Uncovering Disaggregated Oil Market Dynamics: A Full-Information Approach to Granular Instrumental Variables*

June 30, 2023

Revised: April 26, 2025

Christiane Baumeister
University of Notre Dame
University of Pretoria
cjsbaumeister@gmail.com

James D. Hamilton
University of California at San Diego
jhamilton@ucsd.edu

ABSTRACT

The world price of oil is determined by the interactions of multiple producers and consumers who face different constraints and shocks. We show how this feature of the oil market can be used to estimate local and global elasticities of supply and demand and provide a rich set of testable restrictions. We develop a novel approach to estimation based on full-information maximum likelihood that generalizes the insights from granular instrumental variables. We conclude that the supply responses of Saudi Arabia and adjustments of inventories have historically played a key role in stabilizing the price of oil. We illustrate how our structural model can be used to analyze how individual producers and consumers would dynamically adapt to a geopolitical event such as a major disruption in the supply of oil from Russia.

JEL codes: Q43, C32, E32

Key words: structural vector autoregression, identification, heterogeneous elasticities, granular instrumental variables, Russian oil supply cut, scenario analysis

*An earlier version of this paper was circulated under the title, “A Full-Information Approach to Granular Instrumental Variables.” We thank Anastasia Allayioti, Marinho Bertanha, Giorgio Calzolari, Jean-Marie Dufour, John Elder, Nida Çakır Melek, and Jamel Saadaoui for helpful comments on earlier drafts of this paper.

1 Introduction

Aggregate economic outcomes result from the interactions of many individual units. Modeling what those individual units have in common and the ways in which they differ can help identify the local and aggregate effects of structural shocks. A popular example is Bartik-type instruments, which use a weighted average of aggregate conditions with weights given by local shares as an instrument to estimate local elasticities; see Bartik (1991), Blanchard et al. (1992), Goldsmith-Pinkham, Sorkin and Swift (2020), and Carlino and Drautzburg (2020). Gabaix and Koijen (2024) demonstrated that in some situations one can use the difference between the arithmetic average and a share-weighted average of observations across units as a valid instrument, which they described as "granular instrumental variables." Qian (2023) extended granular instrumental variables to allow for heterogeneous responses of individual units. Banafti and Lee (2022) considered granular instrumental variables in large panels and Sarto (2022) in panel vector autoregressions. Another nice illustration of the underlying idea is Caldara, Cavallo and Iacoviello (2019), who used known exogenous shortfalls in oil production in certain countries as an instrument to estimate the response of producers in other countries to an exogenous increase in the price of oil.

In all these applications, the focus has been on developing a valid instrument for purposes of estimating a particular elasticity of interest. In this paper, we adopt a broader systemwide approach based on modeling how individual units interact to produce the aggregate outcome. We show that while such an approach could be used to motivate instrumental-variable estimation used in other studies, the underlying assumptions in fact provide a full characterization of the joint determination of local and aggregate magnitudes. The first-order conditions for maximum likelihood estimation can be given an instrumental variable interpretation. We show in a simple example that the usual granular instruments are just one part of the information exploited by maximum likelihood and that using all the available information can result in substantially improved estimates.

Our method can also be viewed as an application of the familiar approach in structural vector autoregressions of interpreting the correlations between the errors in forecasting individual observed variables as arising from an underlying set of structural shocks. The dynamic effects of structural shocks on local and aggregate variables can then be consistently estimated. Our approach typically produces a rich set of overidentifying restrictions that can be tested against the data.

We illustrate our approach with an analysis of the world oil market. Our framework allows us to estimate country-specific supply and demand elasticities jointly, which allows us to assess adjustment dynamics of major players in the market in response to common or country-specific disturbances. The need to gain a better understanding of the propagation of shocks originating in one particular country has been highlighted by recent geopolitical events in Russia, which we use as one illustration of the usefulness of our proposed modeling approach.

We model oil market dynamics using a small-scale vector autoregression that includes production of oil from the three largest producers over 1974-2023 (the United States, Saudi Arabia, and Russia), consumption of oil by the three largest historical consumers (the United States, Japan, and Europe), and aggregate magnitudes. Our estimates imply a global short-run price elasticity of oil supply of 0.08, consistent with the estimates of Caldara, Cavallo and Iacoviello (2019, Table 3) of 0.05 to 0.08, Balke, Jin and Yücel's (2024) estimates of 0.07 to 0.10, and Baumeister and Hamilton's (2019) estimate of 0.15. Unlike any previous study, our estimates are also consistent with all the individual elasticities that we also report. We estimate a short-run supply elasticity of 0.25 for Saudi Arabia and 0.02 to 0.07 for other countries. A few studies have estimated separate supply elasticities for different countries or regions, and where this has been done these earlier estimates are consistent with our findings. Alonso-Alvarez, Di Nino and Venditti (2022) arrived at separate supply elasticities of 0.20 for OPEC and 0.06 for non-OPEC. When Caldara, Cavallo and Iacoviello (2019) estimated elasticities separately for different groups of countries, their estimates were 0.21 for Saudi Arabia, 0.19 for other OPEC countries, and essentially zero for non-OPEC countries, again in line with our findings. A substantially larger price elasticity for OPEC versus non-OPEC countries is also supported by the analysis in Almutairi, Pierru and Smith (2023) and Balke, Jin and Yücel (2024).

We estimate the global short-run price elasticity of the demand for oil to be -0.12, which is again consistent with the conclusions that earlier studies arrived at using very different methods from ours. Caldara, Cavallo and Iacoviello (2019, Table 3) estimated the short-run price elasticity of world petroleum demand to be -0.03 to -0.08, similar to the -0.05 estimate of Pierru, Smith and Zamrik (2018). Slightly larger estimates were obtained by Balke, Jin and Yücel (2024) (-0.17), Alonso-Alvarez, Di Nino and Venditti (2022) (-0.28), and Baumeister and Hamilton (2019) (-0.35). Meta-analyses of hundreds of earlier studies estimated short-run gasoline demand elasticities from -0.25 to -0.34 and short-run elasticities for the demand for crude oil from -0.05 to -0.07 (Hamilton (2009*b*, Table 3)). Again our global elasticity is calculated by aggregating the estimated demand responses for individual countries, which range from -0.001 for Japan to -0.20 for Europe.

Ours is the first study to simultaneously estimate demand and supply elasticities that differ across all countries. We do this in a unified statistical framework motivated by the principle that the price of oil equilibrates global supply and demand. Estimates like ours could prove helpful in calibrating theoretical models of oil market dynamics such as Bornstein, Krusell and Rebelo (2023) and Balke, Jin and Yücel (2024).

We use our model to analyze the effects of different local and global shocks to supply or demand. We find that Saudi Arabian production is an important factor in stabilizing the price of oil, consistent with the conclusions of Almutairi, Pierru and Smith (2023). We further conclude that inventories are typically drawn down in response to a temporary increase in price and that this plays a critical role in smoothing out temporary price shocks. This supports the conclusions of Knittel and Pindyck (2016) and is in contrast to the claim sometimes made that speculative inventory accumulation accentuates price increases. We find that in the absence of an ability to draw down or accumulate inventories, most of the short-run adjustment to shocks would come in the form of changes in demand rather than changes in supply.

As a case study, we use our model to analyze what would happen in response to a 50% cut in Russian oil production arising from exogenous political factors unique to that country. Our model predicts that about 1.9 million barrels a day of the 5.3 mb/d shortfall would be met by increased production from other countries and the rest by lower consumption.

The plan of the paper is as follows. The data are described in Section 2. Section 3 introduces the model of market equilibrium that underlies our structural analysis. Section 4 explains how we can use heterogeneous observations on production, consumption, and the price of oil to identify and estimate supply and demand elasticities. Section 5 presents our empirical results. Section 6 develops case studies illustrating how the model can be used to learn about dynamics at the disaggregate level and answer counterfactual questions. Section 7 uses the framework to analyze how the oil market has changed over time. Section 8 briefly concludes.

2 Data

The U.S. Energy Information Administration publishes monthly data on the production of crude oil in a number of different countries going back to 1973 and consumption of petroleum products for a different set of countries going back to 1982. For purposes of this study, we used historical published issues of the *Monthly Energy Review* to extend the consumption data back to 1973 for a small number of countries.

2.1 Measuring growth rates

There is a strong seasonal component to petroleum consumption for some countries. For this reason we use year-over-year growth rates, which we measure as

$$q_{it} = \frac{Q_{it} - Q_{i,t-12}}{0.5(Q_{it} + Q_{i,t-12})} \tag{1}$$

$$c_{jt} = \frac{C_{jt} - C_{j,t-12}}{0.5(C_{jt} + C_{j,t-12})}. \tag{2}$$

Here Q_{it} is the quantity of oil produced in country i in month t and C_{jt} is the quantity of oil consumed in country j in month t, both measured in millions of barrels per day. This is preferable to alternative measures of the growth rate such as $(Q_{it} - Q_{i,t-12})/Q_{i,t-12}$ or $\log(Q_{it}) - \log(Q_{i,t-12})$. The latter gets arbitrarily large when either Q_{it} or $Q_{i,t-12}$ get small, and would diverge to minus or plus infinity in the case of the complete cessation and subsequent resumption of production, as happened for example in Iraq and Kuwait in 1990. Expression (1) can be viewed as a first-order Taylor approximation to the function $\log Q_{it} - \log Q_{i,t-12}$ where the approximation is taken at a point halfway between Q_{it} and $Q_{i,t-12}$. The approximation is almost exact as long as Q_{it} is not less than half the size of $Q_{i,t-12}$ and not more than twice the size of $Q_{i,t-12}$; see Figure A.1. For larger changes, expression (1) is less extreme than $\log(Q_{it}) - \log(Q_{i,t-12})$, and is bounded between ± 2 for all values of Q_{it} and $Q_{i,t-12}$.

Our production data describe countries $i = 1, 2, ..., n$ where country n is defined as "rest of world" so that $\sum_{i=1}^{n} Q_{it}$ exactly equals total measured world oil production in month t. Our empirical application uses $n = 4$ where $i = 1, 2$, or 3 correspond to the three largest producing countries over our historical sample, which were the U.S., Saudi Arabia, and Russia. A key magnitude is the average share of country i in total world production,

$$s_{qi} = T^{-1} \sum_{t=1}^{T} \frac{Q_{it}}{Q_t},$$

which we collect in an $(n \times 1)$ vector $\mathbf{s}_q$. In our data set, $\mathbf{s}_q = (0.12, 0.12, 0.15, 0.61)'$. We will approximate the annual growth in global production using the share-weighted average of individual country growth rates:

$$\frac{Q_t - Q_{t-12}}{0.5(Q_t + Q_{t-12})} = \frac{\sum_{i=1}^{n} Q_{it} - \sum_{i=1}^{n} Q_{i,t-12}}{0.5(Q_t + Q_{t-12})} = \sum_{i=1}^{n} \left[\frac{Q_{it} - Q_{i,t-12}}{0.5(Q_{it} + Q_{i,t-12})} \frac{(Q_{it} + Q_{i,t-12})}{(Q_t + Q_{t-12})} \right]$$

$$\simeq \sum_{i=1}^{n} \left[\frac{Q_{it} - Q_{i,t-12}}{0.5(Q_{it} + Q_{i,t-12})} s_{qi} \right] = \sum_{i=1}^{n} s_{qi} q_{it}. \tag{3}$$

Similarly, our consumption data describe countries $j = 1, 2, ..., m$ where country m is defined as "rest of world" so that $\sum_{j=1}^{m} C_{jt}$ exactly equals total measured world petroleum consumption in month t. Our procedure does not require n to equal m nor does it require the producers to be the same countries as the consumers. Our baseline results use the historically three largest consuming countries (the U.S., Japan, and Europe) so that $m = 4$. Average consumption shares are summarized by the $(m \times 1)$ vector $\mathbf{s}_c = (0.25, 0.07, 0.08, 0.60)'$. We approximate the year-over-year growth rate of global consumption as

$$\frac{C_t - C_{t-12}}{0.5(C_t + C_{t-12})} \simeq \sum_{j=1}^{m} s_{cj} c_{jt}. \tag{4}$$

It is not the case in EIA reported data that global oil consumption C_t is the same number as global oil production Q_t. There are three reasons for this. First, there are conceptual differences in definition. Production is measured in the number of barrels of oil taken out of the ground. One barrel of oil produces more than one barrel of refined product used by consumers, and additional consumable product comes from biofuels and processing of natural gas. For these reasons, measured global consumption exceeds measured global production. Second, consumption and production numbers are collected from different underlying data sources and there are acknowledged errors in measuring all of these variables. Third, global production could be greater or less than global consumption in a given month t if there is an increase or decrease in global oil inventories. We will take all these factors into account in the model developed below.

Although oil is produced and consumed in different locations around the world, it is a world market for oil in which the quality-adjusted product sells for essentially the same price everywhere in the world. We measure the global real price of oil in month t (denoted P_t) as the dollar price of a barrel of Brent crude oil deflated by the U.S. consumer price index.[1] We convert this to monthly growth rates $p_t = \log(P_t) - \log(P_{t-1})$. The observed data for month t are summarized by the $[(n + m + 1) \times 1]$ vector $\mathbf{y}_t = (\mathbf{q}_t', \mathbf{c}_t', p_t)'$ consisting of the growth rates of production and consumption for each country in the world along with the world price of oil.

2.2 Data during the COVID-19 pandemic

The pandemic shut-downs in 2020 completely disrupted both oil supply and demand; for a description and analysis of these disruptions see Baumeister (2023). Events of

[1]The CPI is from FRED (https://fred.stlouisfed.org/series/CPIAUCSL). The Brent spot price is available from EIA since May 1987 and was extended back to 1973 by Baumeister, Korobilis and Lee (2022) using the growth rate of the refiners' acquisition cost for imported oil.

2020 also show up very dramatically in a broad range of other economic indicators. A number of approaches for dealing with this structural break have been proposed. Lenza and Primiceri (2022) suggested we could treat the underlying structural relations as unchanged but allow for a big increase in the magnitude of structural shocks. Ng (2021) argued that these disruptions were an entirely new shock that we could model using direct observations on measures of hospitalization, positive cases, or deaths. In our paper we adopt the more general view that potentially all the structural relations and structural shocks were different during the pandemic, implying that structural and reduced-form parameters during this episode should be estimated separately from the rest of the sample. Since there are not enough observations during the pandemic to estimate a full set of parameters over this short episode, in practice this means dropping these observations from the sample and pooling post-COVID and pre-COVID observations into a single sample. Schorfheide and Song (2024), Lenza and Primiceri (2022), and Hamilton (2025) noted that this is what researchers might often want to do, and this is the approach followed in this paper.

Year-over-year growth rates of oil production and consumption are profoundly impacted for 2020:M3 through 2021:M2. Since we use twelve lags of these as explanatory variables, we therefore drop two years of data associated with the pandemic. The left-hand variable in our equations covers observations from 1975:M1 through 2020:M2 and 2022:M2 through 2023:M2, for a total of $T = 555$ observations. For notational convenience, we will write $\sum_{t\in\{\text{1975:M1-2020:M2} \,\cup\, \text{2022:M2-2023:M2}\}}$ simply as $\sum_{t=1}^{T}$ where T is the total number of observations on the dependent variable in our pooled sample. We obtained very similar results if we just end the sample in 2019:M12.

3 Market equilibrium

The production of oil from country i is presumed to be governed by the structural equation

$$q_{it} = \phi_{qi} p_t + \mathbf{b}'_{qi}\mathbf{x}_{t-1} + u_{qit} \tag{5}$$

for $i = 1, ..., n$. Here $\mathbf{x}_{t-1} = (1, \mathbf{y}'_{t-1}, ..., \mathbf{y}'_{t-r})'$ is a $(k \times 1)$ vector consisting of a constant term and r lags of the production and consumption of every country in the world along with the world price; thus $k = 1 + r(n + m + 1)$. The term u_{qit} represents factors other than the lags $\mathbf{x}_{t-1}$ and the current price p_t that influence the production in country i. The term $\mathbf{b}'_{qi}\mathbf{x}_{t-1}$ governs the dynamic behavior of oil supply in country i. We assume that structural dynamics are incorporated in the definition of $\mathbf{b}_{qi}$ so that u_{qit} can be regarded as serially uncorrelated; see Appendix A.2 for more discussion. The appendix also provides examples of how equation (5) could arise with perfect competition, market power, or forward-looking behavior by producers. In our empirical analysis we take $r = 12$. Note that although we are measuring q_{it} in year-over-year growth rates, the inclusion of lags means that ϕ_{qi} represents the response of supply to an unanticipated change in price. Thus ϕ_{qi} should be interpreted as the within-month price elasticity of supply for country i.

Likewise the structural demand equation for country j takes the form

$$c_{jt} = \phi_{cj} p_t + \mathbf{b}'_{cj} \mathbf{x}_{t-1} + u_{cjt} \tag{6}$$

for $j = 1, ..., m$. Here ϕ_{cj} is the short-run demand elasticity in country j and u_{cjt} is a shock to country j demand.

Let v_t denote the difference between global production and global consumption:

$$v_t = \textstyle\sum_{i=1}^{n} s_{qi} q_{it} - \sum_{j=1}^{m} s_{cj} c_{jt}. \tag{7}$$

If production and consumption data were completely accurate, v_t would represent the change in global oil inventories. We interpret it here as the change in inventories plus measurement error. We allow v_t to respond to prices and be serially correlated:

$$v_t = \phi_v p_t + \mathbf{b}'_v \mathbf{x}_{t-1} + u_{vt}. \tag{8}$$

We interpret ϕ_v as the short-run price elasticity of inventory demand and u_{vt} as a combination of measurement error and a shock to inventory demand.[2] Note that inclusion of a constant term in (8) allows for systematic average differences between measured production and consumption that have no implications for the price of oil. We discuss the relation between our measure of changes in inventories and those available from other data sources in Appendix A.3.

The equilibrium price is determined by equations (5)-(8). It is helpful to rewrite these in vector form as[3]

$$\underset{(n\times 1)}{\mathbf{q}_t} = \underset{(n\times 1)}{\boldsymbol{\phi}_q} \, p_t + \underset{(n\times k)}{\mathbf{B}_q} \, \mathbf{x}_{t-1} + \underset{(n\times 1)}{\mathbf{u}_{qt}} \tag{9}$$

[2]An earlier version of this paper allowed for the possibility of separate measurement error terms in individual country-level production and consumption figures. The empirical results were similar to those obtained in the current version, which interprets all measurement error as entering through u_{vt}.

[3]Equation (9) is obtained by stacking the n equations in (5) defining $\mathbf{q}_t = (q_{1t}, ..., q_{nt})'$, $\boldsymbol{\phi}_q = (\phi_{q1}, .., \phi_{qn})'$, $\mathbf{B}_q = \begin{bmatrix} \mathbf{b}_{q1} & \mathbf{b}_{q2} & \cdots & \mathbf{b}_{qn} \end{bmatrix}'$, and $\mathbf{u}_{qt} = (u_{q1t}, ..., u_{qnt})'$. Equation (10) is likewise obtained by stacking the m equations in (6).

$$\underset{(m\times 1)}{\mathbf{c}_t} = \underset{(m\times 1)}{\boldsymbol{\phi}_c} \, p_t + \underset{(m\times k)}{\mathbf{B}_c} \, \mathbf{x}_{t-1} + \underset{(m\times 1)}{\mathbf{u}_{ct}} \tag{10}$$

$$\mathbf{s}_q'\mathbf{q}_t - \mathbf{s}_c'\mathbf{c}_t = \phi_v p_t + \mathbf{b}_v'\mathbf{x}_{t-1} + u_{vt}. \tag{11}$$

The structural model consists of equations (9), (10), and (11). Identification comes from restrictions on the covariances between the structural shocks and possible restrictions on the elasticities $\boldsymbol{\phi}_q$ and $\boldsymbol{\phi}_c$.

