Rethinking Markups and Inventories over the Business Cycle*

Lukasz A. Drozd
Federal Reserve Bank of Philadelphia

Marina M. Tavares
International Monetary Fund

March 2, 2025

ABSTRACT

Existing theories of inventory dynamics require countercyclical markups to explain why inventory stocks fall *less* than sales during demand recessions. In this paper, we revisit this link and argue that the lagging response of inventories to falling sales in recessions points to the presence of other frictions—rather than markups—that delay the unwinding of inventories and cause the inventory-to-sales ratio to rise. The key mechanism is customer capital irreversibility and the complementary role of inventories in attracting customers. Our analysis has implications for the New Keynesian interpretation of existing inventory models as supportive of the proposition that markups are countercyclical, and, consequently, that sticky prices—rather than sticky wages—play a dominant role in breaking monetary neutrality.

Keywords: markups, inventory, inventory-to-sales ratio, business cycle, customer capital

JEL codes: E22, E31, E32, E52, E58

*We thank Ben Lester, Urban Jermann, Joachim Hubmer, Joao Gomez, Rasmus Lentz, Christian Mosner, Roc Armenter, Diego Kanzig, Mathias Trabandt, Kim Ruhl, and Jonas Arias for insightful comments. All remaining errors are our own. Drozd (corresponding author): Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106 (email: lukaszadrozd[@]gmail[.]com). Tavares: International Monetary Fund, 700 19th St NW, Washington, DC 20431 (email: marinamendestavares[@]gmail[.]com).

1 Introduction

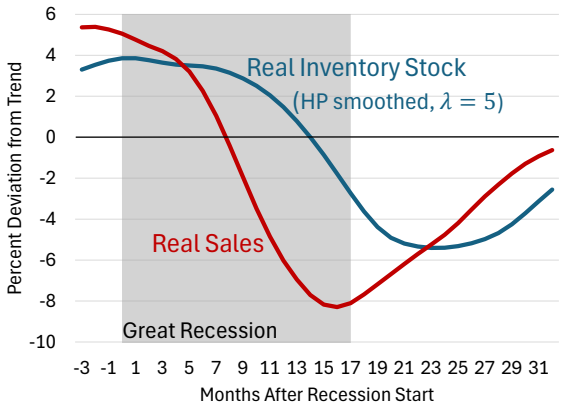
Inventory investment, though small, is highly volatile and arithmetically accounts for a sizable portion of the decline in final sales during recessions (Ramey and West, 1999). Since the early days, inventories have garnered attention due to their potential role in propagating business cycles (Metzler, 1941; Abramovitz, 1950). However, as with any theory, introducing new data to existing models can lead to broader insights. In this vein, inventory theory has been interpreted as supportive of the New Keynesian proposition that markups are countercyclical due to sticky prices, and that sticky prices, rather than sticky wages, play a dominant role in breaking monetary neutrality in the data (Kryvtsov and Midrigan, 2012). Fundamentally, this interpretation owes to the fact that, unless markups are countercyclical, existing theory struggles to account for the fact that inventory stocks fall *less* than sales during recessions, and that the inventory-to-sales ratio is countercyclical.[1]

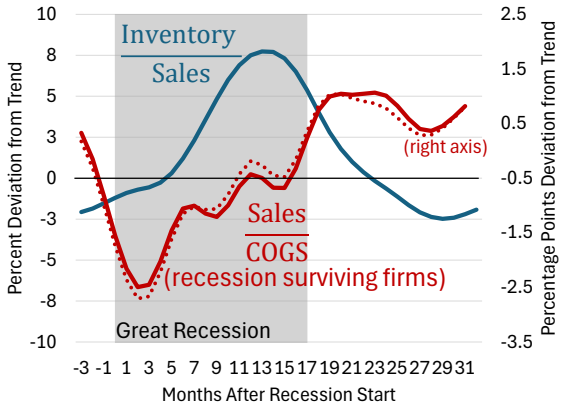
In this paper, we revisit the link between inventory dynamics and markups and argue that the lagging response of inventories to falling sales in recessions points to the presence of other frictions—rather than markups—that delay the unwinding of inventories and cause the inventory-to-sales ratio to rise. We enrich the existing theory with a notion of customer-hailing intangible capital to model this delay. The key mechanism is investment irreversibility in customer-hailing capital and the complementary role of inventories in attracting customers. The proposed mechanism flips the conventional logic on its head: the more markups fall, the longer the lag and the more countercyclical the inventory-to-sales ratio becomes. As a result, inventory behavior is no longer informative about the cyclical properties of markups.

The development of existing theory has been shaped by two robust empirical regularities. The first is that inventory stocks persistently decline during recessions. This pattern—shown in Figure 1 (left panel) and further analyzed in Section 2 using a proxy SVAR—was initially considered puzzling and led to the rejection of models that view inventories as a mere buffer stock shielding production from

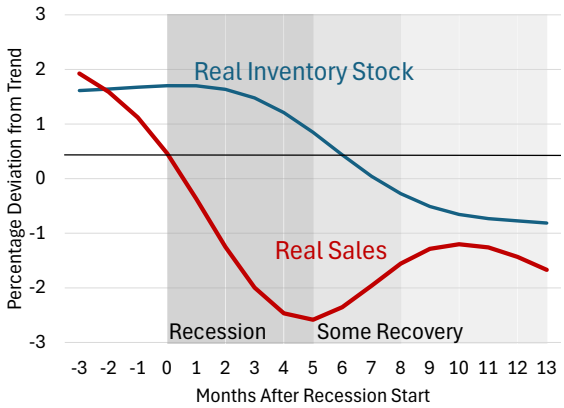
[1] Kryvtsov and Midrigan (2012) develop this argument by evaluating several prominent models of inventory dynamics within the New Keynesian (NK) framework and in response to monetary policy shocks. They show that substantial price rigidity is necessary to match the data and that sticky wages are less relevant because they imply counterfactual inventory dynamics within all existing frameworks. The focus on monetary policy shocks is essential. As shown by Khan and Thomas (2007b), responses to productivity shocks are consistent with standard RBC theory enriched by the S-s model of inventories proposed by Khan and Thomas (2007b).

A. Great Recession:





B. Average for U.S. recessions, 1979–2007:



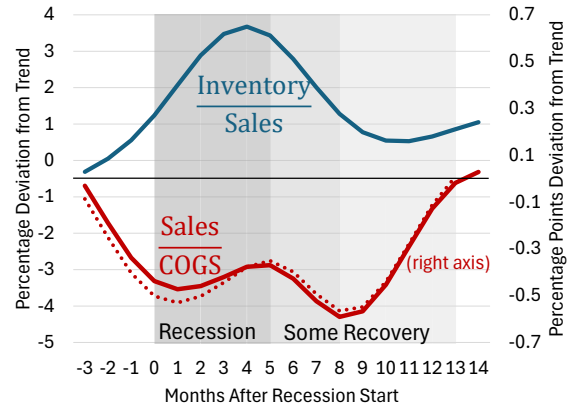


Figure 1: Inventory, sales, and gross profit margins in the U.S. across recession periods, 1979–2011.

Notes: The figure shows deviations from HP trend using a smoothing parameter of $\lambda = 10,000$—with additional smoothing applied to deviation (HP filter with $\lambda = 5$). Real inventory and sales data are plotted for the manufacturing and trade industries (wholesale and retail sectors). The sales-to-COGS ratio is derived from Compustat Quarterly Fundamentals (North America) as described in Sesction 2 and includes firms in the manufacturing and trade sectors with positive sales right before *each* recession and one year after each recession (recession-surviving firms). Dotted lines incorporate output elasticities to map margins onto implied markups accordingly to the methodology developed by De Loecker et al. (2020). See next section for more details. Shaded areas represent the duration of each recession. A detailed list of data sources and the same figure for all sectors can be found in the Online Appendix.

fluctuating demand or costs (hereafter, the production-smoothing motive).[2] First, inventory holdings are too large in the data to be justified by such a motive alone, and second, it implies that inventories should *rise* rather than *fall* during recessions that lower marginal production costs. This discrepancy spurred the development of the current generation of models, in which inventories are either produc-

[2]See Blinder et al. (1998) and Fitzgerald (1997) for a comprehensive review of the early literature on inventory dynamics and the basic data patterns.

tive in generating sales or producing output.[3] However, by tying inventories to current sales, these augmented models struggle to account for the fact that inventory stocks fall *less* than sales during recessions (Figure 1, left panel), and that the inventory-to-sales ratio is strongly countercyclical in the data (Figure 1, right panel). This second regularity links the existing theory to markup dynamics and, hence, the New Keynesian proposition that markups are countercyclical due to sticky prices.

As first noted by Bils and Kahn (2000), when inventories are productive in generating sales or producing output, markups determine profits that firms *concurrently* expect to earn on sales generated by additional inventory holdings. Since firms maximize profits, unless markups rise in recessions, firms unwind inventory stocks in proportion to sales, leaving the inventory-to-sales ratio roughtly unchanged. Supply-side shocks may disrupt this prediction, as shown by Khan and Thomas (2007b) in the context of the RBC setup used by Bils and Kahn (2000). However, as pointed out by Kryvtsov and Midrigan (2012), similar patterns also emerge in response to monetary-policy-induced demand shocks, and so rising markups are needed to account for the conditional response of inventories to demand shocks.[4]

While countercyclical markups address the issue, the countercyclicality of markups is a controversial proposition. The controversy stems from the challenges of measuring marginal costs directly from the data and the fact that the canonical model of production, broadly adopted in macroeconomics, ties markups to gross profit margins (sales revenue over the *average* variable production cost, which in the data we associate with cost of goods sold).[5] Unlike markups, margins and profits are measurable, and both are strongly procyclical in the data—as shown in Figure 1 (right panel) and further documented in Section 2. On the other hand, a major departure from this framework is needed to meaningfully flip this correlation. As we discuss, this has broad-based consequences for economic modeling, and it currently lacks firm empirical support.

[3]This understanding of inventories has been advanced by Kahn (1987), Bils and Kahn (2000), Khan and Thomas (2007b), and Kryvtsov and Midrigan (2012), among others.

[4]A combination of the production-smoothing motive and sales-generating motives could, in principle, explain both facts under acyclical or procyclical markups, but it fails quantitatively. Inventory stocks are too large, and the sales-generating motive dominates in the calibrated models.

[5]By canonical model of production, we mean a setup in which firms have access to a Cobb-Douglas production technology and take variable input prices as given. The link between markups and margins in such a setup follows from the static cost minimization problem and is independent of the price-setting mechanism. This argument, originating from Hall (1988), forms the basis for measuring markups from microdata in De Loecker et al. (2020). We review it in Section 2.

Christiano et al. (1997) were the first to highlight that the mechanism through which sticky prices break monetary neutrality relies on countercyclical profits. Building on this insight, Bilbiie and Kanzig (2023) demonstrate that introducing *procyclical* profits into the canonical New Keynesian model fundamentally alters its implications, rendering recessionary demand shocks *inflationary* rather than *deflationary*. Extending this line of inquiry, Broer et al. (2019) show that incorporating distributional considerations within a Heterogeneous Agent New Keynesian (HANK) framework nullifies the real effects of sticky prices even when profits are *countercyclical*—a mechanism they find implausible to begin with because households do not get wealthier in recessions. The reason is that profits accrue to high-income households who supply little labor, and the wealth effect of rising profits during recessions is central to the transmission mechanism in the sticky price models. In particular, the wedge between the marginal product of labor and the marginal rate of substitution between consumption and leisure plays a negligible role under standard preferences.[6]

A potential resolution of this conundrum is that marginal cost curves are steeper at the aggregate level than those postulated by the canonical model, due to features such as overtime pay or other frictions in adjusting variable inputs (Bils, 1987; Rotemberg and Woodford, 1999; Galí et al., 2007). If true, this could affect the correlation between markups and margins predicted by the standard theory. However, as discussed by Nekarda and Ramey (2020), the claim that such features are *sufficiently* pronounced to *flip* the correlation is debatable. Moreover, even in that case, one cannot "have the cake and eat it too." First, such an approach still undermines the transmission mechanism under sticky prices, as it is *profits*, rather than *markups*, that are central to the real effects of sticky prices. Second, sufficiently steep marginal cost curves introduce new modeling challenges to existing models; for example, they dampen employment volatility, which is difficult to account for to begin with (Kehoe et al., 2022). Finally, as we discuss in Section 2, existing estimates of the production functions-based markups do not seem to indicate such an inversion.

Amid these controversies, our paper contributes to the literature by demonstrating that countercyclical

[6]As Broer et al. (2019) note, "With the kind of preferences used in macroeconomic literature (King et al., 1988) and without profit income accruing to workers, the income and substitution effects from changes in the wage level cancel out the distributional effects of demand shocks in HANK models. (...) The reduced-form link (...) relies on a transmission mechanism that is implausible: output falls in response to a monetary tightening because markups and total profits rise, increasing the representative working household's income and thus her demand for leisure."

markups are not necessary to explain inventory dynamics once the sluggish or delayed response of inventories is accounted for. The fact that inventories and sales bottom out at different times during recessions suggests the presence of such a delay in the data (Figure 1, left panel). The proposed mechanism broadly reaffirms the validity of existing approaches to modeling inventory dynamics because the implied delay is independent of the specific motive for holding inventories.

To model the delayed response of inventories to declining demand, we draw on the extensive macroeconomic literature highlighting the costs firms incur in building demand and attracting customers. Firms allocate substantial resources to selling output, much of which is reflected in SG&A expenses on income statements and, to some extent, in R&D investments aimed at enhancing product appeal or differentiation (He et al., 2024). Several studies have shown that treating these expenses as investments in customer-hailing capital can help explain a broad range of macroeconomic phenomena. The idea that firms invest in demand, and that non-price mechanisms play a crucial role in building demand, is also consistent with the descriptive evidence on the types of investments and strategies firms employ to enter new markets or expand their existing market shares. A notable feature of many of these models is the sluggish response of customer-hailing capital to shocks, which here we attribute to a microfounded mechanism of investment irreversibility.[7]

In our setup, firms attract searching customers by establishing atomless customer-hailing units referred to as *outposts*. The stock of outposts constitutes capital to firms, and setting up an outpost involves a sunk cost. Crucially, outposts hail customers but also segment production and distribution of goods; that is, they are associated with differentiated products and involve distinct production processes. This fundamental friction, stemming from the economy-wide taste for variety, limits the economy's ability to reallocate productive resources, inventories, and customers across outposts. It also generates a random order flow at the individual outpost level, giving inventories a productive role as firms strive to avoid costly stockouts. To generate real effects of demand shocks, our model assumes downwardly rigid wages and fully flexible prices.

[7]See, for example, Bai et al. (2024), Drozd and Nosal (2012), Gourio and Rudanko (2014), and Crouzet and Eberly (2023). The friction introduced here is conceptually similar to the reduced-form frictions considered in this literature, which have been shown to perform well on standard macroeconomic moments. For instance, the friction in Drozd and Nosal (2012) aligns closely with ours and succeeds in matching key macroeconomic patterns. On the empirical side, Argente et al. (2024) provide direct evidence that firms incur substantial non-price costs building market shares.

Following a transient decline in aggregate demand, the key mechanism is that the stock of outposts falls gradually because profit-maximizing firms opt not to liquidate outposts whose useful life exceeds the duration of the shock that lowers demand for their goods. Consequently, at the onset of a recession, firms find themselves holding an excess of customer-hailing capital, which they then utilize by holding inventory so that these units still attract customers. Amid declining sales per outpost and the slow natural attrition of outposts, this leads to a persistent rise in the inventory-to-sales ratio. Declining sales and faster replenishment of inventory amid lower sales rates further amplifies this effect.

The same frictions that drive the inventory-holding motive lead to endogenously variable and *pro-cyclical* markups. The underlying mechanism is novel and extends beyond the context of durable goods.

Amid the economy's limited inventory-holding capacity, as demand falls in recessions, firms lower markups because this increases the time their outposts spend producing output rather than awaiting customers. As we show analytically, markups in such an environment depend on a *modified* demand elasticity that *additionally* involves the times required to produce a good and sell the good by finding a customer for it. The key prediction of this mechanism is that price hikes (discounts) and stockout frequency are positively (negatively) correlated within narrowly defined product categories. As shown by Cavallo and Kryvtsov (2023), such a correlation was prominently featured during the post-pandemic inflation surge across individual products.[8]

We calibrate the model to the data and compare its quantitative predictions to the empirical impulse responses obtained from an estimated proxy SVAR that isolates the response of inventories and markups to a monetary policy shock. To estimate the SVAR, we follow Gertler and Karadi (2015). We find that our model can replicate the empirical response within the 90 percent confidence bands of the proxy SVAR—with the exception of markups, which exhibit too little volatility relative to the data.

Two features are central to our model's ability to quantitatively match the data. First, depreciation and amortization expenses in the data are primarily associated with "maintenance" costs that sustain the

[8]At the aggregate level, our model generates a negatively sloped price Phillips curve and links markup (price) dynamics to observable variables such as backlogs, delivery delays, and stockouts. During the recent inflation outbreak, diffusion indices for these variables exhibited comovement that qualitatively aligns with the predictions of our model. This apparent from the correlation of series published by the Institute for Supply Management (ISM).

durability of outposts, as opposed to capital depreciation that typically offsets the impact of investment irreversibility. A simple way to think about this feature is that, in practice, many complementary but distinct types of capital must work together to enable sales. In such cases, if, say, a unit of capital of type A requires physical replacement while its complementary units of type B and C do not, the firm is compelled to replace depreciating unit A to avoid scrapping the still-functional units B and C. This effectively mutes depreciation as a viable adjustment margin for firms.

Second, while aggregate inventory holdings implied by our model match the inventory-to-sales ratio (monthly sales) of 1.5 in the data, our calibration targets lean inventories on the outpost level. This assumption is necessary for both the stockout motive to arise and markups to be variable, and it is supported by the data.[9] First, the notion of lean inventories is consistent with with the just-in-time (JIT) delivery model widely adopted in manufacturing since the 1980s (Ortiz, 2022). The JIT model relies on precise timing of deliveries to minimize inventories of raw material, which are the majority of inventories in the data. Second, lean inventory holdings are also consistent with the anecdotal evidence suggesting that inventories play a minimal role in production smoothing. For example, according to the firm survey by Blinder et al. (1998) (pages 96 and 277), 67 percent of inventory-holding firms consider inventories "totally unimportant" for smoothing demand fluctuations.[10] Through the lens of our theory, similar volatility of gross output and sales seen in the data implies that inventories per outpost cannot be too large because firms would use them to smooth gross output fluctuations vis-à-vis sales.

Related Search Theory Literature.— Our modeling builds on several key contributions in the search literature. Most notably, it draws on the insights of Wolinsky (1986) and Anderson and Renault (1999), who introduced taste heterogeneity to random search models to address the Diamond paradox. We adopt a similar approach to model product differentiation and market power implied by search

[9] As implied by the results in Khan and Thomas (2007a), this is generally necessary for stockout models to succeed. The presence of Dixit-Stiglitz varieties plays a similar role in Kryvtsov and Midrigan (2012).

[10] Among surveyed firms, approximately three-quarters of the transactions are business-to-business (B2B), and 86 percent for manufacturing firms. While delays may result in lost sales for retailers, predictable delays in B2B relations are acceptable to business partners operating under a JIT model. However, firms that adopt the JIT model have little flexibility in allowing for longer or shorter delay because their production processes heavily relies on timely deliveries of supplies and raw materials.

frictions, and similarly avoid the nonlinearities associated with the Diamond paradox. Our approach is closely related to the recent information-theoretic models in Cheremukhin and Restrepo-Echavarria (2020) and Cheremukhin et al. (2020), which similarly feature targeted search and endogenous information acquisition. The main distinction between their approach and ours lies in the mechanism for information acquisition. In our model, information acquisition arises from explicitly modeled repeated random searches and an assumed distribution of match-specific taste shocks. In contrast, their information-theoretic framework postulates a unit cost of targeting "rare" sellers, as implied by the Kullback-Leibler divergence of distributions.[11] Other relevant contributions include Menzio (2007) and Lester (2011), and on the labor side, Lentz et al. (2024).[12]

2 Data

This section documents the key data patterns using a proxy SVAR à la Gertler and Karadi (2015)—extending the unconditional evidence shown in Figure 1 and discussed in Section 1.

2.1 Methodology

We begin by stating the key identifying assumption that focuses attention on the class of models that our analysis applies to.

2.1.1 Identification Restriction

Let inventory-holding firms in the economy be indexed by $i = 1, 2, 3, \ldots$, and suppose firms take prices of variable inputs as given and produce output according to a production function $Y_i = \mathscr{Y}(\mathscr{V}(V_{1i}, V_{2i}, \ldots), F_{1i}, F_{2i}, \ldots)$, where $V_{1i}, V_{2i}, \ldots$ are variable inputs aggregated into a composite

[11] In a related contribution, Matějka and McKay (2015) develop information theory-based microfoundations for the commonly assumed logit structure in discrete choice models.

[12] In addition, the analysis herein is related to the applied literature that studies the role of inventories in explaining the response of trade to large shocks and the response of prices and output to post-pandemic supply chain disruptions (Alessandria et al., 2010, 2011, 2023).

bundle via a constant returns to scale aggregator $V_i := \mathscr{V}(.)$, and $F_{1i}, F_{2i}, \ldots$ are fixed inputs over the business cycle. Assume the static cost minimization problem is well defined, and the variable production cost of each firm i is $C_i(Y_i) := \min \sum_j (v_j V_{ji})$, subject to $Y_i = \mathscr{Y}(\mathscr{V}(V_{1i}, V_{2i}, \ldots), F_{1i}, F_{2i}, \ldots)$, where v_j is the input price of a variable factor $j = 1, 2, ...$ that the firm takes as given.

Define the following cost-weighted objects derive from this framework:

$$\text{Markup } \mu := \frac{C_i(Y_i)}{\sum_i C_i(Y_i)} \frac{P_i}{C'(Y_i)} \tag{1}$$

$$\text{Gross Margin } \bar{\mu} := \frac{C_i(Y_i)}{\sum_i C_i(Y_i)} \frac{P_i Y_i - C_i(Y_i)}{C_i(Y_i)} \approx \log\left(\frac{\sum_i P_i Y_i}{\sum_i C_i(Y_i)}\right) \tag{2}$$

and

$$\text{Log Output Elasticity } \varepsilon := \frac{C_i(Y_i)}{\sum_i C_i(Y_i)} \log\left(\mathscr{Y}_{V_i}(V_i, F_{1i}, F_{2i}, \ldots) \frac{V_i}{Y_i}\right), \tag{3}$$

where $C_i'(Y_i)$ is the marginal cost derived from the cost minimization problem above. Using the first-order condition and the envelope condition implied by the above cost minimization problems, it can be shown that—regardless of the demand structure determining the price P_i at which each firm sells its output—the markup and margin in this framework are linked through the following relation:[13]

$$\mu = \bar{\mu} + \varepsilon. \tag{4}$$

In light of this relation, our paper focuses on economic models which imply that the impulse response function (time path) of $\bar{\mu}(t)$ and $\mu(t)$ are positively related after a monetary policy shock. Specifically, for any t, there exists a consistent $c(t) > 0$ such that $\bar{\mu}(t) = c(t)\mu(t)$ for sufficiently many periods (to be specified in the context of a specific impulse response function). Based on this assumption, in what follows, we identify the conditional response of $\bar{\mu}$ to a monetary policy shock in the data and assume that the underlying theory relates this observation to the implied dynamics of markups.

A Cobb-Douglas production function obeys this restriction in a strict sense (the last term is a constant). However, as shown by Nekarda and Ramey (2020), as far as the sign of the correlation goes, it applies

[13]This result comes from Hall (1988) and forms the basis of the production-based markup estimation approach developed by De Loecker et al. (2020) and De Loecker and Warzynski (2012). See De Loecker et al. (2020) for derivation of this condition.

to a broader class of models. As we discussed in Section 1, this is the essence of the challenge of accounting for profit dynamics using models featuring countercylical markups.

