Monetary Policy Shocks: A Case Against Countercyclical Markups*

Lukasz A. Drozd
Federal Reserve Bank of Philadelphia

Marina M. Tavares
International Monetary Fund

June 23, 2024

Abstract

Existing business cycle theories of inventory dynamics have been used argued that markups are countercyclical in response to aggregate demand shocks because the inventory-to-sales ratio is. Here we challenge this conclusion by developing a new search model of inventory dynamics consistent with endogenously procyclical markups and positive correlation with the inventory-to-sales ratio. The key element of our theory is that market access is costly and serves as a long-lived complement to inventories in generating sales, which affects both the dynamics of the inventory-to-sales ratio and markups. Our model is quantitatively consistent with how profit margins move in the data—a fact that the standard theory struggles to account for in the parametric cases when inventory dynamics become informative about markup dynamics.

Keywords: markups, monetary policy, demand shocks, inventories, search, business cycle

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

*We thank Ben Lester for his comments and access to his search theory lecture notes. We thank Jonas Arias for his invaluable help in estimating the SVAR and access to additional codes for his work with Dan M. Waggoner and Juan F. Rubio-Ramirez. Drozd (corresponding author): Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106 (email: lukaszadrozd@gmail.com). Mendes M. Tavares: marinamendestavares@gmail.com.

1 Introduction

Are markups procyclical, countercyclical, or acyclical? The answer to this question is consequential for macroeconomics. For example, if markups are procyclical, sticky goods prices cannot be all that important in breaking monetary neutrality in New Keynesian models because they implicate and rely on countercyclical markups to break monetary neutrality. Markup dynamics also affect the interpretation of measured total factor productivity, and hence the importance of demand shocks in accounting for business cycle fluctuations. The post-pandemic outbreak of inflation has spurred renewed interest in how markups respond to aggregate demand shocks and their potential to generate what is known as the "price-price spiral."[1]

The debate on whether markups are procyclical or countercyclical is far from settled due to the difficulties of measuring marginal costs (Nekarda and Ramey, 2013). Despite these difficulties, the consensus view in New Keynesian literature is that markups are countercyclical in response to demand shocks. One of the leading arguments supporting this view is based on the idea that the countercyclical response of the inventory-to-sales ratio to monetary policy (MP) shocks informs how markups must move, indicating that, measurement issues aside, markups are countercyclical.

Intuitively, according to business cycle theories of inventories, inventories must contribute productively to aggregate sales to explain procyclical inventory dynamics in the data (Bils and Kahn, 2000; Fitzgerald, 1997; Blinder and Maccini, 1991). But if that is the case, the marginal value of inventory then largely derives from the markups earned on the sales that inventory directly enables. Since the inventory-to-sales ratio rises during demand recessions, the models predict that markups must high during recessions, and hence the conclusion. Kryvtsov and Midrigan (2012) (KM, henceforth) articulate this point in a New Keynesian menu cost model using monetary policy shock-based identification of demand shocks. Their work builds on the

[1]Blanchard and Blanchard and Bernanke (2023) and Weber and Wasner (2023) argue that market power and markups contributed to inflation. This topic has also garnered attention from policymakers in the context of rising prices. For instance, the House Energy and Commerce Subcommittee on Consumer Protection and Commerce held a hearing titled "Pandemic Profiteers: Legislation to Stop Corporate Price Gouging" on February 2, 2022, addressing concerns about corporate pricing behavior during the pandemic. Similarly, the Senate Committee on the Budget conducted a hearing on April 5, 2022, titled "Corporate Profits are Soaring as Prices Rise: Are Corporate Greed and Profiteering Fueling Inflation?"

earlier work by Bils and Kahn (2000) (BK, henceforth).[2]

But the thesis that markups are countercyclical and positively correlated with the inventory-to-sales ratio is not without controversy. In particular, it clashes with the cyclical properties of profits and profit margins seen in the data vis-à-vis the predictions of NK models featuring countercyclical markups and the standard assumptions on the production function (Figure 1 shows the unconditional correlation). Christiano et al. (1997) were the first to point this out, and their findings have been broadly echoed by the literature. For example, Broer et al. (2019) conclude that sticky wage models perform better because the effects of monetary policy are more plausible. According to Broer et al. (2019), the distributional effects of sticky goods price models imply "a transmission mechanism that is implausible: output falls (...) because markups and total profits rise, increasing (...) demand for leisure (. . .)."

These issues are not surprising. As first shown by Hall (1988), breaking the link between markups and profit measures requires a departure from the standard Cobb-Douglas production function and/or the assumption of linear pricing of labor—as we explain in the Data Section.[3] Part of the empirical literature argues that such a departure is warranted (Bils, 1987; Rotemberg and Woodford, 1999), but the extent to which this is the case is subject to debate (Nekarda and Ramey, 2013). Since most New Keynesian models do not feature these departures, countercyclical markup dynamics implies implausible profit dynamics in these models.

In this paper, we are motivated by this controversy and challenge the proposition that inventory dynamics inform markup dynamics. To that end, we explore the logical alternative to KM's argument: markups are procyclical or acyclical, as suggested by the conventional Cobb-Douglas identification of markups implied by the models, but the existing theory of inventories fails to account for the countercyclical dynamics of the inventory-to-sales ratio. We call this the *markup-inventory puzzle*, quantify it using a proxy SVAR, and seek to resolve it by proposing a new theory.

[2]Conditioning on MP shocks is important to control the effects of interest rates on the discount factor. As shown by Khan and Thomas (2007), the conclusion does not extend to productivity shocks in the standard RBC framework featuring an Ss model of inventories. However, unless there is a direct link between markups and interest rates (the discount factor), the conclusion reached by KM likely extends to other types of aggregate demand shocks. Monetary policy shocks are thus used as a litmus test for what may be true more broadly.

[3]As we show in the Data Section, markups are gross margins to the extent that costs of goods sold only capture variable production costs.

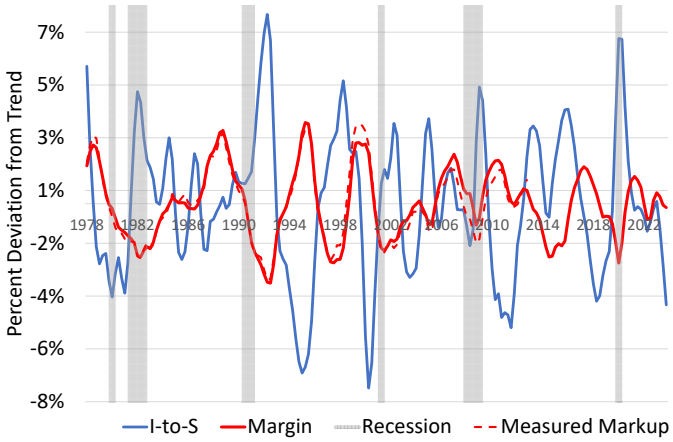

Figure 1: Gross profit margin and I-to-S ratio of inventory holding publicly traded firms.

