary over time, x_{it}. Then, the generalized version of $\check{a}_{it} = a_{it} - a_{Et}$ in Step 1 becomes $\check{a}_t = Q^{X,W} a_t$ where $Q^{X,W}$ is defined in (23) (if $X = \iota$, then we recover $\check{a}_{it} = a_{it} - a_{Et}$ for $W = I$, and $\check{a}_{it} = a_{it} - a_{\tilde{E}t}$ for $W = (V^u)^{-1}$.

[30] We refer to Fan et al. (2016) for a related approach in the context of principal components analysis.

4. Estimate $\left(\alpha^{\text{Endo}}\right)^e := \frac{\left(\alpha^{\text{Endo}}M\right)^e}{M^e}$. Similarly, we can recover from $\Lambda_S^{\text{Endo}}\alpha^{\text{Endo}}$ from $M = \frac{1}{1-\Lambda_S^{\text{Endo}}\alpha^{\text{Endo}}}$. In both cases, M^e and $\left(\alpha^{\text{Endo}}M\right)^e$ need to be sufficiently precisely estimated. This is analogous to the weak instruments problem in the context of IV estimators.

When the idiosyncratic shocks are heteroskedastic, we may be able to improve the finite-sample properties. We start from $E_i = \frac{1}{N}$. After Step 2, we use the residuals to obtain an estimate of $\sigma_{u_i}^2$ and update the weights to $\tilde{E}_i = \sigma_{u_i}^{-2} / \sum_{i=1}^N \sigma_{u_i}^{-2}$. In all cases, our estimators are consistent for large N and T. However, to obtain consistency with finite N and large T, we need the precision-weighted $\tilde{E}$.[31] This implies a standard bias-efficiency trade-off if estimating volatilities reduces the efficiency of the estimator. If the idiosyncratic volatilities are related to size, we can parameterize them and estimate

$$\ln \sigma_{u_i}^2 = \sigma_0 + \sigma_1 \ln S_i + \epsilon_i,$$

and use $\sigma_{u_i}^2 = \exp\left(\sigma_0\right) S_i^{\sigma_1}$.

First and second stages Let us see how to estimate α^f in that first and second stage language. The *first stage* is the regression:

$$y_{St} = bz_t + \beta_\eta^{ys}\eta_t^e + \beta_C^{ys}C_{St} + \varepsilon_t^{ys}, \tag{59}$$

where we regress on z_t, using our recovered factors η_t as controls. From the model, we know that the regression coefficient identifies $b = M$, the multiplier.

The *second stage* is the regression:

$$F_t^f = \alpha^f\left(b^e z_t\right) + \beta^{Ff}\eta_t^e + \varepsilon_t^{Ff}, \tag{60}$$

which gives the estimator for α^f. Alternatively, we can regress F_t^f on z_t (with controls) and the coefficient is $\alpha^f M$. When estimating the influence coefficient, we can also view the second stage as estimating

$$y_{Et} = \gamma M^e z_t + \beta_\eta^{ys}\eta_t^e + \beta_C^{ys}C_{St} + \varepsilon_t^{ys},$$

which gives γ.

3.3 A formal identifiability result

We encourage the reader to skip this section at the first reading. We provide here formal conditions for identification, completing the simpler case of Section 2.

[31] Here we are talking about consistency in the estimation of M and $\alpha^f M$. It is achieved even if we do not consistently estimate the underlying factors η_t. This may be surprising, but this is already the case in the simple supply and demand case of Section 2.

We study the "semi-parametric" case. We have some characteristics x_{it} of actors (e.g. countries or and firms): for instance, depending on the application we know that the loading is an affine function of log market capitalization, or the stock market beta of a bank, or OPEC membership. We also have a priori knowledge that $\lambda_{it}^f = \lambda_0^f + \lambda_1^f x_{it}^f$, for some parameters λ_1^f in the parametric case (51), and something similar in the non-parametric case (52). This is consistent with the practice in modern finance in which risk exposures (betas) align with characteristics (see e.g. Fama and French (1993)), so that parametric approaches are preferred, in particular because they are more stable than non-parametric approaches. Section 10.1 develops the full non-parametric version, estimating the factors. We don't have a priori information about the η_t, nor their covariance V^η.

To obtain identification, we shall make the following two assumptions.

Assumption 1 (Condition for identification with GIV) *The vector $V^u S$ is not spanned by the factors loadings λ^f (where V^u is the covariance matrix of u_t).*

Assumption 2 (Restriction on the admissible variance-covariance matrix of residual u_t) *The variance-covariance on u_t is diagonal.*

Assumption 1 is essential and could not be relaxed. It ensures that the GIV is not identically 0. Economically, this assumption seems like a mild restriction. It is generically satisfied.[32]

Assumption 2 could be relaxed in number of ways.[33] Other sufficient condition for identification might be that V^u is $k-$sparse, e.g. has at most k non-zero off-diagonal elements, for some k, e.g. $N - r^2$ (see also Zou et al. (2006)). Another is to allow for some correlation that depends on the distance between entities i and j, perhaps via Gaussian processes (Rasmussen and Williams (2005)). We conjecture that this proposition could be generalized in a number of ways, including in the large T, N domain, using material such as Bai and Ng (2006). Doing this would however take us too far afield.

We assume that all shocks are i.i.d. over time, though this would be easy to relax.

We next state a formal identification result, which is proven in Section 9.

Proposition 4 (Sufficient condition for identification with GIV) *Consider the factor model above, when N is fixed but $T \to \infty$, and makes Assumptions 1 and 2. Then, the parametric (and semi-parametric) procedure of Section 3.2 identifies M, α^f by GIV. Furthermore, the standard errors on M and $\alpha^f M$ returned by OLS in this procedure are valid.*

This completes our "abstract" development of GIV. We now turn to two initial applications.

[32] One case that does prevent this assumption to hold is the case where the variance would be inversely proportional to size: then, GIV would fail, as then $V^u S = a\iota$ for some scalar a. Fortunately, in most contexts, variance may decay a bit with size S_i, but less violently than in $1/S_i$ (see e.g. Lee et al. (1998) and the discussion in Gabaix (2011)).

[33] However, relaxations of Assumption 2 will still need to ensure some restrictions on the space of variance-covariances allowed.

4 An empirical model of sovereign yield spillovers

We study spillovers in sovereign yield markets as a first application of GIVs. We focus on the transmission and amplification of idiosyncratic shocks during the European sovereign debt crisis.

4.1 Data

We use daily data on 10-year zero coupon yields from Bloomberg.[34] The countries and Bloomberg tickers that we use are listed in Table 7 in Appendix 11.4.[35] We use data on general government gross debt for each country from Eurostat.[36] The sample is from July 2009 to May 2018.[37]

4.2 An empirical model of sovereign yield spillovers

Section 11 provides a model of sovereign yield spillovers and we summarize its empirical counterpart here.[38] The main idea of the model is that losses in one country will be partially shared with other countries, implying that shocks to sovereign yields in one country spill over to other countries and vice versa. We index countries by i. We define the yield spread, y_{it}, as the yield in country i relative to Germany's yield.[39] The model implies that relative changes in yield spreads satisfy the following empirical model

$$r_{it} = \gamma r_{St} + \lambda_i' \eta_t + u_{it}, \tag{61}$$

where

$$r_{it} = \frac{\Delta y_{it}}{y_{i,t-1}}. \tag{62}$$

The key message is that the spillover impact is such that $\frac{\Delta y_{it}}{y_{i,t-1}}$, rather than Δy_{it}, depends linearly on γr_{St}. This means that a country with almost no default risk should have almost no sensitivity of its yield Δy_{it} (as there is no risk in the first place), which makes sense. Hence, we think that alternative models are likely to yield a similar functional form.

Empirically, we use $r_{it} = \frac{\Delta y_{it}}{0.01 + y_{i,t-1}}$ to avoid problems when spreads get close to zero, $y_{it} \simeq 0$. In

[34] We use Bloomberg's price variable PX_LAST.

[35] The tickers that we use for different countries are the ones used by European Insurance and Occupational Pensions Authority (EIOPA) to construct the regulatory yield curves of insurance companies and pension funds in the European Union. For the final construction of the curves, EIOPA combines data on zero yields and swap curves, while we only use the zero yields.

[36] https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&plugin=1&language=en&pcode=teina225.

[37] We remove days in which markets are closed, which is when none of the yields change on a given day, and holidays.

[38] A natural extension would to be add the banking sector of each country. In addition, it would be interesting to model the level of yields (i.e., the German yield) as well, which should go down as a result of safety effects.

[39] Spillovers in sovereign bond markets may also operate via intermediaries. For instance, if losses in one country impact the intermediaries' constraints, then this can impact the pricing of bonds in other countries in which the intermediaries are active.

addition, the model implies that the size weights are defined as

$$S_{i,t-1} = \frac{B_{i,t-1} y_{i,t-1}}{\sum_j B_{j,t-1} y_{j,t-1}}, \tag{63}$$

where B_{it} denotes the outstanding government debt of country i.

It is essential to control for common factors η_t. It is well understood, see for instance Forbes and Rigobon (2002), that omitted factors and endogeneity impact measures of spillovers and contagion.[40]

4.3 Estimation procedure

We estimate the model using the standard GIV procedure, accounting for heteroskedasticity.

1. We compute the rolling variance of relative changes in yield spreads using the trading days of the last two months, and lag it by a day, $\mathrm{Var}_{t-1}(r_{it})$. We then define

$$\sigma^2_{i,t-1} = \max\left(\mathrm{Var}_{t-1}(r_{it}), m_{t-1}\right), \tag{64}$$

where $m_{t-1} = \mathit{median}\left(\mathrm{Var}_{t-1}(r_{it})\right)$, that is, the cross-sectional median at time $t-1$. We define the $\tilde{E}-$weights as usual as

$$\tilde{E}_{i,t-1} = \frac{1/\sigma^2_{i,t-1}}{\sum_i 1/\sigma^2_{i,t-1}}.$$

We apply the max-operator in (64) to avoid that the $\tilde{E}-$weights put too much weight on a single country if yield spreads for that country happen to be stable and close to zero. The main objective of adjusting for heteroskedasticity is to put less weight on extremely volatile countries.

2. We compute $\check{r}_{it} = r_{it} - r_{\tilde{E}t}$ and adjust $\check{r}_{it}$ for heteroskedasticity, $n_{it} = \frac{\check{r}_{it}}{\sigma_{i,t-1}}$. We use PCA based on n_{it}, $n_{it} = \check{\lambda}_i' \eta_t + \check{u}_{it}$, to estimate the factors, η_t^e.

3. We estimate the multiplier $M = \frac{1}{1-\gamma}$ via the regression

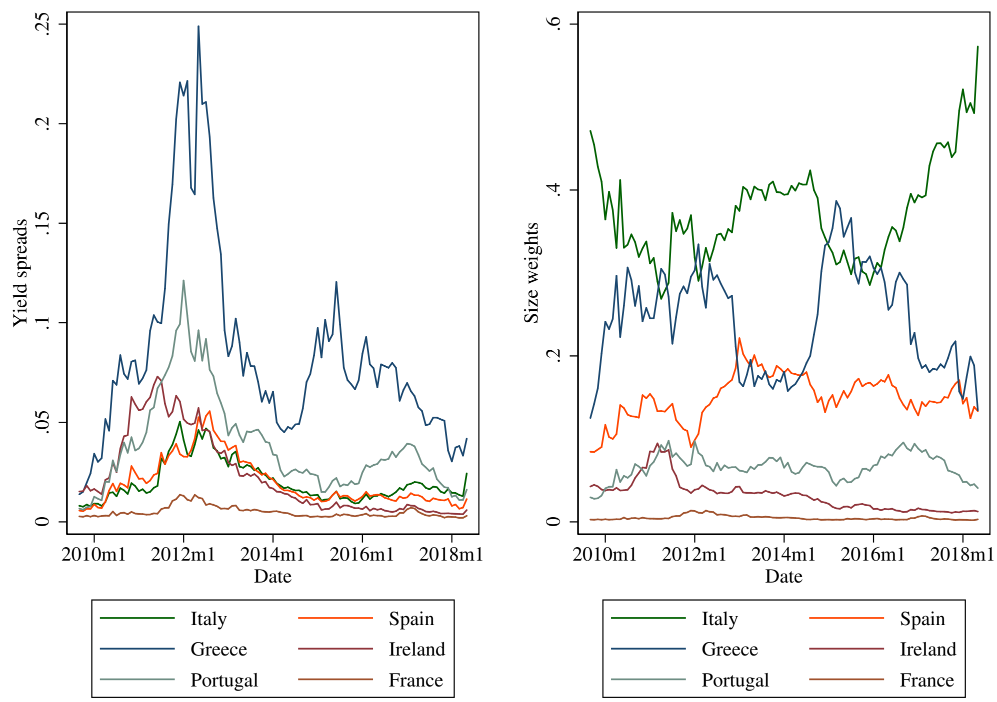
$$r_{St} = k + M r_{\tilde{\Gamma}t} + \lambda_S' \eta_t^e + e_t. \tag{65}$$

To identify the largest shocks and to verify narratively that the shocks are truly idiosyncratic, we run the weighted panel regression

$$r_{it} - r_{\tilde{E}t} = c + \lambda' \eta_t^e + u_{it},$$

[40] Caporin et al. (2018) study spillovers in European sovereign debt markets and show that quantile regressions can be used to test for contagion if contagion is defined as a change in interlinkages. Our definition of contagion (captured by a nonzero γ in 61) is very different from theirs.

Figure 1: The dynamics of sovereign yield spreads and size weights. The figure reports the yield spreads, relative to Germany, for Italy, Spain, Greece, Ireland, Portugal, and France in the left panel from September 2009 to May 2018. The spreads are based on 10-year zero-coupon bonds and are constructed using data from Bloomberg. The right panel displays the size weights based on the definition in (63) for the same countries and the same sample period.



using the size weights. In this panel regression, u^e_{St} is identical to the residual of the regression $r_{\tilde{\Gamma}t} = c + \lambda'_{\tilde{\Gamma}}\eta^e_t + u^e_{St}$. We discuss the largest $|u^e_{St}|$ in detail in Section 4.6.

4.4 Empirical results

We plot the dynamics of spreads, y_{it}, in the left panel, and size weights, using the definition in (63), in the right panel of Figure 1 for France, Greece, Ireland, Italy, Portugal, and Spain. The sample is from September 2009 to May 2018. We distinguish three broad periods. First, from 2010 to 2012, the yield spread dynamics are driven by the European sovereign debt crisis. During 2015, yield spreads in Greece widen once again, but the low-frequency dynamics in other countries are more muted and spreads tighten in most countries. This period is characterized by political turmoil in Greece related in part to negotiations of a bailout deal. During the last months of our sample, there is a jump in Italian yields following political uncertainty about budget plans following the general election. We will revisit these episodes in more detail when analyzing the largest and most influential idiosyncratic shocks in Section 4.6.

Table 1 reports the estimates of the multiplier, M. The first column regresses r_{St} on $Z_t = r_{\tilde{\Gamma}t}$.

Table 1: Multiplier estimates of sovereign yield spillovers. The table reports the estimates of the multiplier in (65). The first to the fourth column include zero to three principal components as controls. In the final column, we re-estimate the model excluding Greece. In this table, $Z_t = r_{\tilde{\Gamma}t}$ and PC_{it} corresponds to the ith principal component that we extract based on $\frac{\tilde{r_{it}}}{\sigma_{i,t-1}}$. The size-weighted average of relative yield spread changes, r_{St}, uses the size weights as defined in (63). The model is estimated using daily data from July 2009 until May 2018.

	r_{St}	r_{St}	r_{St}	r_{St}	r_{St} (excluding Greece)
Z_t	1.632	1.456	1.433	1.488	1.426
	(73.90)	(44.09)	(45.42)	(46.48)	(29.68)
PC_{1t}		0.00215	0.00230	0.00192	0.00215
		(7.13)	(8.00)	(6.67)	(5.47)
PC_{2t}			-0.00332	-0.00330	-0.00160
			(-14.85)	(-14.95)	(-6.35)
PC_{3t}				0.00193	0.00249
				(7.53)	(6.50)
N	2264	2264	2264	2264	2264
R^2	0.707	0.714	0.739	0.745	0.752

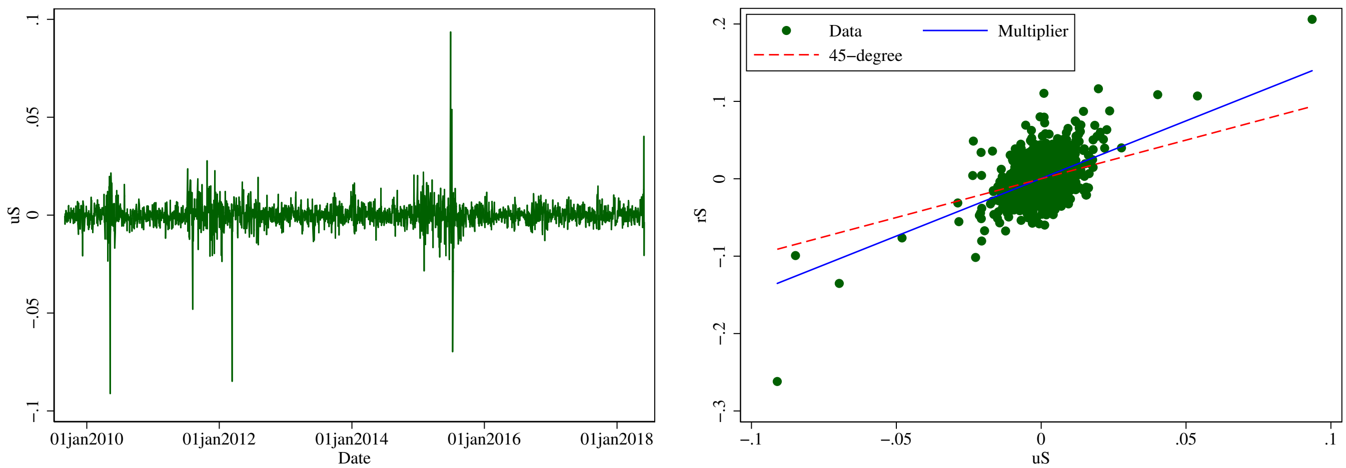
t statistics in parentheses

The second to the fourth column add principal components. The multiplier estimate drops after adding the first principal component from 1.63 to 1.46, but then stabilizes and adding more principal components does not impact the estimate of the multiplier in an economically meaningful way. In the final column, we omit Greece, which plays an important role during this period. However, using the shocks from other countries does not impact the estimates in an economically meaningful way.

The high R-squared in the first column does not estimate the fraction of the variation in aggregate yield spread changes that is due to idiosyncratic shocks, as $r_{\tilde{\Gamma}t}$ is correlated with η_t^e. To estimate the importance of idiosyncratic shocks, we regress r_{St} on u_{St}^e, which provides exactly the same point estimate of the multiplier as in the final column of Table 1. The R-squared of this regression is 24%, implying that a quarter of the variation in aggregate yield spread changes is due to idiosyncratic shocks.

The idiosyncratic shocks to relative changes in yield spreads are fat-tailed, as can be seen from the left panel of Figure 2, which plots the time series of u_{St}^e. The right panel of the same figure plots u_{St}^e (horizontal axis) against r_{St} (vertical axis). If there are no spillovers, $\gamma = 0$, then the multiplier is zero and the points fall along the 45-degree line (the red dashed line). The fact that the estimated slope is steeper, as indicated by the blue solid line, implies that there are significant yield spillovers.

Figure 2: Idiosyncratic shocks over time and aggregate yield changes. The figure shows the time-series dynamics of u^e_{St} in the left panel. We construct u^e_{St} as the residual of a regression of $r_{\tilde{\Gamma}t}$ on η^e_t. The right panel shows a scatter plot of u^e_{St} (horizontal axis) against r_{St} (vertical axis). The size-weighted average of relative yield spread changes, r_{St}, uses the size weights as defined in (63). The series are constructed using daily data from July 2009 until May 2018.



4.5 Interpretation of the coefficients

We find a multiplier $M = \frac{1}{1-\gamma} \simeq 1.5$ and hence a spillover parameter $\gamma \simeq \frac{1}{3}$.[41] To interpret this estimate, suppose that the average yield spread is 1%, and that there is a "primitive" shock to all countries that multiplies their yields by 1.4 (so, $u_{St} = 0.4$). If there were no contagion, the average yield would increase to 1.4%. But as there is contagion, the average yield will increase to 1.6% (as in $0.6 = M \times 0.4$). At the same time, a country with 0 yield spread still still keep a 0 yield spread (as it is and was riskless), while for a country with an initial a yield of 5%, its own yield will increase to 8%.

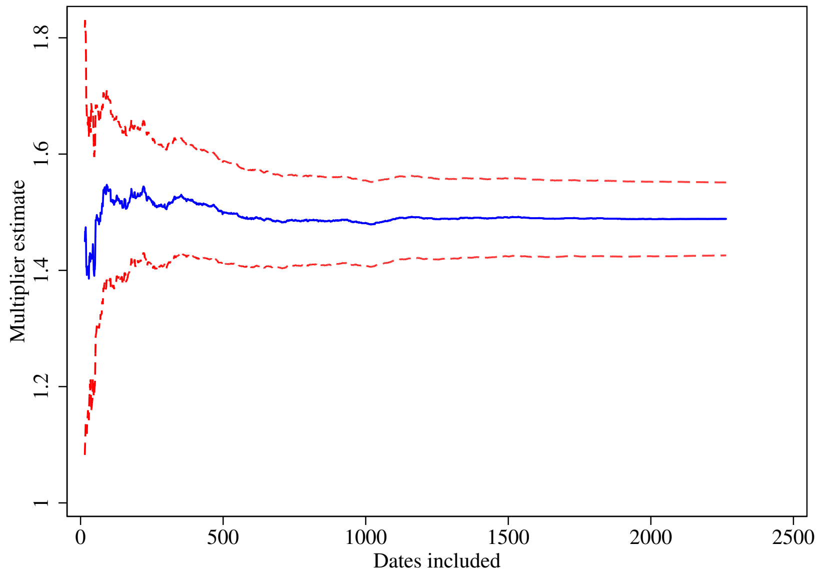
To get some more intuition for the spillover, consider that Italy, near the peak of the crisis, has a relative size of 0.4. Suppose an idiosyncratic shock to Italy makes the Italian yield double ($u_i = 1$ for $i =$ Italy), i.e. the Italian yield spread goes from 2% to 4%. That makes the other yields go up by a relative value $\gamma M \times S_i \times u_i = 0.5 \times 0.4 \times 1 = 0.2$, so that the average yield increases from 1% to 1.20%. In other terms, as the Italian yield spread goes up by 200bp, the other countries' yield spread goes up by 20bp, implying a "pass-through" of 0.1.

4.6 Narrative GIVs

To further inspect the variation that the GIVs are exploiting to estimate the multiplier, we narratively check the largest shocks in Table 2. In particular, we order the dates based on the size of $|u^e_{St}|$. To illustrate the relevance of the largest shocks, we re-estimate the model that includes three

[41] Here we use, omitting aggregate shocks, $r_{it} = \gamma M u_{St} + u_{it}$, with $r_{it} = \frac{\Delta y_{it}}{y_{i,t-1}}$, and $r_{St} = M u_{St}$.

Figure 3: Multiplier estimates using an expanding window. The figure reports the estimates of the multiplier in (65) using three principal components as controls. We estimate the model using an expanding sample where the data are ordered by $|u^e_{St}|$, that is, the magnitude of the idiosyncratic shocks. The number of dates included is depicted on the horizontal axis, starting with 15 observations. The solid blue line corresponds to the point estimate and the dashed red lines to the 95%-confidence interval. The model is estimated using daily data from July 2009 until May 2018.



principal components, that is, the last column of Table 1, using only the days with the largest k shocks. In Figure 3, we show the multiplier estimate, alongside the 95%-confidence interval, where we indicate the number of dates included on the horizontal axis (starting at 15 observations). The estimate is stable for different samples but obviously standard errors tighten as the sample expands.

Panel A in Table 2 reports the yield changes on the 10 days with the largest realization of $|u^e_{St}|$. In Panel B, we scale the yield changes by $0.01 + y_{i,t-1}$. In Panel C, we provide the narratives. If we inspect some of the largest shocks in Table 2, then is is clear that most of them are truly idiosyncratic shocks. Examples include the decision by Greece to close all banks or the outcome of the referendum. There are two shocks, however, on May 10, 2010 and August 8, 2011 that involve actions by the ECB and hence are more likely aggregate shocks as opposed to idiosyncratic shocks. Removing these dates does not impact our estimates, but illustrates the empirical relevance of sporadic factors during times of crisis (see Section 2.8). Most of the shocks are associated with Greece, although the final date corresponds to Italy.[42]

[42] Also following the end of our sample, many of the idiosyncratic shocks in recent months happened in Italy.