To provide a partial validation of this restriction, note that the last term can be identified using the state-of-the-art estimates of *output elasticity* due to De Loecker et al. (2020). While the annual frequency of these estimates is a limitation, given the persistence of margin dynamics in the data, significant movements in these elasticities should still be observed if the *output elasticity* reverses the correlation between the average margin $\bar{\mu}$ and the average markup μ. This is not the case, as implied by the linear interpolation of these elasticities based on the adjacent annual values shown in Figure 1 (red dotted series include estimated elasticities).[14]

2.1.2 Measurement and Empirical Framework

We follow Gertler and Karadi (2015) and extract the conditional impulse responses of the key variables shown in Figure 1 to monetary-policy-induced demand shocks. The SVAR serves as the target for our model to replicate.

The original Gertler and Karadi (2015) specification includes the CPI, the one-year nominal T-bill yield (GS1), the Gilchrist and Zakrajsek (2012) excess bond premium (EBP), and industrial production as a measure of gross output. Our formulation replaces industrial production with real sales in manufacturing and trade industries (BEA) and adds real inventory series (BEA) as well as the gross margin series $\bar{\mu}$—which are derived from firm-level data as described below.[15] Real sales and gross output are closely related, in that sales is gross output plus the change of inventory stock (up to measurement discrepancies). As we show in the Online Appendix, replacing industrial production by real sales has little impact on the estimated impulse responses.

All variables are in logs, and where applicable, we multiply logged series by 100 to express them as percentages (percentage deviations for gross margin). Following Gertler and Karadi (2015), we

[14]The time series for these elasticities are sourced from the replication package for De Loecker et al. (2020). They are estimated using Compustat Annual Fundamentals and based on annual estimates of the production functions on a two-digit NAICS level.

[15]Inventory and sales series are sourced from the Federal Reserve Bank of St. Louis: Real Manufacturing and Trade Industries Inventories [INVCMRMT] and Sales [CMRMTSPL], retrieved from FRED, Federal Reserve Bank of St. Louis.

include 12 monthly lags.[16]

We follow De Loecker et al. (2020) by associating the cost of the variable input bundle $C_i(Y_i)$ in (2) with the cost of goods sold (Cogs) and revenue P_iY_i with the reported sales on the income statements (Sales). Cogs is an accounting measure of production costs meant to capture variable production costs, and typically it includes production labor costs, raw materials, and rent equivalent of production facilities. Applied to the data, we use the following formula for the gross margin:

$$\text{Gross Margin } \bar{\mu}_t := \log\left(\frac{\sum_i \text{Sales}_{it}}{\sum_i \text{Cogs}_{it}}\right) \times 100. \tag{5}$$

To obtain Sales and Cogs for a large cross-section of US firms, we use the S&P's Compustat Quarterly Fundamentals (North America) database. This data compiles 10-Q filings of all publicly traded companies in the U.S. We restrict our sample to two-digit SIC sectors 92–96 (manufacturing, wholesale, and retail).[17] We aggregate firms so as to minimize the impact of entry and exit of firms in and out of public trading. In particular, in Figure 1, we use a balanced panel of firms that survive each recession (i.e., firms reporting strictly positive sales one quarter before and one year after each recession). Since firms which exit suffer a larger decline of profits, this approach biases the results against finding procyclical margins. For SVAR, we mimick this approach by using a balance panel within two consecutive quarters. Specifically, for any firm i and attribute $Z_{it} \in \{\text{Sales}_{it}, \text{Cogs}_{it}, \ldots\}$, we calculate the growth rate $g_{it} := (Z_{it+1} - Z_{it})/Z_{it}$ on the firm level, while requiring $Z_{it} > 0$ and $Z_{it+1} > 0$, and use the obtained weights $w_{it} := Z_{it}/\sum_{i\in\mathcal{I}_t} Z_{it}$ to construct a chain-weighted aggregate index using the recursion: $Z_0 = 100$, $Z_{t+1} = \left(1 + \sum_{i\in\mathcal{I}_t} w_{it} g_{it}\right) Z_t$, for $t = 1, 2, 3, \ldots$.[18] The obtained this way time series imply the same patterns as the series in Figure 1—which is reassuring.

One limitation of the S&P Compustat dataset is that it is based on quarterly earnings reports that

[16] Coefficients are estimated over the period 1979:M7 to 2012:M6, with monetary policy shocks identified using the FF4 instrument series, covering 1990:M1 to 2012:M6. Since our VAR is estimated in logs, the inventory-to-sales ratio would be collinear with sales and inventory, which are included as separate series. To derive the predicted impulse response for the ratio, we compute it from the impulse responses of inventory and sales, then bootstrap the ratio to obtain the corresponding confidence bands. The results are identical when we replace the inventory stock with the ratio, as expected. For brevity, we omit the impulse responses of the one-year T-bill rate and EBP, though these variables are included separately in the SVAR, and the T-bill rate serves as the policy instrument. Additional details are provided in the Online Appendix.

[17] Extended results for all firms are available in the Online Appendix and are identical.

[18] We seasonally adjust the final series (in logs) using the LOESS-STL method; that is, we seasonally adjusts the obtain index for Cogs and Sales.

are reported at a lower frequency than our monthly SVAR and thus lag the included macroeconomic series. The lag arises because firms report earnings based on their own fiscal quarters that typically do not align with the calendar quarters. The lag is i.i.d. across firms, as it depends on the incorporation date. Accordingly, it adds a month a half lag to the aggregated series effectively reduces the frequency of the data. To obtain monthly series for gross margins and adjust the timing of earnings releases at the same time, we use a simple approach interpolating monthly values based on the adjacent quarterly values. The Online Appendix explores an alternative and more standard approach: the Chow-Lin interpolation method based on payroll employment as an auxiliary variable, in which case we use three-month forward moving averages to adjust the timing. The results are similar. The Appendix also shows that including margin series has little impact on the estimated impulse responses for all other variables.[19]

2.2 Results

The impulse responses implied by the SVAR, shown in Figure 2, closely mirror the unconditional patterns in Figure 1.[20] The gross margin and the inventory-to-sales ratio exhibit a strong negative correlation, and the inventory-to-sales ratio persistently rises after a negative monetary policy shock. Crucially, as is the case in Figure 1, the response of inventories similarly lags behind sales by approximately 10 months. This lagged response is significant at the 90 percent confidence level.

[19]To implement Chow-Lin method, we use the replication package to Quilis (2018) and interpolate monthly values using payroll employment in manufacturing and trade industries. An alternative approach would be to aggregate all data to quarterly frequency, following the approach of Bilbiie and Kanzig (2023) that maximizes the informativeness of the instrument. We have not tested their approach on our data, and the timing adjustments for earnings releases would still need to be addressed. Bilbiie and Kanzig (2023) show that, for aggregate corporate profits, the switch to quarterly frequency has little impact on the cyclicality of profits.

[20]This is not surprising given how consistent the results shown in Figure 1 are across recessions. For example, similar patterns were observed during the 2001 recession (not shown), when total factor productivity continued to grow along its pre-recession trend (Jorgenson et al., 2004).

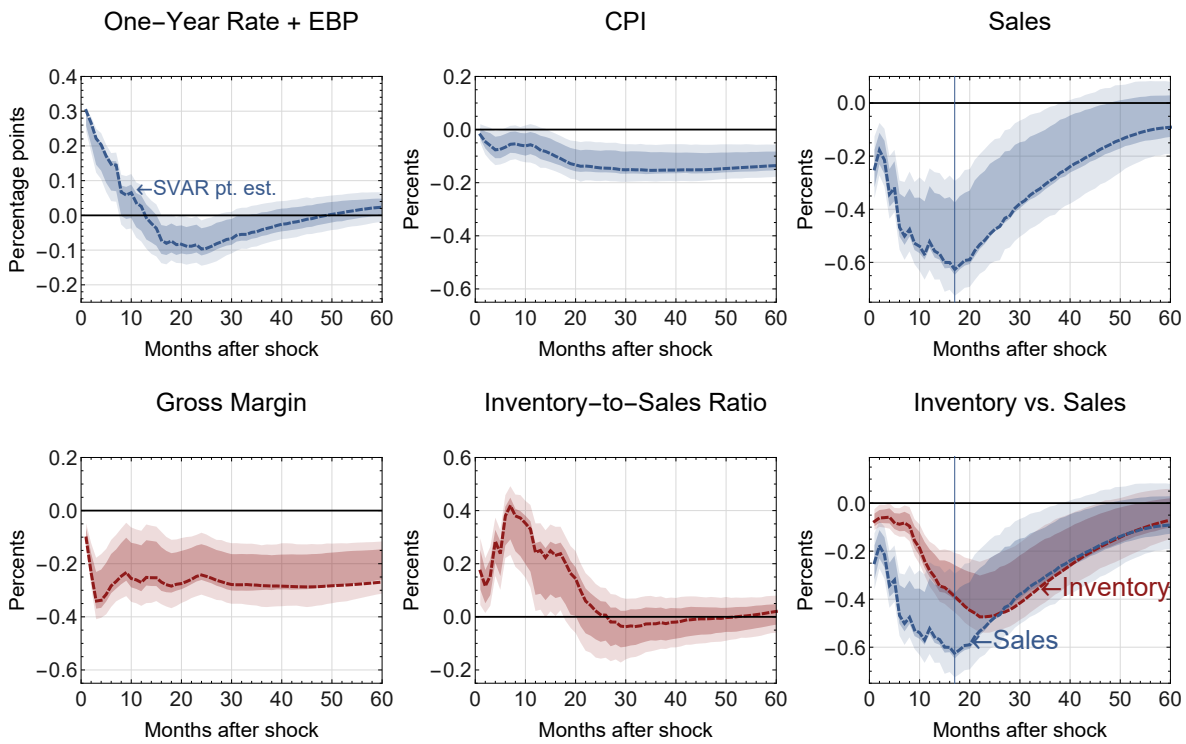


Figure 2: SVAR impulse responses to a monetary policy shock.

Notes: The figure shows the impulse response to a monetary policy shock from the SVAR, which includes the one-year T-bill rate (GS1), excess bond premium (EBP), CPI, real sales, real inventory, and average margin (measured markup) constructed as described in the Data section. The impulse response for the inventory-to-sales ratio is derived from the component series and bootstrapped separately. Light shaded areas represent the 90 percent confidence bands, while dark shaded areas represent the 68 percent bands. The dotted line is the point estimate. A detailed list of data sources is provided in the Online Appendix.

3 Theory

Our model is set in partial equilibrium and considers an "MIT" monetary policy shock under perfect foresight. This setup facilitates the comparison of the model to the SVAR from Section 2.

3.1 Environment

Time is continuous and indexed by $t \in [0, \infty)$. The economy evolves along a path determined by an exogenous *MIT monetary policy shock* that occurs unexpectedly at $t = 0$ and determines the trajectory of the discount rate $\{\rho_t\}_t$ and the aggregate (retail) demand schedule $\{D_t(P_t) := D_{0t}P_t^{-\varepsilon}\}_t$, where P_t is the price level, $\varepsilon > 0$ is aggregate demand elasticity, and D_{0t} determines the level of aggregate

demand.[21]

There are three types of agents in the economy: distributors, shoppers (distributors' representatives), and producers. Producers are all identical and manufacture *intermediate goods* and sell them to shoppers through atomless units referred to as *outposts*. Shoppers are sent out by the distributors and match with outposts to bring intermediate goods back to the distributors (one good per shopper). Distributors are all identical and produce final consumption goods and sell them in a centralized, Walrasian *retail market*. The market-clearing retail price, denoted by P_t, determines the value of the final good and, by extension, the value of intermediate goods and matches for the distributor.

Labor is the only input to production and it is supplied freely at a rigid wage $v > 0$. The wage serves as the numéraire, but we retain the symbolic notation for clarity and carry v throughout. The equilibrium concept in the decentralized market with search frictions restricts attention to pure strategies and assumes symmetric behavior across identical agent types.

3.1.1 Distributors and Search Frictions

At each instance of time t—defined as time interval $(t, t + dt]$ of vanishing length for clarity of exposition—the representative distributor sends a mass of atomless shoppers who search on her behalf, match with producers' atomless outposts, and bring a unit of an intermediate good each. Shoppers must follow the distributor's policy, and operating shopping technology involves both a fixed cost, $\chi v > 0$, and a variable cost—which we detail after describing the shopper's search technology. Distributors bear these costs either through arm's-length transactions with affiliated shoppers or because the shopping sector operates competitively under free entry. Matches are transient and anonymous. Representative distributor is small enough to ignore the impact of her decisions on the the aggregate state of the economy.

Intermediate goods are differentiated in that a given match delivers a unit of intermediate good and yields $1 + \eta$ units of the final good to the distributor, where η is a match-specific idiosyncratic random

[21] Since the transmission of monetary policy to aggregate demand is not the focus of the paper, we abstract from it. The reason is that the standard frictionless household problem predicts an immediate "jump" in consumption demand in response to interest rate hike. This is at odds with the gradual decline seen in the data (our SVAR) and it would distance our model from the data for reasons that are unrelated to our mechanism.

variable referred to as the *preference shock*. The shock η is assumed exponentially distributed with mean $\eta_0 > 0$, and we denote its probability density by $g(\eta) = \eta_0^{-1} \exp\left(-\eta_0^{-1}\eta\right)$ and its cumulative distribution by $G(\eta) = 1 - \exp\left(-\eta_0^{-1}\eta\right)$.

The distributor produces $Q_t\left(1 + \mathbb{E}_\pi\left\{\eta\right\}\right)$ units of the final good, where $\mathbb{E}_\pi\left\{\eta\right\}$ depends on the distributor's policy function π—which we detail below. Accordingly, the distributor's expected revenue from a match is $P_t Q_t\left(1 + \mathbb{E}_\pi\left\{\eta\right\}\right)$, and the distributor's costs include the prices paid by the shoppers and search costs. The retail price P_t of a unit of the final good clears the retail market, implying

$$Q_t\left(1 + \mathbb{E}_\pi\left\{\eta\right\}\right) = D_t(P_t) \equiv D_{0t} P_t^{-\varepsilon}. \tag{6}$$

Since aggregate demand level D_{0t} is assumed exogenous, from the point of view of the wholesale trade, we abuse the language and directly refer to Q_t as the *aggregate demand*.

Consider now a single atomless shopper who brings a good characterized by a quoted price $\tilde{p}_t$ and a shock η. The distributor's gross surplus generated by the match delivered by this shopper is

$$\underbrace{P_t(1+\eta) - \underbrace{\tilde{p}_t}_{\text{quoted price}}}_{\text{distributor's surplus}} = P_t - \underbrace{\left(\tilde{p}_t - \eta P_t\right)}_{p_t\text{, effective price}}. \tag{7}$$

Assuming the distributor cares only about the surplus and ignores any spurious characteristics of the match, the above expression tells us that $p_t := \tilde{p}_t - \eta P_t$ is the sufficient statistic summarizing the match for the distributor. We refer to p_t as the *effective price* of the good.

The distributor's policy function is an indicator function on the set of measurable tuples $(\tilde{p}, \eta)$, and hence on the effective prices p under the above assumption. As we show next, the distributor's *optimal* policy can be represented as *reservation effective price* $\bar{p}_t$ that the shopper cannot exceed, or equivalently, as *search precision* $\pi_t := \Pr\left(p \geq \bar{p}_t\right)$ that the shopper must obey while looking for the lowest price. The on-to-one relationship between these two representations is given by the following identity:[22]

[22] $\Pr\left(p \geq \bar{p}_t\right)$ is strictly monotone in our model, but this not necessary for this result to hold. If the probability function is weakly monotone—which is always the case—we can define the reservation price as the *lowest* price yielding the same probability $\Pr\left(p \geq \bar{p}_t\right)$ to guarantee bijective mapping. Such a restriction does not affect the distributor's surplus and

$$\pi_t := \Pr\left(p_t \geq \bar{p}_t\right) = \Pr\left(\tilde{p}_t^* - \eta P_t \geq \bar{p}_t\right) = \Pr\left(\eta \leq \frac{\tilde{p}_t^* - \bar{p}_t}{P_t}\right) = G\left(\frac{\tilde{p}_t^* - \bar{p}_t}{P_t}\right), \tag{8}$$

where the last equality follows from the fact that in equilibrium all outposts, by being identical, quote the same price $\tilde{p}_t^*$. (All Proofs are in the Online Appendix unless otherwise noted. Throughout, $\Pr(.)$ denotes the context-relevant probability function.)

Lemma 1. *The distributor's optimal policy can be represented either by a reservation effective price $\bar{p}_t$—the highest effective price a shopper is allowed to pay—or by search precision $\pi_t := \Pr\left(p_t \geq \bar{p}_t\right)$, with $0 \leq \pi_t \leq 1$, that the shopper must follow while searching for the lowest effective price.*

Shopper Search.— Shoppers have access to a technology that enables them to pull (identify) a *random* effective price from all quoted prices within a time interval of order $(dt)^2$ at a resource cost $c_0 v > 0$. Since dt is infinitesimal, shoppers can pull as many quotes as they wish within an instance of time. As a result, essentially all shoppers return with a good by $t + dt$ as long as there is a strictly positive mass of prices admitted by the distributor's policy.

The lemma below derives the expected search cost and the expected surplus under an arbitrary distributor's policy π. The intuition is straightforward. Shopper search is a geometric process because the probability of finding a good priced below the reservation price is fixed at $1 - \pi$. Accordingly, the expected number of searches corresponds to the mean of the standard geometric distribution. Ignoring technical considerations related to discounting, multiplying the expected number of searches by the unit search cost c_0 gives the expected total search cost $c(\pi)$ stated in the lemma. The expected surplus—$s(\pi)$ in the lemma—is cumbersome to derive, but conceptually it corresponds to the conditional mean effective price on $\bar{p}$. Since all producers quote the same price $\tilde{p}^*$ in equilibrium, this conditional mean is effectively determined by the conditional mean of the preference shock η.

Lemma 2. *Let $0 < \pi_t \leq 1$ be the search precision associated with some feasible reservation effective price $\bar{p}_t$. In equilibrium with single quoted price $\tilde{p}_t^*$, the expected search cost is*

$$c\left(\pi_t\right) = c_0\left(1 - \pi_t\right)^{-1}, \tag{9}$$

hence constitutes an equivalent representation of her policy function.

the expected surplus for the distributor is

$$s_t(\pi_t) := (1-\pi_t)^{-1} \int_{[0,\bar{p}_t]} (P_t - p)\, \boldsymbol{Pr}(dp \mid p \leq \bar{p}_t) = P_t(1+\eta_0) - \tilde{p}_t^* - \eta_0 P_t \log(1-\pi_t) \quad (10)$$

and the associated reservation price is[23]

$$\bar{p}_t = \tilde{p}_t^* - P_t G^{-1}(\pi) = \tilde{p}_t^* + \eta_0 P_t \log(1-\pi_t). \quad (11)$$

Distributor Problem.— The distributor chooses policy π to maximize the expected *net* surplus from a match: $s(\pi) - c(\pi)v$, where $s(\pi)$ and $c(\pi)v$ are given by the expressions in Lemma 2. We assume zero profits in distribution, and given the fixed cost of shopper search is χv, the equilibrium net surplus must satisfy

$$y_t := \max_{0\leq\pi\leq 1} \{s_t(\pi) - c(\pi)v\} = P_t - \tilde{p}_t^* - \eta_0 P_t \log\left(\frac{c_0}{\eta_0}\frac{v}{P_t}\right) = \chi v. \quad (12)$$

The implied optimal search precision is

$$\pi_t = 1 - \frac{c_0}{\eta_0}\frac{v}{P_t}, \quad (13)$$

and the associated reservation effective price is given by (11).

Outpost-level Demand.— Shoppers accept matches with a single producer outpost at a Poisson rate that depends on the effective price associated with the quoted price $\tilde{p}_t$, the preference shock realization η, and the distributor's policy $\bar{p}_t$. Since all producer outposts quote the same price in equilibrium, and shoppers almost surely match within a given instance of time, the equilibrium arrival of shoppers who accept a match with a given atomless producer outpost is

$$\Lambda_t = \frac{Q_t}{M_t}, \quad (14)$$

[23] The formula for $\mathbb{E}_\pi\{\eta\}$ is not needed here, but it immediately follows from (10) in the proof (see Remark 1).

where M_t is the mass of producer outposts available for matching at a given instance of time, and we refer to Λ_t as the *outpost-level demand*.

Since producers have market power and act as price setters, to determine the equilibrium quoted price $\tilde{p}_t^*$, we need to known how the outpost-level demand changes when an outpost deviates from the equilibrium price to some other price $\tilde{p} \neq \tilde{p}_t^*$—assuming all other outposts quote $\tilde{p}_t^*$. Since such a deviant knows that a shopper accepts its quote when her realized preference shock η satisfies $p := \tilde{p} - \eta P_t \leq \bar{p}_t$, the good is sold if

$$\eta \geq \frac{\tilde{p} - \bar{p}_t}{P_t} \equiv \frac{\tilde{p} - \tilde{p}_t^* + \tilde{p}_t^* - \bar{p}_t}{P_t}, \tag{15}$$

as opposed to this $\eta \geq \frac{\tilde{p}_t^* - \bar{p}_t}{P_t}$ for the equilibrium price. Accordingly, the arrival rate of shoppers accepting the deviant's price gets augmented by the conditional probability that the preference shock lies above the cutoff associated with the deviant's quoted price but below the cutoff associated with the equilibrium quoted price. Hence, the arrival rate to the deviant outpost is

$$\lambda_t\left(\tilde{p}, \tilde{p}_t^*\right) := \underbrace{\frac{Q_t}{M_t}}_{\Lambda_t\text{, aggregate demand}} \underbrace{\frac{1 - G\left(\frac{\tilde{p} - \tilde{p}_t^*}{P_t} + \frac{\tilde{p}_t^* - \bar{p}_t}{P_t}\right)}{1 - G\left(\frac{\tilde{p}_t^* - \bar{p}_t}{P_t}\right)}}_{\text{price impact functional}}. \tag{16}$$

In equilibrium, since $\tilde{p} = \tilde{p}_t^*$, demand elasticity is

$$\frac{\partial_{\tilde{p}} \lambda_t\left(\tilde{p}, \tilde{p}_t^*\right)\big|_{\tilde{p} = \tilde{p}_t^*}}{\lambda_t\left(\tilde{p}_t^*, \tilde{p}_t^*\right)} \tilde{p}_t^* = -\frac{g\left(\frac{\tilde{p}_t^* - \bar{p}_t}{P_t}\right)}{1 - G\left(\frac{\tilde{p}_t^* - \bar{p}_t}{P_t}\right)} \frac{\tilde{p}_t^*}{P_t} = -\eta_0^{-1} \frac{\tilde{p}_t^*}{P_t}, \tag{17}$$

where $\partial_{\tilde{p}}(\cdot)$ denotes partial derivative with respect to the sub-scripted variable. As we shall see, we chose the exponential distribution of the shock because the implied by it elasticity results in constant markup pricing.

3.1.2 Producers and Outposts

Producers set up atomless outposts to match with searching shoppers. Producers own these outposts and make pricing and production decisions for them. As with the distributors, producers are small enough to ignore the impact of their decisions on the aggregate state of the economy. In what follows, we consider the problem of a representative producer.

Not all producer outposts attract shoppers. Specifically, we assume that an outpost attracts shoppers if and only if it has inventory stock available for sale (either raw materials for production or finished goods, depending on the sector). Furthermore, we assume that inventory holding costs are *sufficiently high* such that outposts *choose* to hold a single (normalized) unit of stock on the equilibrium path. For clarity, we state this assumption formally at the end of the section and proceed. Accordingly, existing outposts transition between two idiosyncratic states: a *production* state (no stock, no search) and a *marketing* state (unit of stock, search).