Notes: The figure shows log deviations of gross profit margin and inventory-to-sales ratio for publicly traded inventory holding firms from HP10000 trend (after seasonal adjustment and after minor smoothing of the resulting series). The data excludes finance, real estate, insurance, and utilities, and the source is S&P Compustat Quarterly Fundamentals. The aggregation procedure controls for entry and exit as described in Section 2. The inventory-to-sales ratio also comes from Compustat, and total inventories have been analogously aggregated to construct an aggregate inventory-to-sales ratio.

The empirical part of our paper establishes four stylized facts about the responses of measured markups under the conventional Cobb-Douglas identification of markups from firm-level gross margins (De Loecker et al., 2020). These facts include (1) a countercyclical response of the inventory-to-sales ratio, (2) a procyclical or acyclical response of measured markups, (3) a scant response of the gross output-to-sales ratio and retail price, and (4) a decline in industrial capacity utilization and labor productivity. To identify the impact of monetary policy shocks, we follow Gertler and Karadi (2015) and estimate a proxy SVAR that includes markups and other relevant business cycle variables for its estimation and our study.

The key novel element of our theory, compared to the BK theory—a version of which our model nests—is the presence of *sunk* market access costs and search frictions that influence how buyers and sellers match and turn production into consumption. Specifically, production and distribution are assumed to be fragmented into long-lasting and costly-to-develop units, referred to as *access posts* (or simply shops). Access posts (shops) are essential for attracting searching customers, who are representative of distributors, and they *fragment* production into

random order streams from the producers' perspective. This fragmentation affects capacity utilization and pricing decisions after demand shocks. Monetary policy has real effects because wages are assumed (downwardly) rigid.

The key mechanism of our model that changes the correlation between markups and the inventory-to-sales ratio in response to MP shocks is that market access posts (shops) are long-lived assets complementary to inventories in generating sales. Following monetary contractions, the most economical adjustment is to allow the stock of access posts to depreciate at the exogenous rate—a fairly small number. Scrapping existing access posts, whose utility outlasts the shock's transient duration, is inefficient because investment in access posts is irreversible. Naturally, since access posts are more productive when utilized, as demand falls, firms' inventory holdings rise relative to sales. In addition, production fragmentation implies that firms lower markups to bring in more customers and increase resource utilization. Formally, this is because firms balance two constraints affecting the sales rate in the model: *production time*, which is positive due to production fragmentation and independent of the price (markups), and *matching time*, which is positive due to search frictions and depends on the price (markups). During demand recessions, matching time gains significance, leading firms to lower markups. Each individual firm deviates little from the prevalent markup set by its competitors, but equilibrium markups fall significantly because search induces complementarity in pricing decisions.

Our modeling of inventories is motivated by anecdotal evidence documented by related empirical literature. Survey data show that 86 percent of transactions in the manufacturing sector (70 percent across all sectors) occur via specialized business-to-business (B2B) supplier relationships (Blinder et al., 1998) and that about 70 percent of firms adopt a just-in-time delivery model (JITD) (Ortiz, 2022).[4] JITD does not imply the absence of inventories but rather that inventories arise as a byproduct of production and marketing. Consistent with this view, Blinder et al. (1998) report that 67 percent of inventory-holding firms consider inventories as "totally unimportant" in smoothing demand fluctuations.[5] Our model embeds this evidence

[4] A logistics strategy that involves maintaining minimal inventories and timing deliveries of intermediate goods precisely. This approach aims to minimize inventory storage costs and enhance the overall efficiency of the supply chain.

[5] See pages 96 and 277 in Blinder et al. (1998), who note "the most salient observation is that the great majority of sales are made not to consumers but to other businesses," and "if they [inventories] are used to

by assuming inventories play little role in smoothing output. Second, firms spend almost 20 percent of their costs on sales-related items, classified as Selling, General, and Administrative expenses on the income statement. Many of these expenses create lasting assets that enhance a firm's ability to find customers. In our model, these costs build the stock of access posts.

Our theory builds on the insights of Diamond (1971) and the subsequent literature on the so-called "Diamond paradox." The contributions by Wolinsky (1986) and Anderson and Renault (1999) are particularly relevant here because they similarly feature taste shocks to avoid the extreme implications of the Diamond paradox. The targeted nature of search in our model is different from but related to the relative entropy approach developed by Cheremukhin and Restrepo-Echavarria (2020). In the context of the business cycle literature, our work draws on the insights of Drozd and Nosal (2012) and Storesletten et al. (2011). Although their models do not incorporate inventories, they share with us the motivation for including search frictions and the common theme of documenting the importance of these frictions to understand a host of data patterns. The enodgenous link between resource utilization and demand is similar in our model to that in Storesletten et al. (2011).

2 Data

This section documents the key facts for our analysis. Our sample combines MP shocks with quarterly Compustat data, which we use to recover empirical impulse responses of markups, inventories and other relevant business cycle variables to MP shocks in a proxy SVAR setup.[6] Our approach of identifying markups (*measured markups*, hereafter) follows closely that in De Loecker et al. (2020). We describe it below.

2.1 Empirical framework and identification assumptions

To identify markups, we assume that firms operate a nested production function $\mathbf{Y}\left(\mathbf{V}\left(V_1, V_2, ...V_n\right), F\right)$, where F is a fixed factor bundle over the business cycle, and V_i are

buffer demand shocks, inventories of finished goods can be a source of price stickiness; when asked about this theory, however, firms reject it resoundingly."

[6]Detailed list of data sources is in the Appendix at the end.

inputs that can be adjusted on business cycle frequency. Assuming $\mathbf{Y}, \mathbf{V}$ are well behaved, and the price of each variable factor is given by some function $v_i(V_i)$ of the input V_i, production cost is

$$c(Y) := \min_{V \geq 0} (V(v) + fF) \text{ s.t. } Y = \mathbf{Y}(V, F), \tag{1}$$

where

$$v(V) := \min_{V_i \geq 0} \left(\sum_{i=1,...,n} v_i(V_i) V_i \right) \text{ s.t. } V = \mathbf{V}(V_1, V_2, ...V_n).$$

Using the first-order condition and the envelope condition, $v'(V) = c'(Y) \mathbf{Y}_V(V, F)$, and multiplying by the price p, it is clear that cost minimization implies the following formula for the markup:[7]

$$\mu := \log \frac{p}{c'(Y)} = \underbrace{\log \frac{pY}{v'(V) V}}_{\text{gross margin}} + \log \underbrace{\frac{\partial \mathbf{Y}(V, F)}{\partial V} \frac{V}{Y}}_{\text{output elasticity } \alpha}. \tag{2}$$

Accordingly, to identify markups, we need firms' revenue pY, as well as the cost of variable inputs measured at the marginal cost of these inputs.