Table 2: Summary of the largest idiosyncratic shocks and narratives. The table reports the properties of yield spread changes, Δy_{it}, on the 10 days with the largest realization of $|u^e_{St}|$ in Panel A. In Panel B, we report the corresponding relative yield changes, $r_{it} = \frac{\Delta y_{it}}{0.01+y_{i,t-1}}$. In Panel C, we provide the narratives associated with these events. The series are constructed using daily data from July 2009 until May 2018.

Panel A: Unscaled idiosyncratic shocks

Date	Austria	Belgium	Finland	France	Greece	Ireland	Italy	Netherlands	Portugal	Slovenia	Spain
10-May-10	-0.08%	-0.24%	-0.07%	-0.08%	-3.87%	-1.18%	-0.43%	-0.06%	-1.62%	-0.30%	-0.62%
8-Aug-11	-0.03%	-0.09%	0.00%	0.08%	0.01%	0.05%	-0.66%	0.00%	-0.10%	-0.16%	-0.73%
26-Oct-11	-0.05%	-0.03%	-0.03%	-0.09%	3.11%	0.01%	-0.02%	-0.03%	-0.15%	0.01%	-0.04%
12-Mar-12	0.04%	0.04%	0.01%	0.05%	-7.32%	0.00%	0.11%	0.01%	0.06%	-0.04%	0.09%
3-Feb-15	-0.01%	-0.01%	-0.01%	0.00%	-1.21%	-0.02%	-0.07%	-0.01%	-0.12%	-0.06%	-0.05%
29-Jun-15	0.05%	0.07%	0.03%	0.07%	3.21%	0.13%	0.36%	0.04%	0.48%	0.11%	0.36%
6-Jul-15	0.02%	0.04%	0.01%	0.03%	2.35%	0.06%	0.16%	0.01%	0.27%	0.03%	0.18%
10-Jul-15	-0.04%	-0.05%	-0.03%	-0.05%	-3.28%	-0.10%	-0.21%	-0.04%	-0.24%	-0.23%	-0.22%
13-Jul-15	-0.01%	0.01%	0.00%	0.01%	-1.11%	0.00%	0.02%	0.00%	-0.03%	-0.01%	0.02%
29-May-18	0.07%	0.04%	0.03%	0.04%	0.36%	0.08%	0.49%	0.03%	0.20%	0.08%	0.17%

Panel B: Scaled idiosyncratic shocks

Date	Austria	Belgium	Finland	France	Greece	Ireland	Italy	Netherlands	Portugal	Slovenia	Spain
10-May-10	-5.2%	-14.2%	-5.2%	-5.9%	-41.6%	-29.5%	-17.5%	-4.8%	-38.0%	-12.8%	-24.1%
8-Aug-11	-1.9%	-3.1%	0.2%	4.4%	0.1%	0.6%	-14.4%	-0.3%	-1.2%	-4.0%	-15.9%
26-Oct-11	-2.6%	-1.0%	-2.1%	-4.3%	19.4%	0.2%	-0.5%	-2.3%	-1.5%	0.3%	-0.8%
12-Mar-12	1.9%	1.6%	0.8%	2.5%	-31.7%	0.0%	2.8%	0.9%	0.5%	-0.9%	2.2%
3-Feb-15	-1.2%	-0.7%	-0.7%	-0.1%	-11.5%	-1.2%	-3.0%	-0.7%	-3.7%	-2.9%	-2.0%
29-Jun-15	3.8%	5.0%	2.8%	5.1%	33.4%	6.9%	16.6%	3.1%	17.4%	4.7%	16.6%
6-Jul-15	1.8%	2.5%	0.7%	2.0%	18.9%	3.3%	6.8%	1.2%	8.6%	1.1%	7.7%
10-Jul-15	-3.0%	-3.5%	-2.2%	-3.7%	-22.1%	-5.0%	-8.8%	-3.0%	-7.6%	-9.1%	-9.0%
13-Jul-15	-0.5%	1.0%	-0.2%	1.1%	-9.6%	0.1%	1.0%	0.3%	-0.9%	-0.5%	1.1%
29-May-18	5.2%	2.6%	2.3%	2.9%	7.2%	5.0%	14.6%	2.3%	7.3%	4.7%	7.7%

Panel C: Narrative analysis

Date	Event
10-May-10	Stock markets leap across Europe as EUR750bn eurozone rescue package is agreed
8-Aug-11	ECB decides to start buying Italian and Spanish bonds as part of the Securities Markets Program
26-Oct-11	EU leaders announced an agreement, including deal with private sector investors to take a 50% loss on Greek bonds
12-Mar-12	Greece Bailout Package Signed Off by EU Leaders
3-Feb-15	Greek government said to retreat from a demand for a debt writedown.
29-Jun-15	Greece closes banks
6-Jul-15	Greece bailout referendum on July 5th where voters reject austerity package
10-Jul-15	The Greek government submitted its highly anticipated plan for the country's economic overhaul to bailout authorities
13-Jul-15	Greek PM Alexis Tsipras conceded to a further swathe of austerity measures and economic reforms
29-May-18	Italian political turmoil (snap election plus new budget plan) cause largest 1-day decline in Italian bonds in 25 years

5 Demand and supply elasticities in crude oil markets

5.1 Data

Our data construction follows the existing literature (Kilian (2009), Caldara et al. (2018), Baumeister and Hamilton (2019), henceforth BH). The data on oil supply and prices are from the U.S. Energy Information Administration (EIA). We observe the monthly oil supply for 20 countries (both OPEC and non-OPEC) from January 1985 until December 2015.[43] As we also observe the total non-OPEC production, we also construct a fictitious country which produces the residual non-OPEC supply. The real oil price series is obtained based on the refiner acquisition cost of imported crude oil and deflated using the US CPI to obtain the real price of oil as in Kilian (2009).

We focus on estimating short-run (monthly) demand and supply elasticities, consistent with the literature. To construct innovations, we use a state vector X_t that includes lagged (i) monthly price changes, (ii) world supply growth, (iii) changes in inventories, and (iv) growth in industrial production.[44] We use the data of BH for the latter two series.

5.2 Model

We model the supply growth of country i in period t as

$$\Delta y_{it} = \phi^s \Delta p_t + \lambda_i \eta_t + u_{it} + \gamma_y' X_{t-1},$$

and model changes in aggregate oil demand (both in use and inventories) as

$$\Delta d_t = \phi^d \Delta p_t + \gamma_d' X_{t-1} + \epsilon_t.$$

Market clearing, $\Delta y_{St} = \Delta d_t$, implies

$$\begin{aligned} \Delta p_t &= \frac{M}{\phi^d} u_{St} + \gamma^{p\prime} \mathcal{C}_t, \\ \Delta y_{St} &= M u_{St} + \gamma^{y\prime} \mathcal{C}_t \end{aligned}$$

where

$$M = -\frac{\phi^d}{\phi^s - \phi^d} \in [0, 1]$$

is the multiplier, and γ^p, γ^y are loadings on $\mathcal{C}_t = (\eta_t, \varepsilon_t, X_{t-1})$, and whose precise value does not matter here.

[43] We follow Caldara et al. (2018) and remove Gabon from the sample due to concerns about data quality. In addition, we scale the supply of the USSR using the ratio of supply of the USSR to the supply of Russia to obtain a continuous series and to avoid a sudden jump in the non-OPEC supply.

[44] The results do not change significantly if we use 12 lags instead of one lag.

Our goal is to estimate the short-run supply and demand elasticities, ϕ^s and ϕ^d (with presumably $\phi^d < 0 < \phi^s$). The equations for aggregate supply and price changes are part of the VAR models that are commonly used in the recent literature on oil prices and their impact on economic growth.

5.3 GIV estimation

The supply changes in some periods are extreme for some countries during supply disruptions, and we therefore winsorize the growth rates at 2.5% and 97.5% across all countries and periods to estimate Δy_{Et}.[45] We use Δy_{it}^W to denote the winsorized supply growth. We then estimate the model using the following steps:

1. Run a panel regression with country and time fixed effects,[46]

$$\Delta y_{it} = k_i + a_t + \check{e}_{it}.$$

2. Use $\check{e}_{it}$ to estimate η_t^x and η_t^{PCA} (as in Section 3.2, Step 2), using which we define a new vector of controls $C_t = \left(\eta_t^x, \eta_t^{PCA}, X_{t-1}\right)'$.

3. Estimate $\frac{M}{\phi^d}$ using (with $Z_t \coloneqq \Delta y_{\Gamma t}$ which is our z_t plus some linear function of C_t, which is anyway controlled for in the regression):

$$\Delta p_t = \frac{M}{\phi^d} Z_t + \beta^{p\prime} C_t + e_t^p, \tag{66}$$

   and $M = -\frac{\phi^d}{\phi^s - \phi^d}$ using

$$\Delta y_{St} = M Z_t + \beta^{y\prime} C_t + e_t^y. \tag{67}$$

4. We can recover the supply and demand elasticities using the estimates of $\frac{M}{\phi^d}$ and M, where $\phi^s = \frac{\phi^d}{M}(M-1)$. However, to get the standard errors on the elasticities as well, we use the 2SLS estimator based on the first stage, which corresponds to (66), and denote the fitted value by $\Delta \hat{p}_t \coloneqq \left(\frac{M}{\phi^d}\right)^e Z_t$. The second stage estimator for the demand elasticity corresponds to

$$\Delta y_{St} = \phi^d \Delta \hat{p}_t + \beta_d' C_t + e_t^d, \tag{68}$$

   and for the supply elasticity to

$$\Delta y_{Et} = \phi^s \Delta \hat{p}_t + \beta_s' C_t + e_t^s. \tag{69}$$

[45] To ensure growth rates are always defined, we set supply to one in case it drops to zero, which happens in seven country-months.

[46] Note that the time fixed effects absorb the controls, X_{t-1}, in this case.

Table 3: Multiplier estimates in the oil market. The first column reports the estimate of M, see (66), and the second column of $\frac{M}{\phi^d}$, see (67). The third column reports the Two Stage Least Square (2SLS) estimate of the demand elasticity ϕ^d, see (68), and the fourth column the 2SLS estimate of the supply elasticity ϕ^s, see (69). We suppress the coefficients on the controls, X_{t-1}, that include lagged (i) monthly price changes, (ii) world supply growth, (iii) changes in inventories, and (iv) growth in industrial production. The t-statistics, which are reported in parentheses, are based on OLS and 2SLS standard errors. The sample is from January 1985 to December 2015.

	y_{St}	Δp_t	y_{St}	y_{Et}
Z_t	0.878	-2.328		
	(15.12)	(-4.42)		
$\eta_{PCA,t}$	-0.117	0.176	-0.0508	-0.126
	(-13.84)	(2.29)	(-1.58)	(-12.54)
$\eta_{OPEC,t}$	0.391	-0.188	0.320	0.401
	(12.91)	(-0.69)	(2.99)	(11.97)
Δp_t			-0.377	0.0524
			(-4.30)	(1.91)
N	370	370	370	370
R^2	0.542	0.263		

5.4 Empirical results

We report the estimation results of the multipliers $M = 0.88$ and $\frac{M}{\phi^d} = -2.3$ in Table 3 alongside both elasticities. We estimate a demand elasticity of $\phi^d = -38\%$ (with a standard error of 9%) and a supply elasticity of $\phi^s = 5\%$ (with a standard error of 3%). Changes in demand also include changes in inventories, which respond more elastically to changes in prices (Kilian and Murphy (2014)).

To put these estimates in perspective, we compare them to recent estimates in the literature. Baumeister and Hamilton (2019) use sign restrictions in combination with a Bayesian estimator to find supply and demand elasticities of 15% and -35%, respectively, with 68% credibility intervals of (9%, 22%) for the supply elasticity and (-51%, -24%) for the demand elasticity. Caldara et al. (2018) use a narrative approach and estimate a supply elasticity of 8% (with a standard error of 3.7%) and a demand elasticity of -8% (with a standard error of 8%). Kilian and Murphy (2014) also combine sign restrictions and a Bayesian estimator with short-run supply elasticities bounded at 2.5%, 5%, and 10%, and corresponding demand elasticities range from -44% to -47%.

We construct our instrument as the residual from a regression of $y_{\Gamma t}$ on X_{t-1} and the two factors and refer to it as $u_{\Gamma t}$. If we regress it on the instrument of Caldara et al. (2018), which is non-zero only during 14 months in this sample, we get a slope coefficient of 0.93, with a t-statistic of 12.7 and an R-squared of 93%. Moreover, if we restrict ourselves to more extreme episodes by only using data when $u_{\Gamma t}$, in absolute value, exceeds a threshold of 0.5% (370 observations), 0.75%, ..., 1.25% (19 observations), then the first-stage estimate declines monotonically from -2.3, with a t-statistic of -4.4, to -5.2, with a t-statistic of 3.5. This highlights that by focusing on the more extreme events, $\frac{M}{\phi^d}$ becomes more negative. Intuitively, in case of more extreme shocks, the role of inventories diminishes and the demand curve becomes more inelastic. This reconciles our estimates with those of Caldara et al. (2018).

If we inspect the largest shocks in terms of contribution to the instrument, $S_{i,t-1}u_{it}$, then many of the extreme shocks are as described in Caldara et al. (2018). However, in some cases, the GIV identifies shocks that are not included in the narrative approach. An example includes a reduction in supply by Saudi Arabia in January 1989. Per the description of Caldara et al. (2018), OPEC agreed upon a reduction in supply in November of 1988 but reports in subsequent months were interpreted as "indicating that the OPEC member country was seriously attempting to cut back production based on the new agreement." One possible interpretation of this shock is that markets learn about the exposure to the common OPEC shock, $\eta_{OPEC,t}$. In January 1989, the real price of oil jumped up by 13.2%. In the context of GIV, those are valid idiosyncratic shocks that can be used as instruments.

In summary, the GIV estimator results in estimates that are in the range of estimates documented in the recent literature, thereby providing some external validation of GIVs as an approach to estimating demand and supply elasticities. At the same time, the GIV procedure arguably requires less domain-specific ingenuity than the previous studies we mentioned.

In future work, granular country-level data on (net) imports and oil consumption can be used to construct a second instrument that can be used to both sharpen the estimates and to test for overidentifying restrictions. This instrument may be particularly powerful given the apparent importance of demand shocks during the last 15 years.

6 Discussion and extensions of the framework

6.1 Extensions

There are many ways to increase the number of setups in which the GIV idea can be applied.

Multidimensional GIV One can handle multidimensional "actions": for instance, a firm could have a shock that affects both productivity and labor demand. A country could have a shock that affects both productivity and oil demand. Formally, the actions y_{it} and shocks u_{it} are now multidimensional. The GIV idea goes through, and this is developed in Section 12.1. We have seen that with one GIV, we can estimate $1 + d_F$ parameters (M, $M\alpha^f$), where d_F is the number of endogenous, observed factors. With q-dimensional actions, we have q GIVs, and we can estimate

$q^2 + qd_F$ parameters, which correspond to M and α^f.[47] So, potentially many parameters can be recovered with multidimensional "actions" by firms or countries.

GIV with different size weights This framework can be extended with size weights that vary across factors, $F_t^f = \eta_t^f + \alpha^f y_{S^f,t} + k^f + \phi^f C_t^F$. Then, we can identify more parameters, as each $z_{S^f t} = y_{S^f t} - y_{Et}$ is an instrument (see Section 12.2), for each distinct and useful weight S^f. Indeed, then we can not only identify M and $M\alpha^f$ as in the regular GIV, but also all the λ^f.[48]

GIV with a more complex matrix of influences The GIV can also be extended to non-homogeneous influences in the context of loops. Suppose a model:

$$y_{it} = \gamma \sum_j G_{ij} y_{jt} + \lambda_i \eta_t + u_{it}, \tag{70}$$

i.e., in vector form

$$y_t = \gamma G y_t + \Lambda \eta_t + u_t, \tag{71}$$

with a given "influence" matrix G (in our baseline model, $G = \iota S'$). We'd like to identify γ, the strength of linkages.

A simple generalization of our GIV is to define a "size" vector $S := G'E$. Then, left-multiplying (71) by E', we get

$$y_{Et} = \gamma y_{St} + \Lambda_S \eta_t + u_{St}.$$

The key moment is still $\mathbb{E}\left[\left(y_{Et} - \gamma y_{St}\right) z_t\right] = 0$, where the GIV is again $z_t = y_{St} - y_{Et}$ in the simple case where $\Lambda = \iota$ and $G\iota = \iota$; see Section 12.6 for the general case. Hence, GIV generalizes to "spatial" models with common shocks (most spatial models do not have latent common shocks).

Bayesian GIV Another extension is a Bayesian interpretation of the GIVs. This way, we can interpret GIVs in a Bayesian framework – see Section 14. In particular, this opens the possibility of marrying GIV estimation with priors on other parameters. In the simplest cases with Gaussian shocks, the maximum likelihood estimate is our GIV – confirming its optimality properties. At the same time, the basic GIV doesn't actually use normality assumptions.

Furthermore, many econometric extensions might be useful, e.g. with stochastic volatility, and various dimensions of autocorrelations. We leave those extensions to future research.

[47] As u_{St} and y_{St} are q-dimensional, $M = \frac{dy_{st}}{du_{St}}$ $q \times q$ dimensional, and each of the $\frac{dF_t^f}{du_S} = M\alpha^f$ is also q-dimensional.

[48] On the other hand, the difference between different size weights may be small to the estimation will be more fragile.

6.2 Comparison with Bartik instruments and other procedures

Comparison with Bartik instruments The GIV estimator shares some similarities with Bartik instruments, also known as "shift-share estimators," that were first introduced in Bartik (1991). To put it simply, Bartik instruments allow to estimate the cross-sectional (or micro) sensitivities to shocks, but not aggregate sensitivities; whereas GIVs are mostly designed to estimate aggregate (or macro) elasticities. Hence, they are complementary.

To see this, let us use the notation established earlier. Shift-share estimators aim to estimate the coefficients λ_1^f in the structural equation $y_{it} = \sum_f (\lambda_0^f + \lambda_1^f x_{it}) F_t^f + \eta_t^y + u_{it}$ using $x_{it} g_t^f$ as an instrument for $x_{it} F_t^f$. In this notation, the "shares" are x_{it} and the "shifters" are g_t^f (for instance g_t^f could be the China shock, and be correlated with η_t^y). Shift-share estimators have been the study of much recent econometric work including Goldsmith-Pinkham et al. (2018); Adao et al. (2018b); Borusyak et al. (2018). Borusyak et al. (2018) lay out sufficient identifying conditions for the shift-share estimator to estimate the structural parameter of interest λ_1^f and show that the key orthogonality condition is that the shifters g_t^f are orthogonal to the share-weighted structural disturbances. That is, the shifters are as-good-as-randomly assigned. Goldsmith-Pinkham et al. (2018) provide alternative identifying conditions for shift-shares but these are less relevant for the GIV estimator.

Returning to the GIV estimator and the notation we established earlier, recall that the shares S_i are either held fixed throughout the analysis or set in the previous period, e.g. $S_i = S_{i,t-1}$ provided that the previous period shares are orthogonal to u_{it}. Therefore, a critical orthogonality condition for the GIV estimator is that the idiosyncratic errors u_{it} are orthogonal to the disturbances in the structural equation of interest. In this sense, the orthogonality condition for the GIV estimator is similar to the condition provided in Borusyak et al. (2018), where we now think of the idiosyncratic errors u_{it} as the shifters. The GIV estimator then provides a very general strategy for constructing valid instruments based upon the underlying granular economic structure and as shown earlier, these granular instruments are optimal instruments. However, this does not fully capture the contribution of the GIV estimator. Shift-share estimators are unable to estimate the mean effect λ_0^f. Moreover, as we also show earlier, the GIV approach identifies multiple parameters in a system of simultaneous equations $(M_t, M_t \alpha^f)$ and therefore it additionally enables the researcher to identify multipliers. This is generally not true in shift-share settings, which typically consider single-equation systems.

Procedures containing elements of GIVs A few papers have explored the idea of using idiosyncratic shocks as instruments to estimate spillover effects, such as Leary and Roberts (2014b) in the context of firms' capital structure choice and Amiti et al. (2019) in the context of firms' price setting decisions. The structure of the estimating equations in these papers is similar to the model that we consider here:[49]

$$y_t = \lambda y_{wt} + mC_t + u_t,$$

where $y_{wt} = w' y_t$ can be equally-weighted (Leary and Roberts (2014b)) or size-weighted (Amiti et al. (2019)), depending on the weights w. Both papers use industry and/or year fixed effects, which can be viewed as a choice of controls or exogenous factors, η_t, to which all firms in a given industry have the same exposure.

There are two main differences compared to GIV. First, both papers use idiosyncratic shocks to another variable than y_t, say g_t, to construct an instrument for y_{wt}. Leary and Roberts (2014b) use idiosyncratic stock returns and Amiti et al. (2019) use shocks to competitors' marginal cost, exchange rates, or export prices. We, instead, propose to use idiosyncratic shocks to y_t rather than another instrument (this way requiring fewer times series). Second, and related, we control for heterogeneous exposures to common factors to extract the idiosyncratic shocks, which is important in asymptotic theory and in practice in realistic samples (see Section 7).

A third difference is specific to Leary and Roberts (2014b). GIVs crucially depend on the difference between size- and equal-weighted averages of variables. If the estimating equation depends on equal-weighted averages, GIV cannot be applied. In most models, however, not all competitors receive equal weight and larger firms, or perhaps firms that are closer in product space, receive a larger weight.

Lastly, the use of model-based idiosyncratic shocks has some similarities with Amiti and Weinstein (2018), who extract bank supply shocks from Japanese data using a panel of fixed effects, and then estimate the sensitivity of aggregate investment to these shocks. However, unlike our model, Amiti and Weinstein (2018) assume a uniform sensitivity to the aggregate shocks ($\lambda_i \eta_t$ with $\lambda_i = 1$ for all i), and do not allow for feedback loops: shocks to banks affect aggregate investment, but aggregate investment does not circle back around to affect individual bank behavior (so, they assume $\alpha^f = 0$ in our notations). This is the key source of endogeneity in many of the models we consider, and by tackling it we are able to estimate a richer set of parameters.

Other methods to estimate aggregate elasticities Rigobon (2003) introduces another method that can be used to estimate spillover effects and aggregate multipliers using time-variation in second moments. If shocks are heteroskedastic and the structural parameters are stable across regimes, then the different volatility regimes add additional equations to the system so that the structural parameters can be identified. GIV does not require heteroskedasticity, but can accommodate it, and is therefore complementary to identification methods that rely on heteroskedasticity.

[49] Amiti et al. (2019) study the price setting decision of firms. In their model, the pricing equation features two endogenous variables, namely the same firm's marginal cost and the size-weighted average of competitors' prices. We focus on the spillover effects of competitors' prices in our discussion in this section.

Influence and the "reflection problem" We finish with another example, known as "contagion" or the "reflection problem" (Manski (1993); Kline and Tamer (Forthcoming)). Suppose that actions follow:

$$y_{it} = \gamma y_{St} + \eta_t + u_{it}, \tag{72}$$

where η_t is uncorrelated with the u_{it}. This equation means that y_{it} is influenced by the aggregate action of other agents (γy_{St}), and in addition by the usual aggregate shocks η_t, and idiosyncratic shocks u_{it} (which we assume to be uncorrelated to η_t). The "influence" or "contagion" parameter γ is of high interest.

The GIV approach works as follows. Taking the size-weighted average of (72), we have $y_{St} = \gamma y_{St} + \eta_t + u_{St}$, so that

$$y_{St} = M\left(\eta_t + u_{St}\right), \qquad M = \frac{1}{1-\gamma}. \tag{73}$$

We form $z_t := y_{\Gamma t}$, which by (72) will give $y_{\Gamma t} = u_{\Gamma t}$. Hence, if we estimate M^e by OLS:

$$y_{St} = M^e y_{\Gamma t} + \varepsilon_t^y,$$

then we have a consistent estimator of the multiplier M, and therefore of γ.[50] We have a simple GIV approach to the "reflection problem". To the best of our knowledge, this approach is new. Indeed, it may seem to contradict earlier impossibility results. Section 12.7 solves the apparent contradiction. The short summary is that Manski (1993) and Bramoullé et al. (2009) do not consider anything like a GIV, as they immediately reason on averages based on observables, eschewing any exploration of the noise.[51] In contrast, GIVs are all about exploring some structure in the noise – the idiosyncratic shocks of large entities.[52]

In a tangentially related recent paper, Sarto (2018) uses factor analysis to extract values of η^f (much as we do when we "recover" a factor η^f). Take the basic example in our paper. Then, Sarto does not identify α: even if η (the aggregate shock to demand) were perfectly identified, that would not allow to estimate p. In the supply and demand example, Sarto would identify the demand elasticity ϕ^d, but not the supply elasticity ϕ^s.

Spatial econometrics. In some applications of GIVs we have considered separately, growth in a region affects that of the other regions. So there is a similarity between our setup and that of spatial econometrics (e.g. Kelejian and Prucha (1999); Blasques et al. (2016); Shi and Lee (2017); Kuersteiner and Prucha (2018)). However, the estimators are quite different. The reason is that spatial econometrics studies the "local" influence (e.g. of neighboring cities on a city), while GIVs

[50] And we will have $\varepsilon_t^y = \eta_t + u_{Et}$.