Let M_t be the measure of outposts in the marketing state, as we introduced in the context of (14), and let N_t be the measure of outposts in the production state. Conceptually, we will think of the marketing state as representing an outpost holding a unit of inventory, either as a finished good available for customer sampling or as raw materials enabling timely production for a prospective customer. In manufacturing sectors, the first interpretation is more applicable, while in retail/wholesale sectors, the second interpretation is more relevant. In either case, M_t represents the inventory stock held in the economy, while $M_t/Q_t \equiv \Lambda_t^{-1}$ corresponds to the inventory-to-sales ratio of the economy, where Q_t is *proportional* to *real* sales (i.e., sales measured at a fixed-period price).

Producer Problem.— The value of an outpost evolves dynamically. In the marketing state, it is described by the HJB equation of the form:

$$\rho_t V_{1t} = -(\zeta_0 + \zeta)v + \max_{\tilde{p}} \left\{ \lambda_t \left(\tilde{p}, \tilde{p}_t^* \right) \left(\tilde{p} + V_{0t} - V_{1t} \right) \right\} - \delta V_{1t} + \dot{V}_{1t}. \tag{18}$$

The left-hand side represents the opportunity cost of holding the outpost and the right-hand side represents the flows and capital gains associated with operating the outpost. These include (from left to

right): i) the fixed cost of maintaining the outpost and its inventory holding capacity (warehousing capacity, hereafter), $\zeta_0 v \geq 0$; ii) the variable cost of holding stock, $\zeta v \geq 0$; iii) the scaled by the Poisson arrival rate of purchasing shopper $\lambda_t\left(\tilde{p}, \tilde{p}_t^*\right)$ revenue from selling a good a the chosen quoted price $\tilde{p}$; iv) the similarly scaled capital loss associated with the state transition implied by successful sale, $V_{0t} - V_{1t}$; iv) the exogenous Poisson destruction event of the outpost, $-\delta V_{1t}$, where $\delta > 0$;[24] and v) the capital gain/loss $\dot{V}_{1t}$ associated with the passage of time and the evolving aggregate state. The quoted price maximizes these flows, satisfying the first-order condition:

$$\tilde{p}_t = V_{1t} - V_{0t} - \frac{\lambda_t\left(\tilde{p}_t, \tilde{p}_t^*\right)}{\partial_{\tilde{p}} \lambda_t\left(\tilde{p}_t, \tilde{p}_t^*\right)}. \tag{19}$$

Analogously, the value of an outpost in the *production* state V_{0t} evolves according to

$$\rho_t V_{0t} = -\zeta_0 v + \max_{0 \leq \hat{\tau} \leq \tau} \left\{\hat{\tau}\left(-v + V_{1t} - V_{0t}\right)\right\} - \delta V_{0t} + \dot{V}_{0t}. \tag{20}$$

Since shoppers ignore such outposts, the value is solely derived from capital gains associated with state transition, $V_{1t} - V_{0t}$, which arrives at a Poisson rate $\hat{\tau} < \tau$. The flows on the right-hand side comprise the previously seen cost of maintaining the outpost and the one-time production cost v incurred when the Poisson production fairy arrives. The production rate $\hat{\tau}$ is chosen by the producer optimally and it is constrained by the maximum production rate $0 < \tau < \infty$; the first-order condition is

$$\hat{\tau}_t = \begin{cases} \tau & \text{if } V_{1t} - V_{0t} \geq v, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Entry and Exit of Outposts.— In managing the stock of outposts, producers choose the entry (creation of new outposts) rate $a_t \geq 0$ and the exit/liquidation rate $d_t \geq 0$. Entry involves a sunk cost ϕv upon the realization of the underlying entry Poisson event that adds an outpost, and it costs $\kappa_a (a_t - \delta)^2 (M_t + N_t)$ whenever $a_t \geq \delta$, and zero otherwise, where $\kappa_a > 0$. Liquidation are for free, but liquidation rate similarly involves a flow c cost $\kappa_d d_t^2 N_t$, where $\kappa_d > 0$. We impose convex adjustment

[24] We assume that stock becomes worthless if the outpost depreciates. This assumption will not be quantitatively important and it saves on notation.

costs on liquidations to simplify the notation, and in reporting all our results we exclusively focus on the limiting case of the model that involves negligible cost $\kappa_d \to 0$. (To simplify notation, we assume liquidations target outposts in $N_t > 0$, and hence $V_1 > V_0$ applies on the equilibrium path.)

Producers maximize choose entry and liquidation rates to maximize the expected value of their portfolio of outposts:

$$\mathcal{W}_t := \max_{a\geq 0, d\geq 0} \left\{ V_{1t} M_t + V_{0t} N_t + |a|(M_t + N_t)(V_{0t} - \phi v) - \kappa_a a^2 (M_t + N_t) - |d| N_t V_{0t} - \kappa_d d^2 N_t \right\}. \tag{22}$$

The optimal entry/liquidation policy satisfies the first order conditions: $d_t = -\frac{1}{2} \min\left\{V_{0t}, 0\right\} / \kappa_d$ and $a_t = \delta \, \mathbf{1}_{\{d_t = 0\}} + \frac{1}{2} \max\left\{V_{0t} - \phi v, 0\right\} / \kappa_a$, where $\mathbf{1}_{\{.\}}$ is an indicator function which equals one when the sub-scripted condition is true and zero otherwise. Accordingly, the stock of outposts is follows the law of motion given by:

$$\dot{M}_t = \hat{\tau} N_t - \Lambda_t M_t - \delta M_t, \text{ and } \dot{N}_t = \Lambda_t M_t - \hat{\tau}_t N_t - (\delta + d_t) N_t + a_t (M_t + N_t). \tag{23}$$

The policy of the producer implies that the stock of outposts evolves without interventions (i.e., without liquidation or entry) as long as $0 < V_{0t} < \phi v$. Since setting up an outpost involves a sunk cost, this is a nonempty set and we refer to it as the *inaction region*. The presence of the inaction region is central to the mechanism of our model.

Inventory Holding Costs.— Inventory holding costs or warehousing capacity costs are assumed sufficiently high to ensure outposts choose not to produce in the marketing state. In our calibration, we choose the lowest cost possible and report it as the calibrated value.

Define auxiliary values $\left(V_{1t}^{+}, V_{2t}\right)$ that capture the payoff from an off-equilibrium behavior. These

values follow the following HJB equations of the form:

$$\rho_t V_{1t}^+ = -\left(\zeta_0 + \sigma\zeta_0 + \zeta\right) v + \max_{\tilde{p},\hat{\tau}\leq\tau} \left\{\hat{\tau}\left(-v + V_{2t} - V_{1t}^+\right) + \lambda_t\left(\tilde{p}, \tilde{p}_t^*\right)\left(\tilde{p} + V_{0t} - V_{1t}^+\right)\right\} - \delta V_{1t}^+ + \dot{V}_{1t}^+, \tag{24}$$

$$\rho_t V_{2t} = -\left(\zeta_0 + \sigma\zeta_0 + 2\zeta\right) v + \max_{\tilde{p}} \left\{\lambda_t\left(\tilde{p}, \tilde{p}_t^*\right)\left(\tilde{p} + V_{1t}^+ - V_{2t}\right)\right\} - \delta V_{2t} + \dot{V}_{2t}.$$

Here, V_{1t}^+ represents the value of an outpost choosing to produce while in the marketing state, and V_{2t} corresponds to the value of the *off equilibrium* marketing state with *two* units ready for sale. The parameter $0 < \sigma \leq 1$ represents the share of the outpost maintenance cost ζ_0 required to increase warehousing capacity to accommodate a deviation to two units, while 2ζ represents the additional variable inventory holding cost—twice as large in the two-unit marketing state corresponding to V_{2t}. We assume the following condition applies in the steady state (denoted by the superscript 'ss' throughout) and hence after a negative demand shock:

Assumption 1 (A1)**.** *In steady state, ζ or σ are such that $V_{1t}^{+,ss} \leq V_{1t}^{ss}$.*

3.1.3 Equilibrium Definition

The definition of equilibrium is standard and it involves a set of paths that satisfy the stated above equilibrium conditions after the shock.

Definition 1. Given the path for the monetary policy shock $\{\rho_t, D_{0t}\}$ and the initial state $\{M_0, N_0\}$, the *perfect foresight equilibrium* comprises: the paths of: i) prices $\{\tilde{p}_t^*, P_t\}$, ii) demand functions $\{\lambda_t(\cdot,\cdot)\}$, iii) distributor's policy $\{Q_t, \pi_t\}$, iv) outpost masses $\{M_t, N_t\}$, iv) outpost values $\{V_{0t}, V_{1t}\}$, and v) producer entry and exit/liquidation policy $\{d_t, a_t\}$, such that conditions (8), (12), (13), (6), (16), (18), (19), (20), (21), (22), and (23) are all satisfied.

Steady state equilibrium is defined analogously but it additionally assumes a constant path for demand $D_{0t} \equiv D^{ss} > 0$ and the discount rate $\rho_t \equiv \rho^{ss} > 0$, and requires a constant path for $N_t = N^{ss}$. The lemma below shows that the steady state equilibrium exists. The parametric restriction in this lemma is needed to ensure that distributors can make strictly zero profits in equilibrium.

Assumption 2 (A2). $c_0 \frac{1-\eta_0}{\eta_0} < \tau^{-1}\left(\phi\left(\delta+\rho\right)+\zeta_0+\tau\right)+\chi$.

Lemma 3. *Steady-state equilibrium exists and is unique.*

3.2 Characterization of Equilibrium

This section analytically characterizes the key mechanism of the model. We discuss what drives variability of markups and why markups correlate negatively with the inventory-to-sales ratio after a recessionary demand shock.

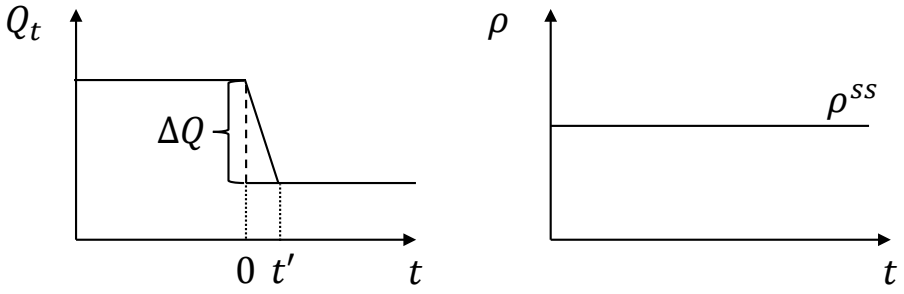


Figure 3: Assumed pure demand shock path in model analytics.

Notes: The figure shows the shock assumed for the purpose of analytic equilibrium characterization (Section 3.2).

We focus on a pure and permanent demand shock, which assumes that—as shown in Figure 3 above—D_{0t} follows a path such that Q_t either experiences a discrete "jump" downward by some $\Delta Q < 0$ at $t = 0$ or it follows a sufficiently steep linear and continuous path approximating this "jump" between $t = 0$ and some $t' > 0$. The path of ρ is fixed at its steady-state level.

For concreteness, Figure 4 plots the impulse responses generated by our calibrated demand shock and our calibrated model that best fits the SVAR-implied impulse response of the inventory-to-sales ratio. In this case, note, the shock is gradual and demand recovers. Nonetheless, it is steep enough to imply a decline in the outpost-level demand Λ_t, which will be key in what follows next. The figure examines two scenarios to decompose the shock to the demand channel (considered here) and the direct interest rate channel (compare 'Model constant ρ') versus 'Model SVAR ρ'). While quantitatively important, the direct effect of the interest rates does not change the picture of how the economy responds to the

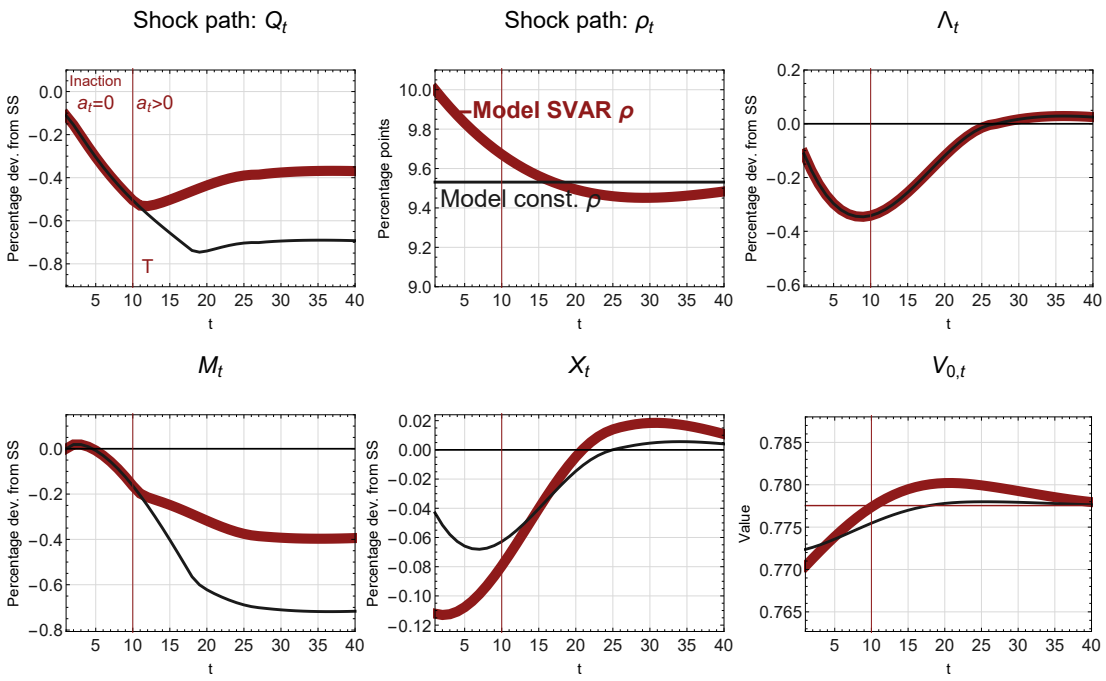


Figure 4: Impulse responses to the calibrated demand shock in the (calibrated) model.

Notes: The figure shows impulse responses implied by the calibrated model to a demand shock (top-left panel). The path of the demand shock, $\{Q_t\}$, is set to match the the smoothed out path of the inventory-to-sales ratio implied by the SVAR, which maps onto Λ_t^{-1} in the model. The figure considers two calibrations: 1) 'Model SVAR ρ' assumes SVAR path for ρ_t. 2) 'Model constant ρ' assumes fixes steady state value for ρ. The response of markups and prices for both models is shown in Figure 6.

shock. The duration of the inaction zone is indicated by T; that is, between $t = 0$ and $t = T$, we know that $0 < V_0 t < \phi v$.

3.2.1 Variable Markups

We begin by characterizing the mechanism that varies markups.

To define the markup implied by the model, it is instructive to define the value of stock as $X_t := V_{1t} - V_{0t}$; subtracting (20) side-by-side from (18), it is clear that this value follows the HJB equation of the form:

$$\rho X_t = -\zeta v + \max_{0 \leq \hat{\tau} \leq \tau, \tilde{p}} \{\lambda(\tilde{p}, \tilde{p}_t^*)(\tilde{p} - X_t)\} - \hat{\tau}(-v + X_t) - \delta X_t + \dot{X}_t. \tag{25}$$

As noted earlier, the policy of an outpost (price/production rate) is a function of the value of stock X_t,

and by the first order condition implied by the above problem we obtain the following formula for the markup (also gross margin):

$$\mu_t := \frac{\tilde{p}_t^* - v}{v} = \frac{X_t - v}{v} + \eta_0 P_t. \tag{26}$$

This first order condition shows that markup variation in our model can arise from only two sources: an additive term linked to the retail price P_t (last term), and the ratio of production cost to the user cost of outpost $X_t - v$. In the steady state, note that this user cost is given by $X^{ss} - v = v\left(\zeta_0 + \phi(\delta + \rho)\right)\tau^{-1}$. As expected, it comprises the operational cost $\zeta_0 v$ incurred over the expected production duration τ^{-1}, and the financing flow costs associated with the sunk expense ϕv and accrued at the gross rate $\rho + \delta$.

Since X and μ are interdependent in the above equation, this formula provides only a partial explanation of the economic forces driving pricing decisions. To uncover these forces, we substitute V_0 from (20) into (18) and use the following equivalent maximization of the outpost value in the marketing state (to ease the exposition, here we assume $\zeta_0 = \zeta = 0$):

$$\rho V_{1t} = (\rho + \delta)^{-1} \max_{\tilde{p}} \frac{\tilde{p} - v\tau\left(\delta + \rho + \tau\right)^{-1} + T(\tilde{p}, \tilde{p}_t^*)\dot{V}_{1t} + \mathbf{T_0}\dot{V}_{0t}}{\mathbf{T_0} + T(\tilde{p}, \tilde{p}_t^*)}, \tag{27}$$

where $T(\tilde{p}, \tilde{p}_t^*) := \left(\lambda_t(\tilde{p}, \tilde{p}_t^*)\right)^{-1}$ and $\mathbf{T_0} = \left(\delta + \rho + \tau\right)^{-1}$. This maximization highlights that, effectively, producers maximize the flow of profits per unit of time. Note that the denominator of the objective function is the total time required to sell a good—which includes the time to find a customer (denoted by $T(.,.)$) and the discount factor-adjusted time to restock (denoted by $\mathbf{T_0}$)—and the numerator is the gross profit from a unit sale, adjusted for depreciation, discounting, and capital gains. The first-order condition derived from this maximization gives

$$\tilde{p} = v\tau\left(\delta + \rho + \tau\right)^{-1} + \varepsilon(\tilde{p}, \tilde{p}_t^*) + \mathbf{T_0}\dot{X}_t, \tag{28}$$

and it involves what we refer to as *modified* demand elasticity that determines pricing:

$$\varepsilon(\tilde{p}, \tilde{p}_t^*) := \frac{\mathbf{T_0} + T\left(\tilde{p}, \tilde{p}_t^*\right)}{\partial_{\tilde{p}} T\left(\tilde{p}, \tilde{p}_t^*\right)} \text{ and hence } \varepsilon(\tilde{p}_t^*, \tilde{p}_t^*) = \eta_0 P_t\left(1 + \Lambda_t \mathbf{T_0}\right). \tag{29}$$

It is clear that the above maximization reduces to the standard monopoly problem when $\mathbf{T_0} = 0$. In

that case, $\tilde{p} = v\tau\left(\delta+\rho+\tau\right)^{-1} + \eta_0 P_t$, and markup is constant and detached from the level of demand Λ_t. However, when $\mathbf{T_0} > 0$, this is no longer the case, and the markup is potentially variable because it depends on the outpost-level demand Λ_t through the *modified* elasticity $\varepsilon(\tilde{p}, \tilde{p}_t^*)$.

Intuitively, this additional term captures the following basic idea. At a higher level of markups and prices, outposts wait longer for customers to arrive. When production takes time ($\mathbf{T_0} > 0$), this is costly because the longer waiting time wastes the time that could otherwise be spent producing output and lower the utilization of outposts in production. Since markups have a greater impact on selling time when demand is low and the wait time for customers is longer to begin with—due to its larger share in the total time required to sell a good—*lower* demand incentivizes producers to *lower* markups—giving rise to the additional term in (29). Despite lower profit margins, this strategy increases profits because it accelerates customer turnover and boosts production rate.

To the best of our knowledge, this is a novel mechanism. Notably, it applies to service sectors as well and does not depend on the durability of the good in question or the presence of inventories. For example, in a nail salon, dentist's office, or restaurant—where serving a customer requires time per unit of fixed capital—the same logic applies because serving a customer takes time.

General Equilibrium Propagation.— General equilibrium linkages between wholesale and retail prices further propagate the response of prices and markups in our model. However, markup variability in general equilibrium is fueled by the partial equilibrium mechanism described above.

While deriving prices in closed form is not feasible, an approximate solution can be obtained by performing a first-order Taylor expansion around the steady state of the last term in (12). The linearization of this term yields:

$$\log\left(\frac{c_0}{\eta_0}\frac{v}{P_t}\right) \approx \log\left(\frac{c_0}{\eta_0}\frac{v}{P_{ss}}\right) + \left(\frac{P_t}{P_{ss}}\right)^{-1} - 1. \tag{30}$$

where P^{ss} is the steady-state retail price (a function of parameters). Using this approximation, the lemma below highlights this key result.

Lemma 4. *To the first order approximation around the steady state, prices are*

$$\tilde{p}_t^* \approx X_t\left(1+\Theta\right)+v\Theta\Gamma, \quad P_t \approx \frac{X_t+v\Gamma}{\eta_0}\Theta, \tag{31}$$

where $\Gamma := \frac{\eta_0 P^{ss}}{v}+\chi>0$ *and* $\Theta := \eta_0\left(1+\eta_0 P^{ss}\log\left(\frac{c_0}{\eta_0}\frac{P^{ss}}{v}\right)\right)^{-1}>0$.

3.2.2 Main Result

We now show that the inventory-to-sales ratio rises and correlates negatively with the variable markup. Since the key friction is the fixed cost $\phi v > 0$, we compare two polar cases: i) the *frictionless case (model)* assuming $\phi = 0$, and ii) the *baseline case* assuming $\phi > 0$ (as noted earlier, all results are reported for the limit case $(\kappa_a, \kappa_d) \to 0$).

i) Frictionless Case.— In the frictionless case, the value of stock is constant, and so $\dot{X}_t = 0$ for all t. This follows from the fact that no sunk cost implies $V_0 = \phi v = 0$, which, by the HJB equation in (20), then implies $X_t = \tau^{-1}\left(\zeta_0+\tau\right)v$.

By Lemma 4, markups are constant. Intuitively, this occurs because, producers essentially control the stock of M_t through the entry and exit margins (a_t, d_t). Simply put, producers can generate an infinite inflow to N_t (and hence M_t) and immediately after liquidate outposts in N_t that are no longer needed. Since all this is for free, the the frictionless model essentially boils down to a microfounded version of Bils and Kahn (2000).

However, the analysis of this case is illuminating because it highlights a more general property: pro-cyclical markups are fundamentally *incompatible* with the countercyclicality of the inventory-to-sales ratio. This key insight underpins the findings of Bils and Kahn (2000) and Kryvtsov and Midrigan (2012) and it follows here from the relationship derived straight from (25) and (31):

$$\Lambda^{-1} = \frac{M}{Q} = \frac{\zeta_0+\tau\left(1+\Gamma\right)}{\tau\left(\delta+\zeta+\zeta_0+\rho\right)+\zeta_0(\delta+\rho)}\Theta. \tag{32}$$

In particular, the above formula implies that, had markups varied in the model due to factors such as

non-constant elasticity of customer demand or sticky prices—which would be manifested in Θ or Γ according to Lemma 4—the inventory-to-sales ratio would end up strictly *positively* correlated with the markup, which is counterfactual, as we documented in Section 2. This formula also connects the frictionless model to the discussion at the beginning of Section 2, namely the identifying restriction that we have imposed.

ii) Baseline Case.— We now turn to the analysis of the baseline model and establish the following main result:

Assumption 3 (A3). $\zeta < \tau + (c_0 + \chi)(\delta + \rho + \tau)$.

Proposition 1. *If $\phi > 0$, the inventory-to-sales ratio and the markup exhibit a strictly negative correlation: the former rises and the latter falls, with both gradually returning to their initial steady state thereafter.*[25]

We first prove this result and then provide the intuition for it.