2.1.1 Identification assumptions and discussion

To identify markups from firm-level data—which from now on we refer to as measured markups—we use the above formula and make the following assumptions:

1. Output elasticity α is independent of monetary policy shocks, firm input choices, and the aggregate state. That is, amid the apparent variation of inputs in the data, the production function is Cobb-Douglas, $\mathbf{Y}(V, F) = V^\alpha F^\beta$, with potentially stochastic parameters $\alpha, \beta > 0$ that are independent of firm choices and monetary policy.

2. The marginal cost of the composite variable input bundle $v'(V)$ is constant; that is, $v'(V) =: v > 0$. The sufficient condition for this to be true is that $\mathbf{V}(\cdot)$ is constant returns to scale production function and factor markets are competitive.

[7] The first order condition is $v = c'(Y) \frac{\partial Y(V,F)}{\partial V}$ after plugging in $c'(Y)$ in place of the Lagrange multiplier by the envelope theorem. Dividing by $c'(Y)$ and multiplying by P/v gives the expression in the text.

In addition, we assume that we can measure these objects using costs of goods sold on reported income statements:

3. The cost of goods sold reported by firms on their income statements identifies $v'(V)V = vV$.

Assumptions 1 and 2 are controversial in light of Bils (1987) and Rotemberg and Woodford (1999). However, as discussed by Nekarda and Ramey (2013), the results of this empirical literature are inconclusive. Furthermore, assumptions 1 and 2 are valid in most macro models, including the very inventory models used to argue that countercyclical inventory-to-sales ratio implies countercyclical markups. Finally, as discussed in Section 2.5, if assumptions 1,2,3 are violated, consistent with Bils (1987) and Rotemberg and Woodford (1999), the marginal cost curves at the firm level are steep so that they flip the correlation between markups and margins in the data. But steep marginal cost curves even in the basic BK model imply a negative correlation between markups and the inventory-to-sales ratio (see discussion in Section 4.5 and BK), which means that the inventory dynamics is no longer informative about the markup dynamics. As for assumption 1, we use output elasticities estimated by DEU at a two-digit industry level and interpolate them to a quarterly frequency. We find that these estimated elasticities do not affect the results and do not pursue further robustness in this respect. This is apparent from Figure 1, which plots both the gross margin and the measured markup that takes into account time-varying elasticities reported by DEU.

2.1.2 Implementation

To construct the aggregate markup under the stated identifying assumption, we use the S&P Compustat Quarterly Fundamentals (NA) dataset and operationalize the following formula based on the above theoretical framework:

$$\text{Measure Markup}_t = \log \sum_{i \in \mathcal{I}_t} \left(\frac{\text{COGS}_{it}}{\sum_{j \in \mathcal{I}_t} \text{COGS}_{jt}} \alpha_{it} \left(\frac{\text{Sales}_{it}}{\text{COGS}_{it}} \right) \right) = \log \frac{\sum_i \alpha_{it} \text{Sales}_{it}}{\sum_j \text{COGS}_{jt}}, \tag{3}$$

where α_{it} is the two-digit NAICS elasticity taken from De Loecker et al. (2020) and interpolated to quarterly frequency, Sales_{it} is firm-level sales (pY in equation 2), and COGS_{it} is firm-level cost of goods sold (vV in equation 2).

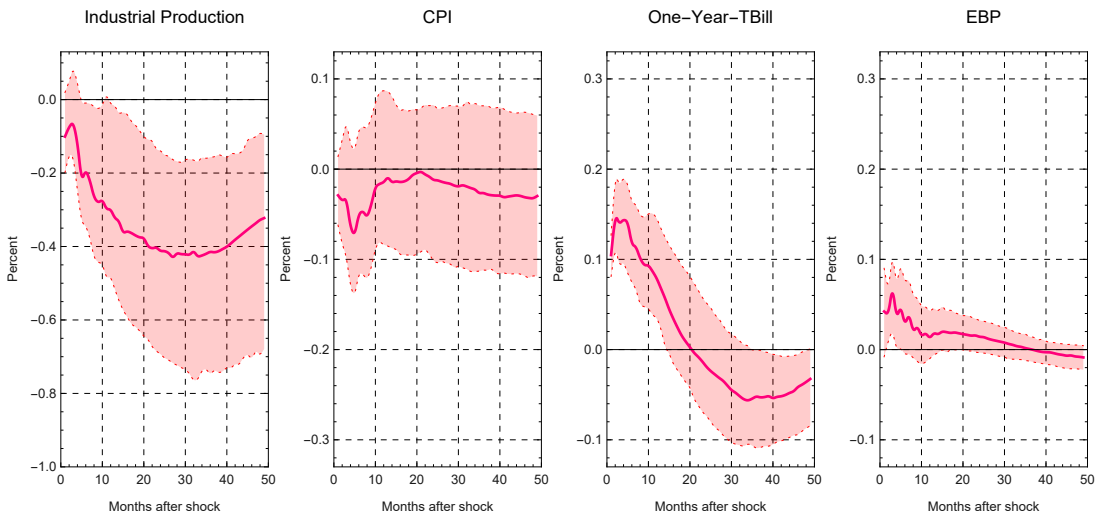

Figure 2: Estimated Impulse Responses of the Baseline Series (SVAR).

To account for the fact that Compustat only includes publicly traded firms and mitigate the impact of entry and exits to that dataset, we construct aggregates from the growth rates of firms in the dataset rather than summing up all existing firms—in the spirit of chain-weighting. This procedure ensures that firms entering or exiting the pool of publicly traded firms affect the aggregates only to the extent that their growth rate differs from the average growth rate of the firms in the sample. At the same time, we avoid the issues with the alternative route of fixing the panel of firms over such a long time horizon. Specifically, given some $Z_{it} \in \{\alpha_{it}\text{Sales}_{it}, \text{COGS}_{it}, ...\}$, or any other aggregate of interest, we construct $\sum_i Z_{it}$ by first calculating growth rates from the formula $g_{it} := (Z_{it+1} - Z_{it})/Z_{it}$, where t is a quarter and i is such that $Z_{it} > 0$ and Z_{it+1} are both in the sample—which defines some subset $\mathcal{I}_t^c \subset \mathcal{I}_t$ of firms—and then use the weight of each such firm $w_{it} := Z_{it}/\sum_{i\in\mathcal{I}_t^c} Z_{it}$ to construct the aggregate index from the recursion:[8] $Z_0 = 100, \quad Z_{t+1} = (1 + \sum_{i\in\mathcal{I}_t^c} w_{it} g_{it}) Z_t, \quad$ for all t.

To identify the response of markups to monetary policy shocks, and their comovement with other business cycle variables, we estimate a proxy SVAR using the framework developed by

[8]We also seasonally adjust the aggregated series by running a regression on quarterly dummies using data from 1979 to 2012. To seasonally adjust these series, we first log the series, use the HP filter with parameter 10000 to remove the low-frequency trends, and use the residuals from a regression on quarterly dummies to remove seasonality from the detrended series. We then add back the trend to these residuals to obtain seasonally adjusted series. We drop firms in the so-called FIRE sectors: finance, insurance, real estate, and public sector.