Let $\mathbf{y}_t = (\mathbf{q}_t', \mathbf{c}_t', p_t)'$ denote the $(N \times 1)$ vector of observed variables with $N = n + m + 1$ and $\mathbf{u}_t = (\mathbf{u}_{qt}', \mathbf{u}_{ct}', u_{vt})'$ the $(N \times 1)$ vector of structural shocks. The structural model can be written more compactly as

$$\mathbf{A}\mathbf{y}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u}_t \tag{12}$$

$$\underset{(N\times N)}{\mathbf{A}} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{nm} & -\boldsymbol{\phi}_q \\ \mathbf{0}_{mn} & \mathbf{I}_m & -\boldsymbol{\phi}_c \\ \mathbf{s}_q' & -\mathbf{s}_c' & -\phi_v \end{bmatrix} \tag{13}$$

$$\underset{(N\times k)}{\mathbf{B}} = \begin{bmatrix} \mathbf{B}_q \\ \mathbf{B}_c \\ \mathbf{b}_v' \end{bmatrix}.$$

Given any value for the structural shocks $\mathbf{u}_t$, equation (12) describes how price and quantities adjust to equate global supply and demand.

4 Identification and maximum likelihood estimation

In this section we discuss how elasticities and other structural parameters can be estimated from the data.

4.1 Granular instrumental variables

To start with a simple example, suppose we were willing to assume that demand shocks are uncorrelated with supply shocks ($E(\mathbf{u}_{qt}\mathbf{u}_{ct}') = \mathbf{0}_{nm}$) and that demand elasticities are the same across countries ($\boldsymbol{\phi}_c = \phi_c \mathbf{1}_m$ for ϕ_c the common demand elasticity and $\mathbf{1}_m$ an $(m \times 1)$ vector of ones). In this case (10) becomes

$$\underset{(m\times 1)}{\mathbf{c}_t} = \underset{(1\times 1)}{\phi_c} \underset{(m\times 1)}{\mathbf{1}_m} \, p_t + \underset{(m\times k)}{\mathbf{B}_c} \, \mathbf{x}_{t-1} + \underset{(m\times 1)}{\mathbf{u}_{ct}}. \tag{14}$$

Gabaix and Koijen (2024, eq. (2)) posited that $\mathbf{u}_{ct}$ admits a factor structure,

$$\underset{(m\times 1)}{\mathbf{u}_{ct}} = \underset{(m\times r)}{\boldsymbol{\lambda}}\underset{(r\times 1)}{\mathbf{f}_t} + \underset{(m\times 1)}{\boldsymbol{\eta}_{ct}},$$

where $\mathbf{f}_t$ is an $(r \times 1)$ vector of common factors in demand shocks (with $r < m$) and $\boldsymbol{\eta}_{ct}$ are mutually uncorrelated idiosyncratic demand shocks. Their approach relies on finding a $(1 \times m)$ vector $\boldsymbol{\Gamma}$ for which $\boldsymbol{\Gamma}\mathbf{1}_m = 0$ and $\boldsymbol{\Gamma\lambda} = \mathbf{0}$. Here we show that even in the absence of any factor structure (so that the covariance matrix of demand shocks $E(\mathbf{u}_{ct}\mathbf{u}'_{ct}) = \mathbf{D}$ is completely unrestricted), as long as the demand shocks $\mathbf{u}_{ct}$ are uncorrelated with the supply shocks $\mathbf{u}_{qt}$, the granular IV idea developed by Gabaix and Koijen (2024) can still be applied.

To demonstrate this, notice that $\mathbf{s}'_c\mathbf{c}_t$ represents global consumption and let $\mathbf{w}'_c\mathbf{c}_t$ denote some other linear combination of individual countries consumption where $\mathbf{w}_c$ is any $(m \times 1)$ vector satisfying $\mathbf{1}'_m\mathbf{w}_c = 1$. Consider premultiplying (14) by $(\mathbf{s}_c - \mathbf{w}_c)'$:

$$\begin{aligned}(\mathbf{s}_c - \mathbf{w}_c)'\mathbf{c}_t &= (\mathbf{s}_c - \mathbf{w}_c)'\phi_c\mathbf{1}_m p_t + (\mathbf{s}_c - \mathbf{w}_c)'\mathbf{B}_c\mathbf{x}_{t-1} + (\mathbf{s}_c - \mathbf{w}_c)'\mathbf{u}_{ct} \\ &= (\mathbf{s}_c - \mathbf{w}_c)'\mathbf{B}_c\mathbf{x}_{t-1} + (\mathbf{s}_c - \mathbf{w}_c)'\mathbf{u}_{ct}. \end{aligned} \tag{15}$$

Note that under these assumptions, $(\mathbf{s}_c - \mathbf{w}_c)'\mathbf{c}_t$ is uncorrelated with the vector of supply shocks $\mathbf{u}_{qt}$. Thus any of the supply equations in (5) could be estimated by instrumental variables using $(\mathbf{s}_c - \mathbf{w}_c)'\mathbf{c}_t$ and $\mathbf{x}_{t-1}$ as instruments. This is the granular instrumental variable idea of Gabaix and Koijen (2024). A popular approach is to take $\mathbf{w}_c = m^{-1}\mathbf{1}_m$. In this case, granular IV uses the difference between a share-weighted average and an unweighted average of consumption across countries as an instrument for price for purposes of estimating supply elasticities.

Similarly, if we assume that supply elasticities are the same across countries ($\boldsymbol{\phi}_q = \phi_q\mathbf{1}_n$ for ϕ_q the common supply elasticity and $\mathbf{1}_n$ an $(n \times 1)$ vector of ones), we could use $(\mathbf{s}_q - \mathbf{w}_q)'\mathbf{q}_t$ (for example, the difference between production weighted by each country's production share and an unweighted average) as an instrument for any of the demand equations.

If we believed that both demand and supply elasticities are homogeneous across countries, we would want to pool the m individual country demand regressions in (10) to estimate the common value of ϕ_c. This raises the questions, what is the optimal way to pool the regressions and what are the optimal choices for $\mathbf{w}_c$ and $\mathbf{w}_q$? Moreover, under these maintained assumptions, is there a way to generalize the approach to produce estimates of ϕ_c and ϕ_q that are more precise than those that would result from granular IV? Our proposal in this paper is to answer these questions by considering estimation from the perspective of full-information maximum likelihood.

4.2 Maximum likelihood estimation

Since structural dynamics are modeled through the lagged coefficients on $\mathbf{x}_{t-1}$ in (5)-(8), the vector of structural shocks $\mathbf{u}_t$ is uncorrelated with its own lags or lags of $\mathbf{y}_t$. If we were willing to further assume that structural shocks have a Gaussian distribution with $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{D})$, the log likelihood function of the observed data $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\}$ conditional on the initial observations $\mathbf{x}_0$ would be given by

$$\begin{aligned}\ell(\boldsymbol{\theta}, \mathbf{B}) = &-(TN/2)\log(2\pi) + (T/2)\log(|\mathbf{A}|^2) - (T/2)\log|\mathbf{D}| \\ &- (1/2)\textstyle\sum_{t=1}^{T}(\mathbf{A}\mathbf{y}_t - \mathbf{B}\mathbf{x}_{t-1})'\mathbf{D}^{-1}(\mathbf{A}\mathbf{y}_t - \mathbf{B}\mathbf{x}_{t-1}).\end{aligned} \tag{16}$$

Here $\boldsymbol{\theta}$ collects unknown elements of $\mathbf{A}$ and $\mathbf{D}$ and the determinant $|\mathbf{A}|$ is the inverse of the Jacobian of the transformation from $\mathbf{y}_t = (\mathbf{q}_t, \mathbf{c}_t, p_t)'$ to $\mathbf{u}_t$. The principle of maximum likelihood estimation calls for choosing as estimates the values of $\boldsymbol{\theta}$ and $\mathbf{B}$ that maximize the likelihood function (16). An important reason to rely on maximum likelihood for estimation of $\boldsymbol{\theta}$ is that when $\mathbf{y}_t$ is stationary, any other way to use the observed data to produce a consistent and asymptotically Normal estimate of $\boldsymbol{\theta}$ cannot have a smaller asymptotic variance than $\hat{\boldsymbol{\theta}}_{MLE}$.

Finding the values of $\boldsymbol{\theta}$ and $\mathbf{B}$ that maximize (16) is greatly facilitated by using OLS regressions to concentrate the likelihood function. Define $\hat{\boldsymbol{\Pi}}$ to be the $(N \times k)$ matrix of coefficients from OLS regressions of each variable on lags of all the variables:

$$\hat{\boldsymbol{\Pi}} = \left[\textstyle\sum_{t=1}^{T}\mathbf{y}_t\mathbf{x}_{t-1}'\right]\left[\textstyle\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right]^{-1}. \tag{17}$$

Let $\hat{\boldsymbol{\epsilon}}_t$ be the $(N \times 1)$ vector of residuals from these OLS regressions and $\hat{\boldsymbol{\Omega}}$ their $(N \times N)$ sample variance-covariance matrix:

$$\hat{\boldsymbol{\epsilon}}_t = \mathbf{y}_t - \hat{\boldsymbol{\Pi}}\mathbf{x}_{t-1} \tag{18}$$

$$\hat{\boldsymbol{\Omega}} = T^{-1}\textstyle\sum_{t=1}^{T}\hat{\boldsymbol{\epsilon}}_t\hat{\boldsymbol{\epsilon}}_t'. \tag{19}$$

Although many of the details for applications like ours are well known,[4] for the reader's convenience we collect some key results in the form of a proposition, which we prove in Appendix A.4.

[4]See for example Hamilton (1994, pp. 331-332).

Proposition 1. *Define*

$$\eta(\boldsymbol{\theta}) = -(TN/2)\log(2\pi) + (T/2)\log\left[\left(\mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v\right)^2\right]$$
$$- (T/2)\log|\mathbf{D}| - (1/2)\textstyle\sum_{t=1}^{T}(\mathbf{A}\hat{\boldsymbol{\epsilon}}_t)'\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\epsilon}}_t). \quad (20)$$

If there are no restrictions on $\mathbf{B}$, *then the values of* $\boldsymbol{\theta}$ *and* $\mathbf{B}$ *that maximize* (16) *are given by*

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max \eta(\boldsymbol{\theta})$$
$$\hat{\mathbf{B}}_{MLE} = \mathbf{A}(\hat{\boldsymbol{\theta}}_{MLE})\hat{\mathbf{\Pi}}.$$

The function $\eta(\boldsymbol{\theta})$ *can alternatively be calculated as*

$$\eta(\boldsymbol{\theta}) = -(TN/2)\log(2\pi) + (T/2)\log\left[\left(\mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v\right)^2\right]$$
$$- (T/2)\log|\mathbf{D}| - (T/2)\text{trace}\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\hat{\mathbf{\Omega}}\right]. \quad (21)$$

Moreover, the value achieved for (16) *at the maximum likelihood estimates is equal to*

$$\ell(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\mathbf{B}}_{MLE}) = \eta(\hat{\boldsymbol{\theta}}_{MLE}).$$

Proposition 1 tells us that maximum likelihood estimates of elasticities are based on the covariances $\hat{\mathbf{\Omega}}$ between the errors we make in forecasting the observed variables, for example, the covariances between the errors in predicting production or consumption in individual countries and the error in forecasting the price of oil.

The structural parameter $\boldsymbol{\theta}$ is said to be identified if there exists a unique value of $\boldsymbol{\theta}$ that maximizes (21).[5] Identification requires prior information about the structure that implies restrictions on the values $\mathbf{A}$ and $\mathbf{D}$ can take on. A necessary requirement for identification is the order condition that the number of unknown elements of $\mathbf{A}$ and $\mathbf{D}$ is no greater than the number of unique elements in $\hat{\mathbf{\Omega}}$, which is $N(N+1)/2$.

[5] See Definition 2 in Rothenberg (1971). Other researchers take identification to mean that the population parameter $\boldsymbol{\theta}$ can be consistently estimated, which requires the population analog of (21), namely

$$\log\left[\left(\mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v\right)^2\right] - \log|\mathbf{D}| - \text{trace}\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t')\right],$$

to have a unique maximum.

4.3 Maximum likelihood estimation assuming homogeneous elasticities

In this subsection we discuss maximum likelihood estimation in the special case when supply shocks are uncorrelated with demand shocks,

$$\underset{(N\times N)}{\mathbf{D}} = \begin{bmatrix} \underset{(n\times n)}{\mathbf{D}_q} & \mathbf{0}_{nm} & \mathbf{0}_{n1} \\ \mathbf{0}_{mn} & \underset{(m\times m)}{\mathbf{D}_c} & \mathbf{0}_{m1} \\ \mathbf{0}_{1n} & \mathbf{0}_{1m} & \underset{(1\times 1)}{\sigma_v^2} \end{bmatrix}, \tag{22}$$

and elasticities are the same across countries: $\boldsymbol{\phi}_q = \phi_q \mathbf{1}_n$ and $\boldsymbol{\phi}_c = \phi_c \mathbf{1}_m$. Partition the OLS forecasting residuals $\hat{\boldsymbol{\epsilon}}_t$ into the errors made in forecasting each country's production, each country's consumption, and the world price:

$$\underset{(N\times 1)}{\hat{\boldsymbol{\epsilon}}_t} = \begin{bmatrix} \underset{(n\times 1)}{\hat{\boldsymbol{\epsilon}}_{qt}} \\ \underset{(m\times 1)}{\hat{\boldsymbol{\epsilon}}_{ct}} \\ \underset{(1\times 1)}{\hat{\varepsilon}_{pt}} \end{bmatrix}.$$

In this special case (20) becomes

$$\begin{aligned}\eta(\boldsymbol{\theta}) = &-(TN/2)\log(2\pi) + (T/2)\log\left[(\phi_q - \phi_c - \phi_v)^2\right] - (T/2)\log|\mathbf{D}_q| - (T/2)\log|\mathbf{D}_c| \\ &- (T/2)\log(\sigma_v^2) - (1/2)\textstyle\sum_{t=1}^T(\hat{\boldsymbol{\epsilon}}_{qt} - \phi_q\mathbf{1}_n\hat{\varepsilon}_{pt})'\mathbf{D}_q^{-1}(\hat{\boldsymbol{\epsilon}}_{qt} - \phi_q\mathbf{1}_n\hat{\varepsilon}_{pt}) \\ &- (1/2)\textstyle\sum_{t=1}^T(\hat{\boldsymbol{\epsilon}}_{ct} - \phi_c\mathbf{1}_m\hat{\varepsilon}_{pt})'\mathbf{D}_c^{-1}(\hat{\boldsymbol{\epsilon}}_{ct} - \phi_c\mathbf{1}_m\hat{\varepsilon}_{pt}) - (1/2)\textstyle\sum_{t=1}^T\dfrac{(\hat{\varepsilon}_{vt} - \phi_v\hat{\varepsilon}_{pt})^2}{\sigma_v^2}\end{aligned} \tag{23}$$

where

$$\hat{\varepsilon}_{vt} = \mathbf{s}_q'\hat{\boldsymbol{\epsilon}}_{qt} - \mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} \tag{24}$$

and $\boldsymbol{\theta}$ consists of $(\phi_q, \phi_c, \phi_v, \sigma_v)$ and the nonredundant elements of the symmetric matrices $\mathbf{D}_q$ and $\mathbf{D}_c$.

Proposition 2. *Define* $\hat{\mathbf{u}}_{qt} = \hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q\mathbf{1}_n\hat{\varepsilon}_{pt}$, $\hat{\mathbf{u}}_{ct} = \hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c\mathbf{1}_m\hat{\varepsilon}_{pt}$, *and* $\hat{u}_{vt} = \hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt}$. *The first-order conditions for maximization of (23) are given by*

$$\hat{\mathbf{D}}_q = T^{-1}\textstyle\sum_{t=1}^T\hat{\mathbf{u}}_{qt}\hat{\mathbf{u}}_{qt}' \quad \hat{\mathbf{D}}_c = T^{-1}\textstyle\sum_{t=1}^T\hat{\mathbf{u}}_{ct}\hat{\mathbf{u}}_{ct}' \quad \hat{\sigma}_v^2 = T^{-1}\textstyle\sum_{t=1}^T\hat{u}_{vt}^2 \tag{25}$$

$$\hat{\phi}_q = \frac{\sum_{t=1}^{T} \tilde{z}_{qt}\tilde{q}_t}{\sum_{t=1}^{T} \tilde{z}_{qt}\hat{\varepsilon}_{pt}} \tag{26}$$

$$\hat{\phi}_c = \frac{\sum_{t=1}^{T} \tilde{z}_{ct}\tilde{c}_t}{\sum_{t=1}^{T} \tilde{z}_{ct}\hat{\varepsilon}_{pt}} \tag{27}$$

$$\hat{\phi}_v = \frac{\sum_{t=1}^{T} \tilde{z}_{vt}\hat{\varepsilon}_{vt}}{\sum_{t=1}^{T} \tilde{z}_{vt}\hat{\varepsilon}_{pt}} \tag{28}$$

$$\tilde{q}_t = \hat{\mathbf{w}}_q'\hat{\boldsymbol{\epsilon}}_{qt} \qquad \hat{\mathbf{w}}_q' = \mathbf{1}_n'\hat{\mathbf{D}}_q^{-1}/(\mathbf{1}_n'\hat{\mathbf{D}}_q^{-1}\mathbf{1}_n) \tag{29}$$

$$\tilde{c}_t = \hat{\mathbf{w}}_c'\hat{\boldsymbol{\epsilon}}_{ct} \qquad \hat{\mathbf{w}}_c' = \mathbf{1}_m'\hat{\mathbf{D}}_c^{-1}/(\mathbf{1}_m'\hat{\mathbf{D}}_c^{-1}\mathbf{1}_m) \tag{30}$$

$$\tilde{z}_{qt} = (\mathbf{s}_c - \hat{\mathbf{w}}_c)'\hat{\boldsymbol{\epsilon}}_{ct} + (\tilde{c}_t - \hat{\phi}_c\hat{\varepsilon}_{pt}) + (\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt}) \tag{31}$$

$$\tilde{z}_{ct} = -(\mathbf{s}_q - \hat{\mathbf{w}}_q)'\hat{\boldsymbol{\epsilon}}_{qt} - (\tilde{q}_t - \hat{\phi}_q\hat{\varepsilon}_{pt}) + (\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt}) \tag{32}$$

$$\tilde{z}_{vt} = (\mathbf{s}_c - \hat{\mathbf{w}}_c)'\hat{\boldsymbol{\epsilon}}_{ct} - (\mathbf{s}_q - \hat{\mathbf{w}}_q)'\hat{\boldsymbol{\epsilon}}_{qt} - (\tilde{q}_t - \hat{\phi}_q\hat{\varepsilon}_{pt}) + (\tilde{c}_t - \hat{\phi}_c\hat{\varepsilon}_{pt}). \tag{33}$$

Estimating the variances of structural disturbances ($\mathbf{D}_q$, $\mathbf{D}_c$, σ_v^2) given estimates of the elasticities ϕ_q, ϕ_c, and ϕ_v comes from the straightforward and highly intuitive expressions in (25). For example, the MLE of the variance of the structural shock for inventory demand σ_v^2 is just the average squared value of $\hat{u}_{vt}^2$.