To that end, guess and verify that the assumed aggregate demand shock induces a transient drop in the outpost-level demand Λ, which evolves according to the following differential equation:

$$\dot{\Lambda}_t \equiv (\Lambda^{ss} - \Lambda_t)h_t \Rightarrow \begin{cases} \dot{\Lambda}_t > 0 & \Lambda_t < \Lambda^{ss}, \\ \dot{\Lambda}_t < 0 & \Lambda_t > \Lambda^{ss}, \end{cases} \tag{33}$$

where $h_t > h > 0$ is some continuous and bounded away from zero function of time. Intuitively, the assumed demand shock lowers demand for products on the outpost level, but demand is expected to recover.

By (25) and (31), recall that the value of stock follows the differential equation

$$\dot{X}_t = X_t \left(\underbrace{\rho + \tau + \delta - \Theta\Lambda_t}_{a_1 > 0} \right) - v \left(\underbrace{\tau - \zeta + \Theta\Gamma\Lambda_t}_{a_0 > 0} \right). \tag{34}$$

[25] The result generalizes to the complementary case of Assumption 3. We omit that case from here because it represents a parameterization that is unlikely to be relevant in applications.

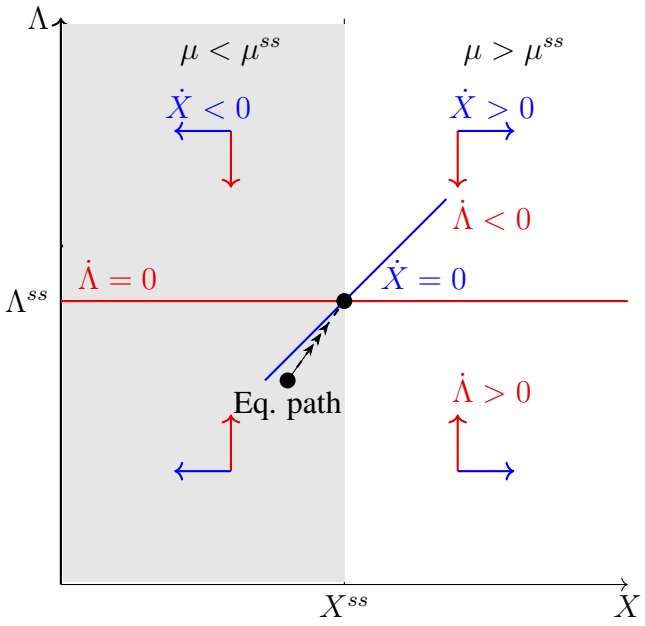


Figure 5: Phase Diagram for Proposition 1.

If ζ is not too high, under Assumption 3, coefficients a_1, a_0 in the above HJB equation are both strictly positive. This is shown in the corollary below.

Corollary 1. *If* $\tau > \zeta - (c_0 + \chi)(\delta + \rho + \tau)$, $a_1 := \rho + \tau + \delta - \Theta\Lambda > 0$ *and* $a_0 := \tau - \zeta + \Gamma\Theta\Lambda > 0$.

Let $\bar{X}(\Lambda)$ be the value of stock such that, for a given value of Λ, $\dot{X} = 0$ in (34). Since in general equilibrium

$$\mu_t \approx \frac{X_t(1+\Theta) - v}{v} + \Theta\Gamma \tag{35}$$

by Lemma 4, Corollary 1 and the above differential equation give

$$\begin{cases} \dot{X}_t > 0 & X_t > \bar{X}(\Lambda), \mu_t^* > \mu^{ss} \\ \dot{X}_t < 0 & X_t < \bar{X}(\Lambda), \mu_t^* < \mu^{ss}. \end{cases} \tag{36}$$

Consider now the phase diagram in Figure 5. The vector field's arrows in this diagram are determined by (36) and the right-hand side of (33); the locus of $\dot{X} = 0$ (referred to as $\bar{X}$ earlier) come from (34), while the locus of $\dot{\Lambda} = 0$ comes from (33). The shaded region indicates the area where the markup falls relative to its steady state value.

As shown in the diagram, the stable manifold is in the bottom-left quadrant relative to the initial state (labeled "eq. path"). Along this path, the markup falls below the steady-state level until the economy fully converges back to the steady state. This establishes the result, and we now turn to validating the assumed path for Λ.

The result is immediate if liquidations are not triggered by the shock. Since $\Lambda_t = Q_t/M_t$, without intervention M_t falls at the rate δ, Λ_t recovers at the same rate and converges back to the steady state. Note that after reaching the steady state $V_0 t = \phi v$, which means that M_t no longer falls by first order conditions implied by the firm value maximization in (22).

If liquidations do occur after the shock, the first-order condition in (22) implies $d_t \propto -\min\{V_{0t}, 0\}$, and hence $V_{0t} < 0$ (recall that $V_0^{ss} = \phi v$). Let $T > 0$ be the first time after the shock when $V_{0T} \geq 0$ and $\dot{V}_{0T} \geq 0$. Such a $T > 0$ must be well defined because the economy cannot liquidate outposts indefinitely and V_{0t} is a continuous and differentiable function of time.[26] Using (20) at time T—rewritten as $(\rho + \delta)V_{0T} = \tau X_{0T} + \dot{V}_{0T}$—and subtracting the steady state version of this equation side-by-side, gives $\tau(X_T - X^{ss}) = (\delta + \rho)(V_{0T} - V_0^{ss}) - \dot{V}_{0T} < 0$, hence $X_T - X^{ss} < 0$ and $\Lambda_T < \Lambda^{ss}$ by the phase diagram in Figure 5. Accordingly, Λ falls as claimed and the economy follows the path shown in the diagram after an initial burst of liquidations.[27]

The intuition behind this result is straightforward. After a transient decline in demand for products, profit-maximizing producers retain the existing outposts because the duration of the shock exceeds their useful life. This "inaction" applies when no liquidations occurs, in which case V_{0t} falls but it remains strictly positive. When liquidations do occur, V_{0t} falls to zero, but liquidations are only part of the response and "inaction" arises thereafter. In both cases producers end up holding an *excess* mass of outposts, which choose to utilize by holding inventory to attract customers. As a result, inventory falls at a rate of at most δ, which is much slower than the decline in sales because the decline in sales is

[26] If perpetual liquidation occurred, M would decline at a rate of at least δ, and since Q_t stabilizes after the initial shock by assumption, Λ_t would have to grow indefinitely at this pace. Eventually, such growth would lead to above steady-state flow profits, causing $V_0 > \phi v$, which contradicts the fact that V_0 must follow a continuous path. We omit the details as they are straightforward from here.

[27] Given that Q remains constant after the initial drop, the law of motion for Λ can be derived by observing that the rate of decline of M is δ. If liquidations do occur, the initial jump in Λ is slightly smaller than the drop in Q and follows the assumed path thereafter. The second part of the proof is more straightforward when Q_t "jumps" discontinuously, but the presented argument covers also the case where Q_t follows a sufficiently steep but continuous path.

additionally driven by the lower arrival rate of shoppers. Moreover, the slowdown in sales causes more outposts to switch from production to marketing than vice versa, and by (23), inventory per outpost increases.

It is worth noting that lower markups—be it via Γ or Θ in (31)—still induce a positive correlation between the markup and the inventory-to-sales ratio in the baseline model. The basic property that markups determine the marginal product of inventory *ceteris paribus* still applies. This is clear from the analog of (32) derived for the baseline model:

$$\Lambda_t^{-1} = \frac{X_t + v\Gamma}{X_t(\delta + \rho + \tau) - \dot{X}_t + v(\zeta - \tau)}\Theta. \tag{37}$$

However, the first term on the right-hand side is now endogenous and it *always* reverses this correlation. This underscores the value of an analytic inspection of the model's mechanism. Note that one could superficially conclude on the basis of this formula that a sufficiently small change in markups is required for this mechanism to work. This is not the case.

While the transient nature of the shock is crucial for the above result, *transient* means that Λ_t and ρ_t are expected to return to their initial values, and not Q_t or M_t. In fact, the decline in Q_t can be permanent, as we assumed here.

4 Quantitative Results

We next calibrate the model and examine whether it quantitatively replicate inventory dynamics implies by the SVAR in the presence of procyclical markups.

4.1 Mapping Model onto the Data

The baseline period of the model ($t = 1, 2, 3$) corresponds to one month. To map model onto the data, we assume (real) sales is Q, real inventory stock is M, the CPI is P, and nominal value added

(nominal GDP) is $P_t Q_t(1 + \mathbb{E}\{\eta\})$.[28] Unlike in the analytic section, to match our data analysis in Section 2, gross margin/markup is measured in logs as $\log(p/v)$. Table 1 lists parameter values.

Table 1: Parameter Values.

η_0	c_0	δ	ϕ	χ	τ	(ζ_0, ζ, σ)	ρ^{ss}	Γ	Θ	(κ_a, κ_d)
.088	.056	4.0×10^{-4}	.78	.7	.5	$(.095, .0, .57)$	7.94×10^{-3}	.87	.081	$(90.9, 0)$

4.1.1 Steady State Calibration Targets

We set the steady state discount rate ρ^{ss} to match the weighted-average real cost of capital (WACC) of 10 percent (annual rate, converted to continuously compounded rate by taking a logarithm). We set τ to match the delivery delay of 60 days in U.S. manufacturing—which corresponds to two (discrete) model periods.[29]

We calibrate the values of χ, η_0, c_0, ϕ, ζ_0 and the auxiliary parameters Θ and Γ to jointly match the following four targets. The mapping between targeted moments and these parameters is analytic and more details can be found in the Online Appendix. Here we only sketch the basic procedure we follows.

The first target is the distribution margin of 38 percent, which in the model corresponds to $(P^{ss} - \tilde{p}^{*ss})/\tilde{p}^{*ss}$. To obtain this target, we use the 1997 domestic-supply-of-commodities input-output (IO) table (15 industries) in conjunction with the use-of-commodities IO table to calculate the corresponding moment in the data.[30] We focus on manufacturing sector/commodity category (out of 15 industries) and cumulate trade and distribution margins reported for manufacturing commodities reported in the use-of-commodities IO table. In particular, this estimate accounts for the re-circulation of man-

[28] $\mathbb{E}_\pi\{\eta\}$ is .26 in our calibration.

[29] For an overview of cost of capital estimates for various industries, see the listing compiled by Aswath Adamodar at https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/wacc.html. Delivery delays in manufacturing come from Deloitte Research Center for Energy & Industrials (2024)—based on the source data from Institute for Supply Management (ISM)—and the report can be found at https://www2.deloitte.com/us/en/insights/industry/manufacturing/manufacturing-industry-outlook.html (see Figure 4). This number has remained steady between 2015 and 2019 according to the earlier reports.

[30] Given the significant transformation of the manufacturing sector during the 2000s and 2010s, much of which lies outside our sample period, we use the 1997 IO tables published by the Bureau of Economic Analysis (BEA). Using more recent tables would increase the targeted moments for markups and distribution margins.

ufacturing goods that are used as intermediate inputs by the manufacturing sector to produce manufacturing commodities. In that case, the reported margins on intermediate goods accrue multiple times.[31] The multiplier implied by re-circulation is 1.38.

Our second target is the gross profit margin of the manufacturing and trade industries of 37 percent, which corresponds to $(\tilde{p}^{*ss} - v)/v$ in the model. We set this target by calculating the ratio of gross surplus to the value added in producer prices for the three sectors in the 1997 use-of-commodities-by-industries IO table.

Our third target is the inventory-to-sales ratio (M^{ss}/Q^{ss}) of 1.5 during the sample period. To calculate this target, we use the series that enter the SVAR.

Our fourth target is the share of expenses on "sales infrastructure" borne by producers, which we set equal to 25 percent and which we associate with $(\phi\delta + \zeta_0(M^{ss} + N^{ss}))/(\tau N^{ss})$ in the model. To set this target, we use the expenses on SG&A relative to gross sales in the Compustat dataset for manufacturing firms—which are about 17 percent during the SVAR's sample period (cost weighted). Since Compustat gross margin pertains to gross sales of an individual firm, and these costs are paid multiple times as manufactured commodities are processed by the manufacturing sector to produce manufacturing commodities, we set it equal to 25 percent to reflect the previously found re-circulation multiplier of 1.38 (see discussion of the first target).[32]

Our fifth target is the average number of three quotes that shoppers look up before they accept an offer, which in the model maps onto $(1 - \pi^{ss})^{-1}$. This is an arbitrary target consistent with best practices, and the results are little changed if we use four or two instead.[33]

Finally, we assume $\zeta = 0$ and support Assumption 1 by selecting an appropriate value for σ. Under this approach, Assumption 1 does not affect the equilibrium path implied by our model, which is why we prefer this approach. The minimal value is $\sigma = 0.64$ and it implies that 2.25 percent of the value

[31] In particular, we do not take into account the manufacturing content of other commodities nor margins on other commodities that enter as intermediate goods into manufacturing sector. This is consistent with our model being a model of production of manufacturing content at some fixed factor cost v.

[32] For an overview of this ratio across industries and countries, see the listings compiled by Aswath Adamodar at https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html.

[33] Search costs in retail prices are 7 percent under our calibration but they depend on all targets. This specific target has little impact on the overall search costs when other targets are fixed. See the discussion of search costs and additional robustness analysis that calibrates the model to lower search costs in the Online Appendix (Section D).

produced by manufacturing and trade industries is required to double the inventory holding capacity of the model economy. This cost reflects the total user cost of structures, land, and the cost of other resources that determine the warehousing capacity in manufacturing and trade industries, which we consider reasonable.

4.1.2 Dynamic Calibration Targets

We set the value of the remaining parameters (δ and κ_a) and calibrate the shock to match the impulse responses generated by the SVAR from Section 2.

The parameter δ governs the rate at which sales and production fall toward the trough of the impulse response in Figure 6 and κ_a governs the rate of recovery. We set the values of these parameters to match the rate of decline and recovery within the 12-month window around the trough in sales predicted by the SVAR.

To calibrate the path for the discount rate $\{\rho_t\}$, we target the impulse response of the sum of the one-year T-bill rate and the excess bond premium (EBP) implied by the SVAR from Section 2. We assume that, after 50 months, the SVAR-implied rate reaches or is expected to converge to the steady state. Using the expectations theory, we back out the implied monthly rates and convert them to continuously compounded rates by taking a natural logarithm. We fit a third-degree polynomial to obtain a continuous function of time as the input to solve the differential equations. (A more detailed description is in the Online Appendix.)

To calibrate the aggregate demand path $\{D_{0t}\}$, we target the SVAR-implied trajectory of the inventory-to-sales ratio, $\Lambda^{-1} = M_t/Q_t$. We fit a two-piece third-degree polynomial to the positive and separately the negative part of the empirical impulse response—assuming continuity at the connection point.[34] The fitted polynomials yield the impulse responses shown in Figure 6. (A more detailed description is in the Online Appendix.)

The paths of Λ and ρ are sufficient to solve the HJB equations in (18) and (20). Given the solution of

[34]The first polynomial covers the positive portion from month $t = 0$ to $t = 25$, and the second covers the negative portion from month $t = 25$ to $t = 60$. We ensure continuity at the transition point at $t = 25$.

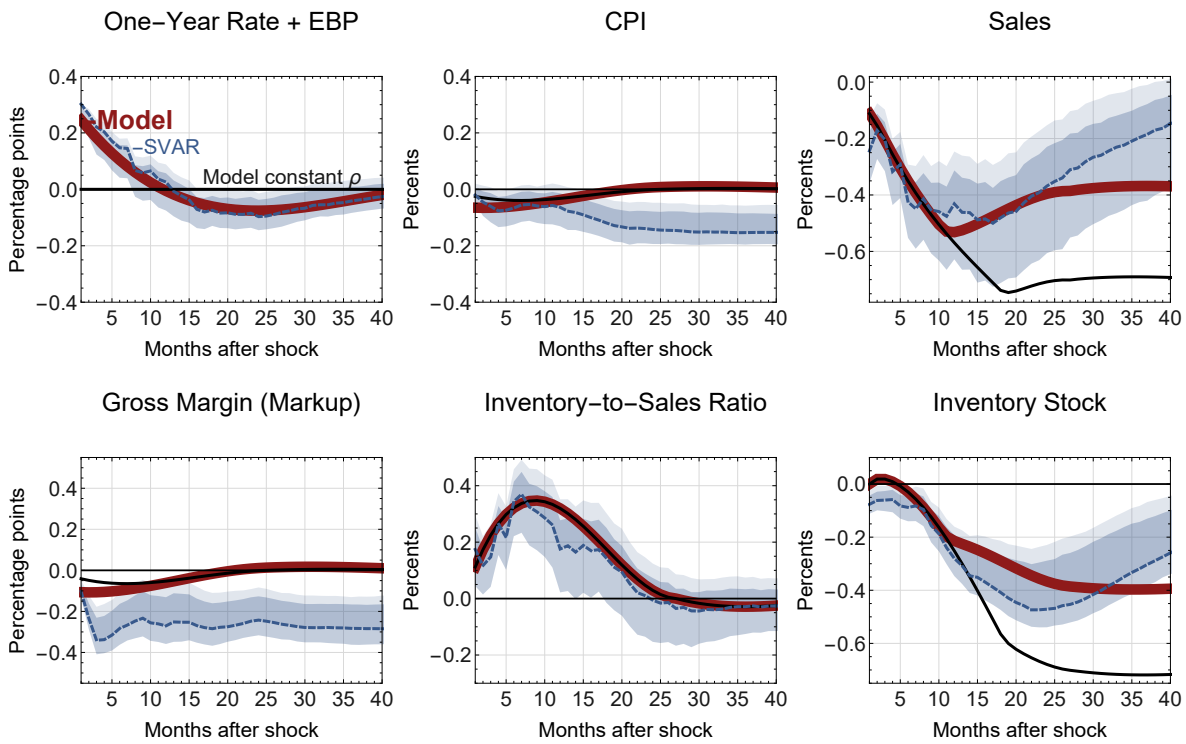


Figure 6: Comparison of model impulse responses to the SVAR from Section 2.

Notes: Notes to Figure 2 apply. The figure shows impulse responses from the calibrated model and compares them to those implied by the SVAR from Section 2.

this system, which we solve forward by assuming the steady state as its terminal value, the differential equations for M and N follow from (23), (21), and firm value maximization in (22). These we solve separately by starting from the steady state value as the initial condition. Since the entry intensity a_t and the recovery speed depend on the value of V_{0t} via the first-order condition $a_t = \delta + \frac{1}{2}\frac{(V_{0t}-\phi v)}{\kappa_a}$, the assumed value of κ_a can be identified.

4.2 Quantitative Results

Figure 6 reports the impulse responses generated by our calibrated model and compares them to the impulse responses generated by the SVAR from Section 2. As we can see, the model implies a 10-month lag in inventory adjustment relative to sales, which is roughly in line with the data. The inventory-to-sales ratio is matched exactly, which is what we targeted in the calibration of the demand shock. As in the data, markups fall and are thus negatively correlated with the inventory-to-sales ratio.

The responses of sales and inventories lie within the 90 percent confidence band of the SVAR. However, the model underpredicts the response of markups. While this underprediction helps match the relatively muted CPI response, the underlying reason is counterfactual and the model-implied CPI would fall twice as much as in the data had we matched the path for markups. Importantly, the variability of the markups in the data within this range would have little impact on the model-implied dynamics because it is fairly modest (the markup is 37 percent in the steady state).

The decomposition of the two channel through which monetary policy drives the depicted responses ('Model SVAR ρ' *versus* 'Model const. ρ') implies that interest rate dynamics notably shortens the duration of inaction (see also Figure 4, with action period being denoted by T). This occurs because higher anticipated path for interest rates significantly lowers the value of stock at the onset of the shock. This property implies that our model would predict a longer delay in response to demand contractions accompanied by falling interest rates (e.g., fiscal shocks).

Our model assumes permanently rigid wages, preventing Q and M from returning to their initial levels. As a result, the shock has a permanent effect and not all variable return to the steady state (Λ_t and ρ_t return to the steady state). The rebound in Q and M is driven by declining interest rates after approximately 20 months, but it is insufficient to restore their initial values. This implies that, according to the model, a downward adjustment in nominal wages is necessary to eventually restore full employment and our assumption of rigid wages prevents this from happening.

5 Robustness, Extensions and Limitations

This section discuss several extensions of our analysis.

Recessions versus Expansions.— We have focused on recessionary demand shocks. In the data, it is recessions that imply the largest spikes in both margins and the inventory-to-sales ratio. However, the cyclical properties of these variables can also be seen during expansions. One might conclude that since the inaction zone is located below the steady-state value for V_0, our model has qualitatively different predictions expansions. This is not the case.

To see this, note that, in response to transient demand expansions, a portion of the value of a new outpost is derived from the future when it is no longer needed. Even if entry is frictionless ($\kappa_a \to 0$), the adjustment through the entry margin is partial and hence similar dynamics arises qualitatively. In addition, in a more general model incorporating firm-level idiosyncratic shocks, some firms would always be in the inaction zone even during expansion due to idiosyncratic demand fluctuations. These firms would react roughly symmetrically because expansionary shocks shorten the duration of inaction period.

Cyclicality of Real Wages.— Our model implies an increase in the real wage because the nominal wage is fixed and the price level falls. This is a well known problem of sticky wage models. In the data real wages are acyclical or countercyclical, as shown by, for example, Christiano et al. (1997). We confirm this result in the Online Appendix using an extended version of our proxy SVAR.

However, as shown in the Appendix, the discrepancy is small quantitatively, and well within the confidence band for real wages given the change in the CPI. On the model side, we also do not take into account the effect of the endogenous labor productivity by relying on a linear production function. In particular, suppose that, instead, $v = wL^{-(1-\alpha)}$ and that w is rigid, as would be the case if a slow-moving physical capital was included in the model. Even if w falls after the shock, v could be approximately constant in that case, as we assumed, simply because labor's marginal product $L^{-(1-\alpha)}$ rises as labor input falls with demand. Measured labor productivity falls in our model due to declining utilization of outposts in production, and this dynamic does not necessarily implicate rising labor productivity in recessions. Furthermore, in the data there may be additional reasons why retail prices do not move as much, which our model over predicts.

Sticky Wages.— Kryvtsov and Midrigan (2012) show that inventory models with sticky wages are now promising because they produce counterfactual predictions for inventory dynamics. Specifically, in their framework, firms that adjust nominal wages during recessions because they anticipate lower future nominal wages and production costs. This prompts them to sell off inventory stocks, exacerbating the model failure. Our model shut down this channel by assuming rigid wages, which warrants a comment.

The reason why we do not consider this channel is because we find its merits debatable in light of the data rather than models. Sticky wages result in inventory reductions because nominal wages decline, but very few recessions lead to declines in nominal wages—which is what is needed here.[35]

Productivity Shocks.— To the extent that productivity shocks are transient, our mechanism remains applicable. As emphasized in the paper, the inaction region is robust to modifications because it is a corner solution. Rising interest rates generally extend the duration of inaction. Furthermore, for reasons related to the mechanism discussed by Bai et al. (2024), productivity moves endogenously in our model and slightly falls during demand recessions. This is driven by falling utilization rate of outposts. As a result, demand shocks in our model can account for a portion of productivity fluctuations in the data for reasons that are unrelated to technology. To the extent that productivity shocks are nearly permanent, our model's predictions are analogous to the frictionless model discussed in Section 3.2.

Measurement of Variable Costs.— Finally, our data analysis assumes that Cogs identifies the variable production costs of a firm. While Cogs primarily focuses on direct, production-related costs, such as materials and labor, these costs may still be influenced by adjustment frictions that bind over the business cycle. This potential mismeasurement could affect the results, but as we discuss next the validity of this argument largely falls under the discussion in Section 1 regarding the hypothesis that margins and markups might be negatively correlated due to a modified production structure.