Gertler and Karadi (2015). The time period of the analysis is that of Gertler and Karadi (2015) for their instrumented estimation and whenever there is an overlap in data usage we use their original series.[9] That is, our SVAR includes the original time series found in Gertler and Karadi (2015)—CPI, industrial production, one-year nominal yield, and the Gilchrist and Zakrajsek (2012) excess bond premium (EBB, hereafter)—and adds additional series for the aggregate markup backed out from (3), inventory-to-sales ratio, capacity utilization, labor productivity, gross-output-to-sales. We construct the inventory-to-sales ratio from the series for real value of sales and inventory in manufacturing and trade industries published by Bureau of Economic Analysis (BEA).[10] We construct the gross output-to-sales ratio from the series and the identity: $Y \equiv S + \Delta I$, where S is real sales and ΔI is the change in real inventory holdings.[11] As shown in Figure 2, the SVAR yields similar results to those found in Gertler and Karadi (2015) for the original variables.

2.2 Key Findings

Figures 2 and 3 plot an impulse response to negative monetary policy shock that results in a normalized decline in the interest rate. We conclude from these responses the following stylized facts about the responses of the included variables to monetary policy shocks. The goal of the next section is to propose a theory consistent with these facts.

[9] Gertler and Karadi (2015) use 1979:M7 to 2012:M6 series to estimate the reduced form VAR, and 1991:M1 to 2012:M6 to identify the shocks using the instrument. Our identification uses 1991:M1 to 2012:M6 for both and used the Bayesian method developed by Arias et al. (2021).

[10] We compared the series with corporate profits to value added ration in national income account and the two series are strongly positively correlated. However, using aggregate corporate profits is not good because they include costs associated with fixed factors underlying the category of "Selling, General and Administrative." This can introduce significant bias, as explained in Section 2.4.

[11] Our estimation is based on the replication codes taken from Arias et al. (2021), which we modify as described in the Online Appendix to add the additional variables. The SVAR uses flat priors and it has 6 lags. All variables except interest rates are in natural logarithms (multiplied by 100%). The SVAR is on monthly frequency, and since both measured and productivity are quarterly series, we interpolate monthly values from their quarterly counterparts. To do so, we use BLS payroll employment (all private) to predict the monthly values spanning each quarter again using the Chow and Lin (1971) method and the MATLAB package taken from Quilis (2018). We have experimented with other interpolation methods, including simple linear interpolation from the endpost and the results are the same. Maintaining monthly frequency has the advantage of using the original MP shock instruments. Additional robustness tests can be found in the Online Appendix.

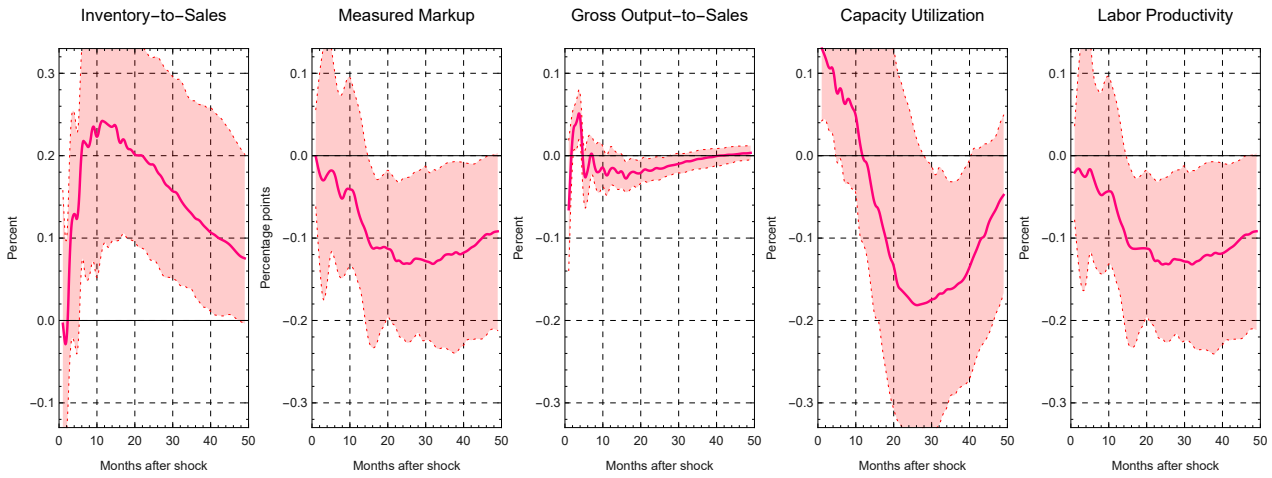

Figure 3: Aggregate Evidence: Impulse Responses of Added Series.

Stylized fact 1. ***Measured markup falls and the inventory-to-sales ratio rises.*** The median response of measured markups to an MP shock is strongly procyclical and the median response of the inventory-to-sales ratio is countercyclical. The result is significant up to the 90th percentile, as shown. Accordingly, the two series exhibit a strong conditional negative correlation.

Stylized fact 2. ***Gross output-to-sales is moves little and slightly rises.*** The gross output-to-sales ratio moves little in the data and is mildly procyclical. This fact indicates that inventories play a scant production smoothing role, echoing earlier findings of the literature (Fitzgerald, 1997; Blinder and Maccini, 1991; Bils and Kahn, 2000).

Stylized fact 3. ***Capacity utilization and labor productivity both fall.*** Capacity utilization falls, albeit with some delay, and labor productivity falls. The initial rise in capacity utilization is not statistically significant and we do not consider it relevant.

Stylized fact 4. ***The impact of MP shocks on retail prices is modest over the horizon of the shock, in part because monetary policy restores the equilibrium.*** The response of prices is muted. MP shock raises one-year T-bill rate for about 20 months and it leads to a significant increase in the excess bond premium—which has been shown to

be important to control for the endogenous response of monetary policy using proxy SVAR (Gertler and Karadi, 2015). During that time prices fall little, and thereafter monetary policy seems to reverse course.

2.3 A comment on mismeasurement of variable costs

Before we go on, it is instructive to discuss the consequences of a mismeasurement of variable production costs (in violation of Assumption 3). While COGS include direct production costs, such as costs of materials, labor costs associated with production, and overhead costs of production facilities, some of these costs may be fixed or sticky on the business cycle frequency. If that's the case, Assumption 3 is violated and the question is how it affects our conclusions. As we argue here, if that is the case, the inventory-markup puzzle is moot and the evidence for countercyclical is the evidence for mismeasurement.