Expression (26) characterizes the MLE of ϕ_q as an IV estimate coming from a regression of $\tilde{q}_t$ on $\hat{\varepsilon}_{pt}$ using $\tilde{z}_{qt}$ as an instrument for the price of oil. The first term in the characterization of the instrument $\tilde{z}_{qt}$ in (31) incorporates the insight of granular instrumental variables, specifying $(\mathbf{s}_c - \hat{\mathbf{w}}_c)'\hat{\boldsymbol{\epsilon}}_{ct}$ as one component of the instrument $\tilde{z}_{qt}$ used by MLE. Note that the principle of maximum likelihood estimation instructs us how to aggregate individual production magnitudes in calculating the left-hand variable in the IV regression, namely $\tilde{q}_t = \hat{\mathbf{w}}_q'\hat{\boldsymbol{\epsilon}}_{qt}$.

But maximum likelihood says that we can do much better than just rely on the standard granular instrument based on consumption across different regions. The assumption that we used to motivate granular IV was that supply shocks $\mathbf{u}_{qt}$ are uncorrelated with the demand shocks $\mathbf{u}_{ct} = \boldsymbol{\epsilon}_{ct} - \phi_c\mathbf{1}_m\varepsilon_{pt}$. Thus a consistent estimate of any individual demand shock $\hat{\varepsilon}_{cjt} - \hat{\phi}_c\hat{\varepsilon}_{pt}$ would also be a valid instrument for estimating the supply elasticity. MLE utilizes the particular linear combination $\hat{\mathbf{w}}_c'(\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c\mathbf{1}_m\hat{\varepsilon}_{pt})$. Finally, the model maintains that $\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt}$ is also a valid instrument. MLE combines all the available sources of information about ϕ_q to produce the consistent estimate $\hat{\phi}_q$ with the smallest possible asymptotic variance.[6]

Note that while equations (25)-(33) provide an analytical characterization of the maximum likelihood estimates, they do not offer a closed-form expression. To use (25) to calculate $\hat{\mathbf{D}}_q$ we need to know $\hat{\phi}_q$. And to use (26) to calculate $\hat{\phi}_q$ we need to know $\hat{\mathbf{D}}_q$. Nevertheless, the analytical expressions suggest a constructive algorithm that could be used to find an estimate $\hat{\boldsymbol{\theta}}$ that simultaneously satisfies all the equations and also helps us understand the nature of the identification behind MLE. We could construct an initial estimate of the supply elasticity ϕ_q using standard granular IV with aggregate production $\tilde{q}_t^{(1)} = \mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt}$ as the dependent variable and $\tilde{z}_{qt}^{(1)} = (\mathbf{s}_c - m^{-1}\mathbf{1}_m)' \hat{\boldsymbol{\epsilon}}_{ct}$ as an instrument for price:

$$\hat{\phi}_q^{(1)} = \frac{\sum_{t=1}^T \tilde{z}_{qt}^{(1)} \tilde{q}_t^{(1)}}{\sum_{t=1}^T \tilde{z}_{qt}^{(1)} \hat{\varepsilon}_{pt}}. \tag{34}$$

This estimate is consistent, but does not efficiently use all the available information. Likewise we could get an initial consistent but inefficient estimate of the demand elasticity using

$$\hat{\phi}_c^{(1)} = \frac{\sum_{t=1}^T \tilde{z}_{ct}^{(1)} \tilde{c}_t^{(1)}}{\sum_{t=1}^T \tilde{z}_{ct}^{(1)} \hat{\varepsilon}_{pt}}. \tag{35}$$

with $\tilde{c}_t^{(1)} = \mathbf{s}_c' \hat{\boldsymbol{\epsilon}}_{ct}$ and $\tilde{z}_{ct}^{(1)} = (n^{-1}\mathbf{1}_n - \mathbf{s}_q)' \hat{\boldsymbol{\epsilon}}_{qt}$. We could now use $\tilde{z}_{vt}^{(1)} = \tilde{c}_t^{(1)} - \hat{\phi}_c^{(1)} \hat{\varepsilon}_{pt}$ as an instrument to estimate the inventory elasticity

$$\hat{\phi}_v^{(1)} = \frac{\sum_{t=1}^T \tilde{z}_{vt}^{(1)} (\mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} - \mathbf{s}_c' \hat{\boldsymbol{\epsilon}}_{ct})}{\sum_{t=1}^T \tilde{z}_{vt}^{(1)} \hat{\varepsilon}_{pt}}.$$

With these initial estimates, we can then form

$$\hat{\mathbf{D}}_q^{(1)} = T^{-1} \textstyle\sum_{t=1}^T \left(\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q^{(1)} \mathbf{1}_n \hat{\varepsilon}_{pt} \right) \left(\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q^{(1)} \mathbf{1}_n \hat{\varepsilon}_{pt} \right)'$$

$$\hat{\mathbf{D}}_c^{(1)} = T^{-1} \textstyle\sum_{t=1}^T \left(\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)} \mathbf{1}_m \hat{\varepsilon}_{pt} \right) \left(\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)} \mathbf{1}_m \hat{\varepsilon}_{pt} \right)'$$

$$\left(\hat{\sigma}_v^2 \right)^{(1)} = T^{-1} \textstyle\sum_{t=1}^T (\mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} - \mathbf{s}_c' \hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_v^{(1)} \hat{\varepsilon}_{pt})^2.$$

[6] The instrument $\tilde{z}_{qt}$ can be further simplified by noting that $\tilde{c}_t = \hat{\mathbf{w}}_c' \hat{\boldsymbol{\epsilon}}_{ct}$ and $\hat{\varepsilon}_{vt} = \mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} - \mathbf{s}_c' \hat{\boldsymbol{\epsilon}}_{ct}$:

$$\begin{aligned} \tilde{z}_{qt} &= (\mathbf{s}_c - \hat{\mathbf{w}}_c)' \hat{\boldsymbol{\epsilon}}_{ct} + (\hat{\mathbf{w}}_c' \hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c \hat{\varepsilon}_{pt}) + (\mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} - \mathbf{s}_c' \hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_v \hat{\varepsilon}_{pt}) \\ &= \mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} - (\hat{\phi}_c + \hat{\phi}_v) \hat{\varepsilon}_{pt}. \end{aligned}$$

We write $\tilde{z}_{qt}$ in the form of (31) to better explain the intuition for why it is a valid instrument and the relation to granular IV.

We thus have a consistent initial estimate of the full vector $\hat{\boldsymbol{\theta}}^{(1)}$.

We can then use this value of $\hat{\boldsymbol{\theta}}^{(1)}$ on the right-hand sides of (29) and (30) to construct improved aggregates $\tilde{q}_t^{(2)}$ and $\tilde{c}_t^{(2)}$. These along with $\hat{\phi}_q^{(1)}, \hat{\phi}_c^{(1)}, \hat{\phi}_v^{(1)}$ give improved instruments $\tilde{z}_{qt}^{(2)}$, $\tilde{z}_{ct}^{(2)}$, and $\tilde{z}_{vt}^{(2)}$ from (31)-(33), and then improved estimates $\hat{\phi}_q^{(2)}$, $\hat{\phi}_c^{(2)}$, and $\hat{\phi}_v^{(2)}$ from (26)-(28). Plugging these estimates into (25) completes the new estimate $\hat{\boldsymbol{\theta}}^{(2)}$. Iterating in this fashion until convergence (that is, until $\hat{\boldsymbol{\theta}}^{(j+1)} = \hat{\boldsymbol{\theta}}^{(j)}$) is a constructive algorithm to arrive at the single value of $\hat{\boldsymbol{\theta}}$ that simultaneously solves equations (25)-(33). This is identical to the value $\hat{\boldsymbol{\theta}}$ that could be obtained by maximizing (21) by numerical search, and also identical to the value of $\hat{\boldsymbol{\theta}}$ that would be obtained by maximizing (16) jointly with respect to $\boldsymbol{\theta}$ and $\mathbf{B}$. This is a form of iterated three-stage least squares as a way to calculate the full-information maximum likelihood estimates.

4.4 Advantages of estimation by maximum likelihood

We have seen that the principle of maximum likelihood allows us to use granular information in an optimal way. It moreover instructs us how to aggregate observations across equations to calculate common parameters, how to use all information efficiently, and how to estimate all structural parameters simultaneously. Furthermore, it does this in a way that is invariant with respect to normalization. For example, if we were to take the goal to be to estimate the inverse of the demand elasticity, the MLE of ϕ_q^{-1} is numerically identical to the inverse of the MLE of ϕ_q. This invariance with respect to parameterization is an advantage of MLE that is not shared by typical instrumental variables applications or GMM. Indeed, the parameterization issue is a well known problem associated with IV and GMM, as shown for example by Yogo (2004).

We have characterized MLE analytically for the special case of homogeneous elasticities and supply shocks that are uncorrelated with demand shocks. But one does not need to know the analytical characterization of the first-order conditions in order to implement maximum likelihood estimation. If we simply maximize (21) by numerical search over allowable values of $\boldsymbol{\theta}$, we will arrive at exactly the same estimates. Doing so of course requires that the model, like the one we analyzed above, is identified. A first step in any application is to verify that the order condition for identification is satisfied ($\dim(\boldsymbol{\theta}) \leq N(N+1)/2$) and that $\hat{\boldsymbol{\theta}} = \arg\max \eta(\boldsymbol{\theta})$ is unique.

Maximum likelihood estimation also allows us to use a likelihood ratio test to check whether proposed restrictions on $\boldsymbol{\theta}$ are consistent with the data. If $\hat{\boldsymbol{\theta}}_0$ denotes the MLE of a restricted model and $\hat{\boldsymbol{\theta}}_1$ the MLE when some of the restrictions are relaxed, we can test the null hypothesis that the restrictions are correct by treating

$$2\left[\eta(\hat{\boldsymbol{\theta}}_1) - \eta(\hat{\boldsymbol{\theta}}_0)\right] \tag{36}$$

as approximately χ^2 with degrees of freedom equal to the number of additional parameters estimated in the less restricted model, namely $\dim(\hat{\boldsymbol{\theta}}_1) - \dim(\hat{\boldsymbol{\theta}}_0)$. In this paper we will typically follow the suggestion of Sims (1980, p. 17) and Hamilton (1994, p. 297) of replacing (36) with the small-sample correction

$$\frac{2(T-k)}{T}\left[\eta(\hat{\boldsymbol{\theta}}_1) - \eta(\hat{\boldsymbol{\theta}}_0)\right]. \tag{37}$$

A leading case is to test jointly all the overidentifying restrictions of a proposed model by comparing the value of the likelihood achieved by the model with that for a completely unrestricted model. A model with no restrictions and a perfect fit to the data would result in $\hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}\left(\hat{\mathbf{A}}'\right)^{-1} = \hat{\boldsymbol{\Omega}}$ or from (21),

$$\eta(\hat{\boldsymbol{\theta}}_{unrestricted}) = -(TN/2)[1 + \log(2\pi)] - (T/2)\log\left|\hat{\boldsymbol{\Omega}}\right| \tag{38}$$

where $\hat{\boldsymbol{\Omega}}$ is the covariance matrix of the OLS forecasting residuals in (19). An unrestricted model imposes no restrictions on how production, consumption, and price could covary, that is, no restrictions on $\hat{\boldsymbol{\Omega}}$. The likelihood ratio test of a proposed model is calculated as

$$\frac{2(T-k)}{T}\left[\eta(\hat{\boldsymbol{\theta}}_{unrestricted}) - \eta(\hat{\boldsymbol{\theta}}_{MLE})\right] \tag{39}$$

which if the restrictions are all valid would have an asymptotic χ^2 distribution with degrees of freedom equal to the number of overidentifying restrictions, that is, degrees of freedom equal to $N(N+1)/2 - \dim(\hat{\boldsymbol{\theta}}_{MLE})$. Our specification strategy is to develop a parsimonious structural model whose restrictions are not rejected by a likelihood ratio test.

We calculate approximate standard errors for $\hat{\boldsymbol{\theta}}_{MLE}$ from the square roots of diagonal elements of

$$\hat{\mathbf{V}} = \left[-\left.\frac{\partial^2\eta(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1}. \tag{40}$$

4.5 Empirical estimates assuming homogeneous elasticities

We begin by reporting an estimate of the demand elasticity ϕ_c that results from a simple application of granular IV. We estimated a regression of $\mathbf{s}_c'\mathbf{c}_t$ on p_t and $\mathbf{x}_{t-1}$ using $(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\mathbf{q}_t$ and $\mathbf{x}_{t-1}$ as instruments,

$$\hat{\boldsymbol{\beta}}_c^{(IV)} = \left[\textstyle\sum_{t=1}^T \mathbf{z}_{ct}^{(IV)}\tilde{\mathbf{x}}_t'\right]^{-1}\left[\textstyle\sum_{t=1}^T \mathbf{z}_{ct}^{(IV)}\left(\mathbf{s}_c'\mathbf{c}_t\right)\right], \tag{41}$$

for $\mathbf{z}_{ct}^{IV} = \left[(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\,\mathbf{q}_t, \mathbf{x}_{t-1}'\right]'$ and $\tilde{\mathbf{x}}_t = (p_t, \mathbf{x}_{t-1}')'$. We first note that the resulting estimate of ϕ_c (the first element of $\hat{\boldsymbol{\beta}}_c^{(IV)}$) is in fact numerically identical to the estimate $\hat{\phi}_c^{(1)}$ that results from the first iteration of our iterated 3SLS procedure described in (35).

Proposition 3. *The coefficient on p_t in the granular IV regression (41) is numerically identical to the estimate $\hat{\phi}_c^{(1)}$ in (35).*

This estimate is $\hat{\phi}_c^{(1)} = -0.11$, which is plausible, but the standard error[7] is twice the size of the coefficient estimate; see Table 1. Thus the estimate $\hat{\phi}_c^{(IV)}$, while asymptotically valid, is in practice not very useful for this sample. Even more striking, if we were to estimate the supply elasticity ϕ_q from a regression of $\mathbf{s}_q'\mathbf{q}_t$ on p_t and $\mathbf{x}_{t-1}$ using $\mathbf{z}_{qt}^{(IV)} = \left[(m^{-1}\mathbf{1}_m - \mathbf{s}_c)'\,\mathbf{c}_t, \mathbf{x}_{t-1}'\right]'$ as instruments (or equivalently, simply use the first iteration from 3SLS in equation (34)), the estimate proves to be completely useless with a standard error of 7.7. The reason is that there is not much difference in these data between the share-weighted and unweighted average of consumption across countries.

Rather than stop with the initial estimates in (35) and (34), we next iterated on

[7] We calculated IV standard errors using the usual asymptotic formula, namely the square root of the first diagonal element of

$$s_c^{(IV)}\left[\textstyle\sum_{t=1}^T \mathbf{z}_{ct}^{(IV)}\tilde{\mathbf{x}}_t'\right]^{-1}\left[\textstyle\sum_{t=1}^T \mathbf{z}_{ct}^{(IV)}\mathbf{z}_{ct}^{(IV)\prime}\right]\left[\textstyle\sum_{t=1}^T \tilde{\mathbf{x}}_t\mathbf{z}_{ct}^{(IV)\prime}\right]^{-1}$$

where

$$s_c^{(IV)} = T^{-1}\textstyle\sum_{t=1}^T\left[\mathbf{s}_c'\mathbf{c}_t - \tilde{\mathbf{x}}_t'\hat{\boldsymbol{\beta}}_c^{(IV)}\right]^2.$$

This again turns out to be numerically identical to

$$s_c^{(1)}\frac{\sum_{t=1}^T\left[(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\hat{\boldsymbol{\epsilon}}_{qt}\right]^2}{\left[\sum_{t=1}^T(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\hat{\boldsymbol{\epsilon}}_{qt}\hat{\varepsilon}_{pt}\right]^2}$$

where

$$s_c^{(1)} = T^{-1}\textstyle\sum_{t=1}^T\left[\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt}\right]^2$$

which is the natural standard error associated with (35).

Table 1: Parameter estimates assuming homogeneous elasticities and that supply shocks are uncorrelated with demand shocks

Parameter	IV	MLE	MLE (restricted)
Demand elasticity ϕ_c	-0.106	-0.130	-0.130
	(0.252)	(0.026)	(0.026)
Supply elasticity ϕ_q	-3.699	0.054	0.053
	(7.717)	(0.009)	(0.009)

Notes to Table 1. Standard errors in parentheses. Last column is based on MLE in which structural shocks follow a factor structure as in (42).

the 3SLS system of equations to find the maximum likelihood estimates. These are reported in the second column of Table 1. The MLE of the demand elasticity ϕ_c is in fact not that different from the IV estimate, but the standard error is much lower.[8] We can have considerably more confidence in the estimate once we combine all the different sources of information that are available to confirm it, namely, all the terms in (32). Using all the available information makes a world of difference for estimating the supply elasticity. The reason MLE is so much better than IV for estimating the supply elasticity is that we can get a fairly good estimate of the demand elasticity $\hat{\phi}_c$ and thus good estimates of $\hat{\boldsymbol{\varepsilon}}_{ct} - \hat{\phi}_c \mathbf{1}_m \hat{\varepsilon}_{pt}$. Using these demand shocks as additional instruments for estimating the supply elasticity (i.e., the second term in (31)) is extremely helpful. Note that the assumption that the demand shocks $\mathbf{u}_{ct}$ are uncorrelated with supply shocks was the key assumption that we used above to verify that the granular difference $(m^{-1}\mathbf{1}_m - \mathbf{s}_c)'\mathbf{c}_t$ could serve as a valid instrument for estimating the supply elasticity. MLE uses all the implications of this underlying assumption to get much better estimates. When we use all the available information, we can say with confidence that the supply elasticity is small but positive (see Table 1).

[8] To calculate standard errors, we parameterized $\mathbf{D}_q = \mathbf{H}_q\mathbf{H}_q'$ and $\mathbf{D}_c = \mathbf{H}_c\mathbf{H}_c'$ with

$$\mathbf{H}_q = \begin{bmatrix} h_{11}^q & 0 & 0 & 0 \\ h_{21}^q & h_{22}^q & 0 & 0 \\ h_{31}^q & h_{32}^q & h_{33}^q & 0 \\ h_{41}^q & h_{42}^q & h_{43}^q & h_{44}^q \end{bmatrix} \quad \mathbf{H}_c = \begin{bmatrix} h_{11}^c & 0 & 0 & 0 \\ h_{21}^c & h_{22}^c & 0 & 0 \\ h_{31}^c & h_{32}^c & h_{33}^c & 0 \\ h_{41}^c & h_{42}^c & h_{43}^c & h_{44}^c \end{bmatrix}.$$

We then let $\boldsymbol{\theta}$ be the (24×1) vector $(\phi_q, \phi_c, \phi_v, \sigma_v, \text{vech}(\mathbf{H}_q)', \text{vech}(\mathbf{H}_c)')$ to calculate second derivatives of (21) with respect to each element in this vector.