To illustrate the effect of this kind of mismeasurement under the assumption we imposed, consider a three-factor Cobb-Douglas production function of the form $Y = V^{\alpha}F^{\beta}F_0^{\zeta_0}$, where α, β, ζ_0 are positive-valued fixed factor shares, V represents a variable factor over the business cycle, and F_0 and F are fixed factors over the business cycle frequency, but F is *erroneously* included in Cogs alongside V. For simplicity, suppose factor prices v, f, f_0 are taken as given by the firm, and markup variation arises solely from business cycle fluctuations in the output price p.

[35] For example, during the Great Recession, the Employment Cost Index for all private workers increased by about 3.5 percent in nominal terms. Over the same period, the PCE chain-weighted price index rose by 1 percent, resulting in a 2.5 percent increase in real wages. Unit labor costs, reflecting productivity-adjusted labor costs, declined only slightly by 1 percent, and so even small inventory holding costs would offset this minuscule benefit. Additionally, SVAR evidence indicates that real wages are acyclical following monetary policy shocks, despite slowing but still positive inflation during most recessions.

Log-linearizing the formula for the measured markup around the steady state, in which all factors can be freely adjusted, gives

$$\log\left(\frac{\text{Markup}}{\text{Markup}^{ss}}\right) = \log\left(\frac{\text{Margin}}{\text{Margin}^{ss}}\right) - \underbrace{\frac{\beta}{\alpha}\frac{1}{\alpha+\beta}\log\left(\frac{Y}{Y^{ss}}\right)}_{wedge}. \tag{38}$$

Accordingly, mismeasurement of this sort introduces a *wedge* between markups and margins and that this *wedge* comoves with real sales.

However, as we stressed in Section 1, the issue is that reversing the correlation between markups and margins calls for significant mismeasurement. For example, using a back-of-the-envelope calculation, flipping the correlation seen in Figure 1 (for the Great Recession), roughly requires $\beta > 0.3$, and this implies $\alpha < 0.4$ instead of $\alpha = 0.7$ in the absence of mismeasurement.[36] This is a problem because the variable marginal cost of production then becomes five times more volatile relative to output—since the standard deviation of the marginal cost to output scales with $(1-\alpha)/\alpha$. As we discussed in Section 1, a number of issues arise when the marginal cost schedules are that steep.

6 Conclusions

We developed a new search theory of inventories to explain the lagged response of inventories to recessionary demand shocks—an aspect of the data that standard inventory models fail to capture. We have shown that this can account for the countercyclical dynamics of the inventory-to-sales ratio under procyclical markups. While a global solution is necessary to incorporate our mechanism into existing business cycle frameworks, a reduced-form implementation of the core idea of our paper can be achieved by using convex adjustment costs instead of irreversible investment—for example, in the spirit of that in Drozd and Nosal (2012) or Gourio and Rudanko (2014).

[36] We consider the following back-of-the-envelope calculation. Assume the stated production function exhibits constant returns to scale ($\alpha+\beta+\zeta = 1$)—after accounting for all fixed factors—and that pure profits are approximately zero. Since (Sales-Cogs)/Cogs$\approx$ 0.7 before the Great Recession for the firms shown in Figure 1, for the stated production function this implies $\alpha = 0.7$ when $\beta = 0$. Given that the gross margin is nearly as volatile as (real) sales in Figure 1, flipping the positive correlation between margins and markups implied by this figure requires $\frac{\beta}{\alpha}\frac{1}{\alpha+\beta}$ is about 1, according to (38). Setting it equal to 1 gives: $\alpha \leq 0.41, \beta \geq 0.29$.

References

ABRAMOVITZ, M. (1950): "Size and Relative Importance of Manufacturers' Inventories," in *Inventories and Business Cycles, with Special Reference to Manufacturers' Inventories*, National Bureau of Economic Research, NBER Chapters, 35–39.

ALESSANDRIA, G., J. P. KABOSKI, AND V. MIDRIGAN (2010): "Inventories, Lumpy Trade, and Large Devaluations," *The American Economic Review*, 100, 2304–2339.

——— (2011): "US Trade and Inventory Dynamics," *American Economic Review*, 101, 303–07.

ALESSANDRIA, G. A., S. Y. KHAN, A. KHEDERLARIAN, C. B. MIX, AND K. J. RUHL (2023): "The Aggregate Effects of Global and Local Supply Chain Disruptions: 2020–2022," Working Paper 30849, National Bureau of Economic Research.

ANDERSON, S. P. AND R. RENAULT (1999): "Pricing, product diversity, and search costs: A Bertrand-Chamberlin-Diamond model," *RAND Journal of Economics*, 719–735.

ARGENTE, D., D. FITZGERALD, S. MOREIRA, AND A. PRIOLO (2024): "How Do Entrants Build Market Share? The Role of Demand Frictions," *American Economic Review: Insights (Forthcoming)*.

BAI, Y., J.-V. RÍOS-RULL, AND K. STORESLETTEN (2024): "Demand Shocks as Technology Shocks," Working Paper 32169, National Bureau of Economic Research.

BEIL, D. R. (2011): *Supplier Selection*, John Wiley & Sons, Ltd.

BILBIIE, F. AND D. KANZIG (2023): "Greed? Profits, Inflations, and Aggregate Demand," *CEPR Discussion Paper DP18385*.

BILS, M. (1987): "The Cyclical Behavior of Marginal Cost and Price," *American Economic Review*, 77, 838–855.

BILS, M. AND J. A. KAHN (2000): "What Inventory Behavior Tells Us about Business Cycles," *American Economic Review*, 90, 458–481.

BLINDER, A., E. CANETTI, D. LEBOW, AND J. RUDD (1998): *Asking About Prices: A New Approach to Understanding Price Stickiness*, New York: Russell Sage Found.

BROER, T., N.-J. HARBO HANSEN, P. KRUSELL, AND E. ÖBERG (2019): "The New Keynesian Transmission Mechanism: A Heterogeneous-Agent Perspective," *The Review of Economic Studies*, 87, 77–101.

BURSTEIN, A. T., J. C. NEVES, AND S. REBELO (2000): "Distribution Costs and Real Exchange Rate Dynamics During Exchange-Rate-Based-Stabilizations," Working Paper 7862, National Bureau of Economic Research.

CAVALLO, A. AND O. KRYVTSOV (2023): "What Can Stockouts Tell Us About Inflation? Evidence from Online Micro Data," *Journal of International Economics*, 146, 103769.

CHEREMUKHIN, A. AND P. RESTREPO-ECHAVARRIA (2020): "Wage Setting Under Targeted Search," *St. Louis Fed Working Paper 2020-041D*.

CHEREMUKHIN, A., P. RESTREPO-ECHAVARRIA, AND A. TUTINO (2020): "Targeted search in matching markets," *Journal of Economic Theory*, 185, 104956.

CHRISTIANO, L. J., M. EICHENBAUM, AND C. L. EVANS (1997): "Sticky Price and Limited Participation Models of Money: A Comparison," *European Economic Review*, 41.

CROUZET, N. AND J. EBERLY (2023): "Rents and Intangible Capital: A Q+ Framework," *The Journal of Finance*, 78, 1873–1916.

DE LOECKER, J., J. EECKHOUT, AND G. UNGER (2020): "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 135, 561–644.

DE LOECKER, J. AND F. WARZYNSKI (2012): "Markups and Firm-Level Export Status," *American Economic Review*, 102, 2437–71.

DELOITTE RESEARCH CENTER FOR ENERGY & INDUSTRIALS (2024): "2024 Manufacturing Industry Outlook," Industry report, Deloitte Research Center for Energy & Industrials.

DROZD, L. A. AND J. B. NOSAL (2012): "Understanding International Prices: Customers as Capital," *American Economic Review*, 102, 364–95.

FITZGERALD, T. (1997): "Inventories and the business cycle: an overview," *Economic Review*, 11–22.

GALÍ, J., M. GERTLER, AND J. D. LÓPEZ-SALIDO (2007): "Markups, Gaps, and the Welfare Costs of Business Fluctuations," *The Review of Economics and Statistics*, 89, 44–59.

GERTLER, M. AND P. KARADI (2015): "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics*, 7, 44–76.

GILCHRIST, S. AND E. ZAKRAJSEK (2012): "Credit Spreads and Business Cycle Fluctuations," *American Economic Review*, 102, 1692–1720.

GOURIO, F. AND L. RUDANKO (2014): "Customer Capital," *The Review of Economic Studies*, 81, 1102–1136.

HALL, R. E. (1988): "The relation between price and marginal cost in US industry," *Journal of Political Economy*, 96, 921–947.

HE, B., L. I. MOSTROM, AND A. SUFI (2024): "Investing in Customer Capital," Working Paper 33171, National Bureau of Economic Research.

JORGENSON, D. W., M. S. HO, AND K. J. STIROH (2004): "Will the U.S. Productivity Resurgence Continue?" *Current Issues In Economics and Finance-Federal Reserve Bank of New York*, 10.

KAHN, J. A. (1987): "Inventories and the Volatility of Production," *The American Economic Review*, 77, 667–679.

KAPLAN, G. AND G. MENZIO (2016): "Shopping Externalities and Self-Fulfilling Unemployment Fluctuations," *Journal of Political Economy*, 124, 771–825.

KEHOE, P. J., P. LOPEZ, V. MIDRIGAN, AND E. PASTORINO (2022): "Asset Prices and Unemployment Fluctuations: A Resolution of the Unemployment Volatility Puzzle*," *The Review of Economic Studies*, 90, 1304–1357.

KHAN, A. AND J. K. THOMAS (2007a): "Explaining Inventories: A Business Cycle Asessment of the Stockout Avoidance and (S,s) Motives," *Macroeconomic Dynamics*, 11, 638–664.

——— (2007b): "Inventories and the Business Cycle: An Equilibrium Analysis of (S, s) Policies," *American Economic Review*, 97, 1165–1188.

KRYVTSOV, O. AND V. MIDRIGAN (2012): "Inventories, Markups, and Real Rigidities in Menu Cost Models," *The Review of Economic Studies*, 80, 249–276.

LENTZ, R., J. MAIBOM, AND E. MOEN (2024): "Directedness in Search," *mimeo*.

LESTER, B. (2011): "Information and Prices with Capacity Constraints," *American Economic Review*, 101, 1591–1600.

MATĚJKA, F. AND A. MCKAY (2015): "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, 105, 272–98.

MENZIO, G. (2007): "A Theory of Partially Directed Search," *Journal of Political Economy*, 115, 748–769.

METZLER, L. A. (1941): "The Nature and Stability of Inventory Cycles," *The Review of Economics and Statistics*, 23, 113–129.

NEKARDA, C. J. AND V. A. RAMEY (2020): "The Cyclical Behavior of the Price-Cost Markup," *Journal of Money, Credit and Banking*, 52, 319–353.

ORTIZ, J. (2022): "Spread Too Thin: The Impact of Lean Inventories," .

QUILIS, E. M. (2018): "Temporal disaggregation of economic time series: The view from the trenches," *Statistica Neerlandica*, 72, 447–470.

RAMEY, V. AND K. WEST (1999): "Inventories," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and M. Woodford, Elsevier, vol. 1, Part B, chap. 13, 863–923, 1 ed.

ROTEMBERG, J. J. AND M. WOODFORD (1999): "Markups and the Business Cycle," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and M. Woodford, Amsterdam: Elsevier, vol. 1, 1051–1135.

WOLINSKY, A. (1986): "True Monopolistic Competition as a Result of Imperfect Information," *Quarterly Journal of Economics*, 101, 493–511.

Online Appendix (Not Intended for Publication)

Contents:

Section A provides omitted proofs from the paper.

Section B details the calibration of model parameters and the solution algorithm.

Sections C and D describe the replication package and list data sources.

(Supporting code and data files can be found in the online replication package. Instructions are provided at the end of this document.)

A Omitted Proofs

This section contains omitted proofs from text.

Proof of Lemma 1:

(It is instructive to read the section describing shoppers' search technology laid out after this lemma. We omit time subscripts to simplify notation.)

In general, the policy function of the distributor is a set-valued indicator function on the space of quoted prices and the taste shock, i.e., feasible tuples $(\tilde{p}, \eta)$. As assumed in the text, the distributor only cares about the implied surplus. Expression in (7) shows that, in that case, it is without loss to restrict attention to policy functions that are identical for any tuple $(\tilde{p}, \eta)$ that maps onto the same effective price p. (The definition of the effective price can be found underneath equation (7). The relevant distribution of effective prices is described by the probability measure Pr(.) generated by $\tilde{p}^*$ and the shock η. When we say probability measure (function) or write Pr(.), we mean this measure.)

Consider the optimal policy (a policy that maximizes distributor's profits), which by the above can be represented as measurable set $\mathcal{A} \subset \mathbb{R}_+$ of effective prices with measure $\pi = 1 - \Pr(\mathcal{A}) > 0$ that the shopper is allowed to accept. This policy requires that the contracted shopper delivers a good within an instance of time t of duration dt at an effective price in set $\mathcal{A}$. As long as there is a positive measure of prices in the market that are consistent with this policy—which is implied by the fact that the contract has been accepted by the shopper—the shopper can deliver a good because she can draw an infinite number of quotes in the limit as $dt \to 0$. Next, we show that the closure of the policy set (i.e., $cl\mathcal{A}$) can be restricted to $I_0(\bar{p}) := \{p : 0 \leq p \leq \bar{p}\}$, or else the implied total search costs can be strictly lowered, which would contradict optimality. We refer to such policies as *reservation effective price policies* (REPP).

By contradiction, suppose the above claim is false. If so, we can pick a reservation policy $\mathbf{I_0} := I_0(\bar{p}_0)$ such that $\pi = 1 - \Pr(\mathbf{I_0})$—which, note, is uniquely defined by $\pi > 0$ because density g has full support, and hence $\Pr[0, \bar{p}]$ is a continuous and strictly increasing function in $\bar{p}$ (intermediate value

theorem).[37] Since shoppers *randomly* pull prices from the population, as described below the statement of the lemma in text, the expected number of searches to draw a member from a set of the same measure is the same. Accordingly, the implied total search cost associated with the policy $\mathbf{I_0}$ is the same as under the original policy because the probability measure of $\mathbf{I_0}$ and $\mathcal{A}$ is the same.

We next show that the expected effective price under $\mathbf{I_0}$ is strictly lower than under the original policy $\mathcal{A}$, which give the contradiction.

Note that the set $E_0 = \mathcal{A}/\mathbf{I_0}$ (prices in $\mathcal{A}$ that exceed $\bar{p}_0$ and hence are not in $\mathbf{I_0}$) must be of a positive probability measure ($\Pr(E_0) > 0$). If this is not the case, $\mathcal{A}$ and $\mathbf{I_0}$ are identical up to a measure zero set, and so there is nothing to prove (which trivially contradicts the hypothesis and we are done). If this is not the case, define the sets $E_1 = \mathcal{A} \cap \mathbf{I_0}$ (prices in $\mathcal{A}$ that are in $\mathbf{I_0}$) and $E_2 = \mathbf{I_0}/E_1$ (prices in $\mathbf{I_0}$ that are not in $\mathcal{A}$), and note that: i) E_0 and E_1 form a partition of $\mathcal{A}$, and ii) E_1 and E_2 form a partition of $\mathbf{I_0}$. Since $\Pr(\mathcal{A}) = \Pr(\mathbf{I_0})$ and $\Pr(E_0 \cup E_1) = \Pr(E_1 \cup E_2)$, it thus must be that $0 < \Pr(E_0) = \Pr(E_2)$. This is a contradiction by the inequality below because, by construction, all prices in the positive measure set E_0 are above $\bar{p}_0$ and all prices in the equal measure set E_2 are below $\bar{p}_0$:

$$\int_{\mathcal{A}=E_0\cup E_1} p \Pr(dp) - \int_{\mathbf{I_0}=E_1\cup E_2} p \Pr(dp) = \int_{E_0} p \Pr(dp) - \int_{E_2} p \Pr(dp) > 0.$$

To see this, note that the left-hand side of this inequality is the difference in the mean price under policy $\mathcal{A}$ and under policy $\mathbf{I_0}$, multiplied by $\Pr(\mathcal{A}) = \Pr(\mathbf{I_0})$. Accordingly, this inequality shows that the expected price conditional on being in set $\mathbf{I_0}$ is *strictly* lower than the expected price conditional on being in set $\mathcal{A}$—a contradiction.

Finally, the fact that $\pi = \Pr(p \geq \bar{p})$ constitutes an equivalent representation of distributor's policy follows from the fact that $\Pr(p \geq \bar{p})$ is monotone in $\bar{p}$ (not necessarily strictly monotone, as is the case in our model). Define the reservation price as the lowest price on the "flat portions" of this function. This is without loss because distributor's surplus remains unchanged. This ensures a bijective mapping. Q.E.D.

Proof of Lemma 2:

The proof considers a discretization of continuous time setup by assuming a fixed period length $dt > 0$ and taking the limit $dt \to 0^+$. We omit time subscripts to simplify notation.

Part 1. By Lemma 1, the relation $1 - \pi = \Pr\left(p \leq \bar{p}\left(\pi\right)\right)$ is well defined and it implies that the event $p \leq \bar{p}\left(\pi\right)$ occurs with probability $1-\pi$ after a single random draw of an effective price by the shopper, after two draws with probability $\left(1-\pi\right)\pi$, and so on and so forth. This defines a geometric process with a fixed success (stopping) probability $\left(1-\pi\right)$. Accordingly, the number of searches until a match is formed corresponds to the mean of the geometric distribution with parameter $\left(1-\pi\right)$.

To calculate the implied search cost, we must take into account discounting with the instance of time of length dt—a tedious technical condition. Since search costs are borne at different points in time on

[37] This result trivially generalizes to distributions that are weakly increasing or do not have a full support.

that interval, the discounted value may be different. Define the highest discount factor on the interval $[t, t+dt)$ as $\gamma_{\text{sup}} \equiv e^{-\left(\sup_{l\in[t,t+dt)}\rho_l\right)d\tau}$, and the lowest discount factor as $\gamma_{\text{inf}} \equiv e^{-\left(\inf_{l\in[t,t+dt)}\rho_l\right)d\tau}$. Note that $\gamma_{\text{sup}}, \gamma_{\text{inf}} \rightarrow_{dt\downarrow 0} 1$. (We assume an MIT shock and we assume that the path for ρ_t is left continuous. Therefore, in a discretization of continuous time considered here there are no "jump" discontinuities on the time interval considered). Note that the mean discounted cost with discount $\gamma = \gamma_{\text{sup}}$ or $\gamma = \gamma_{\text{inf}}$ corresponds to the infinite summation defined by the recursion:

$$Sum_\gamma := c_0 (1-\pi)\gamma + \pi c_0 \gamma + \pi\gamma \left(\underbrace{c_0 (1-\pi)\gamma + \pi c_0 \gamma + \pi\gamma (c_0 (1-\pi)\gamma + ...)}_{=Sum_\gamma} \right). \tag{39}$$

This is similar to the standard proof of the mean value of a geometric distribution but it additionally takes discounting into account. As noted under the expression, the recursion boils down to solving $c_0 (1-\pi)\gamma + \pi\gamma c_0 + \pi\gamma Sum_\gamma = Sum_\gamma$, and hence $Sum_\gamma = c_0\gamma (1-\pi\gamma)^{-1} \rightarrow_{\gamma\uparrow 1} c_0 (1-\pi)^{-1}$. Since the actual search cost is bounded by $Sum_{\gamma_{\text{sup}}} \leq c(\pi) \leq Sum_{\gamma_{\text{inf}}}$, the result follows with the discount factor.

Part 2. To calculate the integral in (10), we change the variable of integration from p to η, as implied by the definition of the effective price $p := \tilde{p}^* - \eta P$. Let $\bar{p}(\pi)$ be the one-to-one correspondence implied by (8) and proven in Lemma 1. Before we take the integral, we note the following properties: i) $dp = -Pd\eta$ by $p := \tilde{p}^* - \eta P$, ii) the implied bounds of integration for η by the integral for $s(\pi)$ are $p := \tilde{p}^* - \eta P$ are

$$\bar{p}(\pi) = \tilde{p}^* - \bar{\eta}(\pi) P \Rightarrow \bar{\eta}(\pi) := \frac{\tilde{p}^* - \bar{p}(\pi)}{P} \tag{40}$$

$$-\infty = \tilde{p}^* - \eta P \Rightarrow \eta = +\infty, \tag{41}$$

iii) the memoryless property of the exponential distribution implies

$$\mathbb{E}[\eta | \eta > \bar{\eta}(\pi)] = \max\{\bar{\eta}(\pi) + \eta_0, 0\}, \tag{42}$$

iv) equation (8) gives $\bar{p}(\pi) = \tilde{p}^* - PG^{-1}(\pi)$ and hence $\bar{p}(\pi) = \tilde{p}^* + \eta_0 P \log(1-\pi)$, and v) the shopper accepts the equilibrium price $\tilde{p}^*$ whenever

$$\eta \geq \bar{\eta}(\pi) := \frac{\tilde{p}^* - \bar{p}(\pi)}{P}. \tag{43}$$

Using the change of variables and the above properties, we now calculate that

$$\begin{aligned} s(\pi) &= \frac{P}{1-\pi}\int_{(-\infty,\bar{p}(\pi)]}\left(\frac{P-p}{P}\right)g\left(\frac{\tilde{p}^*-p}{P}\right)dp \\ &=_{[p=\tilde{p}^*-\eta P,\, dp=-Pd\eta]} \frac{P}{1-\pi}\int_{[\bar{\eta}(\pi),\infty)}\left(\frac{P-\tilde{p}^*}{P}+\eta\right)g(\eta)\,d\eta \\ &= P-\tilde{p}^*+P\mathbb{E}\left[\eta|\eta>\bar{\eta}(\pi)\right] \\ &= P-\tilde{p}^*+PG^{-1}(\pi)+\eta_0 P \\ &= P(1+\eta_0)-\tilde{p}^*-\eta_0 P\log(1-\pi). \end{aligned} \tag{44}$$

The calculation assumes $\bar{\eta}(\pi)+\eta_0>0$, which is equivalent to imposing $\pi\geq 0$ by the above expression. This constraint is imposed in text as a feasibility restriction in the distributor problem.

Remark 1. *The expected preference shock follows from the above result and it is* $\mathbb{E}_\pi\left[\eta|\eta>\bar{\eta}(\pi)\right]=PG^{-1}(\pi)+\eta_0 P=\eta_0\left(P+\log(1-\pi)^{-1}\right)$.

Q.E.D.

Proof of Lemma 3:

We suppress the 'ss' notation. Any variable without an explicit time subscript refers to the steady-state value of that variable. For example, when we write X instead of X_t, we mean X^{ss}, not X_t as used in the text.

(*) In equilibrium, $\lambda(\tilde{p}^*,\tilde{p}^*)=\Lambda$. Accordingly, (23) in steady state requires that the steady-state values of M and N satisfy:

$$0=\tau N-\Lambda M-\delta M, \tag{45}$$
$$0=\Lambda M-\tau N-\delta N+a(M+N), \tag{46}$$

which necessitates $a=\delta$. From the producer problem in (22), we infer $V_0=\phi v$ and hence $d=0$. This follows from the fact that $a\geq\delta$ and $d>0$ cannot be both true under the firm value maximization in (22).

This system establishes a unique steady state ratio M/N, which can be satisfied for any value $M>0$. Since $\Lambda=Q/M$ and $0<D(P)=Q\left(1+\mathbb{E}_\pi\{\eta\}\right)$, where D is exogenous and time-invariant demand in the steady state. A unique value of $\Lambda>0, P>0, \pi$ determines a unique value of $M>0$, in which case N follows. We need $M>0$ or else retail market cannot clear for $D(P)>0$ because there is no production in the steady state.