To examine the effect of such a form of mismeasurement, consider a three factor Cobb-Douglas production function of the form: $Y = V^{\alpha}F^{\beta}F_0^{\zeta}$, where α, β, ζ are positive valued factor shares, V is a variable factor over the business cycle frequency, and suppose that, while F_0 and F are both fixed factors over the business cycle frequency, F is "erroneously" included in COGS alongside the variable factor V. For simplicity, assume factor prices v, f, f_0 are fixed and the variation of markups comes solely from the business cycle fluctuations in the price of output p.

Log-linearizing the implied formula for the measured markup and the "true" markup (markup, hereafter) around the long-run steady state shows that the mismeasurement of variable gods introduces a cyclical wedge

$$\log\left(\frac{\text{Measured Markup}}{\text{Measured Markup}^{ss}}\right) - \log\left(\frac{\text{Markup}}{\text{Markup}^{ss}}\right) = \frac{\beta}{\alpha}\frac{1}{\alpha+\beta}\log\left(\frac{Y}{Y^{ss}}\right).$$

Accordingly, if real gross output Y falls (rises) during recessions (booms), measured markup may fall (rise) even if the actual markup rises (falls).

While this is a problem to identify markup cyclicality from margins, to flip the correlation between the measured markup and the markup, note that β must be large enough—about half

of α the volatility of gross output and measured markup in our data.[12] Since $\alpha + \beta$ is less than .7 (the share of COGS in total costs across in our sample), β must be 1/3 or higher, in which case α is lower than 1/3. If so, flipping the correlation also implies that the marginal cost curves over the business cycle become quite steep, as implied by the log-linearized formula for the marginal cost:

$$\log\left(\frac{MC}{MC^{ss}}\right) = \frac{1-\alpha}{\alpha}\log\left(\frac{Y}{Y^{ss}}\right).$$

As shown by BK, and as we explicitely show in Section 4.5 using our model, with a strongly procyclical marginal cost the inventory-to-sales ratio becomes countercyclical even when markups are procyclical or acyclical. As a result, the inventory-markup puzzle no longer arises. Intuitively, if the marginal costs fall during recessions, firms expect future costs to be higher, and building up inventory after a negative MP shock is a profitable strategy even when markups are fixed. As discussed by Nekarda and Ramey (2013), departures from Assumptions 1 and 2 have a similar effect on the slope of the marginal cost curves and an analogous argument applies. In conclusion, the literature's point about inventories implying countercyclical markups is only valid when there is no mismeasurement, and only in that case what we do here is of interest. This justifies the focus on the marginal cost curves that are consistent with the standard Cobb-Douglas setup or even more flat.

3 Theory

Time is continuous, and the horizon is infinite. The economy consists of a representative household, a representative distributor, a representative producer, and a monetary authority that serves as a source of demand shocks. Goods are traded in a centralized (Walrasian) retail market and a decentralized wholesale market with search frictions.

The household consumes goods purchased from the distributor in the retail market, trades state noncontingent bonds affected by monetary policy, and supplies labor at a nominally rigid wage $w \equiv 1$ that serves as the numeraire. The rigid wage implies that the labor supply is equal to the labor demand at w.

[12] The variance of sales is about twice the variance of the measured markup.

To obtain goods for resale, the distributor dispatches shopping automata (shoppers) at intensity Q to the wholesale market, where shoppers match with access posts set up by the producer by searching for the lowest price posted by the producer's many access posts. Shoppers can haul $1+\eta$ units of the good, where η is an i.i.d. random variable interpreted as a match-specific taste shock only observable to the shoppers upon receiving a price quote. Matches are anonymous, transient, and formed instantly by the shoppers.

The producer employs a linear technology to produce goods from labor, implying marginal cost $v \equiv w \equiv 1$ (although $v \equiv w \equiv 1$, we keep symbols for clarity). The producer sets up access posts to attract shoppers at a sunk cost ϕv. Each access post holds at most one unit of output, contributing to search, and producing output takes τ^{-1} periods of time. Access posts depreciate at a fixed rate and hence represent the stock of marketing capital of the producer. We interpret M as inventory holdings across the economy and interpret M/Q as the inventory-to-sales ratio in the economy.

Throughout, we focus on a pure symmetric strategy equilibrium. Accordingly, all access points quote the same wholesale price $\tilde{p}^*$, and all shoppers employ the same search strategy dictated by the distributor. The aggregate state is represented by some vector $\mathbf{S}$ that evolves according to a known law of motion. All variables are functions of $\mathbf{S}$ and, hence, indirectly, functions of time. To simplify notation, we drop the time subscript t from all variables throughout. We use "dot" notation for time derivatives; that is, for any generic variable x, we write $\dot{x}$ in place of $dx(\mathbf{S}_t)/dt$. Figure 4 provides a schematic representation of the model setup.

3.1 Search, matching, and distribution

Let $M > 0$ be the measure of *stocked* access posts that are *visible* to the shoppers. In equilibrium, all access posts quote the same price,[13] and hence matches form at the rate

$$\Lambda := \frac{Q}{M}, \tag{4}$$

[13]It is helpful to rewrite this condition as follows: $\Lambda M dt = Q dt$. The left-hand side is the flow of matched producer access points, and the right-hand side is the flow of shoppers into search and perhaps from previous searches.

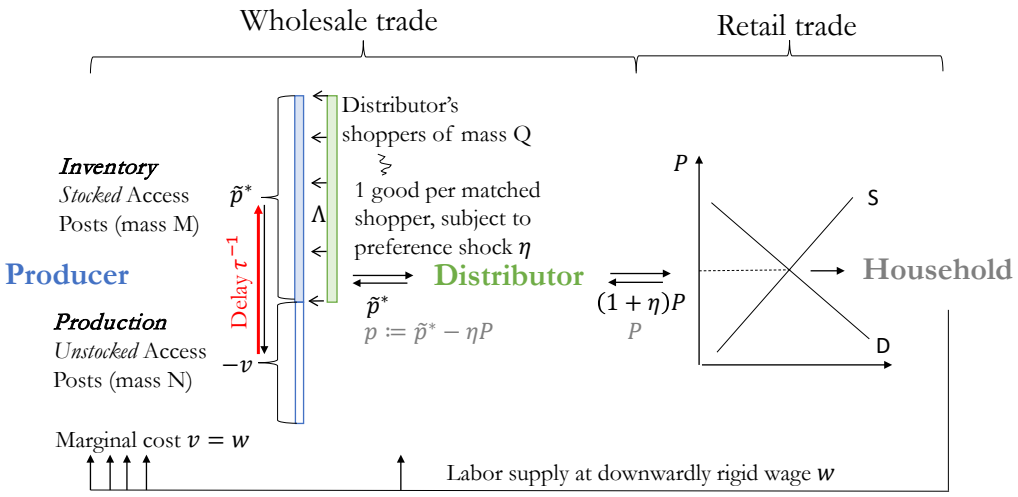

Figure 4: The Model: A Diagrammatic Exposition.

which we refer to as the *aggregate demand*. (We later describe how the arrival rate of shoppers differs for deviant producers who charge a different price.)