4.6 Specification tests

Maximum likelihood also give us convenient tools for testing any proposed specification. For example, in the model described above we assumed that the variance-covariance matrices for supply shocks $\mathbf{D}_q$ and demand shocks $\mathbf{D}_c$ were completely unrestricted. We could consider a more parsimonious specification if we conjecture that the covariances between the supply shocks in different regions have a factor structure:

$$\mathbf{D}_q = \mathbf{\Sigma}_q + \boldsymbol{\gamma}_q \boldsymbol{\gamma}_q' \tag{42}$$

where $\mathbf{\Sigma}_q$ is a diagonal $(n \times n)$ matrix and $\boldsymbol{\gamma}_q$ is an $(n \times 1)$ vector. The unrestricted general model of $\mathbf{D}_q$ used in (22) requires $n(n+1)/2 = 10$ parameters, whereas (42) needs only the n diagonal elements of $\mathbf{\Sigma}_q$ and the n elements of $\boldsymbol{\gamma}_q$, a total of 8 parameters. Assuming a similar factor structure for the demand shocks eliminates two more parameters. Comparing the model with unrestricted covariances with a model in which covariances admit a factor structure leads to a likelihood ratio test statistic in (37) of 2.48. Comparing this with a $\chi^2(4)$ distribution gives a p-value of 0.65. The hypothesis that shocks admit a factor structure is consistent with the observed data. Imposing a factor structure turns out to result in virtually no change in the maximum likelihood estimates; see the last column of Table 1.

More importantly, we can directly test the assumptions that elasticities are homogeneous and that demand shocks are uncorrelated with supply shocks by testing that specification against an unrestricted model. Unfortunately, the likelihood ratio test (39) produces a χ^2 statistic of 142.9 with $45 - 24 = 21$ degrees of freedom– the data overwhelmingly reject the specification. What feature of the observed data leads to this conclusion? The specification requires that there exist scalars ϕ_q and ϕ_c such that the 16 elements of the matrix

$$E(\boldsymbol{\epsilon}_{qt} - \phi_q \mathbf{1}_n \varepsilon_{pt})(\boldsymbol{\epsilon}_{ct} - \phi_c \mathbf{1}_m \varepsilon_{pt})' = \mathbf{0}_{nm}$$

are all zero. We can estimate the elements of $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t')$ sufficiently accurately from $\hat{\mathbf{\Omega}} = T^{-1} \sum_{t=1}^{T} \hat{\boldsymbol{\epsilon}}_t \hat{\boldsymbol{\epsilon}}_t'$ to conclude that no values of ϕ_q, ϕ_c exist that could make

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{nm} & -\phi_q \mathbf{1}_n \end{bmatrix} \hat{\mathbf{\Omega}} \begin{bmatrix} \mathbf{0}_{nm} \\ \mathbf{I}_m \\ -\phi_c \mathbf{1}_m' \end{bmatrix}$$

plausibly close to zero.

We can generalize the model by allowing each country to have different demand and supply elasticities. If demand shocks are nevertheless still uncorrelated with supply shocks, the specification implies that there exist an $(n \times 1)$ vector $\boldsymbol{\phi}_q$ and an $(m \times 1)$ vector $\boldsymbol{\phi}_c$ such that

$$E(\boldsymbol{\epsilon}_{qt} - \boldsymbol{\phi}_q \varepsilon_{pt})(\boldsymbol{\epsilon}_{ct} - \boldsymbol{\phi}_c \varepsilon_{pt})' = \mathbf{0}_{nm}.$$

This still involves some testable overidentifying assumptions, in that there are $m+n = 8$ parameters in $\boldsymbol{\phi}_q$ and $\boldsymbol{\phi}_c$ to be chosen to set $nm = 16$ observed moments of the data to zero. This hypothesis also turns out to be rejected by the data. We conclude that to describe the data, one needs to allow for some correlation between supply and demand shocks.

4.7 A structural model of the world oil market that is consistent with the observed data

We allow for correlation between supply and demand shocks by hypothesizing the existence of a global shock f_t that influences both supply and demand. For example, a global economic downturn might affect both producers and consumers. In addition, we allow for a second global shock f_{ct} that affects consumption in every country and a third global shock f_{qt} that affects production in every country. Under this specification, the structural disturbances in equations (9) and (10) are characterized by

$$\mathbf{u}_{qt} = \mathbf{h}_q f_t + \boldsymbol{\gamma}_q f_{qt} + \boldsymbol{\eta}_{qt} \tag{43}$$

$$\mathbf{u}_{ct} = \mathbf{h}_c f_t + \boldsymbol{\gamma}_c f_{ct} + \boldsymbol{\eta}_{ct}. \tag{44}$$

Here $\mathbf{h}_q$ is an $(n \times 1)$ vector whose ith element summarizes how the global shock f_t affects production in country i. Since f_t is not observed directly and only affects the observed data through the product $\mathbf{h}_q f_t$, a normalization is needed for the units in which f_t is measured. We follow a common normalization in factor models of specifying that f_t has unit variance. Likewise, $\boldsymbol{\gamma}_q$ is an $(n \times 1)$ vector whose ith element summarizes how the global supply shock f_{qt} affects production in country i with $E(f_{qt}^2)$ again normalized at one. We also follow the standard representation that the factors f_t and f_{qt} are uncorrelated with each other. This is again just a normalization; given a proposed second factor $\tilde{f}_{qt}$ that is correlated with f_t, one can define a new factor f_{qt} as the residual from a projection of $\tilde{f}_{qt}$ on f_t with the same observable implications. Appendix A.7 describes how we implement this normalization in practice. Finally, we hypothesize that the shocks $\boldsymbol{\eta}_{qt}$ are purely idiosyncratic supply shocks that are uncorrelated across countries: $E(\boldsymbol{\eta}_{qt}\boldsymbol{\eta}'_{qt}) = \boldsymbol{\Sigma}_q$, a diagonal matrix. Similar assumptions for the terms in (44) imply a variance-covariance matrix of structural shocks given by

$$\mathbf{D} = E(\mathbf{u}_t\mathbf{u}'_t) = \begin{bmatrix} \mathbf{h}_q\mathbf{h}'_q + \boldsymbol{\gamma}_q\boldsymbol{\gamma}'_q + \boldsymbol{\Sigma}_q & \mathbf{h}_q\mathbf{h}'_c & \mathbf{0}_{n1} \\ \mathbf{h}_c\mathbf{h}'_q & \mathbf{h}_c\mathbf{h}'_c + \boldsymbol{\gamma}_c\boldsymbol{\gamma}'_c + \boldsymbol{\Sigma}_c & \mathbf{0}_{m1} \\ \mathbf{0}_{1n} & \mathbf{0}_{1m} & \sigma_v^2 \end{bmatrix} \tag{45}$$

with $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_c$ diagonal matrices. Note this is a restriction relative to (22) in that it imposes a factor structure on $\mathbf{D}_q$ and $\mathbf{D}_c$ but also generalizes (22) by allowing for nonzero correlations between $\mathbf{u}_{qt}$ and $\mathbf{u}_{ct}$ arising from the common global factor f_t. Our favored specification also allows each country to have different supply and demand elasticities with $\boldsymbol{\phi}_q$ and $\boldsymbol{\phi}_c$ unrestricted vectors.

The model thus has $(n+m+1)$ elasticity parameters represented by $\{\boldsymbol{\phi}_q, \boldsymbol{\phi}_c, \phi_v\}$, $(n+m+1)$ parameters that capture the variances of idiosyncratic shocks (namely σ_v^2 and the diagonal elements of $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_c$), $(n+m)$ parameters in $\mathbf{h}_q$ and $\mathbf{h}_c$ that capture the effects of the global factor on supply and demand, $(n-1)$ parameters in $\boldsymbol{\gamma}_q$, and $(m-1)$ parameters in $\boldsymbol{\gamma}_c$. Note that only $(n-1)$ parameters are needed to describe the $(n \times 1)$ vector $\boldsymbol{\gamma}_q$ since we estimate the model subject to the orthogonality normalizations $\mathbf{h}'_q\boldsymbol{\gamma}_q = \mathbf{h}'_c\boldsymbol{\gamma}_c = 0$ as described in Appendix A.7. The vector $\boldsymbol{\theta}$ thus contains $4(n+m)$ parameters, or 32 parameters in the case when $n = m = 4$. For comparison, an unrestricted model describes $\hat{\boldsymbol{\Omega}}$ perfectly with $(n+m+1)(n+m+2)/2 = 45$ parameters. The proposed model thus has $45 - 32 = 13$ overidentifying restrictions that can be tested.

We maximized the likelihood (21) with $\mathbf{A}$ given by (13) and $\mathbf{D}$ given by (45) by numerical search over values of $\boldsymbol{\theta}$. The resulting value for $\eta(\hat{\boldsymbol{\theta}}_{MLE})$ is $-12,636.25$. This compares with a likelihood for an unrestricted model of $\eta(\hat{\boldsymbol{\theta}}_{unrestricted}) = -12,625.38$ and results in a value for the test statistic (37) of 17.47. Comparing this with a $\chi^2(13)$ distribution gives a p-value of 0.18. We conclude that the restrictions in (45) are consistent with the observed data.

The third factor f_{qt} does not contribute much to the fit. Setting $\boldsymbol{\gamma}_q = \mathbf{0}$ frees up three more degrees of freedom but still achieves almost as high a log likelihood $(-12,637.54)$. The statistic (37) for testing all the overidentifying restrictions of this further restricted specification generates a $\chi^2(16)$ test statistic with a p-value of 0.24. The specification (45) with $\boldsymbol{\gamma}_q = \mathbf{0}$ is used for the results reported in Section 5. We obtained very similar results to those reported below when we estimated a three-factor model that did not impose $\boldsymbol{\gamma}_q = \mathbf{0}$.

4.8 Estimating the contemporaneous effects of structural shocks on impact

An alternative to the structural model developed above is to start with simple equations to forecast each variable based on the r most recent values of all the variables:

$$\mathbf{y}_t = \mathbf{\Pi}\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t. \tag{46}$$

A reduced-form approach to estimate the forecasting coefficients $\mathbf{\Pi}$ is by the OLS regressions in (17), which minimize the average squared forecast errors. Our structural model of the oil market implies that the forecast errors $\boldsymbol{\epsilon}_t$ can be interpreted as resulting from linear combinations of the structural shocks $\mathbf{u}_t$. Specifically, premultiplying (12) by $\mathbf{A}^{-1}$ puts (12) in the form of (46) with

$$\boldsymbol{\epsilon}_t = \mathbf{A}^{-1}\mathbf{u}_t.$$

From equations (43) and (44) we have

$$\mathbf{u}_t = \begin{bmatrix} \mathbf{h}_q f_t + \boldsymbol{\gamma}_q f_{qt} + \boldsymbol{\eta}_{qt} \\ \mathbf{h}_c f_t + \boldsymbol{\gamma}_c f_{ct} + \boldsymbol{\eta}_{ct} \\ u_{vt} \end{bmatrix}$$

and for the matrix $\mathbf{A}$ given by (13) it turns out that

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{I}_n - \alpha\boldsymbol{\phi}_q\mathbf{s}_q' & \alpha\boldsymbol{\phi}_q\mathbf{s}_c' & \alpha\boldsymbol{\phi}_q \\ -\alpha\boldsymbol{\phi}_c\mathbf{s}_q' & \mathbf{I}_m + \alpha\boldsymbol{\phi}_c\mathbf{s}_c' & \alpha\boldsymbol{\phi}_c \\ -\alpha\mathbf{s}_q' & \alpha\mathbf{s}_c' & \alpha \end{bmatrix} \tag{47}$$

$$\alpha = \left(\mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v\right)^{-1} \tag{48}$$

as can be verified by observing that the product of (13) with (47) is indeed the identity matrix.

We can use these equations to calculate the effect on impact of any structural shock u_{kt} on the $(n + m + 1) \times 1$ vector of observed variables $\mathbf{y}_t$ at time t using $\partial\mathbf{y}_t/\partial u_{kt} = \partial\boldsymbol{\epsilon}_t/\partial u_{kt} = \mathbf{A}^{-1}\partial\mathbf{u}_t/\partial u_{kt}$. For example, a one-percent increase in the supply from country i alone resulting from a one-unit increase in η_{qit} would increase

$\mathbf{u}_{qt}$ by $\mathbf{e}_i^{(n)}$, the ith column of $\mathbf{I}_n$. Thus

$$\frac{\partial \boldsymbol{\epsilon}_t}{\partial \eta_{qit}} = \mathbf{A}^{-1} \begin{bmatrix} \mathbf{e}_i^{(n)} \\ \mathbf{0} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{e}_i^{(n)} - \alpha \boldsymbol{\phi}_q \mathbf{s}_q' \mathbf{e}_i^{(n)} \\ -\alpha \boldsymbol{\phi}_c \mathbf{s}_q' \mathbf{e}_i^{(n)} \\ -\alpha \mathbf{s}_q' \mathbf{e}_i^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_i^{(n)} - \alpha \boldsymbol{\phi}_q s_{qi} \\ -\alpha \boldsymbol{\phi}_c s_{qi} \\ -\alpha s_{qi} \end{bmatrix} \tag{49}$$

where s_{qi} is the share of country i in global production. In equilibrium the increase in supply lowers the world price by $-\alpha s_{qi}$. This leads to higher consumption in country j of $-\alpha \phi_{cj} s_{qi}$ where $\phi_{cj} < 0$ is the demand elasticity of country j and lowers production of country $k \neq i$ by $-\alpha \phi_{qk} s_{qi}$ where $\phi_{qk} > 0$ is the supply elasticity of country k. Similarly, a 1% increase in the demand for country j alone has an effect on impact of

$$\frac{\partial \boldsymbol{\epsilon}_t}{\partial \eta_{cjt}} = \mathbf{A}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_j^{(m)} \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \boldsymbol{\phi}_q s_{cj} \\ \mathbf{e}_j^{(m)} + \alpha \boldsymbol{\phi}_c s_{cj} \\ \alpha s_{cj} \end{bmatrix}$$

for $\mathbf{e}_j^{(m)}$ the jth column of $\mathbf{I}_m$ and s_{cj} the share of country j in world consumption. For example, the increased consumption raises the world price by αs_{cj}. A one-standard-deviation increase in the global demand factor f_{ct} raises $\mathbf{u}_{ct}$ by $\boldsymbol{\gamma}_c$ with equilibrium impact effects given by

$$\frac{\partial \boldsymbol{\epsilon}_t}{\partial f_{ct}} = \mathbf{A}^{-1} \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\gamma}_c \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \boldsymbol{\phi}_q \mathbf{s}_c' \boldsymbol{\gamma}_c \\ \boldsymbol{\gamma}_c + \alpha \boldsymbol{\phi}_c \mathbf{s}_c' \boldsymbol{\gamma}_c \\ \alpha \mathbf{s}_c' \boldsymbol{\gamma}_c \end{bmatrix}. \tag{50}$$

Equation (49) assumes that when the supply curve for country i shifts by 1%, in response to the induced αs_{qi} percent drop in global price, supply from i then decreases by $\phi_i \alpha s_{qi}$ as we move back down along the new supply curve, so that the net change in country i's production is only $1 - \phi_i \alpha s_{qi}$. In some cases we might instead be interested in the effects of a net increase in country i's production of 1% without the endogenous response of country i to the change in price. In this case, the world price would fall by $\alpha_i s_{qi}$ where

$$\alpha_i = \frac{1}{\sum_{k=1,k\neq i}^{n} \phi_{qk} s_{qk} - \sum_{j=1}^{m} \phi_{cj} s_{cj} - \phi_v}. \tag{51}$$

The change in quantity consumed by country j would in this case be $-\alpha_i s_{ci} \phi_{cj}$ and the change in quantity produced by country $k \neq i$ would be $-\alpha_i s_{qi} \phi_{qk}$.

4.9 Dynamic effects of structural shocks

If r lags are enough to capture forecasting dynamics, one can use the reduced-form OLS coefficients $\hat{\mathbf{\Pi}}$ to summarize how new information about any variable causes us to revise a forecast of what will happen s periods later:

$$\mathbf{\Psi}_s = \frac{\partial E(\mathbf{y}_{t+s}|\boldsymbol{\epsilon}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-m})}{\partial \boldsymbol{\epsilon}_t'}.$$

The value of $\mathbf{\Psi}_s$ can be estimated from the value of $\hat{\mathbf{\Pi}}$ using well-known recursions (see for example Hamilton (1994, eq (10.1.19))). We can then calculate a structural dynamic multiplier interpreted as the answer to the following question: if there is a structural shock to u_{kt}, how does this cause us to change our forecast of $\mathbf{y}_{t+s}$:

$$\frac{\partial E(\mathbf{y}_{t+s}|\mathbf{y}_t, \mathbf{y}_{t-1}, ...\mathbf{y}_{t-r+1})}{\partial u_{kt}} = \mathbf{\Psi}_s \frac{\partial \boldsymbol{\epsilon}_t}{\partial u_{kt}}.$$

To get this, we just plug in one of the above expressions for $\partial \boldsymbol{\epsilon}_t / \partial u_{kt}$. For example,

$$\frac{\partial E(\mathbf{y}_{t+s}|\mathbf{y}_t, \mathbf{y}_{t-1}, ...\mathbf{y}_{t-r+1})}{\partial f_{ct}} = \mathbf{\Psi}_s \mathbf{A}^{-1} \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\gamma}_c \\ 0 \end{bmatrix}. \tag{52}$$

We can calculate confidence bands for these estimates as described in Appendix A.8.

5 Empirical results

To summarize, our observations for month t are represented by the $N = 9 \times 1$ vector $\mathbf{y}_t$ consisting of production growth rates in our four regions, consumption growth rates in our four regions, and the percent change in the real price of oil. The dependent variable is observed from $t =$ 1975:M1 through 2020:M2 and 2022:M2 through 2023:M2 for a total number of observations of $T = 555$. The vector of predictor variables $\mathbf{x}_{t-1}$ is a $k = 109 \times 1$ vector consisting of a constant term and 12 lags of each of the nine variables. Our structural model is given by equations (9)-(11). Identification is achieved by modeling the correlations between structural shocks using the factor structure

$$\mathbf{D} = E \begin{bmatrix} \mathbf{u}_{qt}\mathbf{u}_{qt}' & \mathbf{u}_{qt}\mathbf{u}_{ct}' & \mathbf{u}_{qt}u_{vt} \\ \mathbf{u}_{ct}\mathbf{u}_{qt}' & \mathbf{u}_{ct}\mathbf{u}_{ct}' & \mathbf{u}_{ct}u_{vt} \\ u_{vt}\mathbf{u}_{qt}' & u_{vt}\mathbf{u}_{ct}' & u_{vt}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{h}_q\mathbf{h}_q' + \mathbf{\Sigma}_q & \mathbf{h}_q\mathbf{h}_c' & \mathbf{0}_{n1} \\ \mathbf{h}_c\mathbf{h}_q' & \mathbf{h}_c\mathbf{h}_c' + \boldsymbol{\gamma}_c\boldsymbol{\gamma}_c' + \mathbf{\Sigma}_c & \mathbf{0}_{m1} \\ \mathbf{0}_{1n} & \mathbf{0}_{1m} & \sigma_v^2 \end{bmatrix}$$

with $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_c$ diagonal matrices. As shown in Section 4.7, this specification implies 16 testable overidentifying restrictions that are not rejected by the data.

The first column of Table 2 reports the parameter values that maximize the likelihood (21).[9] We estimate a short-run price elasticity of oil supply of 0.25 for Saudi Arabia and 0.02 to 0.07 for other countries. These are all estimated to be positive without the need to impose sign constraints. Using equation (3), these estimates imply an overall world oil short-run supply elasticity of

$$\phi_q = \textstyle\sum_{i=1}^{n} s_{qi}\phi_{qi} = 0.077. \tag{53}$$

This is close to the estimate of 0.054 that we obtained in Table 1 when we assumed that the supply elasticity was the same across all countries. Our estimate of 0.077 is also very similar to the estimates of Caldara, Cavallo and Iacoviello (2019, Table 3) of 0.05 to 0.08 and of Balke, Jin and Yücel (2024) of 0.07 to 0.10, but a little below the estimate in Baumeister and Hamilton (2019) of 0.15. Our country-specific elasticities further support the conclusions of previous researchers including Pierru, Smith and Zamrik (2018), Caldara, Cavallo and Iacoviello (2019), Alonso-Alvarez, Di Nino and Venditti (2022), Almutairi, Pierru and Smith (2023), and Balke, Jin and Yücel (2024) that OPEC production is much more responsive to price than production outside of OPEC.