The values $V_0=\phi v$ and Λ, P, π are the links between the remaining equilibrium conditions and the conditions (variables) above. Thus, from now on, we focus on the remaining equilibrium conditions, while assuming $V_0=\phi v$ and seeking $\Lambda>0$.

We proceed in two steps. First, we show that if the steady state exists, it is unique (part 1). Then, we demonstrate that the steady state exists under the conditions stated in the lemma (part 2).

Part 1 (uniqueness). Consider the HJB equation for V_0 in (20). In the steady state, production must be positive to meet positive retail demand. Therefore, we know $\hat{\tau} = \tau$ and hence $X = V_1 - V_0 \geq v$ must be true. This follows from steady state existence, which here we assume.

The entry condition $V_0 = \phi v$ and (20) evaluated in the steady state implies

$$\rho\phi v = -\zeta_0 v + \tau\left(-v + X\right) - \delta\phi v, \tag{47}$$

and hence

$$X = \tau^{-1}\left(\left(\rho + \delta\right)\phi + \zeta_0 + \tau\right)v. \tag{48}$$

Remark 2. *(For later use) Note that the condition* $X = \tau^{-1}\left(\left(\rho + \delta\right)\phi + \zeta_0 + \tau\right)v \geq v$ *is satisfied for the default ranges of model parameters.*

Next, consider the distributor's zero profit condition in (12) rewritten as follows:

$$P = \tilde{p}^* + \chi v + \eta_0 P \log\left(\frac{c_0 v}{\eta_0 P}\right). \tag{49}$$

Given the first order condition for the wholesale price in (26), which yields

$$\tilde{p}^* = X + \eta_0 P, \tag{50}$$

given the above formula for X, we obtain the following fixed point for the retail price P:

$$(1-\eta_0)P + \eta_0 \log\left(\frac{\eta_0}{c_0 v}P\right)^P = X + \chi v = \tau^{-1}\left(\left(\rho + \delta\right)\phi + \zeta_0 + \tau\right)v + \chi v. \tag{51}$$

(Note: In stating this condition, we used the fact that $P\log\left(\frac{c_0 v}{\eta_0 P}\right) = -\log\left(\frac{\eta_0}{c_0 v}P\right)^P$.)

If the steady state exists, as assumed, $\frac{\eta_0}{c_0 v}P = (1-\pi)^{-1} > 1$, since this term comes from the distributor's profit maximization in (12) and feasibility requires $0 \leq \pi < 1$. Accordingly, for the range of values that P may take in the steady state, we know that the left-hand side of (51) is strictly increasing in P. Thus, if a solution exists, it must be unique because the right-hand side is invariant with respect to P. The steady-state value of Λ can be derived from (25) by substituting the steady-state value of X given by (48), setting $\dot{X} = 0$, and substituting $\tilde{p}^*$ using (50). After some manipulations, we derive:

$$\Lambda = \frac{v}{\eta_0 P}\left(\tau^{-1}\left(\delta + \rho + \tau\right)\left(\phi\left(\delta + \rho\right) + \zeta_0 + \tau\right) + \zeta - \tau\right). \tag{52}$$

We have now shown that there is at most a single set of equilibrium values that satisfy the steady state equilibrium conditions implied by Definition 1: $a, d, M, N, Q, \Lambda, X, V_0, V_1, P, \tilde{p}^*, \pi, \bar{p}$. (The unique value for $\bar{p}$ is stated in Lemma 2 and recall that, by definition, $V_1 \equiv V_0 + X$.)

Part 2 (existence). As discussed in the beginning (see *), the values of Q, M, N, a, d are uniquely pined down by $\Lambda > 0, P > 0, 0 \leq \pi < 1$, which requires $V_0 = \phi v$ as an equilibrium condition. The following equilibrium conditions are then satisfied: the law of motion (23), producer value maximization in (22), and retail market clearing.

To obtain the remaining variables and ensure the remaining equilibrium conditions are satisfied, we start from the requirement that $X \geq v$ for production to be positive—and hence for the retail market to clear for $D > 0$. The steady state value of X is given by (48) (see part 1)—which, recall, embeds $V_0 = \phi v$ and the HJB equation for V_0 in (20). As noted in Remark 2 (see part 1), this value satisfies $X > v$ for the default ranges of model parameters. What needs to be shown is that there exists $\Lambda > 0$ such that, for $\dot{X} = 0$ and X given by (48), the HJB equation in (25) is satisfied (note that this automatically ensures HJB equation for V_1 in (18) is also satisfied). We have derived the implied by that steady value relation for Λ in equation (52) (see part 1). By that equation, if $P > 0$, for all positive parameter values $\Lambda > 0$ because the expression numerator, $\tau^{-1}(\delta + \rho + \tau)(\phi(\delta + \rho) + \zeta_0 + \tau) + \zeta - \tau$, is positive for the default ranges of model parameters. (Note that τ^2 cancels out after bringing this expression to a common denominator τ. Since we are only left with a summation of positive parameters, the expression is positive.)

Next, we show that we can can find $P > 0$ that solves (51) (see part 1) and satisfies the feasibility restriction implied by (13): $0 \leq \pi = 1 - \frac{c_0}{\eta_0}\frac{v}{P} < 1$. Given the steady state value of X in (48), $\tilde{p}^*$ from (50), and assuming we solve for $P > 0$, we have satisfied all remaining equilibrium conditions (remaining relative to *): the entry condition $V_0 = \phi v$, steady state relations implied by the HJB equations in (20) and (18), the first order condition for $\tilde{p}$, positive production requirement $\hat{\tau} = \tau$, we are sure that $\Lambda > 0$, and distributor's zero profit condition is ensured by (51) (which we need to show has a solution).

To show that we can solve for the retail P, note that we can rewrite the above requirement $\frac{\eta_0}{c_0 v}P = (1-\pi)^{-1} > 1$ for P that solves (51), and hence we need to show $P > \frac{c_0 v}{\eta_0}$.

Define $\underline{P} = \frac{c_0 v}{\eta_0}$ so that $\frac{\eta_0}{c_0 v}\underline{P} = 1$, and note that the following properties apply to the second term on the left-hand side of (51): 1) $\log\left(\frac{\eta_0}{c_0 v}\underline{P}\right)^{\underline{P}} = 0$, and 2) $\log\left(\frac{\eta_0}{c_0 v}\underline{P}\right)^{\underline{P}}$ is strictly increasing in P on the restricted domain $[\underline{P}, \infty)$. Accordingly, if at $P = \underline{P}$ we can be sure that the left-hand side of (51) is strictly lower than the right-hand side, the result follows: a unique solution of (51) exists because the left-hand side of (51) is strictly increasing and unbounded in P, lower than the right hand side at $\underline{P}$, and the right-hand side is independent of P. Conversely, if this is not the case, there is no solution, and so this is an 'if and only' relation (recall here that $\underline{P}$ is a tight lowest value, or else $\pi < 0$). Plugging in $P = \underline{P}$ to (51), we obtain the inequality in Assumption (2):

$$(1-\eta_0)\underbrace{(c_0/\eta_0)}_{\underline{P}/v} + \eta_0 0 < \underbrace{\tau^{-1}\left((\delta+\rho)\phi + \zeta_0 + \tau\right)}_{X/v} + \chi. \tag{53}$$

Remark 3. *To gain some intuition, rewrite the above condition as*

$$\underline{P} = \frac{c_0 v}{\eta_0} < v + \tau^{-1}\left((\delta+\rho)\phi v + \zeta_0 v\right) + \chi v + c_0 v.$$

The terms on the right-hand side represent all the costs associated with producing and selling a good when the search precision π is zero. These include the production cost, operational costs, and sunk costs (recall that producing a good takes τ^{-1} units of time). This scenario represents the worst case for the distributor's surplus, since any equilibrium retail price that higher yields a greater surplus. Consequently, the distributor's zero-profit condition cannot be satisfied if this inequality is violated (for reasons discussed in Step 1).

We have now shown that (51) has a solution on the admissible domain $P \in (\underline{P}, \infty)$, and as we noted above, all equilibrium conditions and domain restriction are satisfied when X is given by (48), $\tilde{p}^*$ is given by (50) (given X), Λ is given by (52), $V_0 = \phi v$, $V_1 \equiv \phi v + X$, π is given by (13), and a, d, M, N, Q are pinned down as discussed in the beginning ($\bar{p}$ can be found using Lemma 2). We have also shown that the restriction of the lemma (A2) is a necessary condition, or else the candidate steady state equilibrium implies $\pi < 0$. Q.E.D.

Proof of Corollary 1:

We use the formula for P^{ss} (steady state value) given by the linearized equation in (31). We substitute the steady state value X^{ss} from (48) (proof of Lemma 3) and plug the obtained expression into the steady state formula for Λ^{ss} in (52) (proof of Lemma 3). This gives:

$$\Lambda^{ss} = \frac{\tau\left(\tau^{-1}(\delta+\rho+\tau)\left(\phi(\delta+\rho)+\zeta_0+\tau\right)+\zeta-\tau\right)}{\Theta\left(\Gamma\tau+\phi(\delta+\rho)+\zeta_0+\tau\right)}. \tag{54}$$

Substituting the above steady state value of Λ into the expressions for a_1 and a_0 in the statement of the corollary, after basic algebraic manipulations to put all terms under common denominator, we obtain:

$$a_0 = \frac{\left(\Gamma(\delta+\rho+\tau)-\zeta+\tau\right)\left(\phi(\delta+\rho)+\zeta_0+\tau\right)}{\Gamma\tau+\phi(\delta+\rho)+\zeta_0+\tau} > 0, \tag{55}$$

$$a_1 = \frac{\tau(\Gamma(\delta+\rho+\tau)-\zeta+\tau)}{\Gamma\tau+\phi(\delta+\rho)+\zeta_0+\tau} > 0. \tag{56}$$

(Note that Θ cancels out.) The inequalities above follow from the fact that $\Gamma = \eta_0 P^{ss}/v + \chi$ in (31), and we know from the proof of Lemma 3 (step 2) that in the steady state $P^{ss} > (c_0 v)/(\eta_0)$; hence, $\Gamma > c_0 + \chi$, and therefore $\tau - \zeta + (c_0+\chi)(\delta+\rho+\tau) > 0$ guarantees that all terms in the formulas for a_1 and a_0 are strictly positive. Q.E.D.

Proof of Lemma 4:

The expressions stated in the lemma follow by substituting the linearized term in text and solving for prices using equations (50) and (49) from the proof of Lemma 3. The fact that $\Gamma > 0$ is trivial and the fact $\Theta > 0$ follows from the proof of Lemma 3. The lemma derives the lower bound on the steady

state retail price P^{ss} (step 2). This bound ensures the logarithmic term in the expression for $\Theta > 0$ is positive. (Online Appendix in Section B.1 provides an alternative derivation of these expressions.)

B Model Numerical Solution and Calibration

This appendix discusses the numerical solution of the model and the calibration of its parameters/shocks. Supporting *Mathematica* notebook can be found in the replication package (solve_Model_vX.nb.)

B.1 Equilibrium System and Steady State

Consider zero profit condition of the distributor in (12); that is

$$P - \tilde{p}^* - \eta_0 P \log\left(\frac{c_0}{\eta_0}\frac{v}{P}\right) = \chi v, \tag{57}$$

where we drop time subscripts since we focus here on calculating the steady state.

The above condition cannot be solved in closed form because of the last term. To solve the model numerically, we resort a linear approximation of the last term:

$$\log\left(\frac{c_0}{\eta_0}\frac{v}{P_t}\right) \approx \frac{P^{ss}}{P} - 1 + \log\left(\frac{c_0 v}{\eta_0 P^{ss}}\right), \tag{58}$$

implying

$$\chi v = P(1-\eta_0) - \tilde{p}^* - \eta_0 P^{ss} - \eta_0 P \log\left(\frac{c_0}{\eta_0}\frac{v}{P^{ss}}\right). \tag{59}$$

We introduce two composite parameters to soak up the unwanted terms (P^{ss} and $\log\left(\frac{c_0 v}{\eta_0 P^{ss}}\right)$):

$$\Gamma = \frac{P^{ss}\eta_0}{v} + \chi, \; \Theta = \frac{\eta_0}{1 - \eta_0 \log\left(\frac{c_0 v}{\eta_0 P^{ss}}\right)}. \tag{60}$$

Substituting into the above expression, we obtain

$$P = \frac{\Theta}{\eta_0}\frac{\tilde{p}^* + \Gamma v}{1+\Theta}. \tag{61}$$

The wholesale quoted price is given by (26), and hence

$$\tilde{p}^* = V_1 - V_0 + \eta_0 P. \tag{62}$$

Accordingly,

$$\tilde{p}^* = (V_1 - V_0)(1 + \Theta) + v\Gamma\Theta,\ P = \frac{\Theta}{\eta_0}(V_1 - V_0 + v\Gamma). \tag{63}$$

Adding the entry condition $V_0 = \phi v$ and the Bellman equations (18) and (20) evaluated in the steady state ($\dot{V}_0 = 0, \dot{V}_1 = 0$), we obtain the following steady state system augmented by the auxiliary parameters Γ, Θ:

$$V_0 = \phi v, \tag{64}$$
$$(\rho + \delta)V_1 = -(\zeta + \zeta_0)v + \Lambda(\tilde{p}^* + V_0 - V_1), \tag{65}$$
$$(\rho + \delta)V_0 = -\zeta_0 v + \tau(-v + V_1 - V_0), \tag{66}$$
$$X = V_1 - V_0, \tag{67}$$
$$\tilde{p}^* = (V_1 - V_0)(1 + \Theta) + v\Gamma\Theta, \tag{68}$$
$$P = \frac{\Theta}{\eta_0}(V_1 - V_0 + v\Gamma). \tag{69}$$

The analytic solution gives

$$V_0^{ss} = v\phi, \tag{70}$$
$$V_1^{ss} = \frac{v(\phi(\delta + \rho + \tau) + \zeta_0 + \tau)}{\tau}, \tag{71}$$
$$X^{ss} = \frac{v(\phi(\delta + \rho) + \zeta_0 + \tau)}{\tau}, \tag{72}$$
$$\tilde{p}^{*ss} = \frac{v(\Gamma\Theta\tau + (\Theta + 1)\phi(\delta + \rho) + \zeta_0\Theta + \zeta_0 + \Theta\tau + \tau)}{\tau}, \tag{73}$$
$$P^{ss} = \frac{\Theta v(\Gamma\tau + \phi(\delta + \rho) + \zeta_0 + \tau)}{\eta_0\tau}, \tag{74}$$
$$\Lambda^{ss} = \frac{\delta\zeta_0 + \phi(\delta + \rho)(\delta + \rho + \tau) + \delta\tau + \zeta_0\rho + \zeta\tau + \zeta_0\tau + \rho\tau}{\Theta(\Gamma\tau + \phi(\delta + \rho) + \zeta_0 + \tau)}, \tag{75}$$
$$\pi^{ss} = 1 - \frac{c_0\tau}{\Theta(\Gamma\tau + \phi(\delta + \rho) + \zeta_0 + \tau)} \tag{76}$$

(These formulas a were automatically generated by Mathematica.)

Lemma 3 shows that the steady exists and is unique.

B.2 Calibration (Extended Version)

This section describes the calibration of the model. The baseline period of the model ($t = 1, 2, 3$) corresponds to one month. To map model onto the data, we assume (real) sales is Q, real inventory stock is M, the CPI is P, and nominal value added is $P_t Q_t(1 + \mathbb{E}\{\eta\})$. Unlike in the analytic section, gross margin/markup is measured in logs as $\log(p/v)$. Parameter values are listed in the table below.

B.2.1 Static Steady State-Based Targets

Consider first the targets that map directly onto model parameters. These include the weighted-average cost of capital (WACC) of 10 percent (annual rate) and the delivery delay of 60 days in U.S. manufacturing.[38] WACC pins down the steady state value of ρ according to the formula that first converts it to a monthly rate and then to a continuously compounded rate:

$$\rho^{ss} = \ln(1 + \text{WACC}/100)^{12}. \tag{77}$$

The average delivery time in the model is τ^{-1}, which for the given target yields $\tau = 0.5$. In the baseline calibration we set inventory holding cost $\zeta = 0$ and assume it is σ that prevents firms from holding more than one unit—so as to satisfy Assumption 1.

We next discuss how we calibrate the values of $\chi, \eta_0, c_0, \phi, \zeta_0$, as well as the auxiliary parameters Θ and Γ. We first describe data targets for these parameters.

Since manufacturing sector went through a major transformation in 2000s and 2010s, and much of this period lies outside of our sample period, we use the 1997 input-output (IO) tables published by Bureau of Economic Analysis (BEA) to set targets for margins and distribution costs. Using later tables would increase the targeted moment for by about 10-20 percent depending on the exact date (before of after GFC). (The raw tables can be found in the replication package, folder IO-Calibration.)

Our first data target is the distribution margin,

$$\mathscr{M}_1 = \frac{P - \tilde{p}^*}{\tilde{p}^*} = .38, \tag{78}$$

which we associate with the total trade and transportation margins borne on a unit of manufactured good and indirect (net) taxes borne by the final users. To obtain this number, we proceed as follows. We use the 1997 domestic-*supply*-of commodities IO table (for 15 industries) and focus on manufacturing commodities (in row of the IO table). We calculate the following ratios using this table (as shown in Table 2 below):

$$\tilde{\mathscr{M}}_1^I = \frac{\text{Total trade and transport margin in manufacturing}}{\text{Total supply of manufacturing commodities in basic prices}}, \tag{79}$$

$$\tilde{\mathscr{M}}_1^F = \frac{\text{Total supply of manufacturing commodities in purchaser prices}}{\text{Total supply of manufacturing commodities in basic prices}}. \tag{80}$$

The first ratio is the share of trade and transportation margins relative to the basic price of manufacturing commodities that are sold as either intermediate good or the final goods. The second ratio includes

[38] For an overview of cost of capital estimates for various industries, see the listing compiled by Aswath Adamodar at https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/wacc.html. Delivery delays in manufacturing come from Deloitte Research Center for Energy & Industrials (2024)—based on the source data from Institute for Supply Management (ISM)—and the report can be found at https://www2.deloitte.com/us/en/insights/industry/manufacturing/manufacturing-industry-outlook.html (see Figure 4). This number has remained steady between 2015 and 2019 according to the earlier reports.

(net) taxes and tariffs, where we associate this particular margin with purchases by the final users, as we shall see. In the second step—shown in Table 3—we calculate the following two ratios: i) the fraction of manufacturing output sold to final users

$$f = \frac{\text{Final use of manufacturing commodities}}{\text{Total use of manufacturing commodities}} \tag{81}$$

and ii) the share of manufacturing commodities sold to manufacturing sector as an intermediate input used by that sector:

$$x = \frac{\text{Use of manufacturing commodities as intermediate inputs by manufacturing sector}}{\text{Total use of intermediate inputs by manufacturing sector}}. \tag{82}$$

Table 2: Trade and transportation margins on manufacturing goods, US 1997 (millions of 1997 $).

Commodity view	C1	C2	C3	$\tilde{\mathscr{M}}_1^I$	$\tilde{\mathscr{M}}_1^F$
(row in supply IO table)	Product supply (basic prices)	Product supply (purchaser prices)	Total trade & transportation margins	$\frac{C3}{C1}$	$\frac{C2-C1}{C1}$
Manufacturing	$4,507,147	$5,914,729	$1,074,854	**24%**	**31%**

Notes: This extract comes from BEA's 1997 domestic-supply-of-commodities IO table.

Table 3: Share of manufacturing goods in production and final demand (millions of 1997 $).

Commodity view	C1	C2	C3	**f**	**x**
(row in supply IO table)	Total use of products	Total use as intermediate inputs	Use as intermediates in manufacturing sector	$\frac{C1-C2}{C1}$	$\frac{C3}{C2}$
Manufacturing	$5,914,730	$2,794,901	$1,656,638	**53%**	**59%**

Notes: This extract comes from BEA's 1997 use-of-commodities IO table.

Using these ratios, we calculate the total distribution margin by cumulating the cycles a manufacturing commodity is used as an intermediate input to produce manufacturing goods, where the outflows from the cycle are purchases of manufacturing commodities by final users (final users or other sectors) that accrue trade and transportation margins. Specifically, the outflows accrues trade and transportation margin $\tilde{\mathscr{M}}_1^F$ in proportion to purchases of manufacturing commodities as intermediate goods and $\tilde{\mathscr{M}}_1^I$ in proportion to purchases as final goods. This procedure gives the following geometric series:

$$\mathscr{M}_1 = \tilde{\mathscr{M}}_1^F f + \tilde{\mathscr{M}}_1^I (1-f) + x(1-f)(\tilde{\mathscr{M}}_1^F f + \tilde{\mathscr{M}}_1^I (1-f) + x(1-f)(\tilde{\mathscr{M}}_1^F f + \tilde{\mathscr{M}}_1^I (1-f) + ...). \tag{83}$$

Plugging in the numbers from Tables 3 and 2, we obtain

$$\mathscr{M}_1 = \frac{\tilde{\mathscr{M}}_1^F f + \tilde{\mathscr{M}}_1^I (1-f)}{1 - x(1-f)} = .38. \tag{84}$$

The key assumption is that similar margins apply along the production chain. If the bulk of trade and transportation margins accrue closer to the end of the supply chain, or the beginning, our procedure may bias the results upwards or downwards, respectively. The so-called PCE Bridge published by BEA sheds light on this issue by reporting separate transportation and distribution margins on final consumption that go into PCE.[39] The PCE bridge data suggests that there is no significant discrepancy of this sort in the data. Given the distribution margins used in the related literature, our number is on the conservative side (Burstein et al., 2000).

Our procedure intentionally ignores trade margins on intermediate goods that are implicitly embedded in the value of intermediate commodities purchased by the manufacturing sector from other sectors, since the goal of this calculation is to get to the margins paid on a unit of manufacturing output.

Our second target is the gross profit margin in manufacturing and trade industries:

$$\mathscr{M}_2 = \frac{\tilde{p}^* - v}{v} = .37. \tag{85}$$

To set this target, we use use-of-commodities IO table and take the ratio of gross surplus to the value added in producer prices for the three sectors (manufacturing/retail/wholesale). The extract of relevant data is in Table 4.

Table 4: Gross surplus in manufacturing and trade industries, US 1997 (millions of $).

		Sector view (column in use IO table)			
		Manufacturing	Wholesale	Retail	**All three sectors**
R1	Total intermediate inputs	$2,514,885	$219,597	$277,373	$3,011,855
R2	Compensation of employees	$788,978	$266,144	$338,975	$1,394,097
R3	Net taxes on production	$26,559	$6,940	$2,994	$36,493
R4	Gross operating surplus	$552,767	$138,413	$134,696	$825,876
R5	**Value added (basic prices) R2+R3+R4**	$1,368,304	$411,497	$476,665	$2,256,466
	Ratio: Gross margin R4/R5	40%	34%	28%	**37%**

Notes: This extract comes from BEA's 1997 use-of-commodities IO table.

Our third target is the inventory-to-sales ratio during our sample period of:

$$\mathscr{M}_4 = \frac{M^{ss}}{Q^{ss}} = \frac{M^{ss}}{\tau N^{ss}} = 1.5. \tag{86}$$

Our fourth target is the share of expenses on sales infrastructure borne by producers, which we set equal to 25 percent:

$$\mathscr{M}_5 = \frac{\phi\delta + \zeta_0(M^{ss} + N^{ss})}{\tau N^{ss}} = 0.3. \tag{87}$$

[39] Available at https://www.bea.gov/industry/industry-underlying-estimates.