[51] Somewhat related, Graham (2008) explores the identification of peer effects using conditional variance restrictions on the outcomes by exploiting differences in the sizes of the peer group. Intuitively, smaller peer group sizes lead to a larger contribution of each individual peer on the peer component.

[52] Economically, the idiosyncratic shocks to "big influencers" (e.g. large firms, or perhaps famous people in the networks) affect the aggregate, hence they allow to estimate the social or economic multiplier. This is why they can be handled with GIVs.

study the global influence. Hence, the sources of variation, identifiability conditions and methods are quite different. Certainly, the spatial literature has not identified, as we do, the GIVs as a simple way to estimate elasticities in contexts such as supply and demand problems, and models with feedback loops from banks to sovereign yields (and vice versa). Still, some of the sophisticated techniques of the spatial literature might be used one day to enrich a GIV-type analysis.

6.3 GIV for differentiated product demand systems

We develop the basic ideas for the logit demand model and extend these ideas to the random-coefficients logit model as in Berry et al. (1995a) in the next subsection.

6.3.1 Logit demand

The utility that household h derives from product i, for $i = 0, ..., N$, is given by[53]

$$\begin{aligned} U_{hit} &= \delta_{it} + e_{hit}, \\ \delta_{it} &= -\gamma p_{it} + \beta' x_{it} + \alpha_i + \xi_{it}, \end{aligned}$$

where e_{hit} follows a Type-1 extreme-value distribution, p_{it} denotes the log price, x_{it} observable characteristics, and $\mathbb{E}[\xi_{it}] = 0$. We refer to $i = 0$ as the outside option and normalize $\delta_{0t} = 0$. This model implies that the probability that a given household selects product i is

$$\mathbb{P}(V_{hit} > V_{hjt}, \forall j \neq i) = \frac{\exp(\delta_{it})}{1 + \sum_{j=1}^{N} \exp(\delta_{jt})},$$

which in this case also equals the market share, s_{it}. Firms set prices to maximize profits and we assume that each product is produced by a single firm, which solves

$$\max_{P_{it}} Q_{it} (P_{it} - C_{it}),$$

where C_{it} equals marginal cost and $Q_{it} = s_{it} Q_t$ with Q_t the total size of the market. The firm optimally sets the price to

$$P_{it} = \left(1 - \frac{1}{\epsilon_{it}}\right)^{-1} C_{it},$$

where $\epsilon_{it} = -\frac{\partial \ln s_{it}}{\partial \ln p_{it}}$, that is, the negative of the price elasticity of demand. The goal is to estimate $\theta = (\beta, \gamma)$.

[53] We use log price, p_{it}, instead of the price, P_{it}, in the formulation of δ_{it} to simplify some of the expressions, but the basic logic extends to the case where δ_{it} depends on P_{it}.

It is convenient to rewrite the model as

$$\log\left(\frac{s_{it}}{s_{0t}}\right) = \alpha_i - \gamma p_{it} + \beta' x_{it} + \xi_{it}.$$

To identify β, it is commonly assumed that $\mathbb{E}\left[x_{it}\xi_{it}\right] = 0$ and we maintain this assumption. However, as prices respond to demand shocks, ξ_{it}, we cannot assume $\mathbb{E}\left[p_{it}\xi_{it}\right] = 0$. There are three common approaches to create instrumental variables in the demand estimation literature. First, variables that capture variation in marginal cost, C_{it}, that is unrelated to demand shocks. Second, Berry et al. (1995a) suggest to use the average of characteristics of other firms

$$z_{it}^{BLP} = \frac{1}{N-1}\sum_{j,j\neq i} x_{jt},$$

which results in valid instruments under some assumptions (see Nevo (2000) and the references therein).[54] Third, one can use panel data for the same firm that operates in different locations. Under the assumption that demand shocks are uncorrelated across locations, prices in other locations of the same firm will be valid instruments. The intuition is that prices across locations share the same marginal cost but the demand shocks are, by assumption, uncorrelated, see Nevo (2001).

GIV provides an alternative by exploiting exogenous variation in markups due to idiosyncratic demand shocks to large firms. We assume that demand shocks follow a factor model,

$$\xi_{it} = \log\left(\frac{s_{it}}{s_{0t}}\right) - \alpha_i + \gamma p_{it} - \beta' x_{it} = \eta_t + u_{it},$$

which can be extended to allow for heterogeneous exposures, $\lambda_n \eta_t$. Also, we assume for simplicity that η_t and u_{it} are i.i.d. over time, but the logic in this section can be extended to persistent demand shocks (see also Sweeting (2013)).

Recall that in this simple model

$$\epsilon_{it} = \gamma(1 - s_{it}),$$

which implies that the direct impact of all idiosyncratic demand shocks to other companies on s_{it}, and hence ϵ_{it}, is

$$\sum_{j,j\neq i}^{N} \frac{\partial s_{it}}{\partial u_{jt}} u_{jt} = -s_{it}\sum_{j,j\neq i}^{N} s_{jt}u_{jt}.$$

Hence, shocks to companies with larger market shares have a larger impact. This suggests a GIV instrument

$$z_{it}^{GIV} = \bar{s}_{i,t-1}\sum_{j,j\neq i}^{N} \bar{s}_{j,t-1}u_{jt},$$

[54] If a firm offers multiple products, the average of characteristics of other products produced by the same firm can be used as well.

where $\bar{s}_{j,t-1}$ is the average market share for product j up to time $t-1$. This allows us to add a moment condition $\mathbb{E}\left[z_{it}^{GIV}\xi_{it}\right] = 0$, which identifies γ.

6.3.2 Random coefficients logit as in Berry, Levinsohn and Pakes (1995a)

Berry, Levinsohn and Pakes (1995a) extend the standard logit model by allowing for random variation in the preference parameters

$$\theta_h = \theta + \nu_h,$$

where $\nu_v = \left(\nu_h^{\beta}, \nu_h^{\gamma}\right)$ and $\nu_h \sim F_\nu\left(\theta_2\right)$. The market share equation modifies to

$$s_{it} = \int_\nu s_{hit} dF_\nu\left(\theta_2\right),$$

where

$$s_{hit} = \frac{\exp\left(\delta_{it} - \nu_h^{\gamma} p_{it} + \nu_h^{\beta\prime} x_{it}\right)}{1 + \sum_{j=1}^{N} \exp\left(\delta_{jt} - \nu_h^{\gamma} p_{jt} + \nu_h^{\beta\prime} x_{jt}\right)}.$$

To estimate the model, Berry (1994) suggests to recover δ_{it} from the market shares using a contraction mapping (see Nevo (2000) for an introduction). With δ_{it} in hand, we form moment conditions as before to estimate (θ_1, θ_2).

To construct a GIV instrument in this model, we can recompute the total impact of idiosyncratic shocks to other firms on the demand elasticity, which is now slightly more involved. The negative of the demand elasticity, which enters into the pricing equation via the markup, is given by

$$\epsilon_{it} = \int_\nu \gamma_h \frac{s_{hit}}{s_{it}} \left(1 - s_{hit}\right) dF_\nu\left(\theta_2\right).$$

An approximation of the model around $\theta_h = \theta$ yields the same weights as before, although it is feasible to numerically calculate the optimal weights by computing

$$\sum_{j, j \neq i} \frac{\partial \epsilon_{it}}{\partial u_{jt}} u_{jt}.$$

6.4 When aggregate shocks are made of idiosyncratic shocks

We now discuss how GIVs extend to economies where aggregate shocks η_t are themselves made of idiosyncratic shocks u_{it}. Take the basic supply and demand model of Section 2.2. In the case without loops, we achieved identification provided that $u_{\Gamma t} \perp \varepsilon_t$: we do not need $u_{\Gamma t} \perp \eta_t$, so aggregate demand shocks can be influenced by idiosyncratic shocks, but not aggregate supply shocks.

If aggregate supply shocks are affected by idiosyncratic shocks, the elementary strategy does not work, but a variant does work. We suppose disaggregated supply and demand data (for the commodity in question, e.g. oil) is available, at least for large countries. We model country i's supply and demand with the following the factor model:

$$y_{it}^k = \phi^k p_t + \lambda_i^k \eta_t^k + u_{it}^k, \tag{74}$$

where $k = s, d$ indicates supply or demand, respectively. We allow $\mathbb{E}\left[u_{it}^s u_{it}^d\right]$ to be nonzero: for instance, if the US has a "fracking shock" that affects both supply and demand, it will be captured by both u_{it}^s and u_{it}^d for $i =$ USA. This is a concrete case in which supply and demand shocks are correlated: this happens via the correlations in country-level shocks. Suppose that this correlation captures the common shocks, so that $\eta_t^s \perp u_{\Gamma^d t}^d$ (where Γ^d are the residual granular weights given by the demand-side relative size): then, we can identify the elasticity of supply, via $u_{\Gamma^d t}^d$. Likewise, if $\eta_t^s \perp u_{\Gamma^s t}^s$ then the GIV $u_{\Gamma^s t}^s$ allows to estimate the supply elasticity ϕ^s. Section 12.5 details this, and gives more variants.

One can also consider an economy as a network. The general GIV for that would be a whole topic – Sections 6.1 and 12.6 detail this. In some cases, one can obviate the network structure, e.g. via aggregation theorems such as Hulten's theorem. This is developed in Section 12.12.

In conclusion, one can often handle cases where aggregate shocks are made of idiosyncratic shocks: then, some more disaggregated data and economic reasoning allows to use a GIV to estimate macro parameters of interest.

7 Simulations

We illustrate the precision of granularly identified parameters depending on the size of the sample (both N and T), the degree of concentration, and the volatility of idiosyncratic shocks relative to aggregate shocks.

7.1 Model

We start from the standard supply, y_{it}^s, and demand, y_t^d, model

$$y_{it}^s = \phi^s p_t + \lambda_i' \eta_t + u_{it}, \qquad y_t^d = \phi^d p_t + \epsilon_t,$$

where $\phi^d < 0 < \phi^s$, implying, with $M = -\frac{\phi^d}{\phi^s - \phi^d}$,

$$p_t = \frac{M}{\phi^d}\left(u_{St} + \lambda_S' \eta_t - \epsilon_t\right), \qquad y_{St}^s = M u_{St} + M \lambda_S' \eta_t + (1 - M)\,\epsilon_t.$$

7.2 Estimators and standard errors

To estimate M and $\frac{M}{\phi^d}$, we can use standard OLS. To estimate M, we use

$$y_{St}^s = a + M y_{\Gamma t} + \theta' \eta_t^e + e_t, \tag{75}$$

and to estimate $\frac{M}{\phi^d}$, we use

$$p_t = a_p + \frac{M}{\phi^d} y_{\Gamma t} + \theta_p' \eta_t^e + e_t^p. \tag{76}$$

All standard OLS results apply if we observe the factors, η_t. However, we often do not directly observe all factors. We consider the case in which we know the factor loadings, λ_i^η, and where the loadings are unobserved and estimated using PCA. To provide a point of reference, we also consider the case where we do not control for factors and impose that $\theta = \theta_p = 0$. In all cases, we report the OLS standard errors to assess to what extent the OLS standard errors need to be adjusted for the fact that we use η_t^e instead of η_t.

To estimate the demand and supply elasticities, we can recover them from the estimates of M and $\frac{M}{\phi^d}$. However, as discussed before, this is equivalent to a 2SLS estimator using $y_{\Gamma t}$ as instrument for price, while controlling in this case for the factors. Hence, the first stage corresponds to

$$p_t = a_p + \xi y_{\Gamma t} + \theta_p' \eta_t^e + e_t,$$

and the second stage to estimate the demand elasticity is, with $\hat{p}_t = a_p^e + \xi^e y_{\Gamma t} + \theta_p' \eta_t^e$,

$$y_t^d = a_d + \phi^d \hat{p}_t + \theta_d' \eta_t^e + e_t^d,$$

and for the supply elasticity

$$y_{Et}^s = a_s + \phi^s \hat{p}_t + \theta_s' \eta_t^e + e_t^s.$$

The standard weak instrument tests can be used to assess whether $y_{\Gamma t}$ is a sufficiently strong instrument for price (Section 2.3). In this case, we report the 2SLS standard errors to assess whether their accuracy is impacted by the fact that we estimate the common factors.

7.3 Calibration

In calibrating the model, we target (i) concentration, as measured by the excess Herfindahl, $h = \sqrt{\sum_i S_i^2 - 1/N}$, and (ii) the ratio of the volatility of idiosyncratic shocks to the volatility of aggregate supply shocks. In all cases, we estimate the number of common factors using the procedure in Bai and Ng (2002) by minimizing their $IC_{p2}(k)$ criterion.

We set $\phi^d = -0.3$, $\phi^s = 0.1$, and $\sigma_\epsilon = 3\%$. The size weights are generated as $k_i = i^{-1/\zeta}$,

Table 4: Cases considered in simulations. We calibrate the supply-and-demand model in Section 7 under seven alternative parameterizations. The parameters are the following: N is the cross-sectional sample size; T is the number of simulated i.i.d. time periods; h is the excess Herfindahl that we target in our simulation of the size weights (as described in Section 7.3); τ is the targeted ratio of the volatility of idiosyncratic shocks to the volatility of aggregate supply shocks; the multipliers $M = -\frac{\phi^d}{\phi^s - \phi^d}$ and $\frac{M}{\phi^d}$ are functions of the elasticities ϕ^d and ϕ^s of demand and supply with respect to price. The final column reports the share of the price volatility that is due to idiosyncratic shocks under each of the seven parameterizations.

Case	N	T	h	τ	M	$\frac{M}{\phi^d}$	% price vol. idiosyncratic
1	25	360	0.2	3	0.75	-2.5	12.6%
2	25	360	0.2	4	0.75	-2.5	20.4%
3	25	360	0.3	3	0.75	-2.5	19.1%
4	25	360	0.3	4	0.75	-2.5	29.5%
5	25	120	0.2	4	0.75	-2.5	20.4%
6	50	120	0.2	4	0.75	-2.5	16.1%
7	50	360	0.2	4	0.75	-2.5	16.1%

$S_i = k_i / \sum_i k_i$, where ζ is chosen so that $h \in \{0.2, 0.3\}$.[55] In the benchmark case, we assume a single common factor, which follows a standard normal distribution, and uniformly distributed loadings. We consider two cases, namely where $Corr\,(\lambda, S) = 0$ and $Corr\,(\lambda, S) = -20\%$. We scale the loadings so that the variance of aggregate supply shocks follows $V\left(\lambda_S' \eta_t\right) = \lambda_S^2 = 0.03^2$. Lastly, we select $\sigma_u = \tau \left(\lambda_S' \lambda_S\right)^{1/2}$ to target the ratio τ of idiosyncratic shock volatility to aggregate shock volatility. We vary $\tau \in \{3, 4\}$, $N \in \{25, 50\}$, and $T \in \{120, 360\}$.

The cases considered are summarized in Table 4. The final column reports the fraction of price volatility that is due to idiosyncratic shocks, which ranges approximately from 10% to 30%, in line with the recent literature on granularity in terms of how much of aggregate fluctuations can be traced back to idiosyncratic shocks.

7.4 Simulation results

The simulation results when $Corr\,(\lambda, S) = 0$ are reported in Table 5. We consider four estimators. In the case of M1, we assume that the loadings are known in estimating the factors; this is an ideal case taken as a benchmark. In the case of M2, we use PCA to estimate the factors. In the case of M3, we control for the factors estimated using the known loadings and PCA. In the case of M4, we use no factors and just use $y_{\Gamma t}^s$ without any factors. Note that we do not advocate M4 in practice: M4 is there simply to illustrate what goes wrong if we don't control for factors. The first four columns correspond to the estimates of M, the next four columns to estimates of $\frac{M}{\phi^d}$, the next four columns to estimates of ϕ^d, and the last four columns to estimates of ϕ^s.

[55] Here ζ is the power law exponent of the size distribution. See Gabaix (2009) and Section 12.4.

Table 5: Simulation results when $Corr\left(\lambda, S\right) = 0$ based on 10,000 replications. The parameters used in the different cases are summarized in Table 4. In particular, the data are generated from a model in which $M = 0.75$, $\frac{M}{\phi^d} = -2.5$, $\phi^d = -0.3$, and $\phi^s = 0.1$. GIV estimators M1,..., M4 are described at the beginning of Section 7.4. For each estimator, we report the median, the mean, and percentiles 2.5% (P2.5) and 97.5% (P97.5) in the simulated distribution of estimates. "Coverage" is the fraction of estimates falling within the 95% confidence intervals constructed using OLS standard errors (columns 1 through 8) or the 2SLS standard errors (columns 9 through 16).

		M				$\frac{M}{\phi^d}$				ϕ^d				ϕ^s			
Case	Statistic	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
	Median	0.75	0.75	0.75	0.75	-2.50	-2.50	-2.50	-2.49	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
	Mean	0.75	0.75	0.75	0.75	-2.50	-2.50	-2.50	-2.50	-0.33	-0.34	-0.32	-0.34	0.11	0.12	0.11	0.11
1	P2.5	0.53	0.49	0.52	0.51	-3.81	-3.90	-3.86	-3.83	-0.65	-0.67	-0.67	-0.65	0.01	0.00	0.00	0.00
	P97.5	0.98	1.01	0.99	0.99	-1.20	-1.11	-1.15	-1.17	-0.17	-0.17	-0.17	-0.17	0.27	0.32	0.28	0.30
	Coverage	0.95	0.93	0.95	0.95	0.95	0.94	0.95	0.95	0.93	0.93	0.93	0.93	0.96	0.94	0.95	0.96
	Median	0.75	0.75	0.75	0.75	-2.51	-2.50	-2.51	-2.51	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
	Mean	0.75	0.75	0.75	0.75	-2.51	-2.51	-2.51	-2.51	-0.31	-0.31	-0.32	-0.31	0.10	0.11	0.10	0.10
2	P2.5	0.60	0.56	0.59	0.58	-3.48	-3.54	-3.50	-3.52	-0.52	-0.52	-0.53	-0.52	0.04	0.02	0.04	0.03
	P97.5	0.90	0.94	0.90	0.92	-1.54	-1.47	-1.51	-1.49	-0.20	-0.19	-0.19	-0.19	0.19	0.22	0.19	0.21
	Coverage	0.95	0.90	0.95	0.95	0.95	0.93	0.95	0.95	0.94	0.94	0.94	0.94	0.96	0.91	0.96	0.95
	Median	0.75	0.75	0.75	0.75	-2.51	-2.50	-2.51	-2.51	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
	Mean	0.75	0.75	0.75	0.75	-2.51	-2.51	-2.51	-2.51	-0.31	-0.31	-0.31	-0.31	0.10	0.10	0.10	0.10
3	P2.5	0.62	0.59	0.62	0.61	-3.27	-3.34	-3.29	-3.29	-0.44	-0.45	-0.45	-0.44	0.05	0.03	0.04	0.04
	P97.5	0.88	0.92	0.88	0.89	-1.74	-1.68	-1.72	-1.72	-0.22	-0.22	-0.21	-0.22	0.17	0.21	0.18	0.18
	Coverage	0.95	0.90	0.95	0.95	0.94	0.93	0.95	0.94	0.94	0.94	0.95	0.95	0.95	0.90	0.95	0.95
	Median	0.75	0.75	0.75	0.75	-2.50	-2.50	-2.50	-2.51	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
	Mean	0.75	0.75	0.75	0.75	-2.51	-2.50	-2.51	-2.51	-0.31	-0.31	-0.31	-0.31	0.10	0.10	0.10	0.10
4	P2.5	0.65	0.63	0.65	0.64	-3.15	-3.18	-3.17	-3.18	-0.42	-0.42	-0.42	-0.42	0.06	0.05	0.06	0.05
	P97.5	0.85	0.87	0.85	0.87	-1.86	-1.81	-1.84	-1.84	-0.22	-0.22	-0.22	-0.22	0.15	0.17	0.15	0.16
	Coverage	0.95	0.90	0.95	0.95	0.94	0.93	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.89	0.95	0.95
	Median	0.75	0.75	0.75	0.75	-2.51	-2.50	-2.50	-2.50	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
	Mean	0.75	0.75	0.75	0.75	-2.50	-2.50	-2.50	-2.50	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
5	P2.5	0.66	0.64	0.66	0.66	-3.01	-3.05	-3.02	-3.02	-0.38	-0.38	-0.39	-0.38	0.06	0.05	0.06	0.06
	P97.5	0.84	0.86	0.84	0.84	-1.99	-1.95	-1.99	-1.98	-0.24	-0.24	-0.24	-0.24	0.14	0.16	0.15	0.15
	Coverage	0.95	0.90	0.95	0.95	0.94	0.93	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.89	0.95	0.95
	Median	0.75	0.75	0.75	0.75	-2.50	-2.50	-2.50	-2.50	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
	Mean	0.75	0.75	0.75	0.75	-2.50	-2.49	-2.50	-2.50	-0.32	-0.33	-0.33	-0.33	0.11	0.11	0.11	0.11
6	P2.5	0.56	0.52	0.55	0.52	-3.71	-3.78	-3.74	-3.77	-0.64	-0.66	-0.65	-0.64	0.02	0.01	0.02	0.01
	P97.5	0.94	0.98	0.95	0.97	-1.24	-1.16	-1.24	-1.20	-0.17	-0.17	-0.17	-0.17	0.22	0.26	0.22	0.26
	Coverage	0.95	0.93	0.95	0.95	0.95	0.95	0.95	0.95	0.93	0.93	0.93	0.93	0.96	0.95	0.96	0.96
	Median	0.75	0.75	0.75	0.75	-2.50	-2.50	-2.50	-2.50	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.10
	Mean	0.75	0.75	0.75	0.75	-2.50	-2.50	-2.50	-2.50	-0.31	-0.31	-0.31	-0.31	0.10	0.10	0.10	0.10
7	P2.5	0.64	0.62	0.64	0.62	-3.21	-3.27	-3.23	-3.24	-0.44	-0.44	-0.45	-0.44	0.06	0.05	0.06	0.05
	P97.5	0.86	0.88	0.86	0.88	-1.78	-1.75	-1.77	-1.75	-0.22	-0.22	-0.22	-0.22	0.15	0.17	0.16	0.17
	Coverage	0.95	0.91	0.95	0.95	0.95	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.91	0.95	0.95

Table 6: Simulation results when $Corr(\lambda, S) = -20\%$ based on 10,000 replications. The parameters used in the different cases are summarized in Table 4. In particular, the data are generated from a model in which $M = 0.75$, $\frac{M}{\phi^d} = -2.5$, $\phi^d = -0.3$, and $\phi^s = 0.1$. GIV estimators M1,..., M4 are described at the beginning of Section 7.4. For each estimator, we report the median, the mean, and percentiles 2.5% (P2.5) and 97.5% (P97.5) in the simulated distribution of estimates. "Coverage" is the fraction of estimates falling within the 95% confidence intervals constructed using OLS standard errors (columns 1 through 8) or the 2SLS standard errors (columns 9 through 16).