Our estimates of the short-run price elasticity of demand range from -0.001 for Japan to -0.20 for Europe. These are all estimated to be negative without imposing any sign constraints, though the Japanese estimate is indistinguishable from zero. From equation (4), our estimates imply a global short-run price elasticity of demand of

$$\phi_c = \textstyle\sum_{j=1}^{m} s_{cj}\phi_{cj} = -0.119. \tag{54}$$

This is almost identical to the estimate of -0.130 that we obtained in Table 1 when we assumed that demand elasticity was the same across all countries. Our estimate concludes that demand is a little more responsive to price than the estimates of -0.03 to -0.08 in Cooper (2003), Caldara, Cavallo and Iacoviello (2019, Table 3), and Pierru, Smith and Zamrik (2018) and a little less responsive to price than the estimates obtained by Balke, Jin and Yücel (2024) (-0.17), Alonso-Alvarez, Di Nino and Venditti (2022) (-0.28), and Baumeister and Hamilton (2019) (-0.35).

Shocks to Saudi Arabian production (whose standard deviation is $\sigma_{q,Saudi} = 6.3$)

[9] An earlier version of this paper allowed for separate measurement errors in the individual country production and consumption observations and arrived at similar estimates to those reported here.

Table 2: Maximum likelihood estimates

	Full sample		Post 2005	
Parameter	MLE	Std err	MLE	Std err
$\phi_{q,US}$	0.021	(0.016)	0.063	(0.032)
$\phi_{q,Saudi}$	0.248	(0.058)	0.134	(0.034)
$\phi_{q,Russia}$	0.034	(0.010)	0.017	(0.012)
$\phi_{q,ROW}$	0.066	(0.020)	0.023	(0.010)
$\phi_{c,US}$	-0.077	(0.025)	-0.063	(0.034)
$\phi_{c,Japan}$	-0.001	(0.031)	-0.029	(0.049)
$\phi_{c,Europe}$	-0.202	(0.037)	-0.247	(0.051)
$\phi_{c,ROW}$	-0.139	(0.038)	-0.145	(0.042)
$\sigma_{q,US}$	2.508	(0.076)	1.949	(0.104)
$\sigma_{q,Saudi}$	6.321	(0.256)	1.023	(0.184)
$\sigma_{q,Russia}$	1.529	(0.046)	0.638	(0.036)
$\sigma_{q,ROW}$	1.331	(0.314)	0.414	(0.032)
$\sigma_{c,US}$	1.935	(0.107)	1.080	(0.075)
$\sigma_{c,Japan}$	3.067	(0.120)	1.743	(0.110)
$\sigma_{c,Europe}$	3.492	(0.152)	1.277	(0.170)
$\sigma_{c,ROW}$	2.460	(0.172)	0.951	(0.079)
$h_{q,US}$	-0.028	(0.142)	0.294	(0.190)
$h_{q,Saudi}$	2.430	(0.452)	1.514	(0.171)
$h_{q,Russia}$	0.125	(0.085)	0.292	(0.062)
$h_{q,ROW}$	1.634	(0.275)	0.375	(0.051)
$h_{c,US}$	-0.120	(0.118)	0.263	(0.113)
$h_{c,Japan}$	-0.298	(0.151)	0.380	(0.161)
$h_{c,Europe}$	-0.167	(0.139)	-0.255	(0.091)
$h_{c,ROW}$	1.061	(0.258)	-0.193	(0.120)
ϕ_v	-0.355	(0.061)	-0.187	(0.040)
σ_v	2.825	(0.322)	1.196	(0.148)
$\gamma_{c,US}$	1.367	(0.425)	0.699	(0.783)
γ_{Japan}	1.495	(0.499)	1.023	(0.622)
γ_{Europe}	1.981	(0.537)	1.621	(0.638)
γ_{ROW}	0.881	(0.321)	0.651	(0.665)
α	1.813	(0.101)	2.811	(0.202)
ϕ_q	0.077	(0.017)	0.041	(0.012)
ϕ_c	-0.119	(0.030)	-0.128	(0.037)

Notes to Table 2. The four elements of $\boldsymbol{\gamma}_c$ were not estimated directly but were calculated from the three elements of $\boldsymbol{\omega}_c$ (not reported in the table) along with the four elements of $\mathbf{h}_c$ reported in the table using equations (A26) and (A25). Standard errors for $\boldsymbol{\gamma}_c$ were calculated by simulating draws from the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ as a byproduct of the algorithm used to calculate confidence bands for impulse-response functions. The values of α, ϕ_q and ϕ_c were not estimated directly but were inferred from $\hat{\boldsymbol{\theta}}$ using equations (48), (53), and (54) with standard errors for $\hat{\alpha}$, $\hat{\phi}_q$, and $\hat{\phi}_c$ obtained by simulation. Full sample: 1975:M1-2023:M2 (excluding COVID); Post-2005: 2006:M1-2023:M2 (excluding COVID).

are significantly larger than those to other producing countries and also larger than any shocks to consumption demand. The variance of shocks to inventory demand is comparable in magnitude to the variances of demand shocks for individual countries, and inventory demand is more responsive to price than is the product demand from any individual country. The feature of the data leading to the estimate of $\hat{\phi}_v < 0$ is the observation that the correlations between price and share-weighted consumption and production are smaller than they would be in the absence of adjustment of inventories. We interpret these estimates as consistent with the view that inventory investment responds to price changes in a stabilizing way. A shortfall in supply from any producing country is partially met by selling out of inventories.

The global factor f_t shows up primarily as an increase in demand from the rest of the world and an increase in production from Saudi Arabia and the rest of the world. One possibility is that this represents a direct response of OPEC to strong global demand that is not mediated through price changes. The coefficients $\boldsymbol{\gamma}_c$ on the global demand factor are similar across countries. A one-standard-deviation increase in f_{ct} leads to a 1-2% increase in oil demand everywhere in the world.

6 Applications

In this section we use our model to analyze the effects of certain structural shocks.

6.1 Example 1: The effects of a global demand shock

We first examine the effects of a one-standard-deviation shock to the global demand factor f_{ct}. This raises demand for country j by γ_{cj}, which is around 1-2% for every country. From the last row of equation (50), this leads to an immediate increase in the price of oil of $\alpha \mathbf{s}_c' \boldsymbol{\gamma}_c$, which equals 2.055%. These immediate impact effects are summarized in column 1 of Table 3.

The change in price in turn induces responses of quantities produced and consumed. Column 2 of Table 3 calculates the size of these responses by multiplying the price change 2.055 by the respective elasticities ϕ_{qi} or ϕ_{cj}. Saudi oil production increases by about 0.5% in response to the higher world demand. The price increase also substantially reduces the effect of the demand increase on realized consumption. The net effect (column 3) is the sum of columns 1 and 2.

Figure 1 plots the dynamic effects of the shock calculated using expression (52). Production of oil from Saudi Arabia and the rest of the world continue to climb in the first few months following the shock. These estimates support the conclusion of Almutairi, Pierru and Smith (2023) that Saudi Arabia and OPEC play a major

Table 3: Impact effects of a global demand shock

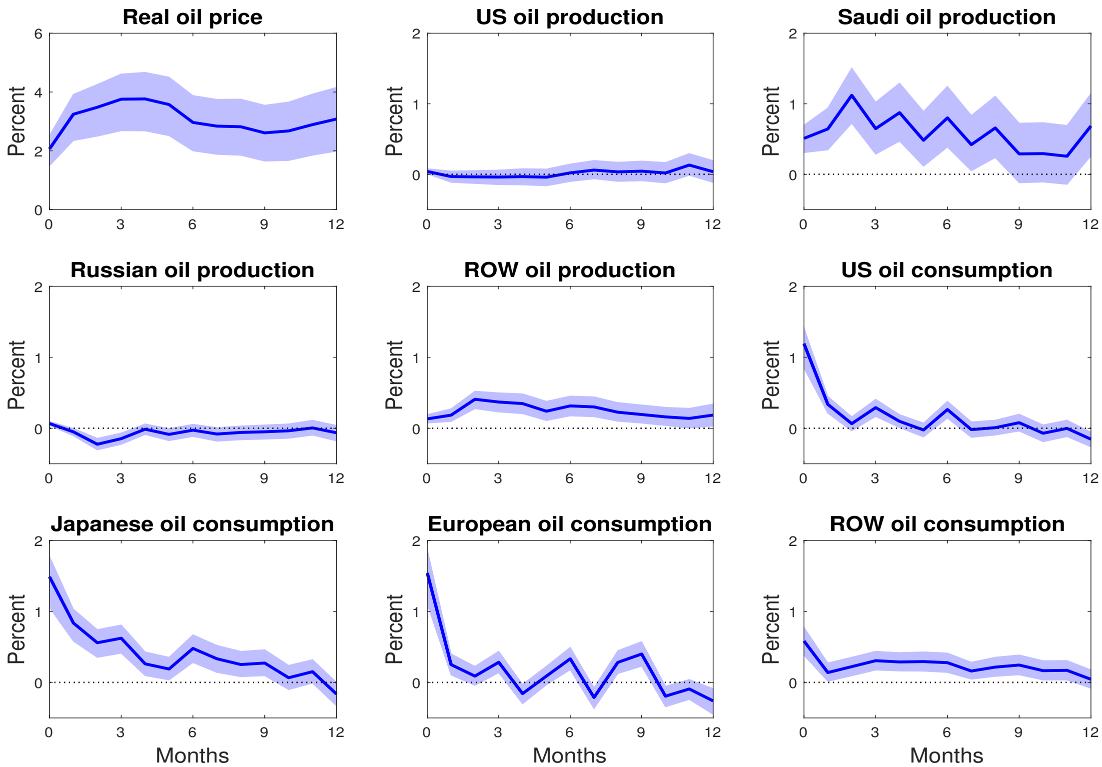
Variable	with estimated ϕ_v				with $\phi_v = 0$			
	as % of country			% of world	as % of country			% of world
	direct effect (1)	response to price (2)	net effect (3)	net effect (4)	direct effect (5)	response to price (6)	net effect (7)	net effect (8)
p	2.055				5.766			
q_{US}	0	0.044	0.044	0.005	0	0.122	0.122	0.015
q_{Saudi}	0	0.509	0.509	0.061	0	1.429	1.429	0.172
q_{Russia}	0	0.070	0.070	0.010	0	0.196	0.196	0.029
q_{ROW}	0	0.135	0.135	0.082	0	0.378	0.378	0.231
q				0.159				0.446
c_{US}	1.367	-0.159	1.208	0.302	1.367	-0.446	0.921	0.230
c_{Japan}	1.495	-0.002	1.493	0.105	1.495	-0.005	1.490	0.104
c_{Europe}	1.981	-0.416	1.565	0.125	1.981	-1.166	0.815	0.065
c_{ROW}	0.881	-0.286	0.595	0.357	0.881	-0.804	0.078	0.047
c				0.889				0.446
v				0.730				0.000

Notes to Table 3. Impact effects of a one-standard-deviation increase in the global demand factor f_{ct} both given the historical average response of inventories (columns 1-4) and under the counterfactual of no adjustment of inventories (columns 5-8).

role in stabilizing the world oil market. The effects of the shock on consumption of individual countries dies out relatively quickly.

Column 4 of Table 3 restates the magnitudes as a percent of the world total by multiplying the entries in column 3 by s_{qi} or s_{cj}. The total initial gains in production (the sum of the first four rows of column 4) only amount to a 0.16% increase in global production, compared with a 0.89% increase in consumption. Thus sales out of inventory play a major role in meeting the temporarily strong demand. Columns 5-8 of Table 3 report what the response to the demand shock would be if there were no changes in inventories, which can be calculated by setting $\phi_v = 0$. The immediate impact on price in that scenario would be $\tilde{\alpha}\mathbf{s}_c'\boldsymbol{\gamma}_c$ where $\tilde{\alpha} = 1/(\boldsymbol{\phi}_q'\mathbf{s}_q - \boldsymbol{\phi}_c'\mathbf{s}_c) = 5.09$. In this counterfactual, the demand increase would lead to a 5.8% increase in prices, almost three times as large as in column 1. If there is no inventory response, the increase in production (0.45% of world supply in column 8 of Table 3) would of necessity exactly equal the increase in world consumption. Comparing column 8 with column 4, most of the balancing in this case comes from the demand side, with the effect of price increases undoing much of the original stimulus to demand.

Figure 1: Dynamic effects of a global demand shock



Notes to Figure 1. Dynamic effects of a one-standard-deviation increase in the global demand factor f_{ct} assuming the historical average response of inventories. First panel plots the cumulative effect on 100 times the log of the real price of oil. Other panels plot year-over-year changes of quantities as a percent of that country's production or consumption. Solid lines are median draws from the asymptotic distribution and shaded regions indicate 68% confidence bands.

6.2 Example 2: The effects of a 50% decrease in Russian production

As a second example we examine the consequences if exogenous political events were to lead to a 50% decline in $u_{q,Russia}$ and that Russia stops responding to price ($\phi_{q,Russia} = 0$). This would represent a loss of over 5 million barrels per day. For this scenario, we use production and consumption shares as of the end of our sample (February 2023).[10] Table 4 summarizes the effect on impact. We first highlight the calculations in columns 5-8 which assume that none of the shock is offset by use of inventory drawdowns ($\phi_v = 0$). The model estimates imply that the price of oil would

[10]These were $\mathbf{s}_{qT} = (0.15,0.12,0.13,0.60)'$ and $\mathbf{s}_{cT} = (0.20,0.04,0.05,0.71)'$.

Table 4: Impact effects of a shock to Russia supply

Variable	with estimated ϕ_v as % of country direct effect (1)	response to price (2)	net effect (3)	in mb/d net effect (4)	with $\phi_v = 0$ as % of country direct effect (5)	response to price (6)	net effect (7)	in mb/d net effect (8)
p	12.712				33.020			
q_{US}	0	0.269	0.269	0.033	0	0.699	0.699	0.086
q_{Saudi}	0	3.152	3.152	0.311	0	8.186	8.186	0.808
q_{Russia}	-50	0.000	-50.000	-5.350	-50	0.000	-50.000	-5.350
q_{ROW}	0	0.834	0.834	0.412	0	2.165	2.165	1.069
q				-4.593				-3.386
c_{US}	0.000	-0.983	-0.983	-0.162	0.000	-2.554	-2.554	-0.420
c_{Japan}	0.000	-0.010	-0.010	0.000	0.000	-0.026	-0.026	-0.001
c_{Europe}	0.000	-2.571	-2.571	-0.106	0.000	-6.679	-6.679	-0.275
c_{ROW}	0.000	-1.772	-1.772	-1.035	0.000	-4.603	-4.603	-2.690
c				-1.303				-3.386
v				3.290				0.000

Notes to Table 4. Impact effects of a 50% cut in Russian oil production both given the historical average response of inventories (columns 1-4) and under the counterfactual of no adjustment of inventories (columns 5-8).

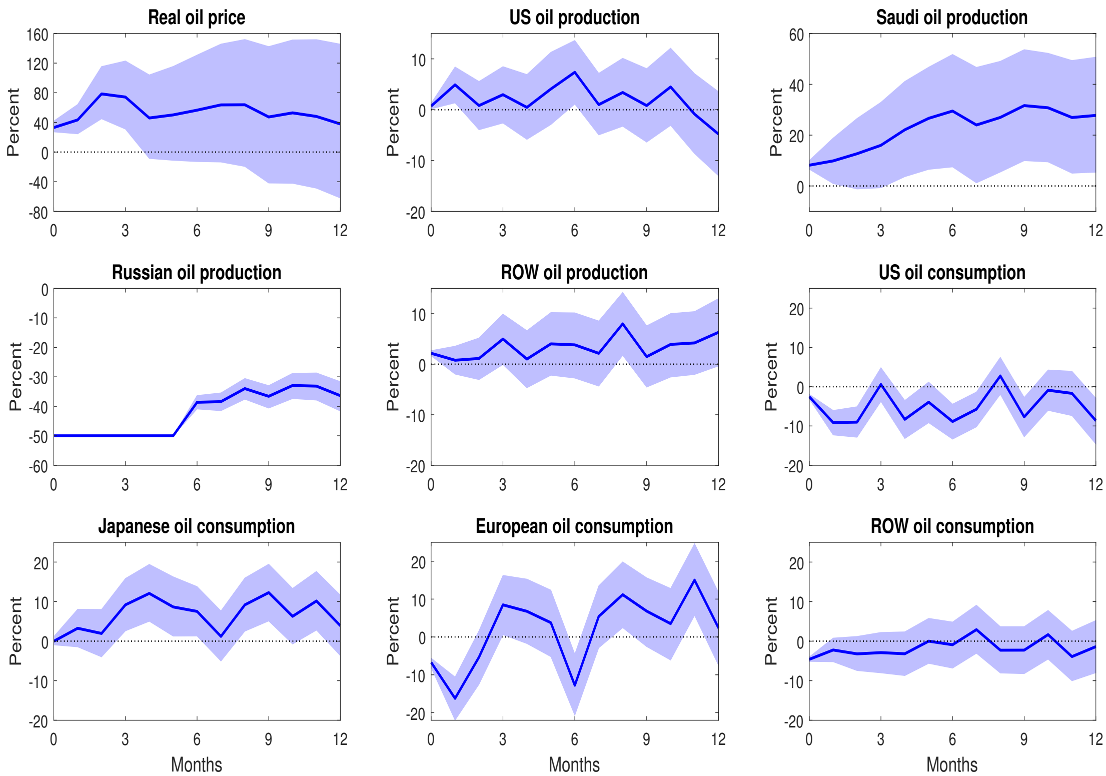
increase by about a third.[11] For convenience we summarize effects on production and consumption in column 8 in units of million barrels per day. This was calculated by multiplying the number reported in column 7 by $s_{qiT}Q_T$ or $s_{cjT}Q_T$ where $Q_T = 82.3$ mb/d is total world oil production in February 2023. Increased production from Saudi Arabia and the rest of the world makes up about 1.9 mb/d of the 5.35 mb/d shortfall. A much bigger part of the adjustment comes from the demand side, with a 400,000 b/d drop in U.S. consumption and a 3.4 mb/d drop in global oil consumption.

Columns 1-4 of Table 4 report the impact response if instead inventories responded to this shock the same way they did to typical historical shocks. This would require drawing down inventories by 3.3 mb/d, or nearly a hundred million barrels in the first month, which clearly is not sustainable.[12] For this reason we emphasize the calculations in column 8 which assume that 100% of the shortfall must be met through a combination of decreased consumption and increased production from other countries.

[11] This was calculated as $-0.5s_{q,Russia,T}/(s_{q,US,T}\phi_{q,US} + s_{q,Saudi,T}\phi_{q,Saudi} + s_{q,ROW,T}\phi_{q,ROW} - \mathbf{s}'_{cT}\boldsymbol{\phi}_c)$.

[12] In January 2022, the U.S. had 415 million barrels of crude oil in commercial inventories and an additional 589 million barrels in the Strategic Petroleum Reserve. As discussed by Baumeister (2023), over the course of the next year, 225 million barrels were released from the SPR.