The resell value of the good for the distributor is $(1+\eta)P$, where $\eta \in \mathbb{R}$ is a match-specific, i.i.d. exponentially distributed taste shock with mean $\mathbb{E}[\eta] = \eta_0$ and variance $\text{Var}[\eta] = \eta_0^2$.[14] The shock is only observable to the shoppers upon forming the match and prices cannot depend on this shock. The probability density function (PDF) of this taste shock is denoted by $g(\eta) = \eta_0^{-1} \exp(-\eta\eta_0^{-1})$, and $G(\cdot)$ denotes its cumulative distribution function (CDF).

To simplify notation, it is convenient to define the *effective price* paid $p := \tilde{p} - \eta P$ by the shopper for a quality-adjusted unit of the good purchased at the quoted price $\tilde{p}$. By definition, in that case the resale price is P, so that the surplus from trade remains the same $P(1+\eta) - \tilde{p} = P - (\tilde{p} - \eta P)$. Since the surplus from trade is all that matters, the effective

[14]The idea of taste shocks builds on the insights of Anderson and Renault (1999) and Wolinsky (1986). As explained by Anderson and Renault (1999): "[The standard search framework without preference shocks] does not account for consumers searching for a product they like. To capture this idea, it is necessary to introduce heterogeneity across products." Our specification is simpler because it assumes that the search is executed according to a priori imposed objectives (reservation prices) by the distributor since search costs are unobservable to the distributor. The idea is that purchasing departments respond to management directives and fulfill them at whatever cost it may entail. Management sets directives based on average costs and does not monitor ongoing searches. Our setup is thus less suitable to describe consumer search, and more suitable to describe delegated B2B search. The presence of this shock in a simple and reduced form way can be thought of as capturing the presence of multidimensional types and relative entropy costs in Cheremukhin and Restrepo-Echavarria (2020).

price is a sufficient statistic to describe the outcome of a match. Given all access posts quote the same price in equilibrium, denoted by $\tilde{p}^*$, note that the CDF of effective prices is generated by the taste shock and given by

$$\Phi(p) := \Pr(\tilde{p}^* - \eta P \leq p) = \Pr\left(\eta \geq \frac{\tilde{p}^* - p}{P}\right) = 1 - G\left(\frac{\tilde{p}^* - p}{P}\right). \tag{5}$$

It is assumed that shoppers can pull a random effective price quote that falls below the reservation price $\bar{p} > 0$ set by the distributor. We provide the microfoundations for this *targeted search technology* below and for now, proceed. The cost of doing so is $c(\pi)v$, where $\pi = 1 - \Phi(\bar{p})$ denotes the required search precision associated with that reservation price, and where $\Phi(p \mid p \leq \bar{p})$ is the CDF of the obtained price.

Given search precision π implied by the reservation price, since the search is random, the surplus from a match is $s(\pi) = \int_{[0,\bar{p}(\pi)]}(P - p)\Phi(dp \mid p \leq \bar{p})$, where

$$\bar{p}(\pi) := \Phi^{-1}(1 - \pi). \tag{6}$$

In equilibrium, all access posts quote the same price; accordingly, the shopper buys the good whenever she draws a shock η that exceeds the cutoff

$$\bar{\eta}(\pi) := \frac{\tilde{p}^* - \bar{p}(\pi)}{P}. \tag{7}$$

Shocks are drawn each time a shopper draws a producer and obtains a new quote. Using the fact that the effective price associated with the shock η is $p := \tilde{p}^* - \eta P$, and the fact that $dp = -P d\eta$, after tedious algebra, it is not difficult to show that the surplus from the match is

$$s\left(\pi\right) = \frac{P - \tilde{p}^* + P\left(\bar{\eta}\left(\pi\right) + \eta_0\right)}{1 - \pi}, \tag{8}$$

where to derive this formula we used the fact that, by definition, $\pi = G\left(\bar{\eta}\left(\pi\right)\right) = 1 - e^{-\frac{\bar{\eta}^*}{\eta_0}}$, and

by the memoryless property of the exponential distribution, we know $\mathbb{E}[\eta \mid \eta > \bar{\eta}] = \bar{\eta}(\pi) + \eta_0$.[15]

The distributor chooses the optimal precision π^* to maximize the expected net surplus from a match, $y = s(\pi) - c(\pi)$. Since sending a shopper costs χv, the equilibrium surplus is

$$y^* := \max_{0 \leq \pi \leq 1} \{s(\pi) - c(\pi)\} = P\left(1 - \frac{\tilde{p}^*}{P} - \eta_0 \log\left(\frac{c_0}{\eta_0}\frac{v}{P}\right)\right) = \chi v. \tag{9}$$

3.1.1 Microstructure of targeted search

To match with producers, shoppers pull effective price quotes until they obtain a quote that falls below the reservation price. The key assumption is that the reservation price is set by the distributor a priori, without knowledge of the search costs incurred by shoppers. Consequently, shoppers must meet this requirement exactly to resell the good to the distributor and cover their expected search costs. This delegated setup offers significant tractability gains compared to a setup where shoppers sequentially decide whether to continue searching after each draw.

Formally, let $d\tau := (dt)^2$ be the duration of a obtaining a single effective price quote and suppose the costs is $c_0 v > 0$. As the period length shrinks to the limit ($dt \to 0$), drawing a price below the reservation price must succeed with probability one because an infinite number of draws is possible[16] Accordingly, as shown in the Appendix, the expected number of searches at precision π is $((1-\pi) + \pi\beta)(1 - \pi\beta)^{-1}$, where $\beta \equiv e^{-\rho d\tau}$, which implies

$$c(\pi) = c_0 (1 - \pi)^{-1}. \tag{10}$$

The expected surplus can be calculated analogously, as also shown in the Appendix, and it is given by (9).[17]

[15] Memoryless property of exponential distribution implies $g_{X|X>x}(z) = g_X(z - x)$, and hence

$$E[X|X > x] = \int z g_X(z - x)\, dz =_{u=z-x, du=dz} \int (u + x)\, g_X(u)\, du = x + \mathbb{E}[X].$$

[16] Individual searches can also be conducted in parallel. Parallel search does not invalidate the assumption of sequential search, as noted by Stahl (1989) (see footnote 1), as long as there is some cost of finalizing the match.