		M				$\frac{M}{\phi^d}$				ϕ^d				ϕ^s			
Case	Statistic	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
	Median	0.75	0.69	0.75	0.56	-2.49	-2.29	-2.50	-1.85	-0.30	-0.30	-0.30	-0.30	0.10	0.13	0.10	0.24
	Mean	0.75	0.69	0.75	0.56	-2.50	-2.30	-2.50	-1.85	-0.33	-0.34	-0.33	-0.30	0.11	0.16	0.11	0.21
1	P2.5	0.52	0.41	0.51	0.29	-3.84	-3.72	-3.87	-3.23	-0.66	-0.74	-0.67	-0.97	0.01	0.01	0.00	0.07
	P97.5	0.98	0.97	0.99	0.82	-1.17	-0.88	-1.13	-0.50	-0.17	-0.16	-0.17	-0.14	0.28	0.47	0.29	0.96
	Coverage	0.95	0.88	0.95	0.67	0.95	0.93	0.95	0.84	0.93	0.93	0.93	0.92	0.96	0.97	0.95	1.00
	Median	0.75	0.74	0.75	0.42	-2.51	-2.46	-2.51	-1.38	-0.30	-0.30	-0.30	-0.30	0.10	0.11	0.10	0.42
	Mean	0.75	0.74	0.75	0.41	-2.51	-2.47	-2.51	-1.39	-0.31	-0.32	-0.32	0.00	0.10	0.11	0.10	-0.12
2	P2.5	0.60	0.56	0.59	0.19	-3.51	-3.50	-3.54	-2.48	-0.52	-0.53	-0.54	-1.05	0.04	0.03	0.04	0.17
	P97.5	0.90	0.92	0.91	0.63	-1.52	-1.43	-1.49	-0.30	-0.19	-0.19	-0.19	-0.12	0.19	0.23	0.19	1.86
	Coverage	0.95	0.91	0.95	0.08	0.94	0.94	0.95	0.41	0.93	0.93	0.94	0.92	0.96	0.94	0.96	0.72
	Median	0.75	0.73	0.75	0.56	-2.51	-2.43	-2.51	-1.86	-0.30	-0.30	-0.30	-0.30	0.10	0.11	0.10	0.24
	Mean	0.75	0.73	0.75	0.56	-2.51	-2.43	-2.51	-1.86	-0.31	-0.31	-0.31	-0.31	0.10	0.12	0.10	0.26
3	P2.5	0.62	0.56	0.61	0.39	-3.29	-3.26	-3.31	-2.68	-0.44	-0.45	-0.45	-0.52	0.04	0.04	0.04	0.12
	P97.5	0.88	0.89	0.89	0.72	-1.73	-1.59	-1.69	-1.02	-0.22	-0.21	-0.21	-0.19	0.18	0.23	0.18	0.51
	Coverage	0.95	0.88	0.95	0.29	0.94	0.93	0.94	0.62	0.94	0.94	0.94	0.94	0.95	0.92	0.95	0.55
	Median	0.75	0.74	0.75	0.51	-2.50	-2.48	-2.51	-1.69	-0.30	-0.30	-0.30	-0.30	0.10	0.10	0.10	0.29
	Mean	0.75	0.74	0.75	0.51	-2.51	-2.48	-2.51	-1.69	-0.31	-0.31	-0.31	-0.31	0.10	0.11	0.10	0.32
4	P2.5	0.65	0.62	0.64	0.34	-3.17	-3.16	-3.19	-2.44	-0.42	-0.42	-0.43	-0.51	0.06	0.05	0.06	0.16
	P97.5	0.85	0.86	0.86	0.66	-1.85	-1.79	-1.82	-0.93	-0.22	-0.22	-0.22	-0.20	0.15	0.18	0.16	0.64
	Coverage	0.95	0.91	0.95	0.07	0.94	0.94	0.95	0.36	0.94	0.95	0.95	0.94	0.95	0.92	0.95	0.11
	Median	0.75	0.74	0.75	0.61	-2.50	-2.45	-2.50	-2.03	-0.30	-0.30	-0.30	-0.30	0.10	0.11	0.10	0.19
	Mean	0.75	0.74	0.75	0.61	-2.51	-2.46	-2.51	-2.03	-0.30	-0.30	-0.30	-0.30	0.10	0.11	0.10	0.20
5	P2.5	0.66	0.62	0.66	0.49	-3.03	-3.02	-3.04	-2.59	-0.39	-0.39	-0.39	-0.41	0.06	0.06	0.06	0.12
	P97.5	0.84	0.85	0.84	0.72	-1.98	-1.89	-1.96	-1.46	-0.24	-0.24	-0.24	-0.23	0.15	0.18	0.15	0.32
	Coverage	0.95	0.88	0.95	0.25	0.95	0.93	0.95	0.57	0.95	0.95	0.95	0.95	0.95	0.89	0.95	0.29
	Median	0.75	0.71	0.75	0.46	-2.49	-2.36	-2.50	-1.53	-0.30	-0.30	-0.30	-0.30	0.10	0.12	0.10	0.35
	Mean	0.75	0.71	0.75	0.46	-2.50	-2.35	-2.50	-1.53	-0.33	-0.34	-0.33	-0.25	0.11	0.14	0.11	0.27
6	P2.5	0.55	0.47	0.55	0.21	-3.73	-3.64	-3.76	-2.84	-0.65	-0.71	-0.67	-1.27	0.02	0.03	0.02	0.12
	P97.5	0.95	0.94	0.95	0.71	-1.22	-1.00	-1.21	-0.21	-0.17	-0.16	-0.17	-0.10	0.22	0.34	0.23	1.84
	Coverage	0.95	0.90	0.95	0.35	0.95	0.94	0.95	0.68	0.93	0.93	0.93	0.91	0.96	0.97	0.96	0.97
	Median	0.75	0.74	0.75	0.46	-2.50	-2.45	-2.50	-1.54	-0.30	-0.30	-0.30	-0.30	0.10	0.11	0.10	0.35
	Mean	0.75	0.74	0.75	0.46	-2.50	-2.45	-2.50	-1.53	-0.31	-0.31	-0.31	-0.31	0.10	0.11	0.10	0.35
7	P2.5	0.64	0.61	0.64	0.30	-3.23	-3.21	-3.24	-2.31	-0.45	-0.45	-0.45	-0.61	0.05	0.05	0.05	0.19
	P97.5	0.86	0.87	0.87	0.62	-1.77	-1.70	-1.76	-0.74	-0.22	-0.21	-0.21	-0.18	0.16	0.18	0.16	0.82
	Coverage	0.95	0.91	0.95	0.03	0.95	0.94	0.95	0.29	0.95	0.94	0.95	0.93	0.96	0.93	0.95	0.16

For each of the estimators, we report the median, the mean, and the 2.5% and 97.5% percentiles. We also compute the fraction of estimates that fall within the 95% confidence intervals constructed using OLS standard errors (columns 1 to 8) or the 2SLS standard errors (columns 9 to 16). We refer to this as the "coverage."

As is clear from all cases, the estimators are mean- and median-unbiased. Moreover, confidence intervals tighten when concentration increases (case 3 relative to case 1 and case 4 relative to case 2) and when the volatility of idiosyncratic shocks increases (case 2 relative to case 1 and case 4 relative to case 3). Naturally, the confidence interval tightens when we increase N and T. The coverage is generally accurate and OLS standard errors only slightly overstate the precision in the case of M2 in estimating M; the 2SLS standard errors are somewhat small in small samples in estimating ϕ^s.

It is tempting to conclude that using $y^s_{\Gamma t}$ as instrument, even without estimating the factors, results in accurate and unbiased estimates of the parameters of interest. However, this is only the case when $Corr\left(\lambda, S\right) = 0$. To illustrate this, we consider a negative correlation between size and exposures, $Corr\left(\lambda, S\right) = -20\%$.

The results are presented in Table 6. Now we find a large bias in case of M4, both in terms of the mean and median. The coverage estimates are also heavily distorted. Intuitively, $y^s_{\Gamma t}$ does not filter out aggregate shocks and the exogeneity restriction is violated. This is why factor estimates are required when loadings may be correlated with size. Even in the case where we have no information about factor loadings (in the case of M2, which relies only on PCA), accounting for common factors removes most of the bias and leads to much improved coverage estimates. When we know the factor loadings (in case of M1), there is no bias and the coverage estimates are accurate. In addition, combining the PCA estimate and the estimate using the known loadings results in almost the same accuracy as M1. This simulation illustrates the importance of accounting for factors in using GIV when loadings correlate with size.

8 Conclusion

We developed granular instrumental variables (GIVs): we remark that idiosyncratic shocks offer a rich source of instruments, and we lay out econometric procedures to optimally extract them from aggregate shocks.

We provided two empirical applications. We plan to put on our web page a series of GIVs, and their control shocks η_t's. They might be useful for empirical work.

Many more applications seem within reach — the introduction listed some. We hope that GIVs will help identifications in new settings and help researchers investigate and understand causal relationships in the economy.

References

Acemoglu, Daron, Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi, "The network origins of aggregate fluctuations," *Econometrica*, 2012, *80* (5), 1977–2016.

Adao, Rodrigo, Michal Kolesár, and Eduardo Morales, "Shift-Share Designs: Theory and Inference," *NBER Working Paper No. 24944*, 2018.

Amiti, Mary and David E Weinstein, "How much do idiosyncratic bank shocks affect investment? Evidence from matched bank-firm loan data," *Journal of Political Economy*, 2018, *126* (2), 525–587.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings, "International shocks, variable markups, and domestic prices," *Forthcoming at the Review of Economic Studies*, 2019.

Andrews, Isaiah, James Stock, and Liyang Sun, "Weak instruments in IV regression: Theory and practice," *Forthcoming at the Annual Review of Economics*, 2018.

Arellano, Manuel and Stephen Bond, "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *The review of economic studies*, 1991, *58* (2), 277–297.

Autor, David H, David Dorn, and Gordon H Hanson, "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, 2013, *103* (6), 2121–68.

Bai, Jushan and Serena Ng, "Determining the number of factors in approximate factor models," *Econometrica*, 2002, *70* (1), 191–221.

Bai, Jushan and Serena Ng, "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Econometrica*, 2006, *74* (4), 1133–1150.

Baqaee, David Rezza and Emmanuel Farhi, "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem," *NBER Working Paper No. 23145*, 2018.

Bartik, Timothy J, *Who benefits from state and local economic development policies?*, W.E. Upjohn Institute for Employment Research, 1991.

Baumeister, Christiane and James D Hamilton, "Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks," *American Economic Review*, 2019, *109* (5), 1873–1910.

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile prices in market equilibrium," *Econometrica*, 1995, pp. 841–890.

Berry, Steven T., “Estimating Discrete-Choice Models of Product Differentiation,” *The RAND Journal of Economics*, 1994, *25* (2), 242–262.

Blasques, Francisco, Siem Jan Koopman, Andre Lucas, and Julia Schaumburg, “Spillover dynamics for systemic risk measurement using spatial financial time series models,” *Journal of Econometrics*, 2016, *195* (2), 211 – 223.

Blundell, Richard and Stephen Bond, “Initial conditions and moment restrictions in dynamic panel data models,” *Journal of econometrics*, 1998, *87* (1), 115–143.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel, “Quasi-experimental shift-share designs,” *NBER Working Paper No. 24997*, 2018.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, “Identification of peer effects through social networks,” *Journal of Econometrics*, 2009, *150* (1), 41–55.

Caballero, Ricardo J and Alp Simsek, “A model of fickle capital flows and retrenchment,” *NBER Working Paper No. 22751*, 2018.

Caldara, Dario, Michele Cavallo, and Matteo Iacoviello, “Oil price elasticities and oil price fluctuations,” *Journal of Monetary Economics*, 2018.

Caporin, Massimiliano, Loriana Pelizzon, Francesco Ravazzolo, and Roberto Rigobon, “Measuring Sovereign Contagion in Europe,” *Journal of Financial Stability*, 2018, *34*, 150–181.

Carvalho, Vasco Miguel and Basile Grassi, “Large Firm Dynamics and the Business Cycle,” *American Economic Review*, 2019.

Chodorow-Reich, Gabriel, “Geographic Cross-Sectional Fiscal Spending Multipliers: What Have We Learned?,” *NBER Working Paper No. 23577*, 2017.

di Giovanni, Julian and Andrei A Levchenko, “Country size, international trade, and aggregate fluctuations in granular economies,” *Journal of Political Economy*, 2012, *120* (6), 1083–1132.

di Giovanni, Julian, Andrei A Levchenko, and Isabelle Méjean, “Firms, destinations, and aggregate fluctuations,” *Econometrica*, 2014, *82* (4), 1303–1340.

Fama, Eugene F and Kenneth R French, “Common risk factors in the returns on stocks and bonds,” *Journal of Financial Economics*, 1993, *33* (1), 3–56.

Fan, Jiangqing, Yuan Liao, and Weichen Wang, “Projected Principal Component Analysis in Factor Models,” *The Annals of Statistics*, 2016, *44* (1), 219–254.

Farhi, Emmanuel and Jean Tirole, “Deadly embrace: Sovereign and financial balance sheets doom loops,” *The Review of Economic Studies*, 2017, *85* (3), 1781–1823.

Forbes, Kristin J. and Roberto Rigobon, "No Contagion, Only Interdependence: Measuring Stock Market Comovements," *The Journal of Finance*, 2002, *57* (5), 2223–2261.

Gabaix, Xavier, "Power Laws in Economics and Finance," *Annual Review of Economics*, May 2009, *1* (1), 255–294.

Gabaix, Xavier, "The granular origins of aggregate fluctuations," *Econometrica*, 2011, *79* (3), 733–772.

Gabaix, Xavier and Matteo Maggiori, "International liquidity and exchange rate dynamics," *The Quarterly Journal of Economics*, 2015, *130* (3), 1369–1420.

Gaubert, Cecile and Oleg Itskhoki, "Granular comparative advantage," *NBER Working Paper No. 24807*, 2018.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift, "Bartik Instruments: What, When, Why, and How," *NBER Working Paper No. 24408*, 2018.

Graham, Bryan S, "Identifying social interactions through conditional variance restrictions," *Econometrica*, 2008, *76* (3), 643–660.

Heckman, James, "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 1978, *46* (4), 931–59.

Kelejian, Harry H and Ingmar R Prucha, "A generalized moments estimator for the autoregressive parameter in a spatial model," *International Economic Review*, 1999, *40* (2), 509–533.

Kilian, Lutz, "Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market," *American Economic Review*, 2009, *99* (3), 1053–69.

Kilian, Lutz and Daniel P Murphy, "The role of inventories and speculative trading in the global market for crude oil," *Journal of Applied Econometrics*, 2014, *29* (3), 454–478.

Kline, Brendan and Elie Tamer, "Econometric Analysis of Models With Social Interactions," in Bryan Graham and Aureo De Paula, eds., *The Econometric Analysis of Network Data*, Academic Press, Forthcoming.

Koijen, Ralph SJ and Motohiro Yogo, "An equilibrium model of institutional demand and asset prices," *Journal of Political Economy*, 2019.

Kramarz, Francis, Julien Martin, and Isabelle Mejean, "Volatility in the small and in the large: the lack of diversification in international trade," 2016.

Kuersteiner, Guido M and Ingmar R Prucha, "Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity," *arXiv preprint arXiv:1802.01755*, 2018.

Leary, Mark T and Michael R Roberts, "Do peer firms affect corporate financial policy?," *The Journal of Finance*, 2014, *69* (1), 139–178.

Lee, Youngki, Luís A Nunes Amaral, David Canning, Martin Meyer, and H Eugene Stanley, "Universal features in the growth dynamics of complex organizations," *Physical Review Letters*, 1998, *81* (15), 3275.

Lewbel, Arthur, "Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models," *Journal of Business & Economic Statistics*, 2012, *30* (1), 67–80.

Long, John B and Charles I Plosser, "Real business cycles," *Journal of Political Economy*, 1983, *91* (1), 39–69.

Manski, Charles F, "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

Nakamura, Emi and Jon Steinsson, "Identification in Macroeconomics," *Journal of Economic Perspectives*, August 2018, *32* (3), 59–86.

Nevo, Aviv, "A practitioner's guide to estimation of random-coefficients logit models of demand," *Journal of economics & management strategy*, 2000, *9* (4), 513–548.

Nevo, Aviv, "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 2001, *69* (2), 307–342.

Onatski, Alexei, "Testing hypotheses about the number of factors in large factor models," *Econometrica*, 2009, *77* (5), 1447–1479.

Pasten, Ernesto, Raphael Schoenle, and Michael Weber, "Price Rigidity and the Origins of Aggregate Fluctuations," *NBER Working Paper No. 23750*, 2017.

Ramey, Valerie A, "Macroeconomic shocks and their propagation," in "Handbook of Macroeconomics," Vol. 2, Elsevier, 2016, pp. 71–162.

Rasmussen, Carl and Christopher Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2005.

Rigobon, Roberto, "Identification through heteroskedasticity," *Review of Economics and Statistics*, 2003, *85* (4), 777–792.

Sarto, Andres Pablo, "Recovering Macro Elasticities from Regional Data," *Working Paper*, 2018.

Shi, Wei and Lung-fei Lee, "Spatial dynamic panel data models with interactive fixed effects," *Journal of Econometrics*, 2017, *197* (2), 323–347.

Stock, James H and Mark W Watson, "Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics," in "Handbook of macroeconomics," Vol. 2, Elsevier, 2016, pp. 415–525.

Stock, James H and Mark W Watson, "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments," *The Economic Journal*, 2018, *128* (610), 917–948.

Stock, James H and Motohiro Yogo, "Testing for weak instruments in linear IV regression," in Donald WK Andrews and James H Stock, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, 2005.

Sweeting, Andrew, "Dynamic Product Positioning in Differentiated Product Markets: The Effects of Fees for Musical Performance Rights on the Commercial Radio Industry," *Econometrica*, 2013, *81* (5), 1763–1803.

Zou, Hui, Trevor Hastie, and Robert Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, 2006, *15* (2), 265–286.

9 Appendix: Proofs omitted in the paper

Variance facts We will repeatedly use the fact that if $(u_i)_{i=1...N}$ is a series of uncorrelated random variables with mean 0 and common variance σ_u^2, then

$$\mathbb{E}\left[u_\Gamma u_E\right] = 0, \tag{77}$$

and

$$\mathbb{E}\left[u_\Gamma^2\right] = \mathbb{E}\left[u_S u_\Gamma\right] = h^2 \sigma_u^2. \tag{78}$$

Hence, the standard deviation of the granular residual $u_{\Gamma t}$ is proportional to the Herfindahl. In the general heteroskedastic case, the quasi-equal weight vector is

$$\tilde{E} = \frac{\left(V^u\right)^{-1} \iota}{\iota \left(V^u\right)^{-1} \iota}.$$

Then, for any Γ such that $\iota' \Gamma = 0$, we have:[56]

$$\mathbb{E}\left[u_\Gamma u_{\tilde{E}}\right] = 0. \tag{79}$$

Proof of Proposition 2 The proof is quite elementary, and uses well-known ingredients. We have

$$\alpha_T^e - \alpha = \frac{\mathbb{E}_T\left[\left(\alpha y_{St} + \varepsilon_t\right) u_{\Gamma t}\right]}{\mathbb{E}_T\left[y_{St} u_{\Gamma t}\right]} - \alpha = \frac{\mathbb{E}_T\left[\varepsilon_t u_{\Gamma t}\right]}{\mathbb{E}_T\left[y_{St} u_{\Gamma t}\right]} = \frac{A_T}{D_T}.$$

Next, the law of large number gives:

$$D_T = \mathbb{E}_T\left[y_{St} u_{\Gamma t}\right] \to^{a.s.} D,$$

with

$$D = \mathbb{E}\left[y_{St} u_{\Gamma t}\right] = \mathbb{E}\left[\left(\eta_t + u_{St}\right) u_{\Gamma t}\right] = \mathbb{E}\left[u_{St} u_{\Gamma t}\right] = \mathbb{E}\left[\left(u_{\Gamma t} + u_{Et}\right) u_{\Gamma t}\right] = \mathbb{E}\left[u_{\Gamma t}^2\right] = \sigma_{u_\Gamma}^2.$$

For the numerator, the central limit theorem gives the convergence in distribution:

$$\sqrt{T} A_T \to^d \mathcal{N}\left(0, \sigma_A^2\right),$$

[56] Here is the proof. We have $\tilde{E} = k \left(V^u\right)^{-1} \iota$ for $k = \frac{1}{\iota (V^u)^{-1} \iota}$. So

$$\mathbb{E}\left[u_\Gamma u_{\tilde{E}}\right] = \mathbb{E}\left[\left(\tilde{E}' u\right)\left(u' \Gamma\right)\right] = \tilde{E}' \mathbb{E}\left[u u'\right] \Gamma = \tilde{E}' V^u \Gamma = k \iota' \left(V^u\right)^{-1} V^u \Gamma = k \iota' \Gamma = 0,$$

as $\iota' \Gamma = 0$.

where

$$\sigma_A^2 = \mathbb{E}\left[\varepsilon_t^2 u_{\Gamma t}^2\right] = \mathbb{E}\left[\varepsilon_t^2\right]\mathbb{E}\left[u_{\Gamma t}^2\right] = \sigma_\varepsilon^2 \sigma_{u_\Gamma}^2,$$

so that

$$\frac{\sigma_A}{D} = \frac{\sigma_\varepsilon \sigma_{u_\Gamma}}{\sigma_{u_\Gamma}^2} = \frac{\sigma_\varepsilon}{\sigma_{u_\Gamma}} =: \sigma_\alpha.$$

Hence,

$$\sqrt{T}\left(\alpha_T^e - \alpha\right) \rightarrow^d N\left(0, \sigma_\alpha^2\right).$$

Then the u_{it}' are i.i.d. across i's, then $\sigma_{u_\Gamma} = h\sigma_u$, see (78).

Proof of Proposition 3 We have

$$\alpha_T^e - \alpha = \frac{\mathbb{E}_T\left[\left(\alpha y_{St} + \varepsilon_t\right) z_t\right]}{\mathbb{E}_T\left[y_{St} z_t\right]} - \alpha = \frac{\mathbb{E}_T\left[\varepsilon_t z_t\right]}{\mathbb{E}_T\left[y_{St} z_t\right]},$$

so the same proof as for Proposition 2 yields the asymptotic error

$$\sigma_\alpha\left(\Gamma\right) = \frac{\sigma_\varepsilon \sigma_z}{\left|\mathbb{E}\left[y_{St} z_t\right]\right|} = \frac{\sigma_\varepsilon \sigma_z}{\left|\mathbb{E}\left[u_{St} z_t\right]\right|} = \frac{\sigma_\varepsilon}{\sigma_{y_S}\left|corr\left(u_{St}, z_t\right)\right|}.$$

So, the best estimator $z_t = u_{\Gamma t}$ maximizes the squared correlation $C\left(\Gamma\right) := corr\left(u_{St}, u_{\Gamma t}\right)^2$:

$$\max_{\Gamma} C\left(\Gamma\right) \text{ subject to } \iota'\Gamma = 0.$$

We next solve this problem.

Call V the variance covariance matrix of the u_i. We have:

$$C^2 var\left(u_{St}\right) = \frac{\mathbb{E}\left[u_{St} u_{\Gamma t}\right]^2}{var\left(u_{\Gamma t}\right)} = \frac{\left(S'V\Gamma\right)^2}{\Gamma V \Gamma}.$$

The problem is invariant to changing Γ into $\lambda\Gamma$ for a non-zero λ. So, we can fix say $S'V\Gamma$ at some value. Given this, we want the minimum value of $\Gamma V \Gamma$. So, we minimize over Γ the Lagrangian

$$\mathcal{L} = \frac{1}{2}\Gamma' V \Gamma - b\Gamma'\iota - c\Gamma' V S \tag{80}$$

with some Lagrange multipliers b, c. The first order condition in Γ' is: $0 = V\Gamma - b\iota - cVS$, i.e.

$$\Gamma = cS + bV^{-1}\iota.$$

Now, using $\iota'\Gamma = 0$ gives $0 = c + b\iota' V^{-1}\iota$, i.e., with $\tilde{E} := \frac{V^{-1}\iota}{\iota' V^{-1}\iota}$,

$$\Gamma = c\left(S - \tilde{E}\right).$$

The factor c doesn't affect the results, (as Γ and $c'\Gamma$ give the same estimator α_T^e), so we may choose $c = 1$.

Proof of Proposition 4: Sketch The full proof is in the online appendix (Section 13.1). Here we provide a proof sketch. For simplicity, we omit the controls C_t. We use a projection matrix Q (defined in (23) with $W = (V^u)^{-1}$) satisfying:

$$Q\Lambda = 0, \tag{81}$$

$$QV^u (I - Q') = 0. \tag{82}$$

Then, premultiplying (54) by Q, we have $Qy_t = Qu_t$. Next, we define $\Gamma := Q'S$, and the GIV as the scalar:

$$z_t := \Gamma' y_t, \tag{83}$$

i.e. $z_t = S'Qy_t = S'Qu_t = \Gamma' u_t$, i.e.

$$z_t = u_{\Gamma t}. \tag{84}$$

Assumption 1 ensures $\Gamma \neq 0$. Assumption 2 ensures that V^u can be recovered from the knowledge of Qu_t.

Recall that we have (56),

$$y_{St} = Mu_{St} + \varepsilon_t$$

for ε_t a shock correlated with the η_t but not with the u_{it}'s. Finally, we have

$$u_{St} = S'u_t = S'Qu_t + S'(I - Q)u_t = z_t + v_t,$$

with $v_t = S'(I - Q)u_t$. Now, (82) ensures $\mathbb{E}[z_t v_t] = 0$. Then, we can write:

$$y_{St} = Mz_t + \varepsilon_t^{yS},$$

with $\varepsilon_t^{yS} := Mv_t + \varepsilon_t$ orthogonal to z_t. Hence, we can estimate the multiplier M by OLS.

Likewise, we have (via (56) and (58))

$$F_t^f = \alpha^f M z_t + \varepsilon_t^f$$

for some shock ε_t^f orthogonal to z_t. Hence, we can estimate $\alpha^f M$ by regressing F_t^f on z_t.

In both regressions, we can add the estimated common shocks η_t^e as controls, which improves the precision. The full proof shows that η_t^e is orthogonal to z_t, so those controls still lead to a consistent estimators of M and $\alpha^f M$.

10 Appendix: Complements and extensions

10.1 The model with heterogeneous loadings on endogenous and exogenous factors

In the main text, we assume for simplicity a homogeneous or parametric sensitivity on endogenous factors, e.g. on the price in the simple supply and demand example. We show how our framework generalizes easily.