Figure 2: Dynamic effects of a 50% Russian supply shock



Notes to Figure 2. The fall in Russian production and inability to use global inventories to offset are assumed to last for six months. First panel plots the cumulative effect on 100 times the log of the real price of oil. Other panels plot year-over-year changes of quantities as a percent of that country's production or consumption. Solid lines are median draws from the constrained asymptotic distribution and shaded regions indicate 68% confidence bands.

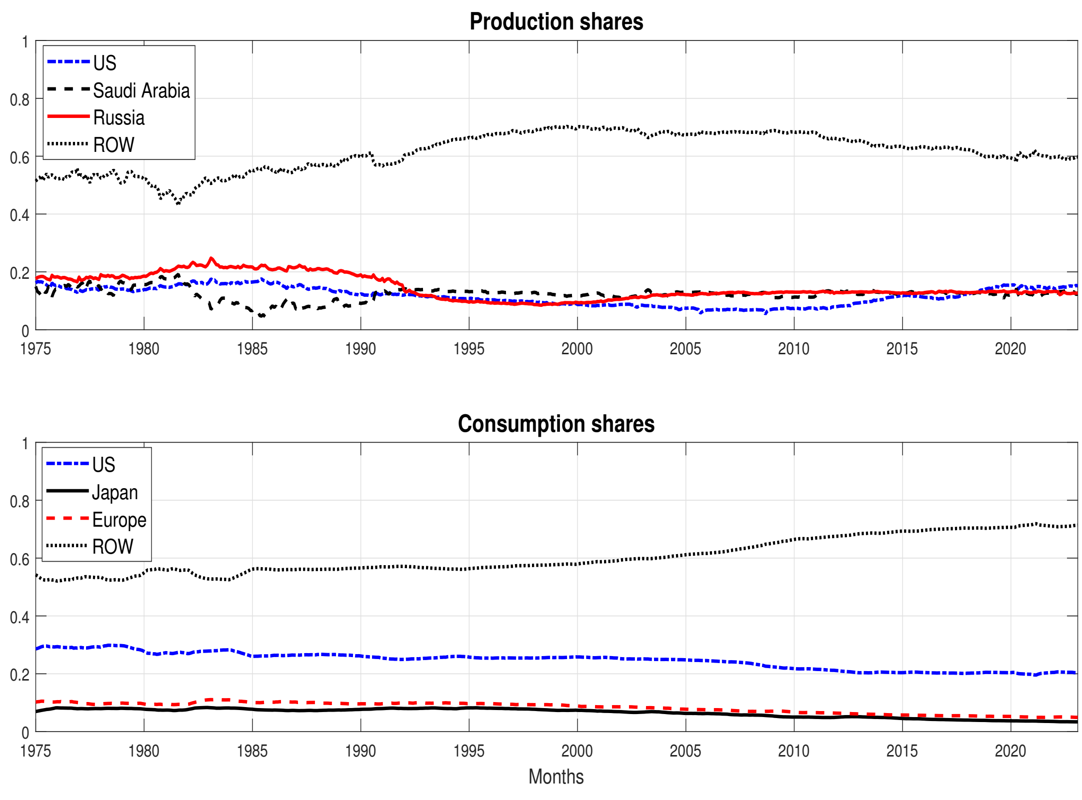
Figure 2 plots the dynamic response under the assumptions that the shortfall in Russian production and inability to draw down global inventories persist for six months. After six months, we assume that the dynamic equations for Russian production and inventory changes revert to their average historical relations.[13] Under this scenario, the price of oil would continue to climb for several months before increasing Saudi production starts to bring the price back down.

[13] Appendix A.8 provides details for how we generate draws for parameters under restricted counterfactual scenarios like this one.

7 Changes over time

The estimates above treated shares and elasticities as constant over time. In practice, shares change very gradually over time. The top panel of Figure 3 shows production shares for each country and each month over our entire sample. There was a period in the 1980s when Saudi Arabia's share was unusually low as the kingdom tried to stabilize the price, and a decline in the U.S. share up until the contributions from shale oil. Consumption shares have strong seasonal patterns across different countries. When we correct for this by looking at 12-month moving averages, consumption shares are even more stable than production shares over time, with a trend up over time in the share of oil consumed by emerging economies.

Figure 3: Monthly production and consumption shares



Notes to Figure 3. Production and consumption shares, 1975:M1-2023:M2. Bottom panel plots 12-month moving averages.

There is much interest in whether changes in how oil is produced and consumed have affected the elasticities of oil supply or demand. To investigate this, we repeated our analysis with a shorter sample beginning in 2005, which marks a major turning point in shale oil production in the United States and oil consumption by emerging economies. Parameter estimates over this shorter sample are reported in the last two columns of Table 2.[14] We find that the elasticity of oil supply from the United States has tripled after the advent of shale oil production compared to its historical value, consistent with the conclusions of Balke, Jin and Yücel (2024) and Aastveit, Bjørnland and Gundersen (2022). But this increase in elasticity from the U.S. was more than offset by a decrease in the elasticity of Saudi production. Subsequent to 1990, Saudi Arabia has made much less effort to stabilize price swings, as discussed for example in Hamilton (2009*a*). We estimate that supply elasticities from Russia and the rest of the world have also declined, with the result that global supply elasticity has fallen from 0.08 to 0.04 in the most recent data, consistent with the conclusions of Baumeister and Peersman (2013).

Another important change has been the huge importance of Chinese oil consumption in the last two decades. This was grouped into the "rest of the world" category in our analysis. While it would be very interesting to break out China as a separate country, we found some striking anomalies in some of the data that are commonly used for Chinese consumption. One market analyst noted that China "has concealed production, import, and export data for years. While tanker tracking companies have popped up and demystified some of the seaborne shipments to and from China, no official China figures exist for any of it–least of all, overall demand."[15] For this reason, we have not attempted the interesting exercise of trying to analyze China as a separate country.

8 Conclusion

The key assumption behind our approach is that correlations between country-specific supply and demand shocks can be summarized with a low-order factor structure. We showed that this assumption allows us to jointly estimate supply and demand elasticities for individual producers and consumers using maximum likelihood estimation of a structural vector autoregression, generalizing the method of granular instrumental variables developed by Gabaix and Koijen (2024). Our method could be applied in any context in which different units interact to determine a market equilibrium. We used this approach to analyze the world oil market. Our estimates of aggregate

[14] For this more recent subsample, the average shares are $\mathbf{s}_q' = (0.10, 0.13, 0.13, 0.64)'$ and $\mathbf{s}_c' = (0.21, 0.05, 0.06, 0.68)'$.

[15] "Oil Markets Are Suffering From A Dearth Of Accurate Data," https://oilprice.com/Energy/Energy-General/Oil-Markets-Are-Suffering-From-A-Dearth-Of-Accurate-Data.html.

elasticities are similar to those obtained by earlier researchers who have used a variety of methods very different from ours. Our approach provides for the first time a characterization of heterogeneity in those elasticities across countries using a unified statistical model of the world oil market. Our estimates imply that variation in Saudi Arabian production and the endogenous adjustment of inventories have historically played a key role in stabilizing the world price of oil.

The method applied here of full-information maximum likelihood is only feasible on a small-scale vector autoregression like the one used above. Larger values of n and m would require estimation of thousands of parameters under our approach. The same framework could be used to analyze much larger systems, though in that case it would be necessary to incorporate some form of regularization to reduce the number of parameters that are allowed to be freely estimated. A promising idea is to use informative Bayesian priors about the lagged VAR coefficients $\mathbf{\Pi}$ as in Chan (2022) or Feldkircher et al. (2022), about the variance matrix $\mathbf{\Omega}$ of forecast errors as in Huber and Koop (2023), and about the elasticities themselves as in Baumeister and Hamilton (2019). We leave this as a subject for future research.

References

Aastveit, Knut Are, Hilde C Bjørnland, and Thomas S Gundersen. 2022. "The Price Responsiveness of Shale Producers: Evidence from Micro Data." *Available at SSRN 4273926*.

Almutairi, Hossa, Axel Pierru, and James L Smith. 2023. "Oil Market Stabilization: The Performance of OPEC and Its Allies." *Energy Journal*, 44(6).

Alonso-Alvarez, Irma, Virginia Di Nino, and Fabrizio Venditti. 2022. "Strategic Interactions and Price Dynamics in the Global Oil Market." *Energy Economics*, 107: 105739.

Balke, Nathan S, Xin Jin, and Mine Yücel. 2024. "The Shale Revolution and the Dynamics of the Oil Market." *Economic Journal*.

Banafti, Saman, and Tae-Hwy Lee. 2022. "Inferential Theory for Granular Instrumental Variables in High Dimensions." *arXiv preprint arXiv:2201.06605*.

Bartik, Timothy J. 1991. *Who Benefits From State and Local Economic Development Policies?* WE Upjohn Institute for Employment Research.

Baumeister, Christiane. 2023. "Pandemic, War, Inflation: Oil Markets at a Crossroads?" *NBER Working Paper*, 31496.

Baumeister, Christiane, and Gert Peersman. 2013. "The Role of Time-Varying Price Elasticities in Accounting for Volatility Changes in the Crude Oil Market." *Journal of Applied Econometrics*, 28(7): 1087–1109.

Baumeister, Christiane, and James D Hamilton. 2019. "Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Demand Shocks." *American Economic Review*, 109(5): 1873–1910.

Baumeister, Christiane, Dimitris Korobilis, and Thomas K Lee. 2022. "Energy Markets and Global Economic Conditions." *Review of Economics and Statistics*, 104(4): 828–844.

Blanchard, Olivier Jean, Lawrence F Katz, Robert E Hall, and Barry Eichengreen. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity*, 1992(1): 1–75.

Bornstein, Gideon, Per Krusell, and Sergio Rebelo. 2023. "A World Equilibrium Model of the Oil Market." *Review of Economic Studies*, 90(1): 132–164.

Caldara, Dario, Michele Cavallo, and Matteo Iacoviello. 2019. "Oil Price Elasticities and Oil Price Fluctuations." *Journal of Monetary Economics*, 103: 1–20.

Carlino, Gerald, and Thorsten Drautzburg. 2020. "The Role of Startups for Local Labor Markets." *Journal of Applied Econometrics*, 35(6): 751–775.

Chan, Joshua CC. 2022. "Asymmetric Conjugate Priors for Large Bayesian VARs." *Quantitative Economics*, 13: 1145–1169.

Cooper, John CB. 2003. "Price Elasticity of Demand for Crude Oil: Estimates for 23 Countries." *OPEC Review*, 27(1): 1–8.

Feldkircher, Martin, Florian Huber, Gary Koop, and Michael Pfarrhofer. 2022. "Approximate Bayesian Inference and Forecasting in Huge Dimensional Multi-Country VARs." *International Economic Review*, 63(4): 1625–1658.

Gabaix, Xavier, and Ralph SJ Koijen. 2024. "Granular Instrumental Variables." *Journal of Political Economy.*

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik Instruments: What, When, Why, and How." *American Economic Review*, 110(8): 2586–2624.

Hamilton, James D. 1994. *Time Series Analysis.* Princeton University Press.

Hamilton, James D. 2009*a*. "Causes and Consequences of the Oil Shock of 2007–08." *Brookings Papers on Economic Activity*, Spring 2009: 215–259.

Hamilton, James D. 2009*b*. "Understanding Crude Oil Prices." *The Energy Journal*, 30(2): 179–206.

Hamilton, James D. 2025. *Vector Autoregressions.* Cambridge University Press.

Huber, Florian, and Gary Koop. 2023. "Subspace Shrinkage in Conjugate Bayesian Vector Autoregressions." *Journal of Applied Econometrics*, 38(4): 556–576.

Knittel, Christopher R, and Robert S Pindyck. 2016. “The Simple Economics of Commodity Price Speculation.” *American Economic Journal: Macroeconomics*, 8(2): 85–110.

Lenza, Michele, and Giorgio E Primiceri. 2022. “How to Estimate a VAR After March 2020.” *Journal of Applied Econometrics*, 37: 688–689.

Ng, Serena. 2021. “Modeling Macroeconomic Variations after Covid-19.” *NBER Working Paper*, 29060.

Pierru, Axel, James L Smith, and Tamim Zamrik. 2018. “OPEC’s Impact on Oil Price Volatility: The Role of Spare Capacity.” *The Energy Journal*, 39(2): 103–122.

Qian, Eric. 2023. “Heterogeneity-Robust Granular Instruments.” Princeton University.

Rothenberg, Thomas J. 1971. “Identification in Parametric Models.” *Econometrica*, 39(3): 577–591.

Sarto, Andres. 2022. “Recovering Macro Elasticities from Regional Data.” New York University.

Schorfheide, Frank, and Dongho Song. 2024. “Real-Time Forecasting with a (Standard) Mixed-Frequency VAR During a Pandemic.” *International Journal of Central Banking*.

Sims, Christopher A. 1980. “Macroeconomics and Reality.” *Econometrica*, 48: 1–48.

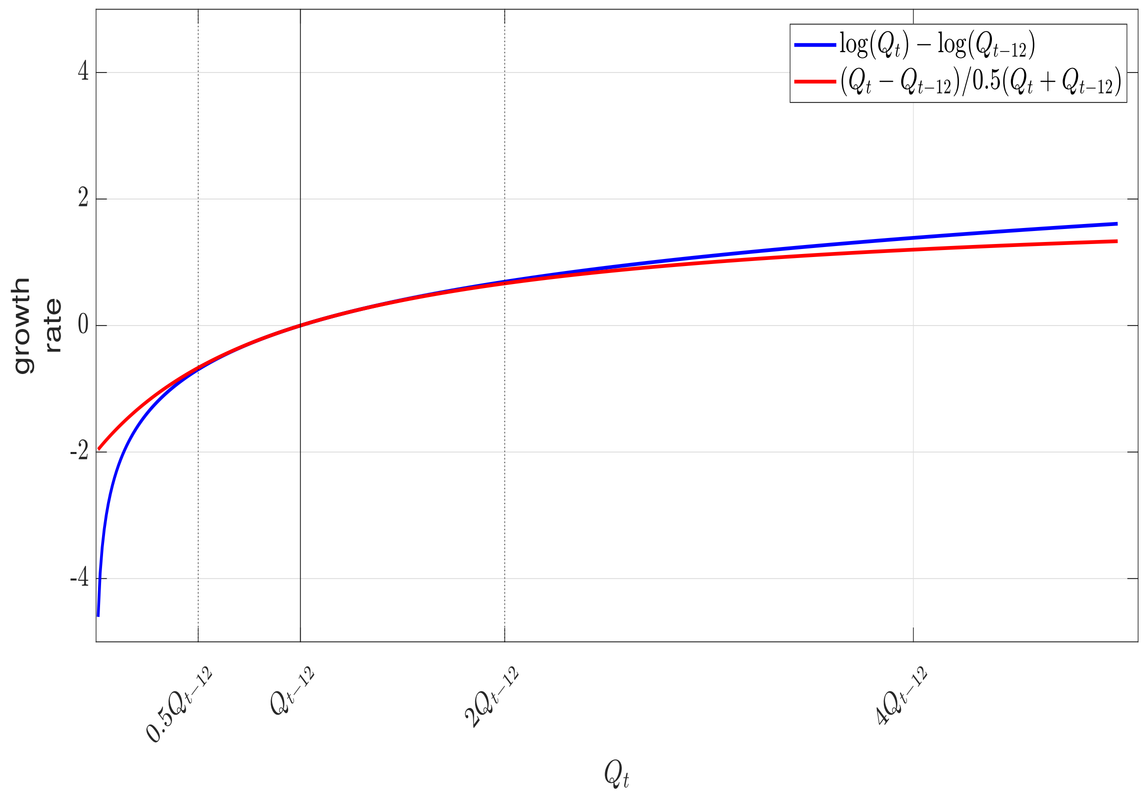
Yogo, Motohiro. 2004. “Estimating the Elasticity of Intertemporal Substitution When Instruments are Weak.” *Review of Economics and Statistics*, 86(3): 797–810.

A Appendix

A.1 A better approach than using log differences for percent change

Figure A.1 compares the log difference function with its first-order Taylor approximation. Differences are small for most of the range but for large changes the latter is much better to use.

Figure A.1: Plot of $\log(Q_t) - \log(Q_{t-12})$ and $(Q_t - Q_{t-12})/0.5(Q_t + Q_{t-12})$ as a function of Q_t



A.2 Derivation of supply equations

Here we describe how optimizing behavior by producers could give rise to supply equations of the form of (5).

Price-taking producer.

Consider first the case of a perfectly competitive supplier i whose cost of producing quantity Q_{it} is given by

$$C_{it}(Q_{it}) = (1 + \psi_i)^{-1} Q_{it}^{1+\psi_i} \exp(w_{it}^*). \tag{A1}$$

Here $\psi_i > 0$ is the elasticity of marginal costs with respect to Q_{it} and w^*_{it} is a cost shock arising from factors such as technological improvements, weather, or geopolitical disruptions. Suppose for illustration that these shocks can be characterized by a first-order autoregression[16]

$$w^*_{it} = \delta_i + \rho_i w^*_{i,t-1} + u^*_{it} \tag{A2}$$

with $u^*_{it} \sim$ i.i.d. $N(0, \sigma_i^{*2})$. A perfectly competitive supplier will choose to produce at the point where marginal cost $\partial C_{it}(Q_{it})/\partial Q_{it}$ equals price P_t:

$$Q_{it}^{\psi_i} \exp(w^*_{it}) = P_t.$$

Denoting $Q_{it} = \exp(q_{it})$ and $P_t = \exp(p_t)$ we can write this condition as

$$\exp(\psi_i q_{it} + w^*_{it}) = \exp(p_t).$$

Taking logs,

$$\psi_i q_{it} + w^*_{it} = p_t. \tag{A3}$$

The same condition holds when t is replaced by $t-1$:

$$\psi_i q_{i,t-1} + w^*_{i,t-1} = p_{t-1}.$$

Multiply the latter equation by ρ_i and subtract the result from (A3):

$$\psi_i q_{it} - \rho_i \psi_i q_{i,t-1} + w^*_{it} - \rho_i w^*_{i,t-1} = p_t - \rho_i p_{t-1}. \tag{A4}$$

Substituting (A2) into (A4) and rearranging gives

$$q_{it} = \psi_i^{-1} p_t - \psi_i^{-1}\delta_i + \rho_i q_{i,t-1} - \psi_i^{-1}\rho_i p_{t-1} - \psi_i^{-1} u^*_{it}. \tag{A5}$$

This is an equation of the form of (5) with $\phi_{qi} = \psi_i^{-1}$, $u_{qit} = -\psi_i^{-1} u^*_{it}$, and

$$\mathbf{b}_i' \mathbf{x}_{t-1} = -\psi_i^{-1}\delta_i + \rho_i q_{i,t-1} - \psi_i^{-1}\rho_i p_{t-1}.$$

The elasticity of supply ϕ_{qi} for this example is the reciprocal of the elasticity of marginal costs with respect to production, the lagged dynamics $\mathbf{b}_i'\mathbf{x}_{t-1}$ result from

[16]More generally, if $w^*_{it} \sim AR(r)$ so that $(1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_r L^r) w^*_{it} = u^*_{it}$ for L the lag operator, we would operate on equation (A3) by $(1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_r L^r)$ to arrive at an equation of the form of (5).

serial persistence of cost shocks, and the structural supply shock u_{qit} is negatively related to the cost shock u_{it}^*; something that increases costs or lowers productivity ($u_{it}^* > 0$) leads to lower production for any given price ($u_{qit} < 0$).

Price-setting behavior.