[17] These formulas apply even in the presence of a diffusion process—even though it is not present in our model. While the variance of a deviation of the Wiener process is of order $\sqrt{t}$, it is still a random walk process

3.1.2 Implied demand for goods

Consider now a deviant access post that quotes a different price from the equilibrium price $\tilde{p} \neq \tilde{p}^*$. Such a deviant knows that the shopper who receives its quote accepts it whenever her shock η obeys the inequality $\tilde{p} - \eta P \leq \tilde{p}^* - \bar{\eta}^* P$, where $\bar{\eta}^* := \bar{\eta}(\pi^*)$ and $\bar{\eta}(\cdot)$ is given by (7) and π^* solves (9). Therefore, the match is formed when the realized shock is such that $\eta \geq \bar{\eta}^* + \frac{\tilde{p}-\tilde{p}^*}{P}$, which happens with probability $G\left(\bar{\eta}^* + \frac{\tilde{p}-\tilde{p}^*}{P}\right)$, which means that the arrival rate of a quote-accepting shopper to that deviant is

$$\lambda(\tilde{p}, \tilde{p}^*) := \Lambda \frac{1 - G\left(\bar{\eta}^* + \frac{\tilde{p}-\tilde{p}^*}{P}\right)}{1 - G(\bar{\eta}^*)}. \tag{11}$$

Since in equilibrium $\tilde{p} = \tilde{p}^*$, the implied elasticity of demand at the equilibrium price is given by

$$\mathcal{E}(\tilde{p}^*) := \frac{\lambda_{\tilde{p}}(\tilde{p}, \tilde{p}^*)\,|_{\tilde{p}=\tilde{p}^*}}{\lambda(\tilde{p}, \tilde{p}^*)} \tilde{p}^* = -\frac{g(\bar{\eta}^*)}{1 - G(\bar{\eta}^*)} \frac{\tilde{p}^*}{P} = -\eta_0^{-1} \frac{\tilde{p}^*}{P}, \tag{12}$$

where $\lambda_{\tilde{p}}(\tilde{p}, \tilde{p}^*) := \frac{\partial \lambda(\tilde{p}, \tilde{p}^*)}{\partial \tilde{p}}$ is the partial derivative with respect to the sub-scripted argument.

3.2 Production and marketing

The producer maximizes the value of its portfolio of access posts, which is $\mathscr{V} = MV_1 + NV_0$, where V_1 is the value of stocked access posts, V_0 is the value of unstocked access points, and M, N are the corresponding masses. The choice variables comprise the price quoted by the representative access post $\tilde{p}$ at each point in time, and the entry and exit of existing access points from the pool of N. Since adding an access point costs ϕv and terminating an access point is free, optimal entry (assuming $N > 0$) requires that

$$0 \leq V_0 \leq \phi v, \; 0 \leq V_1. \tag{13}$$

and its expected innovation is zero at any point in time. Regarding the drift, since it is of order dt, it cannot impact the outcome of searches that take $\tau = (dt)^2$ to complete. Formally, let the average number of searches be n. Note that the time it takes to complete them is $n(dt)^2$, which is a fraction $\frac{n(dt)^2}{dt} = ndt$ of the period length that is normalized to 1. Since the choice of n is independent of the period length, the conclusion follows.

The producer will add or terminate access posts at an infinite rate on that boundary, and hence entry and exist are described by a Dirac delta function x^* at the top and δ^* at the bottom; formally,

$$(\delta^*, x^*) = \begin{cases} (d(V_0), 0) & \text{if } V_0 \leq 0 \\ (0, d(V_0 - \phi v)) & \text{if } V_0 \geq \phi v, \end{cases} \tag{14}$$

where $d(.)$ denotes a Dirac delta function.[18] The policy function thus feature an *inaction region (zone)* and a Dirac control on the boundary, and the measures M, N evolve according to the law of motion

$$\begin{aligned} \dot{M} &= \hat{\tau} N - \lambda(\tilde{p}, \tilde{p}^*) M - \delta M + xN \\ \dot{N} &= \lambda(\tilde{p}, \tilde{p}^*) M - \hat{\tau} N - (\delta + \delta^*) N + x^* N, \end{aligned} \tag{15}$$

where $\hat{\tau} N$ is the endogenous inflow of stocked units such that $0 \leq \hat{\tau} \leq \tau$ and $\lambda(\tilde{p}, \tilde{p}^*) M$ is the matching rate.

The price is set to maximize the value of a stocked access point, which evolves according to the Hamilton-Jacobi-Bellman (HJB) equation of the form

$$\rho V_1 = -\zeta v + \max_{\tilde{p} \geq 0} \{\lambda(\tilde{p}, \tilde{p}^*)(\tilde{p} + V_0 - V_1)\} - \delta V_1 + \dot{V}_1. \tag{16}$$

The left-hand side represents the opportunity cost of operating a stocked access post—the interest flow ρV_1 associated with its fair market value. The right-hand side represents all net flows and capital gains implied by holding the access post. In particular, $\zeta v \geq 0$ is the flow cost of maintaining the access post, and $\max_{\tilde{p} \geq 0} \{\lambda(.)(\tilde{p} + V_0 - V_1)\}$ represents the flow from selling the good, which comprises revenue $\tilde{p}$ from sales and the capital loss associated with the state transition stocked to unstocked state captured by the term $V_1 - V_0$. Finally, $\dot{V}_1$ is the capital gain associated with the evolving aggregate state. Accordingly, the quoted price $\tilde{p} \geq 0$

[18] We omit the law of motion when $N = 0$ as it will not be relevant. Exerting control on the boundary is independent of the differential equation (20). We use Dirac delta for analytic transparency, but this can be replaced by a large value of $x^* > \delta$ to approximately satisfy (13). Alternatively, convex cost of control can be added to HJB equation. This would not affect our results.

satisfies

$$\lambda_{\tilde{p}}\left(\tilde{p},\tilde{p}^{*}\right)\left(\tilde{p}+V_{0}-V_{1}\right)+\lambda\left(\tilde{p},\tilde{p}^{*}\right)=0. \tag{17}$$

An unstocked access post does not contribute to the search, and the production ferry arrives at a Poisson rate τ. Accordingly, the value of an unstocked access point V_0 is given by

$$\rho V_{0}=\max_{0\leq\hat{\tau}\leq\tau}\left\{-\zeta v\left(1-\hat{\tau}\tau^{-1}\right)+\hat{\tau}\left(-v+V_{1}-V_{0}\right)\right\}-\delta V_{0}+\dot{V}_{0}. \tag{18}$$

The producer may choose to idle the unit by picking $\hat{\tau}<\tau$, but in that case the maintenance cost $-\zeta v\left(1-\hat{\tau}\tau^{-1}\right)$ must still be paid. The term $-v+V_1-V_0$ is the production cost and capital gain implied by the arrival of the production ferry. It is clear that the policy function is bang-bang, and typically will involve $\hat{\tau}=\tau$:

$$\hat{\tau}=\begin{cases}\tau & \text{if } \left(-v+V_{1}-V_{0}\right)-\zeta v\left(1-\tau^{-1}\right)\geq 0\\ 0 & \text{otherwise.}\end{cases} \tag{19}$$

In the case the above HJB equation collapses to $\rho V_0=\tau\left(-v+V_1-V_0\right)-\delta V_0+\dot{V}_0$.