The model of Section 3.1 implies the representation:

$$y_{it} = \theta_i u_{St} + \lambda_i \eta_t + u_{it}, \tag{85}$$

where $\theta_i = \frac{\sum_f \lambda_i^f \alpha^f}{1-\sum_f \lambda_S^f \alpha^f}$ is the sensitivity to endogenous factors, η_t is a vector of exogenous factors, and λ_i is a vector of factor loadings, both $r-$dimensional. The new difficulty is to estimate a heterogeneous set of θ_i — in our more basic case we considered the case of a common θ_i. We focus on the case where the shocks are homoskedastic, $V^u = \sigma^2 I$.

To motivate the procedure, assume that we know $\Lambda = (\lambda, \theta)$ and we estimate the residuals, $v_t = Qu_t$, where $Q = Q^{\Lambda,W}$ was defined in (23), with $W = (V^u)^{-1} = \sigma^{-2} I$. Then it holds[57]

$$y_{it} = \theta_i v_{St}^e + \lambda_i \eta_t^e + v_{it}, \tag{86}$$

with η_t^e the estimate of η_t. Then define

$$z_t = v_{St}, \qquad z_{it} = z_t - \Gamma_i^u v_{it}, \tag{87}$$

where

$$\Gamma_i^u := \frac{\mathbb{E}\left[v_{it} z_t\right]}{\mathbb{E}\left[v_{it}^2\right]}$$

which in this case equals

$$\Gamma_i^u = \frac{\Gamma_i}{Q_{ii}} \tag{88}$$

where we define $\Gamma := Q'S$. This implies that $\mathbb{E}\left[v_{it} z_{it}\right] = 0$. So z_{it} is like the traditional GIV, but it is uncorrelated with v_{it}. Morally (and in the case where the u_{jt} are uncorrelated), it is made of the idiosyncratic shocks of the actors (e.g. firms of countries) other than i.

Then, given (86), it holds

$$y_{it} = \theta_i z_{it} + \lambda_i \eta_t + \left(1 + \theta_i \Gamma_i^u\right) v_{it}, \tag{89}$$

[57]Note that $(v_{St}^e, \eta_t^e)' = (\Lambda'\Lambda)^{-1} \Lambda' y_t = (u_{St}, \eta_t)' + (\Lambda'\Lambda)^{-1} \Lambda' u_t$.

and we can estimate θ_i via OLS of y_{it} on z_{it}. We rewrite this equation as

$$y_{it} - \theta_i \Gamma_i^u v_{it} = \theta_i z_{it} + \lambda_i \eta_t + v_{it}.$$

This suggests the following iterative procedure. We call the set of parameters to be estimated $\omega = (\eta_t, \lambda_i, \theta_i)_{i,t}$. At round 0, before any estimation, we initialize $v_{it}^{(0)} = y_{it} - y_{Et}$. At round $n \geq 1$ of estimation, we define:

$$w_{it}^e(\omega) \coloneqq y_{it} - \theta_i \Gamma_i^u v_{it}^{(n-1)} - \theta_i z_{it}^{(n-1)} - \lambda_i \eta_t, \tag{90}$$

and want to minimize

$$\min_{\omega} \sum_t \sum_i w_{it}^e(\omega)^2. \tag{91}$$

More concretely, empirically, we first estimate θ_i using OLS regression

$$y_{it} - \theta_i \Gamma_i^u v_{it}^{(n-1)} = \theta_i z_{it}^{(n-1)} + e_{it}^{(n)}, \tag{92}$$

and then estimate λ and η via PCA on the residuals from the OLS regression, $e_{it}^{(n)}$. The residuals from the PCA step, $e_t^{(n)} - \lambda^{(n)} \eta_t^{(n)}$, is an estimate of the idiosyncratic shocks, $v_t^{(n)}$. We iterate until convergence.

Once the model is estimated, we can get z_t and use it to estimate the sensitivity α^f of the endogenous factors via OLS on z_t, like in the GIV with homogeneous sensitivity to endogenous factors $F_t^f = \alpha^f M z_t + \lambda^f \eta_t + \varepsilon^f$.

Remarks Empirically, a more accurate approach to introduce heterogeneity in loadings is to model $\theta_i = \Theta' x_{it}$ for some vector x_{it} of characteristics. Then, we estimate Θ in the PCA-OLS step.

In the presence of heteroskedasticity, or correlated innovations, there are two potential approaches. First, it may be possible to directly estimate $\Gamma_i^u = \frac{\mathbb{E}[v_{it} z_t]}{\mathbb{E}\left[v_{it}^2\right]}$. Second, call $V = V^u$ the variance-covariance matrix of the u_{it}, and we may take the theoretical value:

$$\Gamma_i^u = \frac{(QV\Gamma)_i}{(QVQ')_{ii}}, \tag{93}$$

where we define:[58]

$$\Gamma := Q'S, \qquad Q = Q^{\Lambda,W}. \tag{94}$$

10.2 When the influence matrix is not proportional to size

10.2.1 Position of the problem

Suppose a model

$$y_{it} = \gamma \sum_j G_{ij} y_{jt} + \lambda_i \eta_t + u_{it}, \tag{95}$$

i.e.

$$y_t = \gamma G y_t + \Lambda \eta_t + u_t, \tag{96}$$

with a given "influence" matrix G. For instance, if we have an "industrial similarity" matrix H with entries H_{ij}, we might set

$$G_{ij} = \frac{H_{ij} S_j}{\sum_k H_{ik} S_k}.$$

In our basic "reflection problem", $G = \iota S'$.

We'd like to identify γ. With $V = \mathbb{E}[u_t u_t']$, we define $\tilde{E} = \frac{V^{-1}\iota}{\iota V^{-1}\iota}$, and the "generalized size vector":

$$S := G' \tilde{E}, \tag{97}$$

which is the analogue of "size" in our simpler setup where $G = \iota S'$.

10.2.2 A simple approach, when the loading on common shocks if uniform

In this subsection we assume that

$$G\iota = \iota, \tag{98}$$

which is satisfied in many examples (Section 12.6 has the general case). Consider some vector $\mathcal{E}$, and define:

$$z_t := \mathcal{E}' (G - I) y_t. \tag{99}$$

[58] The justification for (93) is as follows. As $v_t = Q u_t$,

$$\mathbb{E}[v_t v_t'] = \mathbb{E}[Q u_t u_t' Q'] = QVQ'$$

$$\mathbb{E}[v_t v_{St}] = \mathbb{E}[v_t v_t'] S = QVQ'S = QV\Gamma$$

so that $\Gamma_i^u := \frac{\mathbb{E}[v_{it} v_{St}]}{\mathbb{E}[v_{it}^2]} = \frac{(QV\Gamma)_i}{(QVQ')_{ii}}$. When we take $W = (V^u)^{-1}$, we have the relation $QVQ' = VQ'$ (see (177)), which implies

$$\mathbb{E}[v_t v_t'] = QVQ' = VQ', \qquad \mathbb{E}[v_t v_{St}] = \mathbb{E}[v_t v_t'] S' = VQ'S = V\Gamma,$$

so that $\Gamma_i^u := \frac{(V\Gamma)_i}{(VQ')_{ii}}$.

Then, we have the key relation:

$$\mathbb{E}\left[z_t\left(y_{\tilde{E}t}\text{-}\gamma y_{St}\right)\right]=0, \tag{100}$$

which allows to identify γ by $\gamma=\frac{\mathbb{E}\left[z_t y_{\tilde{E}t}\right]}{\left[z_t y_{St}\right]}$.

Relation (100) works for any z_t of the type (99). It is sensible to take $\mathcal{E}=\tilde{E}$ (one can show that this is the optimum choice in the sense of minimizing the asymptotic error). Then, the GIV is again: $z_t=y_{St}-y_{\tilde{E}t}$.

Derivation of (100) Indeed,

$$y_{\tilde{E}t}-\gamma y_{St}=\tilde{E}'\left(I-\gamma G\right)y_t=\tilde{E}'\left(u_t+\eta_t\iota\right).$$

Given $G\iota=\iota$, and $(G-I)$ and $(I-\gamma G)^{-1}$ commute, we have the useful relation:

$$\left(G-I\right)\left(I-\gamma G\right)^{-1}\iota=0. \tag{101}$$

As a result,

$$z_t=\mathcal{E}'\left(G-I\right)y_t=\mathcal{E}'\left(G-I\right)\left(I-\gamma G\right)^{-1}\left(u_t+\eta_t\iota\right)=\mathcal{E}'\left(G-I\right)\left(I-\gamma G\right)^{-1}u_t.$$

Hence,

$$\begin{aligned} a &:= \mathbb{E}\left[z_t\left(y_{\tilde{E}t}-\gamma y_{St}\right)\right]=\mathbb{E}\left[\mathcal{E}'\left(G-I\right)\left(I-\gamma G\right)^{-1}u_t\left(u_t+\eta_t\iota\right)'\tilde{E}\right] \\ &= \mathcal{E}'\left(G-I\right)\left(I-\gamma G\right)^{-1}V\tilde{E}=\mathcal{E}'\left(G-I\right)\left(I-\gamma G\right)^{-1}\frac{\iota}{\iota V^{-1}\iota}=0. \end{aligned}$$

Online Appendix for
"Granular Instrumental Variables"

Xavier Gabaix and Ralph S.J. Koijen

July 30, 2019

This online appendix gives complements to the theory, the underlying models, and the empirical examples. It gives also additional proofs.

11 Microfoundations for the model of sovereign spillovers

In this model, spillovers happens because debt defaults are partially mutualized. This is a stand-in for potentially much richer economics. For instance, contagion might work via GDP spillovers, or the limited risk capacity of specialized arbitrageurs. Still, the specification that this model delivers might be broadly similar, as we shall see.

11.1 Model setup

We make a number of simplifying assumptions. The safe interest rate is normalized to 0, and pricing is risk neutral. Time is continuous in $[0,T]$. We neglect the $O(dt)$ terms, which are irrelevant for the regression analysis we are interested in, i.e. will write $df(X_t) = f'(X_t)\,dX_t$.[59],[60]

Payoffs are realized at a date T, which should be thought about as faraway. Country i's outstanding debt is B_i, and the value of the debt (per unit of face value) is thus:

$$Q_{it} = \mathbb{E}_t\left[1 - L_{iT}^{+}\right] = e^{-(T-t)y_{it}}, \tag{102}$$

where $x^+ := \max(x,0)$, y_{it} is the yield spread over the the safe interest rate (which we normalized to 0), and L_{iT} is the relative "vulnerability" of the government's bonds, defined as

$$L_{iT} = \frac{F_{iT}}{B_i}, \tag{103}$$

where F_{iT} is the value of potential losses from government defaults (in euros). We assume that F_{iT} follows:

[59] Formally, we write all the differential expressions $dY_t = a_t dZ_t$ modulo an equivalence by terms $b_t dt$ (or, to be pedantic, we quotient by the ring of expressions of the type $b_t dt$ where b_t is an adapted function). So, $df(X_t) = f'(X_t)\,dX_t$ modulo dt, where we keep the "modulo dt" implicit.

[60] We only care, for the regressions, about the "dZ_t" terms, that depends on innovation to underlying Brownian shocks dZ_t, as those are the loading detected by the regressions.

$$F_{iT} = \psi_{iT} G_{iT}, \tag{104}$$

where $\psi_{iT} \in [0,1]$ is a propensity to pass on raw government fiscal losses G_{iT} to bondholders. A financially virtuous country (say Germany) has ψ_{iT} close to 0, and a laxer country has a high ψ_{iT}. To gain intuition, it is useful to think that most variation in yield spreads comes from the political willingness to not pay bondholders, ψ_{iT}.

This raw position G_{iT} is in turn:

$$G_{iT} = V_{iT} - \phi F_{iT}^{+} + \phi m_i F_T, \tag{105}$$

where V_{iT} is a stochastic "latent loss", and the total amount lost on bonds is:

$$F_T = \sum_i F_{iT}^{+}. \tag{106}$$

Debts are partially mutualized with intensity $\phi \in [0,1]$: a fraction ϕ of the loss F_{it}^{+} is passed on to other countries, with a share m_i to country i ($\sum_i m_i = 1$, $m_i \geq 0$). This mutualization creates the sovereign yield spillovers.

To simplify the analysis, we assume that V_{iT} is strictly positive with probability 1, so that F_{iT}, G_{iT} and L_{iT} are all strictly positive with probability 1. This is less restrictive that it may appear: losses could be very small. This is simply to make the analysis very tractable.

11.2 Model solution

Solving the model,

$$\begin{aligned} L_{iT} &= \frac{\psi_{iT}}{B_i}\left(V_{iT} - \phi F_{iT} + \phi m_i F_T\right) \\ &= \frac{\psi_{iT}}{B_i}\left(V_{iT} - \phi B_i L_{iT} + \phi m_i B L_T\right), \end{aligned}$$

with $B = \sum_i B_i$ and $L_T = \frac{F_T}{B}$, i.e.

$$L_T = \sum_i \frac{B_i}{B} L_{iT}. \tag{107}$$

We call $\rho_i = \frac{m_i}{B_i/B}$, the ratio between country i's mutualization share m_i and its debt share.[61],[62].

[61] The ECB's capital key, which defines the equity shares of member states in the ECB, is defined using 50% of GDP shares and 50% of population shares. However, we do not focus exclusively on spillovers that operate via the ECB and there may be other effects via trade linkages, demand shocks from investors, et cetera. We maintain the assumption that the losses, or exposures, to Eurozone-wide losses are proportional to GDP. Alternatively, we could change the measure m_i to be a function of both population and GDP shares.

[62] One can imagine $\rho_i \simeq 1$ as a simple baseline where most variations come from the political willingness ψ_{it}.

This leads to:

$$L_{iT} = \frac{\psi_{iT}}{1+\phi\psi_{iT}}\left(\frac{V_{iT}}{B_i} + \phi\frac{m_i}{B_i}BL_T\right).$$

So, if we define

$$\Psi_{iT} = \frac{\psi_{iT}}{1+\phi\psi_{iT}}, \tag{108}$$

we have:

$$L_{iT} = \Psi_{iT}\left(\frac{V_{iT}}{B_i} + \phi\rho_i L_T\right). \tag{109}$$

This shows the "contagion" in the space of vulnerabilities, L_{iT}.

To move to yields, we do a Taylor expansion for small yield spreads, so that (102) gives:

$$y_{it} = a_t \mathbb{E}_t\left[L_{iT}\right], \tag{110}$$

where

$$a_t = \frac{1}{T-t} \tag{111}$$

is a slowly-varying parameter (as T is far from the interval of times t under study – so we'll take the approximation $da_t \simeq 0$). We define $\Psi_{it} = \mathbb{E}_t\left[\Psi_{iT}\right]$, $v_{it} = a_t \mathbb{E}_t\left[\frac{V_{iT}}{B_i}\right]$. Also, we place ourselves in the "quasi-static" regime, where all noises are small—see Section 11.3 for details. Hence, (109) becomes, in yield space:

$$y_{it} = \Psi_{it}\left(v_{it} + \phi\rho_i y_{St}\right), \tag{112}$$

where

$$y_{St} = \frac{\sum_i B_i y_{it}}{B}. \tag{113}$$

This shows that the yield spread depends on a country-specific fundamental v_{it} and a "spillover" proportional to ϕ. At the same time, for a very financially virtuous country with $\Psi_{it} \simeq 0$, the yield spread is close to 0, so that yield contagion is close to 0: as the country is quite safe anyway, external disruptions cannot move the yield much away from 0.

We have

$$\frac{dy_{it}}{y_{it}} = \frac{d\Psi_{it}}{\Psi_{it}} + \frac{dv_{it}}{v_{it}+\phi\rho_i y_{St}} + \frac{\phi\rho_i y_{St}}{v_{it}+\phi\rho_i y_{St}}\frac{dy_{St}}{y_{St}},$$

hence

$$\frac{dy_{it}}{y_{it}} = dw_{it} + \gamma_{it}\frac{dy_{St}}{y_{St}} \tag{114}$$

for $dw_{it} := \frac{d\Psi_{it}}{\Psi_{it}} + \frac{dv_{it}}{v_{it}+\phi\rho_i y_{St}}$ and for a coefficient $\gamma_{it} := \frac{\phi\rho_i y_{St}}{v_{it}+\phi\rho_i y_{St}} \in [0,1]$. In the simple benchmark where all countries have a similar v_{it} (fundamental government finances) but differ mostly in Ψ_{it} (the propensity to absorb the shocks rather than pass it on to debt holders by defaulting) and $\rho_i = 1$, we have $\gamma_{it} = \frac{\phi y_{St}}{v_t+\phi y_{St}}$.

Written another way, call

$$\tilde{y}_{it} := \ln y_{it}. \tag{115}$$

Then, we have

$$d\tilde{y}_{it} = dw_{it} + \gamma_{it} d\tilde{y}_{\tilde{S}t}, \tag{116}$$

where

$$\tilde{S}_{it} = \frac{B_i y_{it}}{\sum_j B_j y_{jt}}, \tag{117}$$

$$d\tilde{y}_{\tilde{S}t} = \sum_i \tilde{S}_{it} d\tilde{y}_{it} = \sum_i \frac{B_i y_{it} \frac{dy_{it}}{y_{it}}}{\sum_j B_j y_{jt}} = \frac{dy_{St}}{y_{St}}. \tag{118}$$

Hence, if we reason in "log yield spread" space, the proper weights are proportional to $B_i y_{it}$, i.e. debt value times yield spread. This is the formulation that motivates our empirical specification (61). In particular, if $\Psi_{it} = 0$, then the change is $dy_{it} = 0$ always. The importance of the spillovers is given by $\sum_j B_j dy_{jt}$, the change in the yield weighted by debt value, summed over all countries.

11.3 Quasi-static regime of stochastic processes

Suppose a stochastic process, governed by some noise size σ, as in $dY_t = \mu(Y_t)\,dt + \sigma v(Y_t)\,dB_t$, where B_t is a Brownian motion. The "quasi-static" regime is the one where σ is very close to 0. Then, things are much simpler to analyze, especially for non-linear processes, provided we accept $O(\sigma^2)$ approximations.

Indeed, consider that vector-valued process Y_t (for $t \leq T$)

$$X_t = \mathbb{E}_t[F(Y_T)] \tag{119}$$

where F is a C^2 function. Then, in the quasi-static regime, we can write

$$X_t = F(\mathbb{E}_t[Y_T]) + O\left(\sigma^2\right) \tag{120}$$

i.e. we swap $\mathbb{E}_t$ and F.[63] So, that, assuming now that Y_t is a martingale,

$$X_t = F(Y_t) + O\left(\sigma^2\right) \tag{121}$$

and

$$dX_t = F'(Y_t)\,dY_t + O\left(\sigma^2\right) \tag{122}$$

[63] We do not formally prove this, as this is purely mathematical as opposed to economic. One could presumably do it, e.g. using the Clark-Ocone formula from the Malliavin calculus.

Table 7: Bloomberg identifiers of countries included in the sovereign yield model.

Country	Government bond ticker ID	Country	Government bond ticker ID
Austria	G0063Z BLC2 Curncy	Ireland	G0062Z BLC2 Curncy
Belgium	G0006Z BLC2 Curncy	Italy	G0040Z BLC2 Curncy
Finland	G0081Z BLC2 Curncy	Netherlands	G0020Z BLC2 Curncy
France	G0014Z BLC2 Curncy	Portugal	G0084Z BLC2 Curncy
Germany	G0016Z BLC2 Curncy	Slovenia	G0259Z BLC2 Curncy
Greece	G0156Z BLC2 Curncy	Spain	G0061Z BLC2 Curncy

or, more informally (as we do in the economic part of this section),

$$dX_t \simeq F'(Y_t)\, dY_t. \tag{123}$$

To work out a concrete example, take $Y_t = \sigma B_t$,and $X_t = \mathbb{E}_t\left[e^{Y_T}\right]$. The exact values are:

$$X_t = e^{Y_t + \frac{\sigma^2}{2}(T-t)}, \qquad dX_t = X_t dY_t \tag{124}$$

and the quasi-static approximation gives

$$X_t = e^{Y_t} + O\left(\sigma^2\right), \qquad dX_t = e^{Y_t} dY_t + O\left(\sigma^2\right). \tag{125}$$

This is particularly useful when Y_t is multidimensional, as in Section 11.2.

11.4 Details on the data

Table 7 describes the tickers of the yields that we use in our empirical analysis.

12 Complements

12.1 Multi-dimensional actions

Suppose now that the action y_{it} is $q-$dimensional, for some $q \geq 1$. For instance, y_{it}'s components might be the growth rate, and the labor share of firms of firm i, and then $q = 2$. Then, the general GIV procedure extends well, as we shall now see.

We call $a \in \{1, \ldots, q\}$ (as in action) a component of y. The model is:

$$\begin{aligned} y_{S^a t} &= \sum_f \lambda^a_{S^a,f} F^f + u^a_{S^a t}, \\ F^f_t &= \eta^f_t + \sum_a \alpha^f_a y^a_{S^a,t}, \end{aligned}$$

Here u_{it} is q dimensional, α is $q \times r$ dimensional matrix, and λ is $r \times q$ dimensional matrix.

We can also estimate M (hence $\sum_f \alpha^f \lambda^f$), the α^f. For ε_t a composite of aggregate shocks,

$$y_{St} = H y_{St} + u_{St} + \varepsilon_t,$$

where

$$H = \Lambda A = \sum_f \alpha^f \lambda^f,$$

with $\Lambda_{af} = \lambda^a_{S^a,f}$ and $A_{fa} = \alpha^f_a$ matrices with dimensions $q \times r$ and $r \times q$ respectively, so that H is $q \times q$, and

$$u_{St} = \left(u^a_{S^a t}\right)_{a=1\ldots q}.$$

This implies

$$y_{St} = M\left(u_{St} + \varepsilon_t\right), \tag{126}$$

there the multiplier M is now a $q \times q$ matrix:

$$M = (I - H)^{-1}.$$

We will form a GIV:

$$z_t = u_{\Gamma t},$$

which is $q-$dimensional:

$$u_\Gamma = \left(u^a_{\Gamma^a}\right)_{a=1\ldots q}.$$

We want, with $E^a = S^a - \Gamma^a$,

$$\mathbb{E}\left[u_{Et} u'_{\Gamma t}\right] = 0$$

i.e., for all $Q^{ab} = 0$, where

$$Q^{ab} := \mathbb{E}\left[u^a_{E^a t} u^b_{\Gamma^b t}\right].$$

Let us focus on the case where u_{it}, u_{jt} are uncorrelated for $i \neq j$, but for a given i, u^a_{it}, u^b_{it} can be correlated (if a firm have a investment boom, it will likely hire more labor, so that the components of its idiosyncratic shock in $y_{it} \in \mathbb{R}^q$ will be correlated.

We have:

$$Q^{ab} = \sum_i E^a_i \Gamma^b_i v^{ab}_i, \qquad v^{ab}_i := \mathbb{E}\left[u^a_{it} u^b_{it}\right]. \tag{127}$$

For simplicity, we will suppose that that there are v^{ab} and σ^2_i such that

$$v^{ab}_i = \sigma^2_i v^{ab}. \tag{128}$$

Hence, we can simply take $E_i = \frac{k}{\sigma^2_i}$ with $k = \frac{1}{\sum_j 1/\sigma^2_j}$ and set, for all a, $E^a_i = E_i$ and $\Gamma^a = S^a - E^a$.

Then,

$$Q^{ab} = \sum_i \frac{k}{\sigma_i^2} \Gamma_i^b \sigma_i^2 v^{ab} = k v^{ab} \sum_i \Gamma_i^b = 0,$$

so that we have achieved our goal that $\mathbb{E}\left[u_{Et} u'_{\Gamma t}\right] = 0$. In the more general case, other Γ_i^a can probably be found.

Given (126), we have

$$y_{St} = M\left(u_{St} + \varepsilon_t\right) = M\left(u_{\Gamma t} + u_{Et} + \varepsilon_t\right),$$

so

$$\mathbb{E}\left[y_{St} z'_t\right] = M \mathbb{E}\left[z_t z'_t\right],$$

hence our estimator is

$$M = \mathbb{E}\left[y_{St} z'_t\right] \mathbb{E}\left[z_t z'_t\right]^{-1}. \tag{129}$$

Finally, we can also estimate $\alpha^f M$ by regressing on z_t:

$$F_t^f = \eta_t^f + \sum_a \alpha_a^f y_{S^a,t}^a = \eta_t^f + \alpha^f y_{St} = \eta_t^f + \alpha^f M\left(u_{\Gamma t} + u_{Et} + \varepsilon_t\right),$$

so $\beta^f = \alpha^f M$ (a row vector) obtains by simply regressing

$$F_t^f = \beta^f z_t + \varepsilon_t^f,$$

and get $\beta^f = \alpha^f M$, $\beta^f = \mathbb{E}\left[F_t^f z'_t\right] \mathbb{E}\left[z_t z'_t\right]^{-1}$.