To set this target, we use the expenses on SG&A relative to sales in the Compustat data for the manufacturing sector, which are about 17 percent. However, these costs are paid multiple times as manufactured goods are processed within the manufacturing sector, our target is 25 percent to reflect the same input-output multiplier of $(1 - x(1 - f))^{-1}$ =1.38 that we used above to cumulate trade and transportation margins in the calculation of distribution margin above (first target).[40]

Our last, fifth target from this group, pertains to search process and assumes that, on average, shoppers seek three quotes from suppliers before making an offer—which is the standard used by purchasing departments:

$$\mathscr{M}_3 = (1 - \pi)^{-1} = 3. \tag{88}$$

Finally, we set $\zeta = 0$ and relate our model to Assumption 1 by considering a hypothetical firm that faces equilibrium conditions but can accumulate inventory in the marketing state. To prevent that firm from doing so, we find $\sigma = 0.64$. Generally, if we set a different value for σ, we would need to impose $\zeta = 0.0638324 - 0.112234\sigma$ on that firm, which quantifies the friction that is needed to support our model's structure. Recall that $\sigma = 0.57$ operates on the cost of maintaining outposts, which is set at $\zeta_0 v = 0.094v$ in the calibration. Accordingly, 2.25 percent of value added in manufacturing and trade industries would be required to double inventory holding capacity. After accounting for the cost of all structures, land, and other resources that contribute to inventory holding costs, it is a large number.

The mapping between parameters and targets is analytic and invertible. The complete formulas are

[40]For an overview of this ratio across industries and countries, see the listing compiled by Aswath Adamodar at https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html. These numbers (relative to sales) range between 15 and 25 percent across sectors. The estimated share of value added in gross output comes from KLEMS for manufacturing and trade industries (value added weighted) can be found at https://www.bea.gov/data/special-topics/integrated-industry-level-production-account-klems (production account tables).

too long to state and we state them here for $\zeta = 0$, which is what we assume in calibration:

$$
\begin{aligned}
c_0 &= \frac{\mathscr{M}_3((\mathscr{M}_2+1)(\delta+\rho)+\mathscr{M}_2\tau)}{\mathscr{M}_3\mathscr{M}_4(\delta+\rho+\tau)+\mathscr{M}_4}, \\
\eta_0 &= \frac{\mathscr{M}_3((\mathscr{M}_2+1)(\delta+\rho)+\mathscr{M}_2\tau)}{(\mathscr{M}_1+1)(\mathscr{M}_2+1)(\mathscr{M}_3(\delta+\rho+\tau)+1)}, \\
\phi &= \frac{\tau(\mathscr{M}_2-\mathscr{M}_3(\delta+\rho))-\zeta_0(\mathscr{M}_3(\delta+\rho+\tau)+1)}{(\delta+\rho)(\mathscr{M}_3(\delta+\rho+\tau)+1)}, \\
c_0 &= \frac{\mathscr{M}_3((\mathscr{M}_2+1)(\delta+\rho)+\mathscr{M}_2\tau)}{\mathscr{M}_3\mathscr{M}_4(\delta+\rho+\tau)+\mathscr{M}_4}, \\
\chi &= \mathscr{M}_1(\mathscr{M}_2+1)+\frac{\mathscr{M}_3\log(\mathscr{M}_4)\left((\mathscr{M}_2+1)(\delta+\rho)+\mathscr{M}_2\tau\right)}{\mathscr{M}_3(\delta+\rho+\tau)+1}, \\
\zeta_0 &= \frac{\tau(\delta+\rho)\left(\mathscr{M}_5\left(\delta\mathscr{M}_3^2\tau-\mathscr{M}_3(\rho+\tau)-1\right)-\delta\mathscr{M}_3(\mathscr{M}_3\tau+1)\right)-\delta\mathscr{M}_2(\mathscr{M}_5-1)\tau(\mathscr{M}_3\tau+1)}{(\mathscr{M}_5-1)\rho(\mathscr{M}_3\tau+1)(\mathscr{M}_3(\delta+\rho+\tau)+1)}, \\
\Theta &= \frac{\mathscr{M}_3((\mathscr{M}_2+1)(\delta+\rho)+\mathscr{M}_2\tau)}{(\mathscr{M}_2+\chi+1)(\mathscr{M}_3(\delta+\rho+\tau)+1)}, \\
\Gamma &= \frac{\mathscr{M}_3((\mathscr{M}_2+1)(\delta+\rho)+\mathscr{M}_2\tau)}{\mathscr{M}_3(\delta+\rho+\tau)+1}+\chi.
\end{aligned}
\tag{89}
$$

(These formulas were automatically generated by Mathematica.)

Table 5 provides the numeric value of the assumed parameters.

Table 5: Parameter Values.

η_0	c_0	δ	ϕ	χ	τ	(ζ_0, ζ, σ)	ρ^{ss}	Γ	Θ
0.088	0.056	4.0×10^{-4}	0.78	0.70	0.5	$(0.095, 0.0, 0.57)$	7.94×10^{-3}	0.871	0.081

B.2.2 Dynamic Impulse Response-Based Targets

We set the remaining parameters and calibrate the monetary policy shock path to match the impulse responses generated by our SVAR.

Calibration of δ and κ_a.— We use the impulse responses to set the values of δ and κ_a. The parameter δ dictates the rate at which sales and production decline toward the trough, while κ_a controls the rate of recovery after exiting the inaction zone and when the monetary authority shifts policy. Our analysis does not assume an "active" recovery; instead, the recovery is driven by the impulse responses we incorporate as shocks. We target these rates over a 12-month window following the trough in sales as implied by the SVAR.

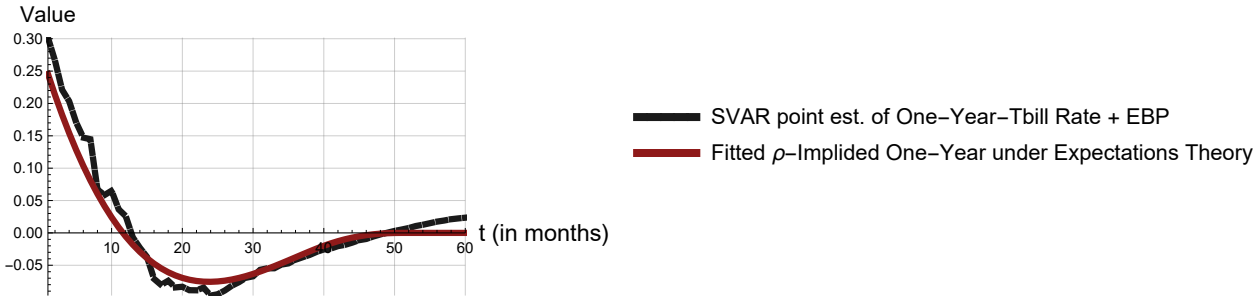


Figure 7: Fitted path of ρ using expectations theory and smooth polynomial approximation.

Notes: The figure shows the implied one-year rate by the fitted path of ρ_t that we assume in the model.

Calibration of the shock.— To calibrate the path of ρ_t, we target the impulse response of the sum of the one-year T-bill rate and the EBP as implied by the SVAR (point estimate). We assume that this rate converges to the steady state after 50 months. Using the expectations theory, we calculate the monthly rates implied by the annual series under perfect foresight. These are then converted to continuously compounded rates by taking the natural logarithm.

Specifically, let i_t^1 represent the monthly rate and i_t^{12} the one-year rate (the sum of the one-year T-bill rate and EBP), where $t = 1, 2, \ldots, 60$ denotes the months of data. To derive the monthly rate, we use the expectations theory under perfect foresight, which implies

$$1 + i_t^{12} = \prod_{j=1}^{12} \left(1 + i_{t+j-1}^1\right), \tag{90}$$

and hence

$$\frac{1 + i_{t+1}^{12}}{1 + i_t^{12}} = \frac{1 + i_{t+12}^1}{1 + i_t^1}, \tag{91}$$

and

$$i_t^1 = \frac{1 + i_t^{12}}{1 + i_{t+1}^{12}} \left(1 + i_{t+12}^1\right) - 1, \tag{92}$$

where, for $t \geq 50$, we assume steady state values: $i_t^1 = i^{ss} := \exp\left(\rho^{ss}\right) - 1$ and $\frac{1+i_t^{12}}{1+i_{t+1}^{12}} = 1$. We fit a third degree polynomial to a natural logarithm of the obtained monthly series $\{i_t^1\}$ to obtain a continuous function for ρ_t that enters the differential equation. Figure 7 shows the result after converting the interpolated function ρ_t back to a one-year forward rate using expectations theory.

To set the path of Q_t, we target the SVAR-implied trajectory of the inventory-to-sales ratio. We achieve this by fitting a third-degree polynomial to the positive portion of the impulse response and a connected polynomial to the negative portion. These two segments are joined where the data's impulse response transitions from positive to negative, ensuring continuity. The estimated polynomials provide a smooth target for the inventory-to-sales ratio, which our model is designed to match precisely by adjusting the path of Q_t. The next section will explain how this is implemented within our model's

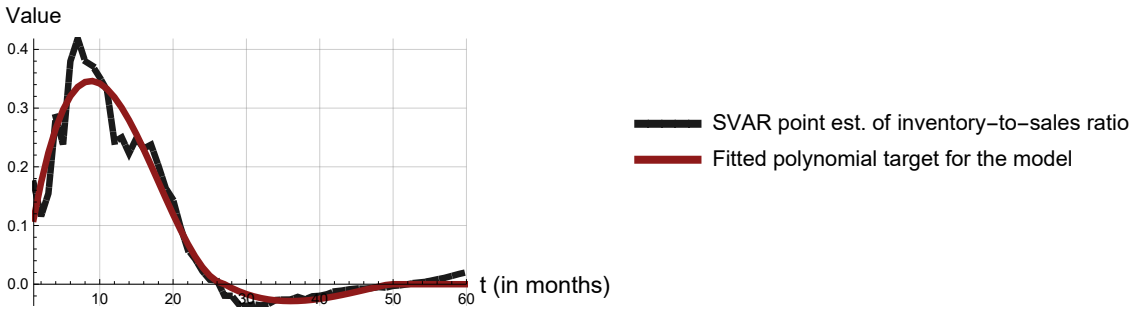


Figure 8: Fitted target for inventory-to-sales ratio.

Notes: The figure shows the fitted smooth target for the inventory-to-sales ratio that our calibrated model must hit via a particular path for Q_t.

solution framework. Figure 8 illustrates the fit of the two-part polynomial.

B.3 Numerical Procedure to Solve the Model

This section outlines the numerical approach we use to solve the model and implement our calibration strategy.

First, we solve the system of Hamilton-Jacobi-Bellman (HJB) equations in (18) and (20), substituting in the equilibrium wholesale price from (31) and the calibrated polynomial paths for ρ_t and Λ_t. We solve this system forward in time, assuming the model converges to the steady state at a distant horizon. The dependence of the HJB equations on ρ_t and Λ_t simplifies the problem, making the calibration of the inventory-to-sales ratio straightforward.

As discussed in the text, convergence to the steady state requires that ρ_t and Λ_t stabilize, but M_t and Q_t do not need to return to their steady-state values—in fact, they do not in our simulation. This is due to the model's assumption of wage rigidity, which allows for permanent changes in labor supply. We do not model how the economy recovers, other than it is implied by the impulse responses we feed (which does bring partial recovery). Importantly, this rigidity does not hinder the convergence of V_{0t} and V_{1t}.

Solving the HJB equations yields the paths for V_{0t}, V_{1t}, and X_t. Using V_{0t}, we identify the time $T > 0$ when the economy exits the inaction region, defined by $0 < V_{0t} < \phi v$. We confirm that V_{0t} remains within these bounds before T and does not reach the lower boundary $V_0 = 0$, which would trigger liquidations—which is never the case. This behavior is evident in the impulse responses shown in Figure 4, which shows how much larger than shock would need to be to trigger liquidations. At time T, $V_{0T} = \phi v$, and for $t > T$, $V_{0t} > \phi v$, indicating that $a_t > 0$ as given by (22). Specifically, for $t > T$, $a_t = \delta + (1/2)(V_{0t} - \phi v)/\kappa_a$, while $a_t = 0$ within the inaction zone (and $d_t = 0$, since the lower boundary is never reached).

Given the path for a_t, and the calibrated path for Λ_t, we next solve the differential equations in (23)

forward, using the steady-state values M_0 and N_0 as initial conditions (the boundary value is the initial value). From the resulting paths for M_t and N_t, we back out $Q_t = \Lambda_t M_t$.

Although we do not explicitly recover the original shock, it can be reconstructed using the relation $D_{0t}P_t^{-\varepsilon} = (1+\mathbb{E}\eta)Q_t$. Detailed implementation can be found in the annotated Mathematica notebook (`solve_Model.nb`) included in the replication package.

C SVAR Robustness and Extensions

In this appendix, we provide extended results for the SVAR analysis in Section 2. We provide alternative interpolation of markups, consider markups for all firms, and provide extended results that are presented in the paper.

C.1 Proxy SVAR Setup

As noted in the paper, we follow Gertler and Karadi (2015) (GK15, hereafter).[41] The SVAR includes 12 lags, a constant, and it takes the following form:

$$\mathbf{Y}_t = \mathbf{C} + \mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{A}_2\mathbf{Y}_{t-2} + \cdots + \mathbf{A}_{12}\mathbf{Y}_{t-12} + \mathbf{u}_t \tag{93}$$

where $\mathbf{Y}_t$ is an $n \times 1$ vector of endogenous variables in month t between 1979:m7 and 2012:m6, $\mathbf{C}$ is an $n \times 1$ set of constants, and $\mathbf{A}_i$ is an $n \times n$ matrix of coefficients for lag i (where $i = 1, 2, \ldots, 12$). We include the following variables:

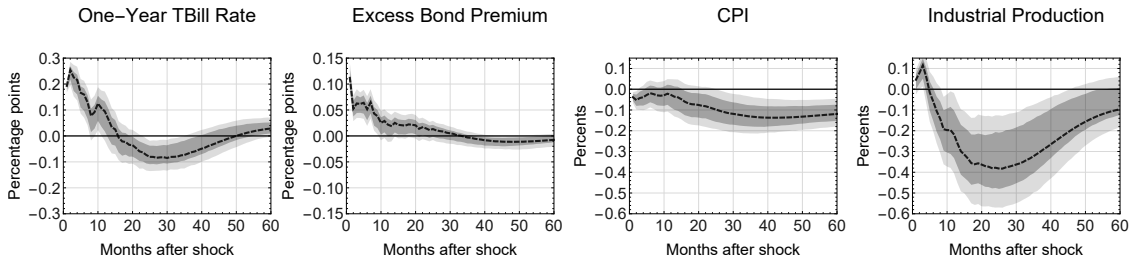
$$\mathbf{Y}_t = \begin{bmatrix} \text{1yr T-Bill Rate (GS1) in percent} \\ \text{Excess Bond Premium (EBP) in percent} \\ \log(\text{Consumer Price Index (CPI)}) \times 100 \\ \log(\text{Real Sales}) \times 100 \\ \log(\text{Real Inventory Stock}) \times 100 \\ \text{Compustat-derived Margin in (5)} \times 100 \end{bmatrix} \tag{94}$$

The first four variables are the same as in GK15, but we replace industrial production by real sales for manufacturing and trade industries given our focus on the inventory-to-sales ratio. This makes little different and the two behave similarly. Omission of mining and utilities is reasonable in the given context.

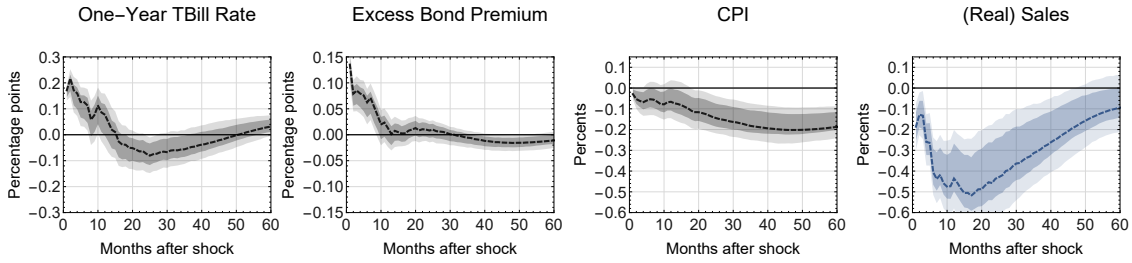
As in GK15, the residuals $\mathbf{u}_t$ are used in conjunction with their instrument (ff4_tc in GK15) to identify monetary policy shocks. The time series for ff4_tc and other variables can be found in the dataM.csv file in the replication package (see last section of this document). The variable 1yr T-Bill Rate serves as the policy indicator and the residuals associated with this variable (u_{1t}) are regressed on the policy

[41] Our Python implementation of GK15 estimation is based on the Matlab VAR Toolbox version of Gertler and Karadi (2015) due to Ambrogio Cesa-Bianchi (https://github.com/ambropo/VAR-Toolbox).

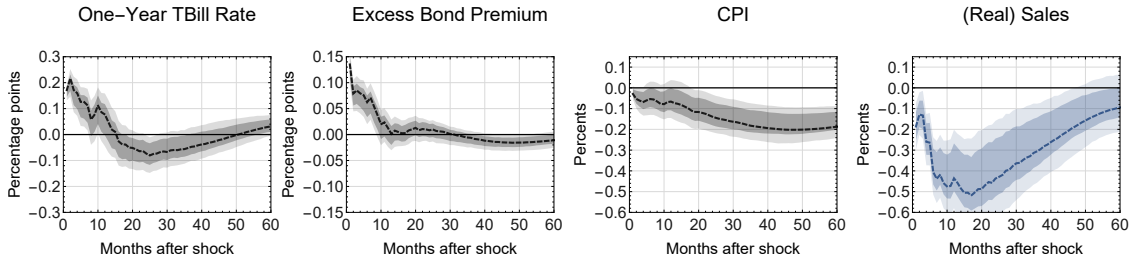
A. Original GK15 specification with industrial production:



B. GK15 specification with real sales replacing industrial production:



C. Baseline SVAR specification with no markup (first 4 variables reported):



D. Baseline SVAR as in the paper (first 4 variables reported):

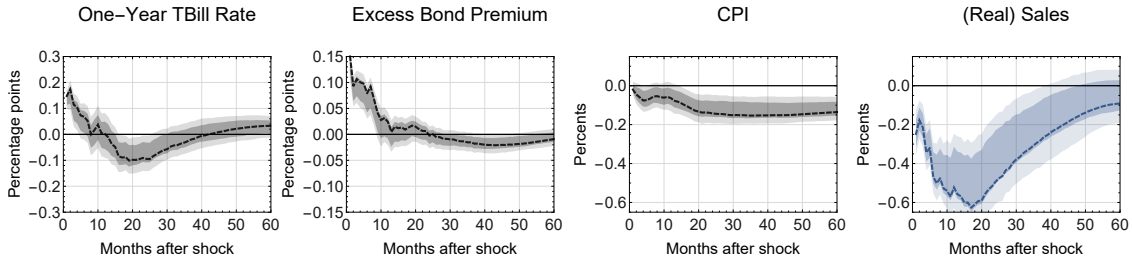


Figure 9: Comparison of Baseline SVAR to GK15 for the original variables.

instrument ff4_tc to obtain the predicted policy shocks. The data range for the instrument ff4_tc is

limited and narrower and ranges from 1990:m1 to 2012:m6. Data for overlapping variables (GS1, EBP and the CPI) are taken from GK15's replication file.

C.1.1 Comparison to GK15

In this section, we show how the results from our SVAR compare to the original GK15 specification. To that end, we report results from three different SVARs and focus on the original set of variables in GK15:

A. The original GK15 specification with industrial production (in place of real sales); specifically, variables Y_1, Y_2, Y_3, and Y_4 replaced by industrial production as in GK15.

B. The original GK15 specification with real sales replacing industrial production; specifically, this SVAR features a limited set of variables: Y_1, Y_2, Y_3, and Y_4 as stated above.

C. The baseline SVAR specification that includes Y_1, Y_2, Y_3, Y_4, Y_5, Y_6 but no gross margin (Y_7) (inventory impulse response for Y_5 and Y_6 is note reported).

D. The baseline SVAR as in the paper. Figure 2 in the paper shows the other impulse responses.

The results of this exercise are shown in Figure 9. As we can see, sales are slightly more volatile, as they exclude the less responsive mining and utilities sectors, but the dynamics are similar across all cases. Figure 9 (panel D) reports individual impulse responses for GS1 and EBP (in the paper we reported 'GS1+EBP', given this is what we used in calibration)

C.1.2 SVAR with Real Wages

This extension additional includes real wages in the SVAR as laid out above. The results are shown in Figure 10.

C.1.3 SVAR with Margin Series Interpolated using ChowLin Method

This extensions replaces gross margin series interpolated linearly in the paper by series interpolated using the ChowLin method. We use payroll series for manufacturing and trade industries to interpolate within quarter monthly values. This does not fix the delayed timing of earnings releases, and to remedy this issue we use a moving average of the obtained series over a 3 month period (average of the current month and the next two months). The results are similar except for size of the response, which is smaller. This is shown in Figure 11.[42]

[42]Replication code for ChowLin interpolation is in the folder TDReplicate (execute the file markup_to_monthly.m). It uses the Matlab replication package for Quilis (2018).

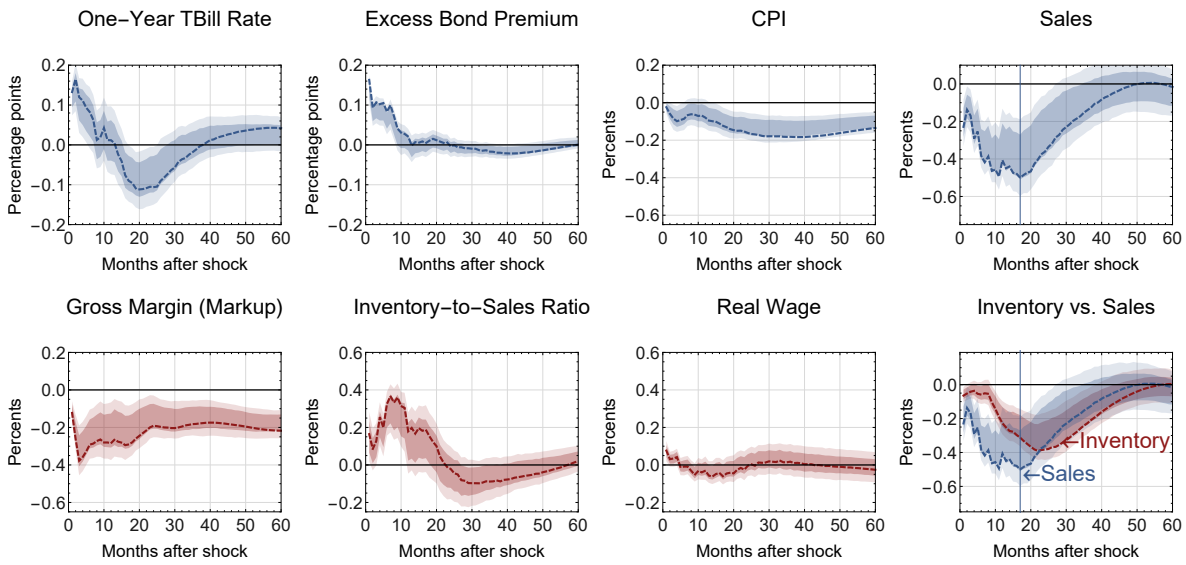


Figure 10: Extended baseline SVAR with real wage.

Notes: Notes to paper's Figure 2 apply.