Consider next a supplier that takes into account the effects of its production decision on the world price. Profit maximization then calls for a production level at which marginal cost equals marginal revenue where marginal revenue is given by

$$MR_{it} = \frac{\partial P_t Q_{it}}{\partial Q_{it}} = P_t + Q_{it}\frac{\partial P_t}{\partial Q_{it}} = P_t\left(1 + \frac{\partial p_t}{\partial q_{it}}\right). \tag{A6}$$

Suppose producer i rationally calculates the consequences of its decision on consumers and other producers as described by equations (9)-(11). Those equations imply that the elasticity of the world price with respect to country i's production is given by

$$\frac{\partial p_t}{\partial q_{it}} = -\alpha_i s_{qi}, \tag{A7}$$

where s_{qi} is country i's share in world production and α_i is given by (51). Thus $MR_{it} = P_t(1 - \alpha_i s_{qi})$ and

$$\log MR_{it} = p_t + \log(1 - \alpha_i s_{qi}). \tag{A8}$$

If the cost of producing takes the same form as in the previous example (A1), then the log of the first-order condition for profit maximization becomes log MC_{it} = log MR_{it} or

$$\psi_i q_{it} + w_{it}^* = p_t + \log(1 - \alpha_i s_{qi}).$$

The same manipulations that led to (A5) now lead to an equation of the identical form of (A5), with the only change being that the constant term is reduced by $\psi_i^{-1}\log(1 - \alpha_i s_{qi}) < 0$. The monopolist produces less than a competitive producer, but since the model implies a constant elasticity $\alpha_i s_{qi}$ of world price with respect to the production of country i, the supply elasticity ϕ_{qi} is still exactly equal to the reciprocal of the elasticity of marginal cost ψ_i^{-1}.

Dynamic optimization.

Consider next a producer whose decisions involve intertemporal considerations arising for example from adjustment costs when $Q_{it} \neq Q_{i,t-1}$ or from time to build. Suppose for illustration that the costs incurred at date t depend on both Q_{it} and $Q_{i,t-1}$: $C_i(Q_{it}, Q_{i,t-1}, u_{it}^*)$. Suppose the producer decides at date t on a sequence

$\{Q_{i,t+\tau}\}_{\tau=0}^{\infty}$ so as to maximize expected profits

$$E_t \textstyle\sum_{\tau=0}^{\infty} \beta^\tau [P_{t+\tau} Q_{i,t+\tau} - C_i(Q_{i,t+\tau}, Q_{i,t+\tau-1}, u^*_{i,t+\tau})]$$

for β a discount rate. The first-order condition for optimal choice of Q_{it} is characterized by

$$\frac{\partial(P_t Q_{it})}{\partial Q_{it}} - \frac{\partial C_i(Q_{it}, Q_{i,t-1}, u^*_{it})}{\partial Q_{it}} - \beta E_t \left[\frac{\partial C_i(Q_{i,t+1}, Q_{it}, u^*_{i,t+1})}{\partial Q_{it}} \right] = 0$$

$$\log \left[\frac{\partial(P_t Q_{it})}{\partial Q_{it}} \right] = \log \left\{ \frac{\partial C_i(Q_{it}, Q_{i,t-1}, u^*_{it})}{\partial Q_{it}} + \beta E_t \left[\frac{\partial C_i(Q_{i,t+1}, Q_{it}, u^*_{i,t+1})}{\partial Q_{it}} \right] \right\}. \quad \text{(A9)}$$

Consider the log-linear approximation

$$\begin{aligned} &\log \left\{ \frac{\partial C_i(Q_{it}, Q_{i,t-1}, u^*_{it})}{\partial Q_{it}} + \beta E_t \left[\frac{\partial C_i(Q_{i,t+1}, Q_{it}, u^*_{i,t+1})}{\partial Q_{it}} \right] \right\} \\ &\simeq d_{i0} + d_{i1} q_{it} + d_{i2} q_{i,t-1} + d_{i3} E_t q_{i,t+1} + d_{i4} u^*_{it}. \end{aligned} \quad \text{(A10)}$$

Substitute (A8) and (A10) into (A9):

$$p_t + \log(1 - \alpha_i s_{qi}) = d_{i0} + d_{i1} q_{it} + d_{i2} q_{i,t-1} + d_{i3} E_t q_{i,t+1} + d_{i4} u^*_{it}. \quad \text{(A11)}$$

In principle, a rational-expectations forecast of $q_{i,t+1}$ would depend on $\mathbf{x}_{t-1}$ and all the variables in $\mathbf{y}_t$. If next period's marginal cost is equal to next period's price, then predicting $q_{i,t+1}$ is the same as predicting next period's price p_{t+1}. In practice, by far the most useful predictors of $q_{i,t+1}$ are going to be p_t, q_{it} and u^*_{it}:

$$E_t q_{i,t+1} \simeq g_{ip} p_t + g_{iq} q_{it} + g_{iu} u^*_{it} + \mathbf{g}'_i \mathbf{x}_{t-1}. \quad \text{(A12)}$$

Equation (A12) would hold exactly if producer i knows the market price p_t but does not know the contemporaneous decisions of other producers and consumers at the time it makes the decision for q_{it}. Substituting (A12) into (A11) and rearranging again results in an expression of the form of (5) where now $\phi_{qi} = (1 - d_{i3} g_{ip})/(d_{i1} + d_{i3} g_{iq})$. Although ϕ_{qi} now reflects a mix of the influence of cost elasticities, adjustment costs, and price and output forecasts, it is still an object of interest to market analysts and policy makers as it summarizes how, as a result of the interaction of all these factors, an unanticipated change in price induces a change in the chosen value of q_{it}.

A.3 Concordance with alternative inventory measures

A key input into our analysis is given in expression (24), which is our proxy for the unanticipated change in world oil inventories in month t. Researchers have used a number of different measures of oil inventories, all of which include significant measurement error. Here we report the correlations between our measure and alternative estimates.

The EIA has published reasonably good data since 1973 on inventories of U.S. crude oil and petroleum products in Table 3.4 of the *Monthly Energy Review*[17] though these only measure inventories for the United States. We fit an $AR(12)$ to monthly changes in the log of U.S. inventories and interpreted the residual as an estimate of the unanticipated change in U.S. inventories. The correlation between the unanticipated change in U.S. inventories and our measure $\hat{\varepsilon}_{vt}$ of the unanticipated change in world inventories over 1975:M1-2020:M2 is 0.17.

Many researchers use a broader measure of monthly petroleum and other liquid stocks that covers all the OECD countries, which is based on the Monthly Oil Survey carried out by the International Energy Agency and is available going back to January 1988.[18] When we fit an $AR(12)$ to monthly log changes of this estimate of OECD inventories, the residuals have a correlation of 0.30 with our global measure $\hat{\varepsilon}_{vt}$.

The EIA also publishes a rough estimate of the monthly change in world inventories, which is available since 1997.[19] When we fit an $AR(12)$ for this series, the residuals have a correlation of 0.65 with our series for $\hat{\varepsilon}_{vt}$.

We conclude that although our estimate $\hat{\varepsilon}_{vt}$ undoubtedly includes some measurement error, it is correctly capturing some of the main movements in world inventories. The big advantage of our measure of the unanticipated change in world inventories is that it reconciles observations on country-level production and consumption data into a unified empirical and theoretical global framework satisfying the accounting identity relating production, consumption, and changes in inventories.

A.4 Proof of Proposition 1

We first show that the determinant of $\mathbf{A}$ in (13) is given by

$$|\mathbf{A}| = \boldsymbol{\phi}_q' \mathbf{s}_{qt} - \boldsymbol{\phi}_c' \mathbf{s}_{ct} - \phi_v. \tag{A13}$$

[17] https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T03.04#/?f=M&start=200001.

[18] https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T03.04#.

[19] This series is labeled "total crude oil and other liquids inventory net withdrawal" and can be downloaded using the STEO data browser for Table 3a https://www.eia.gov/outlooks/steo/data/browser/#/?v=6&f=M&s=0&ctype=linechart&maptype=0.

To verify (A13), define

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{nm} & \mathbf{0}_{n1} \\ \mathbf{0}_{mn} & \mathbf{I}_m & \mathbf{0}_{m1} \\ -\mathbf{s}_q' & \mathbf{s}_c' & 1 \end{bmatrix}$$

and notice that

$$\mathbf{A}^*\mathbf{A} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{nm} & -\boldsymbol{\phi}_q \\ \mathbf{0}_{mn} & \mathbf{I}_m & -\boldsymbol{\phi}_c \\ \mathbf{0}_{1n} & \mathbf{0}_{1m} & \mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v \end{bmatrix}.$$

Since $\mathbf{A}^*$ is lower triangular with ones along the principal diagonal, its determinant is one and the determinant of the product $\mathbf{A}^*\mathbf{A}$ is the same as the determinant of $\mathbf{A}$. Since the product $\mathbf{A}^*\mathbf{A}$ is upper triangular, its determinant is the product of the terms on its principal diagonal. Hence

$$|\mathbf{A}| = |\mathbf{A}^*\mathbf{A}| = \mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v \tag{A14}$$

as claimed in (A13).

Next note that for any value of $\mathbf{A}$ and $\mathbf{D}$, the value of $\mathbf{B}$ that maximizes (16) is found from an OLS regression of $\mathbf{A}\mathbf{y}_t$ on $\mathbf{x}_{t-1}$:

$$\hat{\mathbf{B}}(\boldsymbol{\theta}) = \left[\textstyle\sum_{t=1}^T \mathbf{A}\mathbf{y}_t\mathbf{x}_{t-1}'\right]\left[\textstyle\sum_{t=1}^T \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right]^{-1} = \mathbf{A}\hat{\boldsymbol{\Pi}}. \tag{A15}$$

We then have

$$\mathbf{A}\mathbf{y}_t - \hat{\mathbf{B}}(\boldsymbol{\theta})\mathbf{x}_{t-1} = \mathbf{A}(\mathbf{y}_t - \hat{\boldsymbol{\Pi}}\mathbf{x}_{t-1}) = \mathbf{A}\hat{\boldsymbol{\epsilon}}_t. \tag{A16}$$

Substituting (A13) and (A16) into (16) gives

$$\begin{aligned} \ell(\boldsymbol{\theta}, \hat{\mathbf{B}}(\boldsymbol{\theta})) = &-(TN/2)\log(2\pi) + (T/2)\log[(\mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v)^2] - (T/2)\log|\mathbf{D}| \\ &- (1/2)\textstyle\sum_{t=1}^T(\mathbf{A}\hat{\boldsymbol{\epsilon}}_t)'\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\epsilon}}_t) \end{aligned} \tag{A17}$$

as claimed in (20).

To verify (21), recall if $\mathbf{Q}$ and $\mathbf{R}$ are matrices for which both $\mathbf{QR}$ and $\mathbf{RQ}$ exist, then trace($\mathbf{QR}$) = trace($\mathbf{RQ}$). Hence

$$\begin{aligned} \textstyle\sum_{t=1}^T(\mathbf{A}\hat{\boldsymbol{\epsilon}}_t)'\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\epsilon}}_t) = &\text{ trace}\left[\textstyle\sum_{t=1}^T\hat{\boldsymbol{\epsilon}}_t'\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\hat{\boldsymbol{\epsilon}}_t\right] = \text{ trace}\left[\textstyle\sum_{t=1}^T\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\hat{\boldsymbol{\epsilon}}_t\hat{\boldsymbol{\epsilon}}_t'\right] \\ = &\text{ trace}\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}T\hat{\boldsymbol{\Omega}}\right] = T\text{ trace}\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\hat{\boldsymbol{\Omega}}\right]. \end{aligned} \tag{A18}$$

Substituting (A18) into (A17) gives (21).

A.5 Proof of Proposition 2

Maximizing (23) with respect to $\mathbf{D}_q$, $\mathbf{D}_c$, or σ_v^2 makes use of well known results (e.g., Hamilton (1994, p. 295)) to produce (25). Setting the derivative of (23) with respect to ϕ_q to zero, and assuming that $\hat{\phi}_q - \hat{\phi}_c - \hat{\phi}_v > 0$, we obtain

$$\left.\frac{\partial \eta(\boldsymbol{\theta})}{\partial \phi_q}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{T}{\hat{\phi}_q - \hat{\phi}_c - \hat{\phi}_v} + \textstyle\sum_{t=1}^{T} \mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} (\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q \mathbf{1}_n \hat{\varepsilon}_{pt}) \hat{\varepsilon}_{pt} = 0. \tag{A19}$$

Multiply both sides of (A19) by $(\hat{\phi}_q - \hat{\phi}_c - \hat{\phi}_v)/T$:

$$1 + T^{-1} \textstyle\sum_{t=1}^{T} \mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} (\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q \mathbf{1}_n \hat{\varepsilon}_{pt})(\hat{\phi}_q \hat{\varepsilon}_{pt} - \hat{\phi}_c \hat{\varepsilon}_{pt} - \hat{\phi}_v \hat{\varepsilon}_{pt}) = 0.$$

Add and subtract $\tilde{c}_t$ and $\hat{\varepsilon}_{vt} = \mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} - \mathbf{s}_c' \hat{\boldsymbol{\epsilon}}_{ct}$ to the last term in parentheses:

$$\begin{aligned} &1 + T^{-1} \textstyle\sum_{t=1}^{T} \mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} (\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q \mathbf{1}_n \hat{\varepsilon}_{pt})[(-\mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} + \hat{\phi}_q \hat{\varepsilon}_{pt}) \\ &+ (\tilde{c}_t - \hat{\phi}_c \hat{\varepsilon}_{pt}) + (\hat{\varepsilon}_{vt} - \hat{\phi}_v \hat{\varepsilon}_{pt}) - \tilde{c}_t + \mathbf{s}_c' \hat{\boldsymbol{\epsilon}}_{ct})] = 0. \end{aligned} \tag{A20}$$

Notice that the first term in the summation in (A20) can be written

$$\begin{aligned} &T^{-1} \textstyle\sum_{t=1}^{T} \mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} (\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q \mathbf{1}_n \hat{\varepsilon}_{pt})(-\mathbf{s}_q' \hat{\boldsymbol{\epsilon}}_{qt} + \hat{\phi}_q \hat{\varepsilon}_{pt}) \\ &= -T^{-1} \textstyle\sum_{t=1}^{T} \mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} (\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q \mathbf{1}_n \hat{\varepsilon}_{pt})(\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q \mathbf{1}_n \hat{\varepsilon}_{pt})' \mathbf{s}_q \\ &= -\mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} \hat{\mathbf{D}}_q \mathbf{s}_q \\ &= -1 \end{aligned}$$

where the first equation follows from $\mathbf{1}_n' \mathbf{s}_q = 1$ and the second equation from (25). Thus (A20) becomes

$$T^{-1} \textstyle\sum_{t=1}^{T} \mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} (\hat{\boldsymbol{\epsilon}}_{qt} - \hat{\phi}_q \mathbf{1}_n \hat{\varepsilon}_{pt})[(\tilde{c}_t - \hat{\phi}_c \hat{\varepsilon}_{pt}) + (\hat{\varepsilon}_{vt} - \hat{\phi}_v \hat{\varepsilon}_{pt}) + (\mathbf{s}_c - \hat{\mathbf{w}}_c)' \hat{\boldsymbol{\epsilon}}_{ct})] = 0.$$

Dividing by $\mathbf{1}_n' \hat{\mathbf{D}}_q^{-1} \mathbf{1}_n$ and using the definitions of $\tilde{q}_t$ and $\hat{\mathbf{w}}_q$ in (29),

$$T^{-1} \textstyle\sum_{t=1}^{T} (\tilde{q}_t - \hat{\phi}_q \hat{\varepsilon}_{pt})[(\tilde{c}_t - \hat{\phi}_c \hat{\varepsilon}_{pt}) + (\hat{\varepsilon}_{vt} - \hat{\phi}_v \hat{\varepsilon}_{pt}) + (\mathbf{s}_c - \hat{\mathbf{w}}_c)' \hat{\boldsymbol{\epsilon}}_{ct})] = 0$$

$$T^{-1} \textstyle\sum_{t=1}^{T} (\tilde{q}_t - \hat{\phi}_q \hat{\varepsilon}_{pt}) \tilde{z}_{qt} = 0.$$

Rearranging this equation gives (26).

Similar operations on the first-order condition for maximization of (23) with respect to ϕ_c give

$$\left.\frac{\partial\eta(\boldsymbol{\theta})}{\partial\phi_c}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{-T}{\hat{\phi}_q - \hat{\phi}_c - \hat{\phi}_v} + \textstyle\sum_{t=1}^T \mathbf{1}_m' \hat{\mathbf{D}}_c^{-1}(\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c \mathbf{1}_m \hat{\varepsilon}_{pt})\hat{\varepsilon}_{pt} = 0$$

$$-1 + T^{-1}\textstyle\sum_{t=1}^T \mathbf{1}_m' \hat{\mathbf{D}}_c^{-1}(\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c \mathbf{1}_m \hat{\varepsilon}_{pt})(\hat{\phi}_q\hat{\varepsilon}_{pt} - \hat{\phi}_c\hat{\varepsilon}_{pt} - \hat{\phi}_v\hat{\varepsilon}_{pt}) = 0$$

$$\begin{aligned} &-1 + T^{-1}\textstyle\sum_{t=1}^T \mathbf{1}_m' \hat{\mathbf{D}}_c^{-1}(\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c \mathbf{1}_m \hat{\varepsilon}_{pt})[(-\tilde{q}_t + \hat{\phi}_q\hat{\varepsilon}_{pt}) \\ &+ (\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c\hat{\varepsilon}_{pt}) + (\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt}) + \left(\tilde{q}_t - \mathbf{s}_q'\hat{\boldsymbol{\epsilon}}_{qt}\right)] = 0 \end{aligned}$$

$$T^{-1}\textstyle\sum_{t=1}^T (\hat{\mathbf{w}}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_v\hat{\varepsilon}_{pt})[(-\tilde{q}_t + \hat{\phi}_q\hat{\varepsilon}_{pt}) + (\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt}) + (\hat{\mathbf{w}}_c - \mathbf{s}_c)'\hat{\boldsymbol{\epsilon}}_{ct} = 0$$

$$T^{-1}\textstyle\sum_{t=1}^T (\tilde{c}_t - \hat{\phi}_c\hat{\varepsilon}_{pt})\tilde{z}_{ct} = 0$$

as claimed in (27). Optimization with respect to ϕ_v likewise leads to

$$\left.\frac{\partial\eta(\boldsymbol{\theta})}{\partial\phi_v}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{-T}{\hat{\phi}_q - \hat{\phi}_c - \hat{\phi}_v} + \textstyle\sum_{t=1}^T \hat{\sigma}_v^{-2}(\hat{\varepsilon}_{vt} - \phi_v\hat{\varepsilon}_{pt})\hat{\varepsilon}_{pt} = 0$$

$$\begin{aligned} -1 + T^{-1}\textstyle\sum_{t=1}^T \hat{\sigma}_v^{-2}(\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt})(\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt} - \tilde{q}_t + \hat{\phi}_q\hat{\varepsilon}_{pt} + \tilde{c}_t - \hat{\phi}_c\hat{\varepsilon}_{pt} + \tilde{q}_t - \tilde{c}_t - \hat{\varepsilon}_{vt}) &= 0 \\ T^{-1}\textstyle\sum_{t=1}^T (\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt})(-\tilde{q}_t - \hat{\phi}_q\hat{\varepsilon}_{pt} + \tilde{c}_t - \hat{\phi}_c\hat{\varepsilon}_{pt} + \tilde{q}_t - \tilde{c}_t - \hat{\varepsilon}_{vt}) &= 0 \\ T^{-1}\textstyle\sum_{t=1}^T (\hat{\varepsilon}_{vt} - \hat{\phi}_v\hat{\varepsilon}_{pt})\tilde{z}_{vt} &= 0 \end{aligned}$$

verifying (28).