Extension: causal estimation of the actor-specific multiplier The following is a refinement. We can also identify causally $\mu_i := \lambda_i \alpha = \sum_f \lambda_i^f \alpha^f$. Indeed, use

$$u_{\Gamma t,-i} := u_{\Gamma t} - S_i^u u_{it}, \tag{130}$$

which is the granular shock purged of a correlation with u_{it}. Then, a shock u_{St} creates an impact $\frac{dF_t}{du_{St}} = M\alpha$, hence an impact

$$\frac{dy_{it}}{du_{St}} = M \lambda_i \alpha.$$

Hence, we can identify μ_i, by regression

$$y_{it} = \mu_i M u_{\Gamma t,-i} + \phi^i \mathcal{C}_t + \varepsilon_{it}^y, \tag{131}$$

with some noise ε_{it}^y. This is the average impact of a causal impact of idiosyncratic shocks of the other entities on entity i.

12.2 Full recovery when different factors have different "size" weights

In the basic model, we can identify α^f, $M = \frac{1}{1-\sum_f \lambda^f \alpha^f}$, but not λ^f.

We give some conditions under which we can actually also identify the λ^f (in addition to α^f and M). We show here that this is the case if we assume that the size S^f differs across all factors f, and this knowledge is given to us (from a model).

Here we take the basic set up as in Section 3.1, in the simplified case where $\lambda_i^f = \lambda^f$ for all "endogenous" factors, i.e. for the factors f such that $\alpha^f \neq 0$, the other exogenous factors η all have an impact of 1:

$$y_{it} = u_{it} + \sum_f \lambda^f F_t^f + \eta_t^y, \tag{132}$$

$$F_t^f = \alpha^f y_{S^f,t} + \eta_t^f. \tag{133}$$

This implies

$$y_t = u_t + \iota \sum_f \lambda^f F_t^f + \iota \eta_t^y = u_t + \iota \sum_f \lambda^f \left(\eta_t^f + \alpha^f S^{f\prime} y_{it} \right) + \iota \eta_t^y.$$

Noting "ε^k" some combination of the various η's, and as usual $M = \frac{1}{1-\sum_f \alpha^f \lambda^f}$,

$$\begin{aligned} y_t &= \left(I - \iota \sum_f \lambda^f \alpha^f S^{f\prime} \right)^{-1} \left(u_t + \iota \varepsilon_t^1 \right) \\ &= \left(I + M \iota \sum_f \lambda^f \alpha^f S^{f\prime} \right) \left(u_t + \iota \varepsilon_t^1 \right) \\ y_t &= u_t + M \iota \sum_f \lambda^f \alpha^f u_{S^f,t} + \iota \varepsilon_t^y, \end{aligned} \tag{134}$$

i.e., since that $F_t^f = \eta_t^f + \alpha^f y_{S^f,t}$ this gives:

$$F_t^f = \alpha^f \left(u_{S^f,t} + M \sum_g \lambda^g \alpha^g u_{S^g,t} \right) + \varepsilon_t^f. \tag{135}$$

Hence, suppose that we extracted the $\check{u}_{it} = u_{it} - u_{Et}$ (following our usual procedure). Then, we form

$$z_{\Gamma^f t} := S^{f\prime} \check{u}_t = u_{S^f t} - u_{Et}. \tag{136}$$

Then, regressing F_t^f on the various $z_{\Gamma^g t}$

$$F_t^f = \sum_g b_g^f z_{\Gamma^g t} + \varepsilon_t^{f,1} \tag{137}$$

(for ε^{f1} some residual noise) yields a regression coefficient:

$$b_g^f = \alpha^f \left(1_{f=g} + M\lambda^g \alpha^g\right). \tag{138}$$

This allows to recover everything, and with several overidentifying restrictions. Indeed,

$$b^f := \sum_g b_g^f = \alpha^f \left(1 + M \sum_g \lambda^g \alpha^g\right) = \alpha^f M,$$

which identifies $\alpha^f M$. Next, for $f \neq g$,

$$\frac{b_g^f}{b^f} = \lambda^g \alpha^g,$$

which gives $\lambda^g \alpha^g$ (and should be equal for all f), hence M. Hence, we obtained $\alpha^f M$, M and $\lambda^g \alpha^g$ — hence all quantities: α^f, λ^f, M.

12.3 Complements to the general procedure

The procedure can be simplified in some cases. *When we have a long time-series.* Recall that

$$y_{St} = \sum_f \lambda_{St}^f F_t^f + u_{St}. \tag{139}$$

Hence, if all factors with λ_{St}^f possibly non-zero are observables and exogenous, we can measure the λ_{St}^f by OLS with the regression above, and get u_{St} to be the residual. This is useful when we have high-frequency data (e.g. daily financial returns), which can give an acceptably small error.[64]

We can aggregate entities into categories. For this discussion, we replace "entity" by "firm". We could aggregate the firms into $K > 1$ sub-categories (e.g. industries – or even an arbitrary categorization like "blue firms" and "red firms") — then the above works, but interpreting the partition i as "aggregate firm category i" rather than "firm i". Indeed, (49) aggregates without problem: if aggregate k is made of firm $i \in I_k$, we just define the aggregate size of category k as $S_{[k]} := \sum_{i \in I_k} S_i$, the relative weight of firm i in category k as $\omega_{[k]i} = \frac{S_i 1_{i \in I_k}}{S_{[k]}}$, and the action factor loading as value-weighted averages ($y_{[k],t} = \sum_i \omega_{[k]i,t} r_{it}$, $\alpha_{[k]}^f = \sum_i \omega_{[k]i} \alpha_i^f$). Then, the model works, using those aggregated categories. What we do need is that categories have non-trivial idiosyncratic shock (so that a "very small firms" category would not be valid, as it would have $var\left(u_{it}\right) \simeq 0$).

[64]Indeed, this time-series regressions gives an $O\left(\frac{1}{\sqrt{T}}\right)$ error, which is good enough for large T. Using the cross section, as in the basic procedure, gives an $O\left(\frac{1}{\sqrt{TN}}\right)$ error.

12.4 Typical size of Herfindahls

The GIV instrument is valid as long as $h > 0$, i.e. as long as there is heterogeneity. However, for it to be strong, we need high Herfindhals. In estimates for firms, we typically have $h \in [0.02, 0.5]$. One can have an a priori estimate of its size (for theory purposes). In practice, many size distributions follow a power law with fat tails, $\mathbb{P}(S > x) \sim kx^{-\zeta}$ for large x, with $\zeta \in (1, 2]$ — something also explained via random growth behavior. In that case one can show that (as in **?**, Proposition 2)

$$h \sim k' N^{-\psi}, \qquad \psi = 1 - \frac{1}{\zeta} \in (0, \frac{1}{2}] \tag{140}$$

for k' a non-zero random variable independent of N. The traditional variance case corresponds to $\zeta = 2$, which confirms $h \sim k' N^{-1/2}$ (and then k' is a constant), a very weak instrument. But when $\zeta \in (1, 2)$, we have a decay in $N^{-\psi}$ with $\psi \in \left(0, \frac{1}{2}\right)$. A fatter tail in the distribution of large firms (lower ζ) creates scaling in N that decays more slowly (ψ is lower) as N grows large. In the limit of Zipf's law (i.e., $\zeta \to 1$), we find $\psi \to 0$ (indeed, one can show that we have $h \sim \frac{k'}{\log N}$), a stronger instrument.

To simulate sizes from a power law distribution with exponent ζ, we can take $V_i = i^{-1/\zeta}$, and $S_i = \frac{V_i}{\sum_j V_j}$.[65] In the case of Zipf's law, that yields $h \sim \frac{\alpha}{\log n}$with $\alpha = \frac{\pi}{\sqrt{6}} \simeq 1.3$.

12.5 When we have disaggregated data for both the demand and the supply side

When we have disaggregated data for both the demand and the supply side, we can refine the "exclusion restriction". So far we assumed that $\mathbb{E}[u_{it}\varepsilon_t] = 0$, i.e. no covariance between idiosyncratic demand and supply shock. If that's not the case, we can also decompose each supply with a factor model:

$$y_{it}^k = \phi^k p_t + \lambda_i^k \eta_t^k + u_{it}^k, \tag{141}$$

for type $k = s, d$ for supply and demand. We allow $\mathbb{E}\left[u_{it}^s u_{it}^d\right]$ to be nonzero: for instance, if the US has a "fracking shock" that affects both supply and demand, it will be captured by both u_{it}^s and u_{it}^d for $i = \text{USA}$.

The price p_t adjusts so that supply equals demand, $y_{S^s t}^s = y_{S^d t}^d$ (where S_i^d (resp. S_i^s) is the average fraction of demand (resp. supply) accounted by country i), i.e.

$$p_t = \frac{u_{S^d}^d - u_{S^s}^s + \lambda_{S^d}^d \eta_t^d - \lambda_{S^s}^s \eta_t^s}{\phi_{S^s}^s - \phi_{S^d}^d} \tag{142}$$

[65] More refined, we can simulate n i.i.d. uniform variables U_i, order them $U_{(1)} \leq \cdots \leq U_{(n)}$, and take $V_i = U_{(i)}^{-1/\zeta}$.

Then, we have two GIVs, based on supply and demand respectively:

$$z_t^k := \Gamma^{k\prime} y_t^k = u_{\Gamma^k t}^k, \tag{143}$$

for $k = s, d$ (with $\Gamma^k = S^k - E^k$ in the basic case $\lambda^k = \iota$, and $\Gamma^k = Q^{\lambda^k} S^k$ in the general case). We can also form the difference:

$$z_t^{d-s} := z_t^d - z_t^s. \tag{144}$$

Now, assume $\mathbb{E}\left[u_{it}^k \eta_t^{k'}\right] = 0$ for k, k' in $\{s, d\}$. Then we have $\mathbb{E}\left[\left(y_{Et}^k - \phi^k p_t\right) z_t\right] = 0$ for z_t equal to either z_t^s or z_t^d, or some combination of them. The optimal instrument is $z_t = z_t^d - z_t^s$, as this is the most correlated with the price (142) (this generalizes the reasoning of Proposition 3). We can also use an overidentification test like in Section 2.6, based on the those two GIVs based on supply and demand.

If we assume only that $\mathbb{E}\left[z_t^k \eta_t^\ell\right] = 0$ for some (k, ℓ), we can identify ϕ^kvia $\mathbb{E}\left[\left(y_{Et}^k - \phi^k p_t\right) z_t^\ell\right] = 0$.

12.6 When the influence matrix is not proportional to size: When the loading on common shocks is not necessarily uniform

Here we complete our discussion in Section 10.2. We now study the more general case where:

$$y_t = \gamma G y_t + \Lambda \eta_t + u_t, \tag{145}$$

where the factor loading Λ (an $N \times r$ matrix) is not necessarily equal to ι (but we keep imposing that the Λ spans ι, i.e. there is a q such that $\iota = \Lambda q$). As before, η_t is a low-dimensional vector of factors. However, we do not assume anymore that $G\iota = \iota$.

First, we suppose that we have a first estimate of γ, which we call γ^e. We will later iterate on it. Then, we form:

$$\tilde{y}_t\left(\gamma^e\right) = \left(I - \gamma^e G\right) y_t. \tag{146}$$

If $\gamma^e = \gamma$, then

$$\tilde{y}_t\left(\gamma\right) = \Lambda \eta_t + u_t. \tag{147}$$

Hence, we run a factor analysis on $\tilde{y}_t\left(\gamma^e\right)$, which recovers Λ and $W = \left(V^u\right)^{-1}$. We introduce Q as in (176) so that $Q\Lambda = 0$ and set

$$\check{u}^e = Q \tilde{y}_t\left(\gamma^e\right),$$

so that at $\gamma^e = \gamma$,

$$\check{u}_t^e = Q u_t. \tag{148}$$

Then, we define (with $E = \frac{W\iota}{\iota' W \iota}$), with $S := G'E$,[66,67]

$$\begin{aligned} z_t &= S'Q\left(1-\gamma^e G\right)^{-1} \check{u}_t^e & (149)\\ &= S'Q\left(1-\gamma^e G\right)^{-1} Q\left(I-\gamma^e G\right) y_t. & (150)\end{aligned}$$

Our key moment is as before, equation (100):[68]

$$\mathbb{E}\left[z_t\left(y_{\tilde{E}t}\text{-}\gamma y_{St}\right)\right] = 0. \quad (152)$$

This yields an estimate of γ. Hence, we simply replace the definition of the GIV (99) by (150). In fact, we can show that when $\Lambda = \iota$, the estimator (150) is equal to the estimator (99). In that sense it is its natural generalization.[69]

We note that we have a fixed point: an initial γ^e gives an estimate of γ; that's then the new estimate γ^e, and we re-iterate the process, until convergence.

[66] A variant is:

$$z_t = S'\left(1-\gamma^e G\right)^{-1} \check{u}_t^e = S'\left(1-\gamma^e G\right)^{-1} Q\left(I-\gamma^e G\right) y_t.$$

The advantage of formulation (99) is that in the simple case of the previous subsection (with $\Lambda = \iota$), then it recovers the estimator of that subsection.

[67] The proof shows that this choice works. This particular choice is heuristically motivated by the analogy with (183) and (97), and the fact that when $\eta_t = 0$, $y_t = (I-\gamma G)^{-1} u_t$, so that $y_{St} = S' y_t = \gamma S' \left(1-\gamma^e G\right)^{-1} u_t$. Hence in some loose sense z_t is a good idiosyncratic-based approximation of y_{St}.

[68] Here is the proof. At the right estimator $\gamma = \gamma^e$,

$$z_t = S'Q\left(1-\gamma^e G\right)^{-1} Q u_t = c'Qu_t, \quad (151)$$

with $c' := S'Q\left(1-\gamma^e G\right)^{-1}$. We also have

$$y_{\tilde{E}t} - \gamma y_{St} = E'\left(I-\gamma G\right) y_t = E'\left(\Lambda \eta_t + u_t\right).$$

This implies that

$$\mathbb{E}\left[\left(y_{\tilde{E}t} - \gamma y_{St}\right) z_t\right] \;=\; \mathbb{E}\left[E' u_t u_t' Q' c\right] = E' V Q' c = \frac{1}{\iota' W \iota} \iota' Q' c = 0,$$

as $Q\iota = 0$.

[69] In the case $\Lambda = \iota$, then $Q = I - \iota E'$, so

$$E'GQ = E'G\left(I - \iota E'\right) = E'G - E' = E'\left(G - I\right),$$

so that the estimator in (150) can be written:

$$z_t = E'\left(G-I\right) y_t = E'GQy_t.$$

On the other hand,

$$\begin{aligned} A &:= GQ\left(1-\gamma^e G\right)^{-1} Q\left(I-\gamma^e G\right) = GQ\left(1-\gamma^e G\right)^{-1}\left(I-\iota E'\right)\left(I-\gamma^e G\right)\\ &= GQ\left(1-\gamma^e G\right)^{-1}\left(I-\gamma^e G\right) - GQ\left(1-\gamma^e G\right)^{-1}\iota E'\left(I-\gamma^e G\right)\\ &= GQ - 0 \text{ as } G\iota - \iota \text{ and } Q\iota = 0\\ &= GQ,\end{aligned}$$

Suppose that instead we use[70]

$$Z_t = S' (1 - \gamma^e G)^{-1} \check{u}_t^e. \tag{153}$$

Suppose that $\gamma^e = \gamma$. Then we can write:[71]

$$y_{St} = Z_t + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t Z_t] = 0, \tag{154}$$

for $\varepsilon_t = S' (I - \gamma^e G)^{-1} ((I - Q) u_t + \Lambda \eta_t)$. Also, we have:

$$y_{Et} = \gamma y_{St} + E' \Lambda \eta_t + u_{Et}.$$

Hence, we can estimate γ by OLS:

$$y_{Et} = \gamma Z_t + \beta' \eta_t + \varepsilon_t^{y_E}. \tag{155}$$

This consistently estimates γ.

Calling z_t the GIV in the basic models (with $G = \iota' S$), $Z_t = \frac{z_t}{1-\gamma^e}$.

12.7 Identification of social interactions and the reflection problem

There seems to be a contradiction between Section 6.2's finding that we do achieve identification, and **?**'s Proposition 2 and **?**'s Proposition 1, which seem also to state the impossibility of identification. **?** analyze social interactions of the type:

$$\boldsymbol{y}_t = \beta \boldsymbol{G} \boldsymbol{y}_t + \gamma \boldsymbol{x}_t + y \boldsymbol{G} \boldsymbol{x}_t + \boldsymbol{\varepsilon}_t \tag{156}$$

so that the estimator in (150) can be written:

$$z_t = E' GQ (1 - \gamma^e G)^{-1} Q (I - \gamma^e G) y_t = E' A y_t = E' GQ y_t = E' (G - I) y_t.$$

Hence, when $\Lambda = \iota$, the estimators in (99) and (150) are identical.

[70]Note that the γ^e in the definition of Z_t need not be the same γ^e used above to construct the $\check{u}_t^e$; i.e., we could have $Z_t = S' \left(1 - \gamma^{e,2} G\right)^{-1} \check{u}_t^e$ for some other $\gamma^{e,2}$. There is still a fixed point though, and in the limit the estimated γ in (155) should also be equal to the γ^e and $\gamma^{e,2}$.

[71]Here is the proof. We saw that $\check{u}_t^e = Q u_t$, so,

$$Z_t = S' (I - \gamma^e G)^{-1} Q u_t,$$

while, with $a' := S' (I - \gamma^e G)^{-1}$,

$$y_{St} = S' (I - \gamma^e G)^{-1} (Q u_t + (I - Q) u_t + \Lambda \eta_t) = Z_t + a' (I - Q) u_t + a' \Lambda \eta_t.$$

From (177), we have $\mathbb{E}[a' (I - Q) u_t z_t] = 0$. Hence we have $\mathbb{E}[\varepsilon_t Z_t] = 0$.

with $\mathbb{E}[\boldsymbol{\varepsilon}_t|\boldsymbol{x}_t] = 0$. In their main result, they conclude that if the matrices I, G, G^2 are not linearly independent, then the system is not identified. However, in our setup $G = \iota S'$ (where ι is a vector of 1's) so that $G^2 = G$ and we satisfy **?**'s condition that seems to guarantee the impossibility of identification. However, we can identify the parameters, as we saw in Section 6.2. How do we solve that seeming contradiction?

The short answer is that **?** and **?** do not consider anything like a GIV, as they immediately reason on averages based on observables, eschewing any exploration of the noise. In contrast, GIVs are all about exploring some structure in the noise — the idiosyncratic shocks of large entities. For instance Manski considers something akin to:

$$\mathbb{E}[\boldsymbol{y}_t|\boldsymbol{x}_t] = \beta \boldsymbol{G} \mathbb{E}[\boldsymbol{y}_t|\boldsymbol{x}_t] + \gamma \boldsymbol{x}_t + y \boldsymbol{G} \boldsymbol{x}_t, \tag{157}$$

where all the noise has been averaged out.

Indeed, we do impose some structure, namely:

$$\varepsilon_{it} = \eta_t + u_{it}, \qquad u_{it} \text{ i.i.d., orthogonal to } \eta_t, \tag{158}$$

and that was helpful to derive the GIV estimator (Section 6.2).

It would be interesting to show weaker conditions, or even necessary and sufficient conditions. We leave a full treatment of that to future research. Still, we offer a few remarks with more general sufficient conditions for identification via GIV.

We can generalize the noise condition (158) (while staying with our setup $G = \iota S'$) to the more general condition:

$$\varepsilon_{it} = \lambda_i \eta_t + \sigma_i v_{it}, \tag{159}$$

where λ_i are scalar and η_t, v_{it} all uncorrelated (including across $i's$). More generally, a "low rank" representation where $\eta_t \in \mathbb{R}^k$ with a low k is admissible too.[72]

Second, we can generalize to the case where $\boldsymbol{G}^2 = \boldsymbol{G}$ (the case where G^2 is a linear combination of G and I is similar[73]), which seems to leads to the impossibility of identification in **?**. This is formalized here (and proved in Section 13, with a constructive identification procedure).

Proposition 5 (Identification achieved in the **?** setup). *Suppose that $G^2 = G$, which is satisfied in our basic setup, but leads to the impossibility of identification in the **?** setup without further assumptions. Suppose also the "simple noise structure" assumption (158). Suppose also the existence of two n-dimensional vectors S and Γ satisfying*

$$G'S = S, \qquad G'\Gamma = 0, \qquad \iota' S \neq 0, \qquad \Gamma' S \neq 0. \tag{160}$$

[72] Informally, this generates $2n$ unknowns (λ_i, σ_{it}), while the variance-covariance matrix has dimensions $\frac{n(n-1)}{2}$.

[73] It can be reduced to that case by rescaling $H = b_0 + b_1 G$ with the right coefficient, with $H^2 = H$.

Then GIV is possible in that setup, i.e. with the GIV $z_t = \Gamma' y_t$, we can identify the coefficients (β, γ, y) (and indeed β, as it was assumed that $\gamma = y = 0$).

In our basic setup, we had S_i the relative sizes, and $G = \iota S'$, $\Gamma = S - \frac{\iota}{N}$. Hence (160) is an abstract generalization of our concrete conditions.

Hence, in many situations of interest we can be quite confident that condition (160) is satisfied.[74]

In conclusion: our GIV approach gives some renewed hope for identification in the context of social influence and reflection problems. Indeed, it provides a way to achieve identification where it seemed impossible. Informally, this is by exploiting the idiosyncratic noise of "large players". Formally, and less intuitively, it is by exploiting a little bit of structure in the noise (so that there is a low-dimensional common noise). Future research might profitably firm up the exact necessary and sufficient conditions for this.

12.8 When only some shocks are kept in the GIV

If we truncate the residuals, i.e. use

$$z_t = \sum_i \tau\left(S_i\left(u_{it} - u_{Et}\right)\right)$$

for the hard thresholding function

$$\tau\left(x\right) = x 1_{|x| \geq b}$$

for some $b > 0$, then everything works too. Indeed, we have that $\check{u}_{it} := u_{it} - u_{Et}$ is orthogonal to u_{Et}. Let us assume that it is independent. In our basic example of Section 2.2, we still have $\mathbb{E}\left[\left(p_t - \alpha y_{St}\right) z_t\right] = 0$, so that the IV procedure (16) still works. Likewise, in the more complex supply and demand case, the IV relations (34) and (36) still hold.

Furthermore, the OLS estimates still hold. The key is that we can write:

$$u_{\Gamma t} = z_t + z_t^{<},$$

where $\tau^{<}\left(x\right) = x 1_{|x| < b}$, and $z_t^{<} = \sum_i \tau^{<}\left(S_i \check{u}_{it}\right)$, so that $z_t \perp z_t^{<}$. Hence, regressing $u_{\Gamma t}$ on this truncated z_t gives a coefficient of 1, and all the analysis goes through.

12.9 When the researcher assumes too much homogeneity

Take the supply and demand example, and imagine that the econometrician assumes a homogeneous elasticity of demand ϕ^d, even though there are in fact heterogeneous elasticities ϕ_i^d. What happens then?

[74] As $G^2 = G$, one can always find vectors Γ, S satisfying the first 3 conditions (provided n is big enough and G is not the identity nor 0), and the last one is rather "generically" easy to satisfy.

The model (26)-(27) becomes, for the demand:

$$y_{it} = \phi_i^d p_t + \lambda_i \eta_t + u_{it},$$

and for the supply

$$s_t = \phi^s p_t + \varepsilon_t.$$

As supply equals demand, $y_{St} = s_t$, which gives the price

$$p_t = \frac{u_{St} + \lambda_S \eta_t - \varepsilon_t}{\phi^s - \phi_S^d}. \tag{161}$$

In this thought experiment, the econometrician assumes $\phi_i^d = \phi_i$. He runs a panel model for $y_{it} - y_{Et}$, and we assume that it's large enough that he can extract η_t, successfully.[75] The GIV (we use the notation Z_t rather than z_t to denote the GIV before controls by η_t) is then

$$Z_t := y_{\Gamma t} = \phi_\Gamma^d p_t + \lambda_\Gamma \eta_t + u_{\Gamma t} = \left(1 + \frac{\phi_\Gamma^d}{\phi^s - \phi_S^d}\right) u_{\Gamma t} + \lambda^Z \tilde{\eta}_t = \frac{1}{\psi} u_{\Gamma t} + \lambda^Z \tilde{\eta}_t,$$

so

$$Z_t = \frac{1}{\psi} u_{\Gamma t} + \lambda^Z \tilde{\eta}_t, \qquad \frac{1}{\psi} = \frac{\phi^s - \phi_E^d}{\phi^s - \phi_S^d}, \tag{162}$$

where $\frac{1}{\psi} = 1$ in the common-elasticity case, $\tilde{\eta}_t = (\eta_t, \varepsilon_t, u_{Et})$ gathers the common shocks, and λ^Z is a vector of loadings.

Hence, when we run the first stage

$$p_t = b^p Z_t + \beta^p \eta_t + \varepsilon_t^p,$$

we will gather

$$b^p = \frac{1}{\phi^s - \phi_E^d}.$$

If we run

$$s_t = b^s Z_t + \beta^s \eta_t + \varepsilon_t^s,$$

we will estimate

$$b^s = \frac{\phi^s}{\phi^s - \phi_E^d}.$$

The ratio of the two coefficients still gives ϕ^s. Likewise, the IV on the elasticity of demand will give ϕ_E^d.