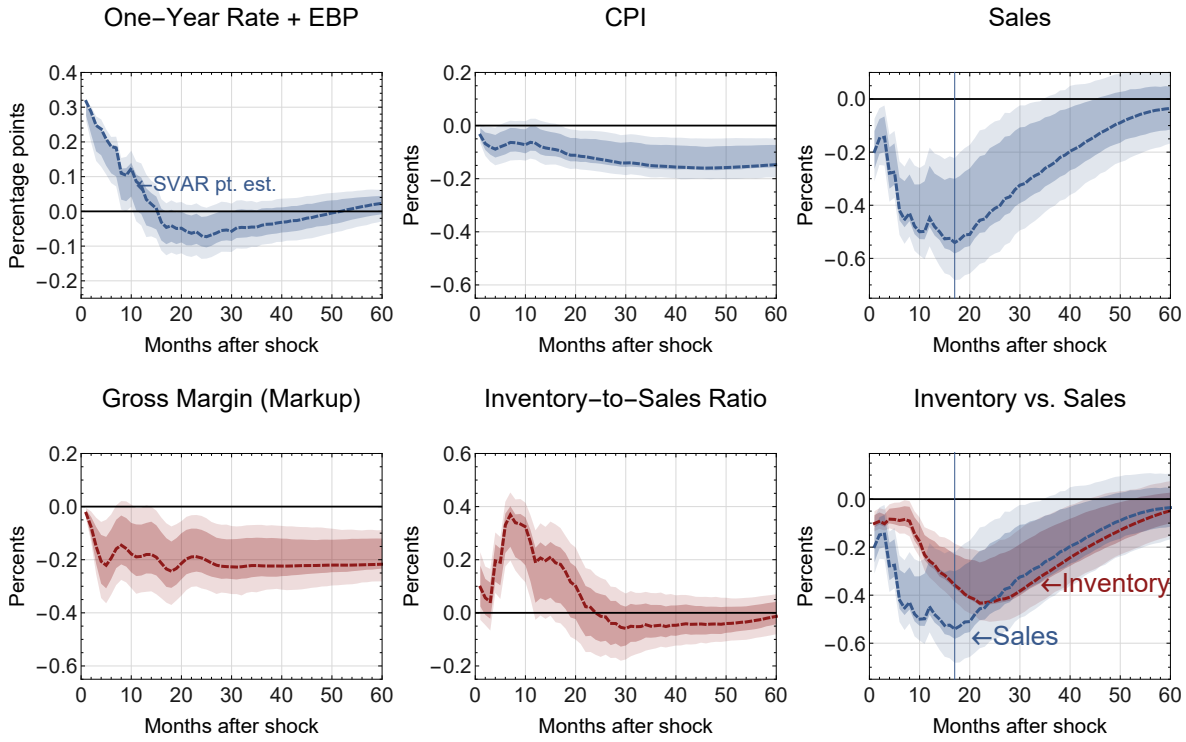


Figure 11: Baseline SVAR with gross margin interpolated using ChowLin method.

Notes: Notes to paper's Figure 2 apply.

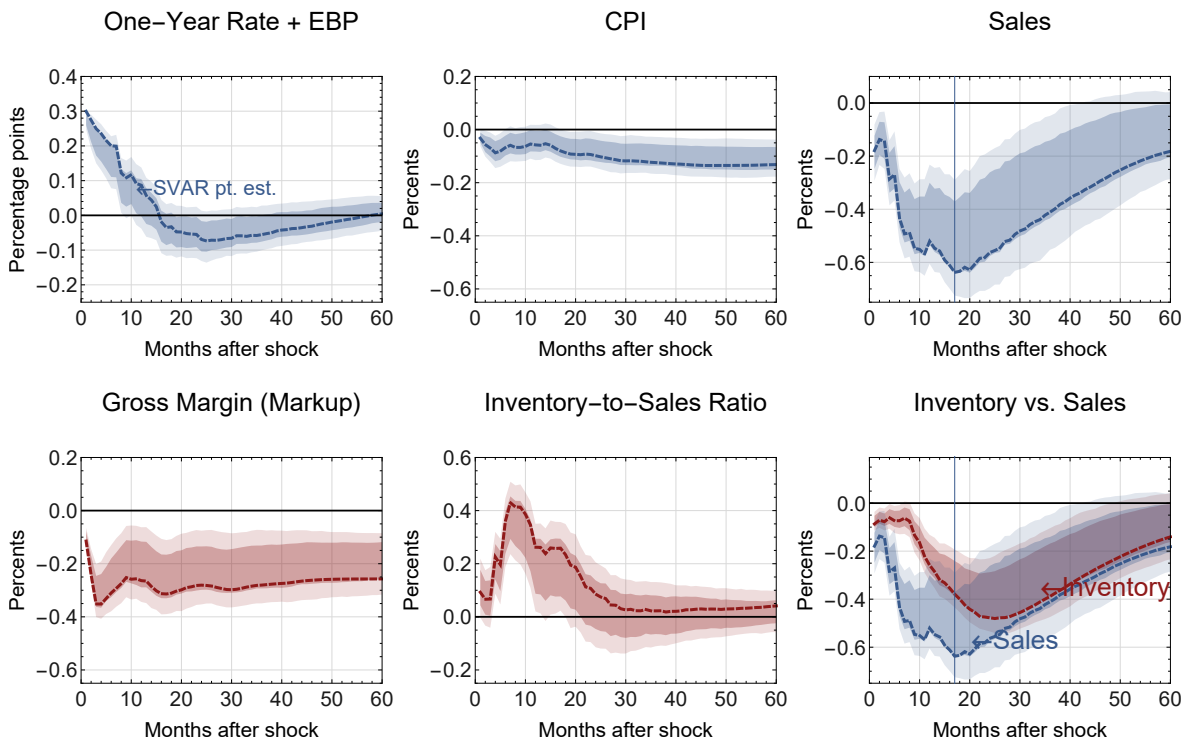


Figure 12: Baseline specification of SVAR with margin series for all firms (ex. FIRE).

Notes: Notes to paper's Figure 12 apply.

C.1.4 SVAR and Figure 1 with Margins for All Firms

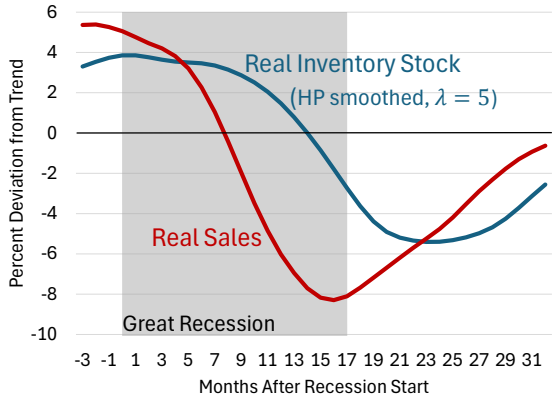
This extensions constructs gross margins using data for all firms, except for firms in finance, insurance and real estate sectors (FIRE). The results are shown in Figure 12 and they are almost identical. Figure 13 shows unconditional evidence as in Figure 1 in the paper for all firms (ex. FIRE).

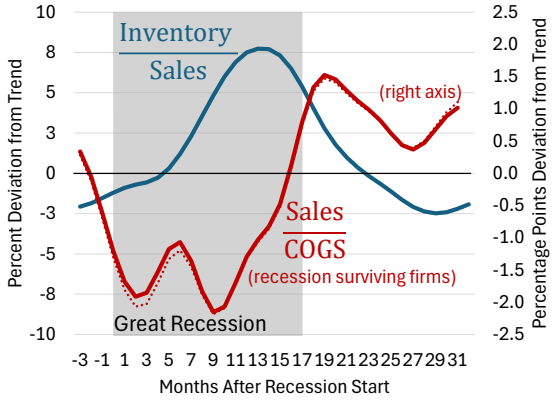
D Model Extension: Low Search Cost Calibration

This section explores the relationship between calibration targets and overall search costs, which amount to 7 percent of the value of produced retail goods in our baseline calibration. To do so, we consider an alternative calibration that excludes the inventory-to-sales ratio target and instead targets search costs at half the value assumed in the paper: 3.5 percent of the final retail price of a good. The assumed parameter values are detailed in Table 6, and the replication code for this extension is available in the *Mathematica* notebook `solve_Model_LS.nb`.

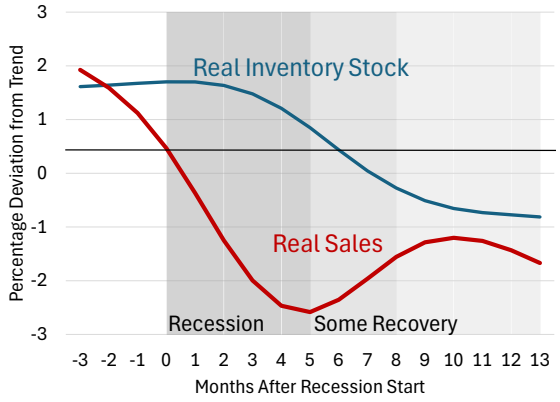
As shown in Figure 14, the results are almost identical. However, since this calibration omits the inventory-to-sales ratio as a target, that ratio drops to 0.5—significantly at odds with the data. This alternative calibration thus highlights a conflict between these two targets. In other words, our model,

A. Great Recession:





B. Average for U.S. recessions, 1979–2007:



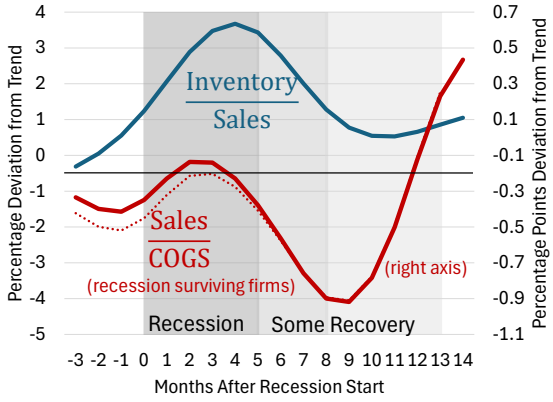


Figure 13: U.S. inventory, sales, and gross profit margins for all firms, 1979–2011.

Notes: The figure shows deviations from HP trend using a smoothing parameter of $\lambda = 10,000$—with additional smoothing applied to deviation (HP filter with $\lambda = 5$). Real inventory and sales data are plotted for the manufacturing and trade industries. The sales-to-COGS ratio is derived from Compustat Quarterly Fundamentals (North America) as described in Sesction 2 and here includes all firms (ex. FIRE) with positive sales right before each recession and one year after each recession (recession-surviving firms). Dotted lines incorporate output elasticities to map margins onto implied markups accordingly to the methodology developed by De Loecker et al. (2020). Shaded areas represent the duration of each recession. A detailed list of data sources is in the last section of this Online Appendix.

as specified, lacks the flexibility to simultaneously target this ratio and deliver lower search costs. We conjecture that introducing a form of congestion that reduces the productivity of search could provide the model with enough flexibility to match both targets.[43] The results shown here and our understanding of the model's mechanism is that such a modification would not affect inventory dynamics.

Are search costs in the range of 7 percent excessive for the entire supply chain? It is difficult to say because search costs cannot be directly measured. However, as a reference, consider the calibration

[43]This conjecture is based on the following reasoning: search costs are necessary to prevent excessive customer searches, ensuring that outposts spend less time in the production state relative to the marketing state. This balance is required to achieve the inventory-to-sales ratio target of 1.5. A reduction in the marginal productivity of search would produce a similar effect, lowering the search costs implied by the model.

Table 6: Parameter Values.

η_0	c_0	δ	ϕ	χ	τ	(ζ_0, ζ, σ)	ρ^{ss}	Γ	Θ
0.040	0.018	4.30×10^{-4}	1.12	0.69	0.5	$(0.137, 0.0, 0.64)$	7.940×10^{-3}	0.77	0.04

in Kaplan and Menzio (2016), who report that consumer search costs at the retail level alone amount to about two percent of the value of the good sold. A number within this range is consistent with other studies that focus on consumer search. Input-output tables indicate that well over two-thirds of transactions in the manufacturing sector are B2B transactions, and the process and risks involved in supplier selection suggest that search costs are substantial across the entire value chain.[44] Scaling the estimate from Kaplan and Menzio (2016) to account for the full supply chain brings search costs close to the level implied by the model. Therefore, while on the high side, we conclude that the 7 percent estimate falls within the range of model-based estimates of search frictions in the literature.

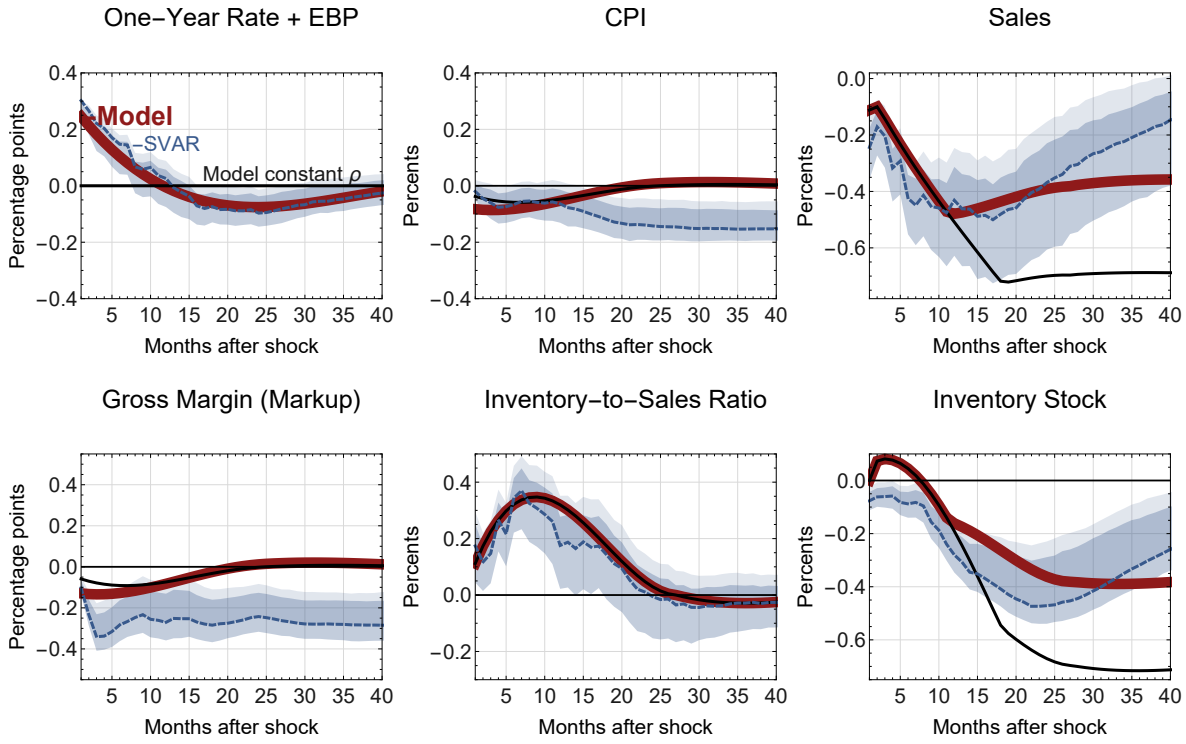


Figure 14: Results from Model with Low Search Costs (3.5 percent).

Notes: Notes to paper's Figure 6 apply.

[44] Beil (2011) provides an overview and discusses best practices in supplier selection.

E Replication Files and Procedures

Supporting codes and data files can be found in the online replication package 'DTreplication.zip'. The replication package uses the following software: *Mathematica*, Stata, Excel, and Python in the Jupyter Notebook environment (version 3.5). Some of the Excel files require HP filter toolbox written by von Kurt Annen and available at https://econpapers.repec.org/software/dgeqmrbcd/165.htm. To run replication files, the user must manually adjust the path to source files as described below. The contents of this package can be used under the standard GNU license GNU license agreement.

E.1 Replication of Main Results

The replication package contains four folders: 'Intro_Figure1', 'Model_Replication', 'IO_Calibration', and 'CL_Replicate'. Below we describe their content and the replication procedures:

Folder: `ModelReplication`

This folder contains codes that replicate all results reported in the paper's **Section 2** (Data), **Section 4** (Quantitative Analysis), and the results in the Online Appendix (Sections C and D). The folder contains the following files with codes/data:

- ***Mathematica* notebook `solve_Model_main.nb`**:

  This file generates the results reported in Sections 2 (Figure 2) and 4 (Figure 4 and 6). The notebook also contains supporting calculations for Sections 3.2, and Appendix Sections A and B. The notebook uses input CSV files found in the `/graphics/` subfolder and places all output files (PDF files) in that folder.

  To run the code, adjust the path at the top of the code so that it points to the `/graphics/` subfolder of the replication package. Input CSV files contain SVAR impulse responses and confidence bands and can be replicated by running `est_SVAR.py` (as described next). **Table 1** (Parameters) is also generated by the notebook (last section of the code).

- **Jupyter notebook `est_SVAR.ipy`**:

  This file estimates the baseline SVAR and generates CSV files found in the `graphics` folder for `solve_Model_main.nb`. Adjust the path in the first input cell so that it points to the `ModelReplication` folder of the replication package. The main data file, `dataM.csv`, contains time series for SVAR estimation except for the gross margin series, which is generated from raw data found in `collapsed_output_inv_csv.csv` (and placed in subfolder `./input/`). This file has been created by Stata do files `prep_data.do` and `proc_data_inv.do` described below.

- **Stata do files `prep_data.do` and `proc_data_inv.do`** (in `/stata_replication/`):

  These files construct the gross margin series as described in Section 2 from the raw S&P Compustat Quarterly Fundamentals dataset (`AllDataQrtly7824.dta`). The raw Compustat dataset file is not provided here due to copyright restrictions. Instructions for obtaining the data from WRDS are detailed in the next section. To replicate results, obtain the source file, rename it to `AllDataQrtly7824.dta` and adjust the paths accordingly.

Folder: `Intro_Figure1`

This folder contains Excel and supporting data files required to generate **Figure 1** in **Section 1**. To regenerate the source data, follow the steps:

1. **Pull Data from Compustat by Running Stata Scripts:**
   - Run `proc_data_inv_intro.do` (must run `prep_data.do` once prior to running proc files), located in the `Model_Replication` folder. Output will be saved in the `./input/` folder within `Intro_Figure1` as `collapsed_output_inv_fig_csv`.
2. **Seasonally Adjust Constructed Data:**
   - Run Jupyter notetebook `SeasonalAdjFigureIntro.ipy` (in `./intro_Figure1/` folder) to apply seasonal adjustments (ensure the path is correctly set). Output file will be placed in the `./output/` subfolder.
3. **Update Excel file `FigureIntro.xls`:**
   - Open the Excel file in `Intro_Figure1`. On the first sheet (`MarkupsFromCompustatRawData`), adjust the path in cell **C1**.
   - Press the **Refresh** button or manually run the macro `LoadDataMarkups` in Module1 (press **Alt+F11** for Visual Basic, or go to the **Developer** tab and select **Macros**)

Folder: `IO_Calibration`

This folder contains miscellaneous files used in calibration (Steady State Targets), including input-output tables for **Appendix Tables 3** and **2**, and BEA's inventory and sales series (`Inventories_Sales.xls`).

E.2 Replication of Supplementary Results

Next, we describe additional files and the replication of results in this Online Appendix. These results are generated by derivatives of the main codes described above:

- ***Mathematica*** **notebook** **`solve_Model_LS.nb`**: This generates results in Appendix Section D (Appendix Figure 14). Run analogously to the main file.

- **Jupyter notebooks** **`est_SVAR_XXX.ipy`**: These generate raw data for Appendix Section C. Output files are placed in `./graphicsXXX/` subfolders. Run analogously to the main file. To replicate Appendix figures, run `gen_SVAR_figs.nb` in each `./graphicsXXX/` subfolder. The subfolders contain the following results:

  - `./graphicsGK/`: Results for the original GK15 SVAR specification with industrial production (Appendix Figure 9—panel A).
  - `./graphicsGKs/`: Results for the GK15 specification with industrial production replaced by real sales (Appendix Figure 9—panel B).
  - `./graphicsE1NoMup/`: Results for the baseline SVAR excluding the margin variable (Appendix Figure 9—panel C). (Panel D is the baseline model and can be replicated by running `gen_SVAR_figs.nb` in the `/graphics/` subfolder.)
  - `./graphicsE1/`: Results for the baseline SVAR with gross margin series interpolated using the Chow-Lin method (Appendix Figure 11).

- **Replicating the Baseline SVAR with Real Wage Series (Appendix Figure 10)**: Run `ext_SVAR` and set `solve_SVAR_version.ipy` to 3 in input cell `In[25]`. Adjust path information in the first input cell. Output files are in `graphics/SVAR_with_real_wage/`. Run `gen_SVAR_figs.nb` in that folder to generate the figure.

- **Replicating the Baseline SVAR with Margin Series for All Firms (Appendix Figure 12)**: Run the Jupyter Notebook `ext_SVAR_all`. Adjust path information as noted. Output files are in the `./graphicsAll/` subfolder. Run `gen_SVAR_figs.nb` in that folder. Source files are in `./inputAll/` and were generated by the Stata do file `proc_data_all.do` in `./stata_replication/`.

To replicate the monthly interpolation using the ChowLin package, download the replication package for Quilis (2018) from the Journal of Monetary Economics data service. Replace the file `run_mainfile.m` and the `./csvfiles/` subfolder of the downloaded package in `./replication_files_jme_public/data` using the contents of the `CL_replication` subfolder in `ModelReplication` folder. Run the file `run_mainfile.m` after adjusting the path to ensure it points to the `./csvfiles/` folder. Output file series are in `/csvfiles/output/monthly_Mup.csv`. The generated series needs to be manually copy-pasted to `ModelReplicate/input/data_M.csv` file (in column `logMup100chowlin`) containing data for the SVAR estimation. The notebook `example_Sec6.nb` details the calculations for the back-of-the envelope discussed in Section 5 ("Mismeasurement of Variable Costs", footnote 36).

F List of Raw Data Sources

Overlapping series are sourced from the replication files of Gertler and Karadi (2015) for consistency. This includes: CPI, Gilchrist-Zakrajsek Excess Bond Premium (EBP), One-Year T-bill rate (GS1), and the monetary policy shock instrument FF4 (ff4_tc). These files were downloaded from the AEA website replication file for Gertler and Karadi (2015). Data on sales (saleq) and COGS (cogsq) come from S&P Compustat Quarterly Fundamentals (North America), accessed via Wharton Research Data Services. Other series were downloaded from the FRED II service of the Federal Reserve Bank of St. Louis in December 2023. These include: (1) real inventories and in manufacturing and trade industries (Real Manufacturing and Trade Inventories, Millions of Chained 2017 Dollars, Monthly, Seasonally Adjusted, and Real Manufacturing and Trade Industries Sales, Millions of Chained 2017 Dollars, Monthly, Seasonally Adjusted), originally sourced from various series published by Bureau of Economic Analysis and compiled by the FRED service of the Federal Reserve Bank of St. Louis to build long time series;[45] (2) payroll employment in manufacturing (All Employees, Manufacturing, Thousands of Persons, Monthly, Seasonally Adjusted), originally sourced from the Bureau of Labor Statistics. Most raw files—except for Compustat-derived series—are included in the replication package (Excel files contain metadata). Processed output pulled from the Compustat dataset is provided, but raw data cannot be shared. To replicate these results, download the Compustat Quarterly Fundamentals (NA) with all main variables for 1970-2015 from Wharton Research Data Services (WRDS) (or similar data outlet) and run the Stata replication code as described in the previous section. Output elasticities come from the online replication package of De Loecker et al. (2020) (downloaded from the personal website of Jan Eckout).

[45] Federal Reserve Bank of St. Louis, Real Manufacturing and Trade Industries Inventories [INVCMRMT] and Sales [CMRMTSPL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CMRMTSPL. The source data are U.S. Bureau of Economic Analysis Real Manufacturing and Trade Industries Sales for various time periods.