A.6 Proof of Proposition 3

This numerical equivalence comes from a similar result to the Frisch-Waugh Theorem. Expression (41) can be rewritten

$$\left[\textstyle\sum_{t=1}^T \mathbf{z}_{ct}^{(IV)}\tilde{\mathbf{x}}_t'\right]\hat{\boldsymbol{\beta}}_c^{(IV)} = \left[\textstyle\sum_{t=1}^T \mathbf{z}_{ct}^{(IV)}\left(\mathbf{s}_c'\mathbf{c}_t\right)\right]$$

$$\textstyle\sum_{t=1}^T \mathbf{z}_{ct}^{(IV)}\left(\mathbf{s}_c'\mathbf{c}_t - \tilde{\mathbf{x}}_t'\hat{\boldsymbol{\beta}}_c^{(IV)}\right) = \mathbf{0}. \tag{A21}$$

Likewise equation (35) can be written

$$\textstyle\sum_{t=1}^T \left[(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\hat{\boldsymbol{\epsilon}}_{qt}\right]\hat{\varepsilon}_{pt}\hat{\phi}_c^{(1)} = \textstyle\sum_{t=1}^T \left[(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\hat{\boldsymbol{\epsilon}}_{qt}\right](\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct})$$

$$\textstyle\sum_{t=1}^T \left[(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\hat{\boldsymbol{\epsilon}}_{qt}\right](\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt}) = 0. \tag{A22}$$

Define

$$\underset{(n\times k)}{\hat{\boldsymbol{\Pi}}_q} = \left[\sum_{t=1}^{T}\mathbf{q}_t\mathbf{x}_{t-1}'\right]\left[\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right]^{-1} \qquad \underset{(m\times k)}{\hat{\boldsymbol{\Pi}}_c} = \left[\sum_{t=1}^{T}\mathbf{c}_t\mathbf{x}_{t-1}'\right]\left[\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right]^{-1}$$

$$\underset{(1\times k)}{\hat{\boldsymbol{\pi}}_p'} = \left[\sum_{t=1}^{T}p_t\mathbf{x}_{t-1}'\right]\left[\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right]^{-1}.$$

Note that the last term in (A22) can be written

$$\begin{aligned}(\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt}) &= \mathbf{s}_c'(\mathbf{c}_t - \hat{\boldsymbol{\Pi}}_c\mathbf{x}_{t-1}) - \hat{\phi}_c^{(1)}(p_t - \hat{\boldsymbol{\pi}}_p'\mathbf{x}_{t-1}) \\ &= \mathbf{s}_c'\mathbf{c}_t - \hat{\phi}_c^{(1)}p_t - \hat{\boldsymbol{\beta}}_x^{(1)\prime}\mathbf{x}_{t-1} \\ &= \mathbf{s}_c'\mathbf{c}_t - \hat{\boldsymbol{\beta}}^{(1)\prime}\tilde{\mathbf{x}}_t\end{aligned}$$

for $\hat{\boldsymbol{\beta}}_x^{(1)\prime} = \mathbf{s}_c'\hat{\boldsymbol{\Pi}}_c - \hat{\phi}_c^{(1)}\hat{\boldsymbol{\pi}}_p'$ and $\hat{\boldsymbol{\beta}}^{(1)\prime} = (\hat{\phi}_c^{(1)}, \hat{\boldsymbol{\beta}}_x^{(1)\prime})$. Substituting this last expression into (A21), if we can verify that

$$\sum_{t=1}^{T}\mathbf{z}_{ct}^{(IV)}(\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt}) = \mathbf{0},$$

then we will have shown that $\hat{\boldsymbol{\beta}}^{(IV)} = \hat{\boldsymbol{\beta}}^{(1)}$ with its first element given by $\hat{\phi}_c^{(1)}$. From the definition of $\mathbf{z}_{ct}^{(IV)}$, the task is to show that

$$\sum_{t=1}^{T}\begin{bmatrix}(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\,\mathbf{q}_t \\ \mathbf{x}_{t-1}\end{bmatrix}(\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt}) = \mathbf{0}. \tag{A23}$$

That the last k terms in (A23) are indeed zero follows from the facts that

$$\sum_{t=1}^{T}\mathbf{x}_{t-1}\hat{\boldsymbol{\epsilon}}_{ct}' = \mathbf{0} \qquad \sum_{t=1}^{T}\mathbf{x}_{t-1}\hat{\varepsilon}_{pt} = \mathbf{0} \tag{A24}$$

since $\hat{\boldsymbol{\epsilon}}_{ct}$ and $\hat{\varepsilon}_{pt}$ are residuals from OLS regressions on $\mathbf{x}_{t-1}$. The first term in (A23) can be written

$$\begin{aligned}\sum_{t=1}^{T}(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\mathbf{q}_t(\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt}) &= \sum_{t=1}^{T}(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'(\hat{\boldsymbol{\Pi}}_q\mathbf{x}_{t-1} + \hat{\boldsymbol{\epsilon}}_{qt})(\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt}) \\ &= \sum_{t=1}^{T}(n^{-1}\mathbf{1}_n - \mathbf{s}_q)'\hat{\boldsymbol{\epsilon}}_{qt}(\mathbf{s}_c'\hat{\boldsymbol{\epsilon}}_{ct} - \hat{\phi}_c^{(1)}\hat{\varepsilon}_{pt})\end{aligned}$$

with the last equality again following from (A24). But this last term is exactly the expression we know to be zero from (A22). Thus the first element in the IV estimate in (41) is exactly the same number as the first iteration estimate $\hat{\phi}_c^{(1)}$ in (35), as claimed.

A.7 Imposing orthogonality of factor loadings

Typical applications of a factor structure such as principal component analysis use a normalization in which the factor loadings are orthogonal to each other. Here we describe how to implement the conditions $\mathbf{h}_q'\boldsymbol{\gamma}_q = \mathbf{h}_c'\boldsymbol{\gamma}_q = 0$ in our algorithm for maximum likelihood estimation.

Let $\mathbf{G}_{q\perp}$ denote the matrix consisting of the first $n-1$ columns of $(\mathbf{h}_q'\mathbf{h}_q)\mathbf{I}_n - \mathbf{h}_q\mathbf{h}_q'$:

$$\underset{n\times(n-1)}{\mathbf{G}_{q\perp}} = \left[(\mathbf{h}_q'\mathbf{h}_q)\mathbf{I}_n - \mathbf{h}_q\mathbf{h}_q'\right]\begin{bmatrix}\mathbf{I}_{n-1}\\ \mathbf{0}_{1,n-1}\end{bmatrix}.$$

Note that $\mathbf{G}_{q\perp}$ is constructed such that each column is orthogonal to $\mathbf{h}_q$:

$$\mathbf{h}_q'\mathbf{G}_{q\perp} = (\mathbf{h}_q'\mathbf{h}_q)(\mathbf{h}_q' - \mathbf{h}_q')\begin{bmatrix}\mathbf{I}_{n-1}\\ \mathbf{0}_{1,n-1}\end{bmatrix} = \mathbf{0}_{1,n-1}.$$

We can then parameterize $\boldsymbol{\gamma}_q = \mathbf{G}_{q\perp}\boldsymbol{\omega}_q$ where $\boldsymbol{\omega}_q$ is an $(n-1)\times 1$ vector of parameters to be estimated. Similarly, we define

$$\underset{m\times(m-1)}{\mathbf{G}_{c\perp}} = \left[(\mathbf{h}_c'\mathbf{h}_c)\mathbf{I}_m - \mathbf{h}_c\mathbf{h}_c'\right]\begin{bmatrix}\mathbf{I}_{m-1}\\ \mathbf{0}_{1,m-1}\end{bmatrix} \tag{A25}$$

and parameterize

$$\boldsymbol{\gamma}_c = \mathbf{G}_{c\perp}\boldsymbol{\omega}_c. \tag{A26}$$

Thus the specification becomes

$$\begin{bmatrix}\mathbf{K}_{qq} & \mathbf{K}_{qc}\\ \mathbf{K}_{cq} & \mathbf{K}_{cc}\end{bmatrix} = \begin{bmatrix}\mathbf{h}_q\mathbf{h}_q' + \mathbf{G}_{q\perp}\boldsymbol{\omega}_q\boldsymbol{\omega}_q'\mathbf{G}_{q\perp}' + \boldsymbol{\Sigma}_q & \mathbf{h}_q\mathbf{h}_c'\\ \mathbf{h}_c\mathbf{h}_q' & \mathbf{h}_c\mathbf{h}_c' + \mathbf{G}_{c\perp}\boldsymbol{\omega}_c\boldsymbol{\omega}_c'\mathbf{G}_{c\perp}' + \boldsymbol{\Sigma}_c\end{bmatrix}.$$

The likelihood function is then maximized with respect to the n elements of $\mathbf{h}_q$, the $(n-1)$ elements of $\boldsymbol{\omega}_q$, the n diagonal elements of the diagonal matrix $\boldsymbol{\Sigma}_q$, the m elements of $\mathbf{h}_c$, the $(m-1)$ elements of $\boldsymbol{\omega}_c$, and the m diagonal elements of the diagonal matrix $\boldsymbol{\Sigma}_c$.

A.8 Confidence bands for impulse-response functions

Let the $(N\times k)$ matrix $\hat{\boldsymbol{\Pi}}$ be the OLS estimate of the reduced-form coefficient matrices in (17), $\hat{\boldsymbol{\pi}} = \text{vec}(\hat{\boldsymbol{\Pi}}')$ the $(Nk\times 1)$ vector from stacking rows of $\hat{\boldsymbol{\Pi}}$ into a single vector, and $\hat{\boldsymbol{\Omega}}$ the OLS estimate of the reduced-form residual variance matrix in (19). We know that the distribution of $\hat{\boldsymbol{\pi}}$ is approximately given by

$$\hat{\boldsymbol{\pi}} \sim N\left(\boldsymbol{\pi}, (\boldsymbol{\Omega} \otimes \left(\textstyle\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}\right) \tag{A27}$$

and that this distribution is asymptotically independent of that of $\hat{\boldsymbol{\Omega}}$. Since the estimates of the structural parameters $\boldsymbol{\theta}$ are a function solely of $\hat{\boldsymbol{\Omega}}$ we can calculate confidence bands as follows. We draw a value for $\boldsymbol{\theta}^{(d)}$ from the distribution $\boldsymbol{\theta} \sim N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\mathbf{V}})$ where $\hat{\boldsymbol{\theta}}_{MLE}$ is the maximum likelihood estimate and $\hat{\mathbf{V}}$ is its estimated variance-covariance matrix from (40). With this draw for $\boldsymbol{\theta}^{(d)}$ we calculate the implied value for $\boldsymbol{\Omega}^{(d)}$:

$$\boldsymbol{\Omega}^{(d)} = [\mathbf{A}(\boldsymbol{\theta}^{(d)})]^{-1}\left[\mathbf{D}(\boldsymbol{\theta}^{(d)})\right][\mathbf{A}(\boldsymbol{\theta}^{(d)})^{-1}]'$$

$$\mathbf{D}(\boldsymbol{\theta}^{(d)}) = \begin{bmatrix} \mathbf{h}_q\mathbf{h}_q' + \boldsymbol{\gamma}_q\boldsymbol{\gamma}_q' + \boldsymbol{\Sigma}_q & \mathbf{h}_q\mathbf{h}_c' & \mathbf{0}_{n1} \\ \mathbf{h}_c\mathbf{h}_q' & \mathbf{h}_c\mathbf{h}_c' + \boldsymbol{\gamma}_c\boldsymbol{\gamma}_c' + \boldsymbol{\Sigma}_c & \mathbf{0}_{m1} \\ \mathbf{0}_{1n} & \mathbf{0}_{1m} & \sigma_v^2 \end{bmatrix}$$

$$[\mathbf{A}(\boldsymbol{\theta}^{(d)})]^{-1} = \begin{bmatrix} \mathbf{I}_n - \alpha\boldsymbol{\phi}_q\mathbf{s}_q' & \alpha\boldsymbol{\phi}_q\mathbf{s}_c' & \alpha\boldsymbol{\phi}_q \\ -\alpha\boldsymbol{\phi}_c\mathbf{s}_q' & \mathbf{I}_m + \alpha\boldsymbol{\phi}_c\mathbf{s}_c' & \alpha\boldsymbol{\phi}_c \\ -\alpha\mathbf{s}_q' & \alpha\mathbf{s}_c' & \alpha \end{bmatrix}$$

$$\alpha(\boldsymbol{\theta}^{(d)}) = \left(\mathbf{s}_q'\boldsymbol{\phi}_q - \mathbf{s}_c'\boldsymbol{\phi}_c - \phi_v\right)^{-1}$$

where for example we have simplified notation by writing $\mathbf{h}_q^{(d)}$ as $\mathbf{h}_q$. We use this value for $\boldsymbol{\Omega}^{(d)}$ to generate a draw for $\boldsymbol{\pi}^{(d)}$ from

$$\boldsymbol{\pi}^{(d)} \sim N\left(\hat{\boldsymbol{\pi}}, \boldsymbol{\Omega}^{(d)} \otimes \left(\textstyle\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}\right).$$

With this pair of $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\Omega}^{(d)}$ we calculate the value of some structural magnitude of interest such as $\partial E(\mathbf{y}_{t+s}|\mathbf{y}_t, \mathbf{y}_{t-1}, ...\mathbf{y}_{t-r+1})/\partial u_{kt}$. We repeat this for draws $d = 1, ..., 10,000$ and calculate the 68% bands for each object of interest.

Drawing parameters subject to restrictions. For the Russian cut-off scenario, we impose $2k$ restrictions on $\boldsymbol{\pi}$ that take the form $\mathbf{R}\boldsymbol{\pi} = \mathbf{r}$ where

$$\underset{(2k\times Nk)}{\mathbf{R}} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{bmatrix} \qquad \underset{(2k\times 1)}{\mathbf{r}} = \begin{bmatrix} \mathbf{e}_{Russia} \\ \mathbf{0}_{k1} \end{bmatrix} \tag{A28}$$

$$\underset{(k\times Nk)}{\mathbf{R}_1} = \begin{bmatrix} \mathbf{0}_{n,2k} & \mathbf{I}_k & \mathbf{0}_{k,(N-3)k} \end{bmatrix}$$

$$\underset{(k\times Nk)}{\mathbf{R}_2} = \begin{bmatrix} \mathbf{s}_q' \otimes \mathbf{I}_k & -\mathbf{s}_c' \otimes \mathbf{I}_k & \mathbf{0}_{k,k} \end{bmatrix}.$$

Here $\mathbf{e}_{Russia}$ is a $(k \times 1)$ vector whose value is unity for the coefficient on the first lag of Russian production and zero for all others. The first k rows of (A28) imply that Russian production stays at its previous value. The next k rows of (A28) force global production to equal global consumption every period.[20] Let $\mathbf{V}^{(d)} = \mathbf{\Omega}^{(d)} \otimes \left(\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}$ and consider the joint distribution of $\mathbf{R}\boldsymbol{\pi}$ and $\boldsymbol{\pi}$ conditional on $\mathbf{\Omega}^{(d)}$:

$$\begin{bmatrix} \mathbf{R}\boldsymbol{\pi} \\ \boldsymbol{\pi} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{R}\hat{\boldsymbol{\pi}} \\ \hat{\boldsymbol{\pi}} \end{bmatrix}, \begin{bmatrix} \mathbf{R}\mathbf{V}^{(d)}\mathbf{R}' & \mathbf{R}\mathbf{V}^{(d)} \\ \mathbf{V}^{(d)}\mathbf{R}' & \mathbf{V}^{(d)} \end{bmatrix}\right).$$

From this we can calculate the distribution of $\boldsymbol{\pi}$ conditional on $\mathbf{R}\boldsymbol{\pi} = \mathbf{r}$:

$$\boldsymbol{\pi}|(\mathbf{R}\boldsymbol{\pi} = \mathbf{r}) \sim N(\boldsymbol{\mu}_r^{(d)}, \mathbf{V}_r^{(d)}) \tag{A29}$$

$$\boldsymbol{\mu}_r^{(d)} = \hat{\boldsymbol{\pi}} + \mathbf{V}^{(d)}\mathbf{R}'(\mathbf{R}\mathbf{V}^{(d)}\mathbf{R}')^{-1}(\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\pi}})$$

$$\mathbf{V}_r^{(d)} = \mathbf{V}^{(d)} - \mathbf{V}^{(d)}\mathbf{R}'(\mathbf{R}\mathbf{V}^{(d)}\mathbf{R}')^{-1}\mathbf{R}\mathbf{V}^{(d)}.$$

The distribution in (A29) is singular. We can use the nonsingular part of this distribution to generate a subset of the elements in $\boldsymbol{\pi}$ and fill in the rest by the restrictions. To do this, let $\mathbf{R}_\dagger$ be the matrix that collects the reduced-form coefficients that are not restricted. Without loss of generality, these could be taken to be: U.S., Saudi, and rest-of-world production; U.S., Europe, and Japan consumption; and the price:

$$\underset{[(N-2)k\times Nk]}{\mathbf{R}_\dagger} = \begin{bmatrix} \mathbf{I}_k & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} \\ \mathbf{0}_{kk} & \mathbf{I}_k & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} \\ \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{I}_k & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} \\ \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{I}_k & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} \\ \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{I}_k & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} \\ \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{I}_k & \mathbf{0}_{kk} & \mathbf{0}_{kk} \\ \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{0}_{kk} & \mathbf{I}_k \end{bmatrix}.$$

We can generate a draw for these seven of the nine equations from a $N(\mathbf{R}_\dagger\boldsymbol{\mu}_r^{(d)}, \mathbf{R}_\dagger\mathbf{V}_r^{(d)}\mathbf{R}_\dagger')$ distribution. Next we set the first own-lag coefficient of $\boldsymbol{\pi}_{q,Russia}$ to 1 and all other coefficients to zero, and set the coefficient vector to predict rest-of-world consumption

[20] Recall that the constant terms don't affect impulse-response functions.

to be

$$\boldsymbol{\pi}_{c,ROW}^{(d)} = \frac{\left[\begin{array}{cccc} \boldsymbol{\pi}_{q,US}^{(d)} & \boldsymbol{\pi}_{q,Saudi}^{(d)} & \boldsymbol{\pi}_{q,Russia}^{(d)} & \boldsymbol{\pi}_{q,ROW}^{(d)} \end{array}\right] \mathbf{s}_q - s_{c,US}\boldsymbol{\pi}_{c,US}^{(d)} - s_{c,Eur}\boldsymbol{\pi}_{c,Eur}^{(d)} - s_{c,Jap}\boldsymbol{\pi}_{c,Jap}^{(d)}}{s_{c,ROW}}.$$

Given this complete draw for the $(Nk \times 1)$ vector $\boldsymbol{\pi}_r^{(d)}$ from the restricted distribution, we can calculate the VAR companion-form matrix $\mathbf{F}_r^{(d)}$ and calculate the IRF for $s = 0, 1, ..., 5$ from $[\mathbf{F}_r^{(d)}]^s$. For $s = 5 + h$ we drop the restrictions $\mathbf{R}$ and use $\left[\mathbf{F}^{(d)}\right]^h [\mathbf{F}_r^{(d)}]^5$ where $\mathbf{F}^{(d)}$ is the companion matrix for a draw from the unrestricted distribution.