In the polar opposite case where η_t cannot be estimated or controlled for, then the simple procedure becomes biased, however, as (162) shows. To fix it, one can estimate the model with

[75] One of the factors, formally, will be p_t. We assume that it is not included in the vector of factors η_t.

non-parametric coefficients (Section 10.1).

12.10 Simple GIV example without feedback loop: estimation by OLS

We go back to the example of Section 2.2.1,

$$y_{it} = \lambda_i \eta_t + u_{it}, \qquad p_t = \alpha y_{St} + \varepsilon_t. \tag{163}$$

and derive the OLS estimation via GIV,

$$z_t = y_{\Gamma t} = u_{\Gamma t}.$$

OLS estimator version One can use a plain OLS estimator with our instrument $u_{\Gamma t}$.[76] Assume for simplicity that the u_{it}'s are uncorrelated and have the same variance (Section 28.17 shows how to relax these assumptions). Given that

$$p_t = \alpha u_{St} + \alpha \eta_t + \varepsilon_t,$$

we can simply estimate α by OLS:

$$p_t = \alpha z_t + \varepsilon_t^0, \tag{164}$$

with $\varepsilon_t^0 := \alpha \eta_t + \varepsilon_t + \alpha u_{Et}$ (indeed, we have $\mathbb{E}\left[u_{\Gamma t} u_{Et}\right] = 0$). Call $\alpha^{G,OLS0}$ that estimator.

An enriched OLS estimator version We can improve the efficiency of the OLS estimator. We assume $\mathbb{E}\left[u_{it}\eta_t\right] = 0$ for all i and t. We use

$$\eta_t^e = y_{Et}$$

as an estimator of η_t. We have $\eta_t^e = \eta_t + u_{Et}$, and $u_{Et} = O\left(\frac{1}{\sqrt{N}}\right)$ can be expected to be small. Hence, regress:

$$p_t = \hat{\alpha} u_{\Gamma t} + \hat{\beta} \eta_t^e + \varepsilon_t^p, \tag{165}$$

which will yield in the limit of large T a consistent estimate of α. [77] In addition, it turns out that it is a more precise estimator (when N is large enough), as we control for η_t.

Precision of GIV estimators The next proposition states when we have a precise estimator.

[76] In estimator 17, the denominator could be of either sign, and close to 0 in finite sample, leading to some instability in the estimate.

[77] However, we will discard $\hat{\beta}$ as it does not estimate β : it is polluted by the correlation with measurement error. Indeed, $\beta = \alpha + b + O\left(\frac{1}{N}\right)$ with a bias $b = \frac{\mathbb{E}[\varepsilon_t \eta_t]}{\mathbb{E}[\eta_t^2]}$

Proposition 6 (Precision of the GIV estimators in this example). *We assume that the u_i are uncorrelated but possibly heteroskedastic. The above estimators based on the granular instrument variable (GIV) δ_Γ all achieve identification of the reaction parameter α. The GIV and two G-OLS estimators have standard errors*

$$\sqrt{T}\left(\hat{\alpha}_T^q - \alpha\right) \sim \mathcal{N}\left(0, \sigma_{\alpha^q}^2\right),$$

where the standard error of estimator q (equal to GIV, G-OLS0 or G-OLS1) has the form

$$\sigma_{\alpha^q} = \frac{\sigma_{\varepsilon(q)}}{\sigma_{u_\Gamma}} \tag{166}$$

with[78]

$$\sigma_{\varepsilon(G-OLS0)}^2 = var\left(\varepsilon + \alpha\eta + \alpha u_E\right), \qquad \sigma_{\varepsilon(IV)}^2 = var\left(\varepsilon\right), \qquad \sigma_{\varepsilon(G-OLS1)}^2 := var\left(\varepsilon^\perp\right)$$

We always have $\sigma_{\alpha^{G-OLS}} > \sigma_{\alpha^{IV}}$, and, if $corr\left(\varepsilon, \eta\right) \neq 0$ and N is large enough, we have $\sigma_{\alpha^{IV}} > \sigma_{\alpha^{G-OLS0}}$.

When the u_i's are IID, we have

$$\sigma_{u_\Gamma} = h\sigma_u. \tag{167}$$

So to have a precise estimate (low σ_α) we need: a few large firms (to have a large excess herfindahl h), and that demand shocks do affect the price importantly, compared to aggregate shocks (large $\sigma_u/\sigma_\varepsilon$).

12.10.1 Proofs

Simple OLS estimator The proof is very similar as in the basic case – and indeed, it is the standard proof of the standard error of the OLS regression (164), $p_t = \alpha z_t + \varepsilon_t^0$, gives $\sigma_{\alpha^{G-OLS0}} = \frac{\sigma_\varepsilon}{\sigma_z}$. We use $z_t = u_{\Gamma t}$ and

$$\sigma_{u_\Gamma} = h\sigma_u. \tag{168}$$

Enriched OLS estimator We have

$$p_t = \alpha u_{St} + \alpha\eta_t + \varepsilon_t = \alpha u_{\Gamma t} + \alpha\left(\eta_t + u_{Et}\right) + \left(\varepsilon_t - \alpha u_{Et}\right).$$

Now, project $\varepsilon_t - \alpha u_{Et}$ on $\eta_t^e := \eta_t + u_{Et}$ and call $\varepsilon_t^\perp$ the residual from projecting $\varepsilon_t - \alpha u_{Et}$ on $\eta_t + u_{Et}$

$$\varepsilon_t - \alpha u_{Et} = \gamma\eta_t^e + \varepsilon_t^\perp.$$

[78] Here $\varepsilon^\perp$ is the residual from projecting $\varepsilon - \alpha u_E$ on $\eta + u_E$. It is detailed in the proof.

We have

$$p_t = \alpha u_{\Gamma t} + (\alpha + \gamma)\, \eta_t^e + \varepsilon_t^{\perp}. \tag{169}$$

Note also that $u_{\Gamma t}$, η_t^e and $\varepsilon_t^{\perp}$ are all uncorrelated. Hence, the OLS formula for standard errors apply: $\sigma_{\alpha^{G-OLS1}} = \frac{\sigma_{\varepsilon^{\perp}}}{\sigma_{u_\Gamma}}$, and we use $\sigma_{u_\Gamma} = h\sigma_u$.

Ranking of precisions We have $\sigma^2_{\varepsilon(G-OLS1)} := var\left(\varepsilon_t^{\perp}\right)$. As $var\left(u_{Et}\right) = \frac{\sigma^2}{N}$, in the large N limit, $\sigma^2_{\varepsilon(G-OLS1)} \to var\left(e_t^{\perp}\right)$, where $e_t^{\perp}$ is the residual from projecting ε_t on η_t ($e_t^{\perp} = \varepsilon_t - \frac{cov(\varepsilon_t,\eta_t)}{var\eta}\eta_t$). Its variance is less than $var\left(\varepsilon_t\right)$, unless ε_t and η_t are uncorrelated.

Proof of Proposition 7 The proof is very much like that of Proposition 2. We have,

$$\hat{\lambda_T} - \lambda = \frac{\frac{1}{T}\sum_t y_{Et} z_t}{\frac{1}{T}\sum_t p_t z_t} - \lambda = \frac{\mathbb{E}_T\left[\left(\lambda p_t + \eta_t + u_{Et}\right) u_{\Gamma t}\right]}{\mathbb{E}_T\left[p_t u_{\Gamma t}\right]} - \lambda = \frac{\mathbb{E}_T\left[\left(\eta_t + u_{Et}\right) u_{\Gamma t}\right]}{\mathbb{E}_T\left[p_t u_{\Gamma t}\right]} = \frac{A_T}{D_T}. \tag{170}$$

Next,

$$D_T = \mathbb{E}_T\left[p_t u_{\Gamma t}\right] \to^{a.s.} D,$$

where, calling $\xi = \frac{1}{\mu - \lambda}$

$$D = \mathbb{E}\left[p_t u_{\Gamma t}\right] = \mathbb{E}\left[\left(\frac{u_{St} + \eta_t - \varepsilon_t}{\mu - \lambda}\right) u_{\Gamma t}\right] = \xi \mathbb{E}\left[u_{St} u_{\Gamma t}\right] = \xi h^2 \sigma_u^2.$$

For the numerator, in the limit of large sample sizes $T \to \infty$,

$$T^{1/2} A_T \to^d \mathcal{N}\left(0, \sigma_A^2\right),$$

where

$$\sigma_A^2 = \mathbb{E}\left[\left(\eta_t + u_{Et}\right)^2 u_{\Gamma t}^2\right] = \mathbb{E}\left[\left(\eta_t + u_{Et}\right)^2\right] \mathbb{E}\left[u_{\Gamma t}^2\right] = \sigma_{\eta + u_E} h^2 \sigma_u^2,$$

where $\sigma^2_{\eta + u_E} = \sigma_\eta^2 + \frac{\sigma_u^2}{N}$.

$$\frac{\sigma_A}{D} = \frac{\sigma_{\eta + u_E} h \sigma_u}{\xi h^2 \sigma_u^2} = \frac{\sigma_{\eta + u_E}}{\xi h \sigma_u} =: \sigma_\lambda.$$

Hence,

$$\sqrt{T}\left(\hat{\lambda_T} - \lambda\right) \sim N\left(0, \sigma_\lambda^2\right).$$

The proof for μ is exactly along the same lines, as $\mu_T^e - \mu = \frac{\mathbb{E}_T[\varepsilon_t z_t]}{\mathbb{E}_T[p_t z_t]}$.

12.11 Link between our initial examples and the general framework

Let us make the link between our initial examples and the general setup of Section 3.1.

The basic example of Section 2.2 was:

$$y_{it} = \eta_t + u_{it}, \qquad p_t = \alpha y_{St} + \varepsilon_t.$$

The factors are p_t and η_t:

$$F_t^1 = p_t, \quad \alpha^1 = \alpha, \quad \lambda^1 = 0, \quad \eta_t^1 = \varepsilon_t,$$

$$F_t^2 = \eta_t, \quad \alpha^2 = 0, \quad \lambda^2 = 1, \quad \eta_t^2 = \eta_t.$$

Factor F_t^1 is endogenous, factor F_t^2 is exogenous. They are both observable (via $y_{Et} = \eta_t + O\left(\frac{1}{\sqrt{N}}\right)$). So the set of controls $\mathcal{C}_t$ is $\mathcal{C}_t = \{\eta_t\}$. The multiplier is $M = 1$.

Let us next make the link with supply-and-demand example of Section 2.3, which was:

$$y_{it} = \phi^d p_t + \eta_t + u_{it}, \qquad p_t = \frac{y_{St}}{\phi^s} - \frac{\varepsilon_t}{\phi^s}.$$

The factors are also p_t and η_t:

$$F_t^1 = p_t, \quad \alpha^1 = \frac{1}{\phi^s}, \quad \lambda^1 = \phi^d, \quad \eta_t^1 = -\frac{\varepsilon_t}{\phi^s},$$

$$F_t^2 = \eta_t, \quad \alpha^2 = 0, \quad \lambda^2 = 1, \quad \eta_t^2 = \eta_t.$$

Here F_t^1 is exogenous and hidden, while F_t^2 is endogenous and observable. So, $\mathcal{C}_t$ is empty.[79] The multiplier is $M = \frac{1}{1-\alpha^1\lambda^1} = \frac{\phi^s}{\phi^s - \phi^d}$, as was estimated in (39).

12.12 Identification of the TFP to GDP multiplier in a production network economy

Suppose a two-period model with a production network, as in **?????**. There are both idiosyncratic TFP shocks $\hat{\Lambda}_{it}$ and a government reform that creates correlated shocks η_t to TFP and change in labor supply $\hat{L}_t$. Utility is $C_t - e^{\eta_t^L} L_t^{1+1/\phi}$, so that ϕ is the Frisch elasticity of labor supply. So, as $C_t = \Lambda_t L_t$, labor supply is $\hat{L}_t = \phi\left(\hat{\Lambda}_t - \eta_t^L\right)$,[80] and GDP is $\hat{Y}_t = \hat{L}_t + \hat{\Lambda}_t$, i.e.

$$\hat{Y}_t = m\hat{\Lambda}_t - \phi\eta_t^L, \qquad m = 1 + \phi \tag{171}$$

We seek to find the "GDP multiplier" $m = 1 + \phi$, so that a TFP of 1 percent translates into a GDP increase of m percent.[81]

[79] This is why we could do only plain OLS regression in (39), without any controls like y_{Et}, unlike in the very simple initial example of Section 2.2

[80] The problem is $\max_{L_t} \Lambda_t L_t - e^{\eta_t^L} L_t^{1+1/\phi}$, which leads to $\left(1 + \frac{1}{\phi}\right) L_t^{1/\phi} = \Lambda_t e^{-\eta_t^L}$,hence the announced expression.

[81] If more than one factor change, m has the broader interpretation of a multiplier between TFP and GDP.

This is potentially a complicated problem, as for instance, in the **?** case, outputs are $\hat{y}_t = (I - A)^{-1} \hat{\Lambda} + \hat{L}$, where A is a the input-output matrix, so that output changes are correlated in complicated ways. However, one can sidestep using this disaggregated production data. We assume that TFP change in industry i is:

$$\hat{\Lambda}_{it} = \lambda_i \eta_t^{\Lambda} + u_{it}. \tag{172}$$

In the neoclassical equilibrium, TFP follows Hulten's theorem, so is $\hat{\Lambda}_t = \sum_i s_i \hat{\Lambda}_{it}$where s_i is the Domar weight (sales of industry i over GDP).

In the simplest case, we assume that industry-level productivities are available, and we get the residuals u_{it}^e. But the same procedure works (with less efficiency) if our data is made of proxies for productivities $\tilde{\hat{\Lambda}}_{it}$ growth (where the tilde indicates that we deal with a proxy). An example could be growth of sales per employee, or even the growth rate of sales. We assume a factor model

$$\tilde{\hat{\Lambda}}_{it} = \tilde{\lambda}_i \tilde{\eta}_t^{\Lambda} + \tilde{u}_{it}. \tag{173}$$

The proxy is of better quality when the proxy's idiosyncratic shock $\tilde{u}_{it}$ has a high correlation with the true idiosyncratic shock u_{it}.

Then, we extract the $\tilde{u}_{it}^e$ from a factor model, form $z_t = \tilde{u}_{St}^e - \tilde{u}_{Et}^e$ (with $S_i = \frac{s_i}{\sum_j s_j}$), and use the moment $\mathbb{E}\left[\left(\hat{Y}_t - m\hat{\Lambda}_t\right) z_t\right] = 0$, which identifies the TFP to GDP multiplier m. Using more general models (e.g. taking into account imperfections as in **?**) would be very interesting, but would be a new paper by itself. Indeed, even in that case z_t is likely to be a useful instrument, even though it won't be the optimal one. In any case, those examples show how GIV, with some economic reasoning, translate to more complex economies where aggregate shocks can be made of idiosyncratic shocks.

13 Proofs omitted in the paper

13.1 Proof of Proposition 4

13.1.1 Parametric identification

We start with the parametric case, deferring the semi-parametric case.

The solution is, with $\lambda_S = S'\Lambda$ a $1 \times r$ vector, $M = \frac{1}{1-\lambda_S \alpha}$,

$$y_{St} = M \left(u_{St} + \lambda_S \eta_t + C_{St} m\right), \tag{174}$$

$$y_t = u_t + \Lambda \left[\alpha M \left(u_{St} + \lambda_S \eta_t + C_{St} m\right) + \eta_t\right] + C_t m. \tag{175}$$

We take the parametric case (the semi-parametric case will then an easy corollary). This is, we have some characteristics x_{it} of actors (e.g. countries or and firms), and a priori knowledge that

$\lambda_{it} = X_{it} R$ for some $r-$dimensional vector $X_{it} = (1, x_{it})$ where x_{it} is a $r-1$ dimensional vector, and R is a $r \times r$ matrix. By rotation-invariance of the η_t (which is an r dimensional vector), we can take the case where $R = I$. Hence, in that sense we know the loadings $\lambda_{it} = x_{it}$ – but don't know the variance-covariance matrix V^{η} of the η_t.

Given a symmetric matrix W of size $N \times N$ (which, later, will optimally be $W = (V^u)^{-1}$, but we don't use that here yet) we define another $N \times N$ matrix:[82]

$$Q^{\Lambda,W} = I - \Lambda \left(\Lambda' W \Lambda\right)^{-1} \Lambda' W, \tag{176}$$

so that $Q = Q^{\Lambda,W}$ satisfies:

$$Q\Lambda = 0, \qquad Q' W \Lambda = 0, \qquad (I - Q) W^{-1} Q' = 0, \qquad Q^2 = Q. \tag{177}$$

Roughly, Q is the projection on the space orthogonal to the Λ, but with a scalar product that depends on W. Hence, (175) implies:

$$Qy_t = Qu_t + QC_t m. \tag{178}$$

Defining, for a vector Y_t,

$$\check{Y}_t := QY_t, \tag{179}$$

we have

$$\check{y}_t = \check{u}_t + \check{C}_t m. \tag{180}$$

The controls C_t^k are all assumed to have non-zero cross-sectional variation: this is what allows to identify their m. A variable that's an "aggregate control" without cross-sectional variation (e.g. a time fixed effect, or maybe the world price of oil if we study the macroeconomics of a small country not affecting it) will be classified as an F_t^f – it's in $\mathcal{F}^{\mathrm{Exo},O}$, the set of observable, exogenous factors.

13.1.2 Estimating multipliers α^f, M by GIV

We assume that we have identified V (up to a multiplicative factor), either because we know for instance that $V^u = \sigma_u^2 I$, or because of the material in Section 13.1.4.

We treat now the more GIV-specific topic of how to estimate the α^f and M. We set $Q = Q^{\Lambda, W^u}$ as in (176). Then, (178) gave

$$Qy_t = Qu_t + QC_t m. \tag{181}$$

Let us define

$$\Gamma := Q' S, \tag{182}$$

[82]For instance, in our basic example with uniform loading $\Lambda = \iota$, $Q = I - \iota E'$, where $E = \frac{W\iota}{\iota' W \iota}$.

and define the GIV to be $z_{\Gamma t} := \Gamma' (y_t - C_t m)$, which gives:

$$z_{\Gamma t} \quad := \quad \Gamma' (y_t - C_t m) = \Gamma' u_t = u_{\Gamma t}. \tag{183}$$

This relation means that we identify $u_{\Gamma t}$ exactly, even though we estimate the η_t with errors.[83]

The GIV is possible if and only if

$$\Gamma \neq 0. \tag{184}$$

This is exactly what motivated Assumption 1 mentioned in Proposition 4.[84]

We define $E = S - \Gamma = (I - Q') S$, so that

$$u_{St} = u_{\Gamma t} + u_{Et},$$

then (177) implies that $\mathbb{E}[u_{Et} u_{\Gamma t}] = S' (I - Q) V^u Q' S = 0$, so that we have the relation:[85],[86]

$$\mathbb{E}[u_{Et} u_{\Gamma t}] = 0. \tag{187}$$

Hence, (174) reads:

$$y_{St} = M (u_{St} + \lambda_S \eta_t + C_{St} m) = (u_{\Gamma t} + u_{Et} + \lambda_S \eta_t + C_{St} m) = M z_t + \varepsilon_t^{yS} \tag{188}$$

for $\varepsilon_t^{yS} = M (u_{Et} + \lambda_S \eta_t)$ uncorrelated with z_t. Hence, M will be consistently estimated by regressing y_{St}:

$$\mathbb{E}[(y_{St} - M z_t) z_t] = 0. \tag{189}$$

Likewise, for f an observable, endogenous control (i.e., one such that α^f need not be 0 a priori), we can regress:

$$F_t^f = \alpha^f M z_t + \varepsilon_t^{Ff}$$

with $\varepsilon_t^F = \eta_t + \alpha M \varepsilon_t^{yS}$ so that we can estimate αM consistently by the OLS regression of F_t on z_t.

[83]This may be surprising, but consider the following simple case to see how this is true: if $y_{it} = \lambda p_t + \eta_t + u_{it}$, then as $\Gamma' \iota = 0$, $y_{\Gamma t} = u_{\Gamma t}$. So we perfectly measure the $u_{\Gamma t}$. This relation with the Q generalizes that simple example with more complex factors.

[84]If $V^u S$ was spanned by the Λ, we could write $S = W \Lambda b$ for some vector b, and we'd have $Q' S = 0$, by (177). Conversely, if S is not spanned by the $V^u S$, then it is easy to check that $\Gamma \neq 0$.

[85]In addition, the value

$$z_{Ct} := C_{St} m \tag{185}$$

is also a valid instrument. Hence, the following is an instrument:

$$z_t = z_{\Gamma t} + z_{Ct}. \tag{186}$$

In this paper we mostly use $z_{\Gamma t}$ as an instrument though, to insist on what is GIV-specific.

[86]As always in this paper, this relation leads to clean relations with finite N, but it relies on doing "generalized least squares" with the proper weight matrix $W = (V^u)^{-1}$. In the general case, $\mathbb{E}[u_{Et} u_{\Gamma t}] = O\left(\frac{1}{\sqrt{N}}\right)$, so that this relation is likely to be approximately true in most cases of interest.

This shows identification even when we do not control for estimated factors η_t^e. To gain statistical power, it is useful to control for estimated η_t^e. We go on to this topic now.

13.1.3 Controlling for other estimated exogenous factors η_t^e

If we have a cross-sectionally important factor η_t^f, we may want to control for it to gain statistical precision. To do so, we first want to extract the factor. For notational simplicity, we assume that we removed all the controls $C_t m$ (i.e., we replace y_t by $y_t - C_t m$).

We define the $r \times N$ matrix[87]

$$L_t \coloneqq \left(\Lambda_t' W \Lambda_t\right)^{-1} \Lambda_t' W, \tag{190}$$

so that

$$L_t \Lambda_t = I_r. \tag{191}$$

Next, using $Q\Lambda_t = 0$ in the factor structure (54) gives:

$$(I - Q)\, y_t = \Lambda_t F_t + (I - Q)\, u_t.$$

Premultiplying this by L_t gives:

$$L_t (I - Q)\, y_t = F_t + L_t (I - Q)\, u_t.$$

Hence, an estimate of the F_t is

$$F_t^e \coloneqq L_t (I - Q)\, y_t. \tag{192}$$

Indeed, we will have

$$F_t^e = F_t + \varepsilon_t^{F^e},$$

where $\varepsilon_t^{F^e} = -L_t (I - Q)\, u_t$ is a small error. In addition, this error is orthogonal to $z_{\Gamma t}$:[88]

$$\mathbb{E}\left[F_t^e z_{\Gamma t}\right] = 0, \tag{193}$$

so that the measurement error in the factors does not introduce a bias when estimating M.

Given our assumptions with $\Lambda_{it} = (1, x_{it})$, with $y_t = \Lambda_t F_t + u_t + C_t^y m$, we can write $F_t = (F_t^1, F_t^x)$, where F_t^1 is the factor multiplying the "1" and F_t^x is the factor multiplying the x_{it}, so that $\Lambda_{it} F_t = F_t^1 + x_{it} F_t^x$. Given this, decompose $F_t = \left(F_t^{\text{endo}}, F_t^{\text{exo}}\right) = (F_t^1, F_t^x)$ with endogenous factors (i.e., affected by u_{it}) and exogenous factors. So here, $F_t^x = F_t^{\text{endo}}$.

[87] We might also take $L_t \coloneqq \left(\Lambda_t' \Lambda_t\right)^{-1} \Lambda_t'$.

[88] Indeed, (183) gives

$$-\mathbb{E}\left[F_t^e z_t\right] = L (I - Q)\, V^u Q' S = L (I - Q)\, W^{-1} Q' S = 0,$$

using (177), $(I - Q)\, W^{-1} Q' = 0$.

We keep $\eta_t^e := F_t^{\text{exo},e}$ as use it as a control, as it satisfies (193). Note that the standard errors returned by OLS will be trustworthy, because of (193) again.

This is all a bit abstract, so to get a concrete sense of the situation, let us take the main example, where $\Lambda_{it} = (1, x_{it})$ with $x_{Et} = 0$ and $y_{it} = \gamma p_t + \varepsilon_t + x_i \eta_t + u_{it}$, and $W = I\sigma_u^{-2}$. So factors are: $F_t = \left(F_t^{\text{endo}}, F_t^{\text{exo}}\right) = (\gamma p_t + \varepsilon_t, \eta_t)$, we have $I - Q = \Lambda L$, so $L(I-Q) = L\Lambda L = L$, hence $Ly_t = \left(y_{Et}, \frac{x_t' y_t}{\|x_t\|^2}\right)$, so that the error is

$$F_t^e - F_t = \varepsilon_t^{F^e} = -L(I-Q)\,u_t = \left(u_{Et}, \frac{x_t' u_t}{\|x_t\|^2}\right), \tag{194}$$

and the standard deviation of its components are $\frac{\sigma_u}{\sqrt{N}}\left(1, \frac{1}{\sigma_x}\right)$. The factor analysis recovers (up to that error) $F_t = \left(F_t^{\text{endo}}, F_t^{\text{exo}}\right) = (\gamma p_t + \varepsilon_t, \eta_t)$, so it recovers

$$\eta_t^e = \eta_t + \frac{x_t' u_t}{\|x_t\|^2} \tag{195}$$

but not ε_t. We can use that η_t^e as a control in the regression.

In conclusion, with a factor model (with known factor loadings Λ, but unknown factor covariance matrix V^η), we have identified V^u, and gotten a GIV, which gave M, αM.

Even though all worked with finite N (but as always, $T \to \infty$), and we don't consistently estimate η_t, we still have a consistent estimator for the GIV.

13.1.4 Estimating the variance-covariance matrix of the residuals, V^u

First, we estimate m, using (180) Basically, we can estimate m by OLS. It's pretty easy, as we have $(N-r) \times T$ effective values to use (where r is the number of factors):

$$m^e = \mathbb{E}_T\left[\check{C}_t' W \check{C}_t\right]^{-1} \mathbb{E}_T\left[\check{C}_t' W \check{y}_t\right]. \tag{196}$$

Next, we estimate V^u. We have $\check{u}_t := Q(y_t - C_t m^e) = Q u_t$, so that:

$$V^{\check{u}} = \mathbb{E}\left[\check{u}_t \check{u}_t'\right] = Q\mathbb{E}\left[u_t u_t'\right] Q. \tag{197}$$

We consider the case where we have a priori knowledge that V^u is diagonal. Let us call $D^u = (V_{ii}^u)_{i=1\ldots N}$ and similarly for $D^{\check{u}} = (V_{ii}^{\check{u}})_{i=1\ldots N}$ (so they are vectors of dimension N) and a new matrix $R_{ij} := \left(Q_{ij}^W\right)^2$. Then:

$$D_i^{\check{u}} := V_{ii}^{\check{u}} = \sum_j Q_{ij}^W V_{jj}^u Q_{ij}^W = \sum_j R_{ij} D_j,$$

i.e., $D^{\check{u}} = R D^u$ so we recover

$$D^{u,e} = R^{-1} D^{\check{u}}. \tag{198}$$

Parametric variant We can do a parametric variant. We parametrize $\ln \sigma_{u_i}^2 = \beta^\sigma x_i^\sigma$ for some vector of characteristics x_i^σ, e.g. log size or log volatility (and x_i^σ has 1 as the first component) — σ is just a superscript here, not an exponent. So, we estimate β^σ by regressing the log estimated variance from (198) on the characteristics:

$$\ln D_i^{u,e} = \beta^\sigma x_i^\sigma + \varepsilon_i,$$

and take the fitted values for the diagonal covariance matrix of the u_i's:

$$D_i^u = e^{\beta^\sigma x_i^\sigma}. \tag{199}$$

Loop over W This gives a consistent estimate of D^u, for any W. Now, Bayesian considerations indicate that the optimum W is

$$W = (V^u)^{-1}. \tag{200}$$

So, we can loop: a good initial W is probably $1/var(y_i)$. This gives an estimate of D^u, and a new, better estimate of $W = diag(1/D_i^u)$. We keep looping until convergence. We have consistently estimated the variance matrix of V^u.

13.1.5 Semi-parametric case

Suppose now the semi-parametric case

$$\lambda_{it}^f = \lambda_0^f + \lambda_1^f x_{it}^f + \zeta_i^f,$$

and the we apply the above parametric procedure. For notational simplicity, we assume that all the control and constants are 0, as they are inessential. So, with $X_{it} = (1, x_{it})$ and $\lambda_X^f = \left(\lambda_0^f, \lambda_1^f\right)$ we have:

$$\Lambda_t = X_t \lambda_X + \zeta \tag{201}$$

and

$$y_t = (X_t \lambda_X + \zeta) F_t + u_t.$$

Recall also that we use $Q = Q^{X,W}$, so $QX = 0$. Then, the GIV will be as in (183)

$$z_{\Gamma t} := \Gamma' y_t = S'Q\left[(X_t \lambda_X + \zeta) F_t + u_t\right] = S'Q\left[\zeta F_t + u_t\right],$$

$$z_{\Gamma t} = u_{\Gamma t} + \zeta_\Gamma \eta_t. \tag{202}$$

Hence, our GIV is partially polluted by a small $\zeta_\Gamma \eta_t$. However, as we will control for η_t^e in the regression, this part $\zeta_\Gamma \eta_t$ will be largely controlled for, and will not impact the results.

We sometimes use the close cousin

$$Z_t := y_{\Gamma t} = u_{\Gamma t} + \lambda_\Gamma \eta_t. \tag{203}$$

Then again, it will be controlled for in the regression, as we control for η_t^e.

Note that there are two "Γ" here. The plain one is $\Gamma^0 = S - E$ (with $E_i = \frac{1}{N}$, if we stay in the homoskedastic case). The other, elaborated in this section, is $\Gamma^Q = Q'S$, where $Q = Q^{X,W}$ (with $W = I$ in the homoskedastic case) as given by (23). In principle, it is better to use Γ^Q than Γ^0, as it "fully purges" the parametric factor in the purely parametric case. By extension it should also be a bit better in the semi-parametric case (as we need to purge a "small" ζ_Γ — which is 0 in the parametric case – rather than a "potentially big" λ_Γ). We advocate the Γ^Q, but in practice using Γ^0 gives similar results.

If we use $\Gamma = \Gamma^Q$, then $\lambda_\Gamma = \zeta_{\Gamma^Q}$ so that

$$Z_t = u_{\Gamma t} + \zeta_{\Gamma^Q}\eta_t = z_t. \tag{204}$$

13.2 Other proofs

Proof of Proposition 5 The identification goes as follows. By rescaling S, we impose $\iota' S = 1$. Define $E := S - \Gamma$ (which is $\frac{1}{N}\iota$ in our framework), and form

$$y_{Et} = E' y_t, \qquad y_{St} := S' y_t,$$

which are our generalized "equal weighted" and "value weighted" averages – for more abstract setting. Then, premultiplying (156) by Γ' and S' gives:

$$z_t := \Gamma' y_t = \gamma x_{\Gamma t} + u_{\Gamma t}.$$

Hence, estimating this by OLS we can obtain γ, and $var\left(u_{\Gamma t}\right)$, so that we obtain also σ_u^2. Next,

$$y_E = \beta y_S + \gamma x_E + y x_S + \eta + u_E,$$

so that

$$\mathbb{E}\left[\left(y_{Et} - \beta y_{St} - \gamma x_{Et} - y x_{St}\right)'(z, x_{St})\right] = \left(\mathbb{E} u_{Et} u_{\Gamma t}, 0\right). \tag{205}$$

The right-hand side is known, as $\mathbb{E} u_{Et} u_{\Gamma t} = E'\Gamma\sigma_u^2$, which is known. So, we have two unknowns β, y and 2 equations: we can solve the system. The condition $\Gamma' S = 0$ ensures that $\mathbb{E}\left[y_{St} z_t\right] \neq 0$.

14 A Bayesian perspective on GIVs

We will see that, under conditions of Gaussianity, our estimators are basically the MLE. As variables may not be Gaussian, we keep the general exposition (showing identification) free of distributional assumptions. If we assume that variables (u_{it}, η_t) are Gaussian, then a Bayesian analysis can be performed. We detail it here.

14.1 The general model: Bayesian version

Here we treat the general model of Section 3.1, in the case where the λ_{it} are the same, and equal to 1, and all factors are observed (except the η_t^y, as in $y_{it} = \sum_f \lambda^f F_t^f + u_{it} + \eta_t^y$). The general case with heterogeneous loadings is done later, in Section 14.5, and uses much the same ideas.

The data D is $D = \left(y_t, F_t^f\right)_{f=1...d_F, t=1...T}$, made of i.i.d. draws from a fixed distribution. To simplify the notations, we'll just denote by f the collection of all variables corresponding to factors (without explicitly mentioning that $f = 1 \dots d_F$).

The solution of the system features:

$$\begin{aligned} y_{St} - My_{\Gamma t} &= b^y \varepsilon_t, \\ F_t^f - \alpha^f M y_{\Gamma t} &= b^f \varepsilon_t, \end{aligned}$$

for some vector b^y, b^f, and $\varepsilon_t := \left(u_{Et} + \eta_t^y, \eta_t^f\right)$.

Hence, we form: $\theta = \left(M, \alpha^f M\right)$; W a parametrization of the relevant variance matrices; $E(W) = \frac{V^u(W)^{-1}\iota}{\iota' V^u(W)^{-1}\iota}$ the corresponding quasi-equal weights vector, $\omega = (\theta, W)$, and form the key quantities:

$$Y_t(\omega) := \left(y_{St} - My_{\Gamma(W),t}, F_t^f - \alpha^f M y_{\Gamma(W),t}\right). \tag{206}$$

We also keep track of

$$\check{y}_{it}(\omega) = y_{it} - y_{E(W),t} \tag{207}$$

and stack those two vectors together as $X_t(\omega)$, which contains all our information:

$$X_t(\omega) = \left(Y_t(\omega), \check{y}_t(\omega)\right). \tag{208}$$

The key "trick to tractability" is to transform the data into that X_t.

There is an invertible matrix $A(\theta)$ such that $D_t = A(\theta) X_t$. Hence, there is no loss of information in using X_t as "conveniently processed" data, rather than the "unprocessed" data D_t. Hence, instead of $\ln \mathbb{P}(D_t|\omega)$, we'll consider

$$\ln \mathbb{P}(X_t|\omega) = \ln \mathbb{P}(D_t|\omega) + \ln |A(\theta)|. \tag{209}$$

The Jacobian $|A(\theta)| := \det A$ is independent of all parameters ω.[89] Hence it can be discarded as a constant in the calculations.

The key simplifying observation is that (under the correct model), $\mathbb{E}\check{y}_t y_{Et} = 0$, so that $Y_t(\omega)$ and $\check{y}_t(\omega)$ have zero covariance. Hence, the log likelihood decouples, and we have

$$-2\ln\mathbb{P}(D_t|\omega) = Y_t'(\omega)\, V^Y(W)^{-1} Y_t(\omega) + \check{y}_t(\omega)'\left(V^{\check{y}}(\omega)\right)^{-1}\check{y}_t(\omega) + \ln\left|V^Y(W)\right| + \ln\left|V^{\check{y}}\right|. \quad (210)$$

As $\check{y}_t$ lives in a space of dimension $N-1$ (as $E'\check{y}_t = 0$), the value of $V^{\check{y}}$ is understood as being of the corresponding dimensions, $(N-1)\times(N-1)$.

Now, imagine that W has already been estimated, and do only the optimization w.r.t. θ. That gives:

$$\min_\theta \mathbb{E}_T\left[Y_t'(\theta,W)\, V^Y(W)^{-1} Y_t(\theta,W)\right]. \quad (211)$$

The first order conditions are:

$$\mathbb{E}_T\left[(y_{\Gamma t},0)\left(V^Y\right)^{-1}Y_t\right] = 0, \qquad \mathbb{E}_T\left[(0,y_{\Gamma t})\left(V^Y\right)^{-1}Y_t\right] = 0,$$

i.e. (given that $0 = \mathbb{E}_T\left[y_{\Gamma t}\left(V^Y\right)^{-1}Y_t\right] = \left(V^Y\right)^{-1}\mathbb{E}_T\left[y_{\Gamma t}Y_t\right]$) we have $\mathbb{E}_T\left[y_{\Gamma t}Y_t\right] = 0$, yielding

$$\mathbb{E}_T\left[y_{\Gamma t}\left(y_{St} - My_{\Gamma t}, F_t^f - \alpha^f M y_{\Gamma t}\right)\right] = 0. \quad (212)$$

Those are precisely the first order conditions of the OLS estimation:

$$\min_M \mathbb{E}_T\left[\left(y_{St} - My_{\Gamma t}\right)^2\right], \qquad \min_{\alpha^f M} \mathbb{E}_T\left[\left(F_t^f - \alpha^f M y_{\Gamma t}\right)^2\right]. \quad (213)$$

Hence, our GIV is also the MLE estimator of $M, M\alpha^f$, when we have Gaussian distributions.

We can also go beyond MLE, and calculate full Bayesian posteriors. Then, the GIV gives an easy way to do finite-sample Bayesian updating. Assuming again for simplicity that we know the variance matrices, we have

$$\ln\mathbb{P}(\theta|D) = \ln\mathbb{P}(\theta) - \frac{1}{2}\sum_t Y_t(\theta)\left(V^Y\right)^{-1}Y_t(\theta) + K(D), \quad (214)$$

where $K(D)$ ensures that the probability sums to 1.

The rest of this section examines instantiations and variants of the general idea we just saw.

[89] First, go from X_t to $\tilde{D}_t = (F_t, y_{Et}, \check{y}_t)$, which is upper triangular with 1 on the diagonals, so has determinant 1; second, go from $\tilde{D}$ to D, which is independent of the ω.

14.2 The supply and demand model of Section 2.3

This model corresponds exactly to the general case, with a factor $F_t^f = p_t$, $p_t = \alpha^f y_{St} + \eta_t^f$ with $\alpha^f = \frac{1}{\mu}$ and $\eta^f = -\frac{\varepsilon}{\mu}$. Then, everything goes through.

14.3 The basic example with self-loop of Section 6.2

We give a Bayesian treatment of this model of Section 6.2:

$$y_{it} = \gamma y_{St} + \eta_t + u_{it}.$$

We are given $D_t = y_t$. We wish to estimate $M = \frac{1}{1-\gamma}$. The vector of parameters of interest is $\theta = M$:

$$Y_t(\omega) = y_{St} - M y_{\Gamma(W),t}.$$

As in the general procedure, we set:

$$\check{y}_{it}(\omega) = y_{it} - y_{E(W),t} \tag{215}$$

and

$$X_t(\omega) = (Y_t(\omega), \check{y}_t(\omega)). \tag{216}$$

In the true model, we have $Y_t = M(u_{Et} + \eta_t)$, so

$$-2\ln \mathbb{P}(D_t|\omega) = \frac{Y_t^2(\omega)}{\sigma_Y^2} + \check{y}_t(\omega)' V^{\check{y}} \check{y}_t(\omega) + \ln \sigma_Y^2 + \ln \left|V^{\check{y}}\right|.$$

Suppose first that we know the variance terms. Then, the MLE is simply to do

$$\max_M \mathbb{E}_T \left[Y_t(\omega)^2\right],$$

which is the identification condition we used, and it corresponds to running the OLS

$$\min_M \mathbb{E}_T \left(y_{St} - M y_{\Gamma(W)t}\right)^2.$$

Next, for the estimation of the variance terms, we optimize on σ_Y^2, $V^{\check{y}}$. Asymptotically, that gives the true values.

14.4 The basic example without loop of Section 2.2.

We now detail the Bayesian version of our example in Section 2.2:

$$y_{it} = \eta_t + u_{it}, \qquad p_t = \alpha y_{St} + \varepsilon_t.$$

We'd like to estimate α especially (or, in a Bayesian context, update our prior on α). This example is actually a bit non-generic, as it endows the economist with a knowledge that $\lambda^f = 0$, which creates some subtle changes: it features the "recovered" factor y_{Et}, used as a regressor.

The data D is a set of $D = (y_t, p_t)_{t=1\dots T}$, assumed to be i.i.d. draws from a fixed distribution.

We call $\theta = (\alpha, \beta)$ the set of "key" model parameters, and W, the variance-covariance matrix $V^{(u+\eta\iota,\varepsilon)}$ (or, it could be some parametrization of it, e.g. if we assume that u is diagonal), the auxiliary parameter, and $\omega = (\theta, W)$ the full set of parameters. The correct value is ω^*.

Given y_t, p_t, we form

$$Y_t(\theta) = p_t - (\alpha y_{\Gamma t} + \beta y_{Et})$$

and $X_t(\theta) = (Y_t(\theta), y_t)$. At the correct parameter ω^*,

$$Y_t(\theta^*) = \varepsilon_t^{\perp},$$

which is defined in the analysis is the "enriched OLS estimator" (Section 13). Hence, at the correct value, Y_t and y_t are uncorrelated. Call $V^X(\omega)$ the variance-covariance matrix of X_t.

We can start the Bayesian analysis:

$$\mathbb{P}(\omega|D) \propto \mathbb{P}(D|\omega)\,\mathbb{P}(\omega)$$

and

$$\ln \mathbb{P}(D|\omega) = \sum_t \ln \mathbb{P}(D_t|\omega)$$

with

$$-2\ln \mathbb{P}(D_t|\omega) = \frac{Y_t(\theta)^2}{\sigma_{\varepsilon^\perp}^2} + y_t' V^y(W)\, y_t + \ln \sigma_{\varepsilon^\perp}^2 + \ln |V^y(W)|, \tag{217}$$

where here $|A|$ is the determinant of a matrix A.

Hence, the MLE estimator maximizes $\sum_t \ln \mathbb{P}(D_t|\omega)$ over $\omega = (\theta, W)$. The problem for θ separates as:

$$\min_{\alpha,\beta} \sum_t Y_t(\theta)^2,$$

i.e.

$$\min_{\alpha,\beta} \sum_t (p_t - (\alpha y_{\Gamma t} + \beta y_{Et}))^2,$$

which is the "enriched GIV-OLS estimator" of Section 12.10. This shows that, with Gaussian distributions, the MLE is just our enriched GIV-OLS estimator.

Maximizing over the other parameters W will allow to recover the variance matrix (including that of ε_t, η_t).

If we have a small sample, we can just update rather than do MLE. The above shows that the "simplifying trick" is to form that statistic $Y_t(\theta)$, which allows for an interpretable updating of the parameters. For simplicity, suppose that we know the value of $V^y(W)$, and $\sigma^2_{\varepsilon^\perp}$.[90] However, we have a prior on $\theta = (\alpha, \beta)$, perhaps Gaussian. Then, our posterior after observing the data D is:

$$\ln \mathbb{P}(\theta|D) = \ln \mathbb{P}(\theta) - \sum_t \frac{Y_t(\theta)^2}{2\sigma^2_{\varepsilon^\perp}} + K(D),$$

where $K(D)$ ensures that the probability sums to 1.

14.5 Heterogeneous loadings

14.5.1 Bayesian model with heterogeneous loadings

Here we extend the basic model of this section to heterogeneous, nonparametric loadings. The model is

$$y_{it} = \sum_f \lambda^f F_t^f + \eta_t^y + \sum_{k=1}^K \lambda_i^k \eta_t^k + u_{it}, \tag{218}$$

where now there are K unobserved factors, with unknown factors η_t^k, and non-uniform loadings λ_i^k. As in the more basic Section 14.1, we assume the existence of a factor η_t^y with uniform loadings (which can be taken to be uncorrelated with η_t), and factors F_t^f are endogenous and observed. More compactly, we can write the model as:

$$y_t = \lambda^F F_t + \iota \eta_t^y + \lambda \eta_t + u_t. \tag{219}$$

We will now see how this case can be reduced to the one of Section 14.1. We define the theoretical object:

$$\tilde{y}_{it} := y_{it} - \sum_{k=1}^K \lambda_i^k \eta_t^k. \tag{220}$$

Then, the results of Section 14.1 apply to $\tilde{y}_t$, conditional on (λ, η_t). Equation (210) becomes:

$$-2\ln \mathbb{P}(D_t|\omega, \lambda, \eta_t) = \tilde{Y}_t'(\omega) V^{\tilde{Y}}(W)^{-1} \tilde{Y}_t(\omega) + \check{\tilde{y}}_t(\omega)' \left(V^{\check{\tilde{y}}}(\omega)\right)^{-1} \check{\tilde{y}}_t(\omega) + \ln\left|V^{\tilde{Y}}(W)\right| + \ln\left|V^{\check{\tilde{y}}}\right| \tag{221}$$

with

$$\tilde{Y}_t(\omega) := \left(\tilde{y}_{St} - M\tilde{y}_{\Gamma(W),t}, F_t^f - \alpha^f M \tilde{y}_{\Gamma(W),t}\right), \tag{222}$$

$$\check{\tilde{y}}_{it}(\omega) = \tilde{y}_{it} - \tilde{y}_{E(W),t}. \tag{223}$$

So, given λ, η_t, the procedure is as in Section 14.1, applied to the tilde variables. In turn, suppose that we have some priors on λ (we'll take them to be diffuse) and on η_t (we'll normalize them to

[90] Otherwise, we can update our knowledge of those, which is standard though tedious to lay out.

be independent standard normals). Then, the full likelihood is:

$$\mathbb{P}\left(D_t|\omega\right) = \mathbb{P}\left(D_t|\omega, \lambda, \eta_t\right)\mathbb{P}\left(\lambda, \eta_t\right), \qquad \mathbb{P}\left(\lambda, \eta_t\right) = \left(2\pi\right)^{-K/2} e^{-\frac{1}{2}\eta_t \eta_t'}. \tag{224}$$

So, we can estimate λ and η_t by Bayesian methods, e.g the E-M method summarized in Section 14.5.2.

What the MLE gives It is worth pausing to see what the MLE does. Consider the MLE estimator of M (keeping the $\alpha^f M$ constant). It is as the analysis of (212) but in $\tilde{y}_t$ space, i.e. it is equivalent to running the OLS regression:

$$\tilde{y}_{St} = M\tilde{y}_{\Gamma t} + e_t, \tag{225}$$

i.e.,

$$y_{St} - \lambda_S \eta_T = M z_t + e_t, \tag{226}$$

where

$$z_t := z_{\Gamma t} + z_{\eta t}, \qquad z_{\Gamma t} := u_{\Gamma t}, \qquad z_{\eta t} := \lambda_{\Gamma t}\eta_t. \tag{227}$$

This means that the MLE uses two primitive sources of shocks for identification (i) $z_{\Gamma t} := u_{\Gamma t}$, which is the "pure" GIV, and (ii) $z_{\eta t} := \lambda_{\Gamma t}\eta_t$, which traces the ripple effects of the aggregate shocks η_t on the aggregate action y_{St}, after controlling for the "direct" effects (this is why $y_{St} - \lambda_S \eta_T$ is on the left-hand side of (226)). Those are two economically very different styles of identification. For economic clarity we find it useful to single out solely the "pure" GIV identification (i.e. regress only on $z_{\Gamma t}$ rather than on $z_{\Gamma t} + z_{\eta t}$).

14.5.2 Maximum likelihood estimation with heterogeneous loadings

We consider the model with heterogeneous loadings

$$y_t = \gamma y_{St}\iota + \lambda\eta_t + u_t,$$

where $u_t \sim N\left(0, V_u\right)$ and $\eta_t \sim N\left(0, 1\right)$. Define $\delta_t(\gamma) = y_t - \gamma y_{St}\iota = \lambda\eta_t + u_t$ and note that the log likelihood contribution of y_t is

$$\mathbb{P}\left(y_t\right) = \mathbb{P}\left(\delta_t\right) + \ln\left(1 - \gamma\right).$$

The likelihood of δ_t can be computed efficiently using the expectation-maximization (EM) algorithm. The steps are as follows, and we refer to **?** for details, where subscripts (n) refer to the n-th iteration of the algorithm.

- Expectation step
  - $\mathbb{E}_{(n)}\left[\eta_t \mid \delta_t\right] = \beta_{(n)}\delta_t$,where $\beta_{(n)} = \lambda'\Sigma^{-1}$ and $\Sigma_{(n)} = \lambda_{(n)}\lambda_{(n)}' + V_{u(n)}$.

- $\mathbb{E}_{(n)}\left[\eta_t^2 \mid \delta_t\right] = 1 - \beta_{(n)}\lambda_{(n)} + \left(\mathbb{E}_{(n)}\left[\eta_t \mid \delta_t\right]\right)^2$.

- Maximization step
  - $\lambda_{(n+1)} = \left(\sum_t \mathbb{E}_{(n)}\left[\eta_t^2 \mid \delta_t\right]\right)^{-1} \sum_t \delta_t \mathbb{E}_{(n)}\left[\eta_t \mid \delta_t\right]$.
  - $V_{u(n+1)} = \frac{1}{T} diag\left\{\sum_t \delta_t \delta_t' - \lambda_{(n+1)} \mathbb{E}_{(n)}\left[\eta_t \mid \delta_t\right] \delta_t'\right\}$.

The log likelihood can be computed as (omitting constants that do not depend on the parameters)

$$\mathbb{P}_{(n)}\left(\delta_t\right) = -\frac{1}{2}\ln\left|\Sigma_{(n)}\right| - \frac{1}{2}\delta_t' \Sigma_{(n)}^{-1} \delta_t,$$

and we iterate until convergence. To initialize the algorithm, we start from estimates of λ and V_u based on PCA.

References for Online Appendix

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile prices in market equilibrium," *Econometrica*, 1995, pp. 841–890.