For later use, note that the sufficient statistic to describe the above dynamic problem is the value of inventory $\mathscr{X}=V_1-V_0$. This is clear by subtracting the two HJB equations side by side and the definition of $\mathscr{X}$, which, assuming $\tau=\hat{\tau}$, gives

$$\rho\mathscr{X}=-\zeta v+\max_{\tilde{p}\geq 0}\left\{\lambda\left(\tilde{p},\tilde{p}^{*}\right)\left(\tilde{p}-v-\mathscr{X}\right)\right\}-\tau\left(-v+\mathscr{X}\right)-\delta\mathscr{X}+\dot{\mathscr{X}}. \tag{20}$$

Using the formula for the access post-level demand, the equilibrium price must satisfy the fixed point:

$$\tilde{p}^{*}=v+\mathscr{X}+\eta_{0}\frac{P}{\tilde{p}^{*}}. \tag{21}$$

3.3 Consumption and monetary policy

The household problem is standard. The household maximizes $\mathbb{E}\int_0^\infty e^{-t\rho^*}\log(C_t)\,dt$ subject to budget constraint $P_tC_t + \dot{B}_t = \rho_t B + wL_t + \Pi_t$, where ρ^* is an exogenous discount factor, B is the stock of nominal bonds (consoles) that pay return $\rho > 0$, L is labor, and Π are aggregate profits. Since the nominal wage is rigid, L is dictated by the labor demand at the rigid wage w rather than chosen by the household. We do not model the details. (Standard no-Ponzi condition applies to B.)

The interest rate ρ is assumed to be set by the monetary authority and follows some (unspecified) policy rule that allows agents to form an accurate expectation of the future path of monetary policy after an MP shock. Market clearing requires that consumption equals production, and hence $C = Q$. The link between the interest rate set by MP and the economy is determined by the Euler equation and given by

$$\rho = \rho^* - \frac{\dot{C}}{C} = \rho^* - \frac{\dot{Q}}{Q}. \tag{22}$$

3.4 Equilibrium definition

We close the model by defining the perfect foresight equilibrium in response to an unexpected time path of an "MIT" MP shock $\{\rho_t\}$.

Definition 1. Given an assumed path for monetary policy $\{\rho_t\}$ and initial conditions $\{M_0, N_0, B_0\}$, the equilibrium comprises the paths of: i) distributor's search intensity and shopper search precision $\{Q_t, \pi_t\}$, ii) masses of producer access posts $\{M_t, N_t\}$, ii) retail and wholesale prices $\{P_t, \tilde{p}_t^*\}$, iii) access post values $\{V_{0,t}, V_{1,t}\}$ (equivalently $\mathscr{X}_t$), and access post stock control policy function $\{\delta_t^*, x_t^*\}$, such that the following conditions are satisfied: 1) HJB equations (16) and (18) given (11), (4) and (12), 2) the boundary condition (14), 3) the Euler equation (22), and 4) the law of motions (15), given policy given (19) and (21). Labor market clearing is replaced by the rigid wage assumption at $v \equiv w \equiv 1$.

The lemma below establishes the existence of a unique steady-state equilibrium. Steady-state equilibrium comprises a constant policy path $\rho_t = \rho^*$ and constant paths of all the

variables in the definition above that satisfy the stated equilibrium conditions assuming steady state values as the initial condition. Given this result, we later ensure the existence of a unique stable manifold in the neighborhood of the steady state. (All proofs are relegated to the Appendix.)

Lemma 1. *For $\eta_0 > 0$ small enough, the deterministic steady state equilibrium exists and is unique.*

3.5 Analytic Results

Here we characterize how aggregate demand affects markups and examine the response of markups and inventories to a demand shock (ceteris paribus). The literature uses monetary policy (MP) to isolate the effect of a demand shock while controlling for the path of interest rates. To that end, our analysis focuses on the case of moving the aggregate demand Λ while assuming constant ρ. This is convenient for two reasons. First, the shock is one dimensional, and second, the path of Λ determines the path of the inventory-to-sales ratio, which, recall, is Λ^{-1}. What we seek to establish is how markups respond to this shock.

3.5.1 Effect of aggregate demand on markups

Lemma 1 implies that the markup in the wholesale market is

$$\mu := \frac{\tilde{p}^* - v}{v} = \frac{\mathscr{X} - v}{v} - \mathcal{E}\left(\tilde{p}^*\right)^{-1} \frac{\tilde{p}^*}{v} = \frac{\mathscr{X} - v}{v} + \eta_0 \frac{P}{v}. \tag{23}$$

The last term ($\eta_0 \frac{P}{v}$) captures the fact that the taste shock is positive on average; that is, the fact that the distributor gets $1+\eta_0$ effective units of the good on average from each match. The important term is thus the first one, namely $\frac{\mathscr{X}-v}{v}$. Observe that the numerator of this term, namely

$$v^m := \mathscr{X} - v = V_1 - V_0 - v,$$

is the user cost of stock (inventory), since $V_1 - V_0 - v$ is the fair market value (cost) of instantly exchanging an unstocked access point for a stocked access point net of production cost v. To

see this in the context of the textbook formula for the user cost, note that in the deterministic steady state its $v^m = v(\delta+\rho)\phi\tau^{-1}$. Accordingly, to produce a unit of output, one needs to "rent" an unstocked access point for τ^{-1} periods of time (on average), which costs $v\phi$ in the steady state to buy or sell, and since the financing cost is $\rho+\delta$, the state user cost follows.[19]

This formula is intuitive and asserts that in equilibrium the markup μv covers the user cost of an access post. But it is not helpful to understand what drives pricing decisions in our model. Note that $\mathscr{X}$ depends on μ. On the boundary (and steady state) $\mathscr{X} = v + \phi v$ and both $\mathscr{X}, \mu$ are determined by the entry condition. But in the inaction zone, which is of interest here, μ and $\mathscr{X}$ are jointly determined and the above formula only tells us the outcome.

To better understand how prices are set in our model, we thus solve the two HJB equation for V_1 by assuming optimal policy function (and $\hat{\tau} = \tau$). After some manipulations, it is not now difficult to derive that

$$V_1 = (\rho+\delta)^{-1} \max_{\tilde{p}\geq 0} \left\{ \frac{\left(\tilde{p} - v\left(\tau(\tau+\rho+\delta)^{-1} + \lambda(\tilde{p},\tilde{p}^*)^{-1}\zeta\right)\right)}{\lambda(\tilde{p},\tilde{p}^*)^{-1} + (\tau+\rho+\delta)^{-1}} \right\} + \mathcal{R}\left(\dot{V}_1, \dot{V}_0\right),$$

where $\mathcal{R}\left(\dot{V}_1, \dot{V}_0\right)$ is some residual of time derivatives with the property that $\mathcal{R}(0,0) = 0$. This is just an equivalent representation of the two HJB equations, no more, but it is more informative.

Ignoring the residual, note that, by the above formula, the producer sets the price to maximize the profit from selling a unit of output (numerator) per unit of time that it takes to sell that units (denominator). In particular, in the numerator, the term $\tilde{p} - v\left(\tau(\tau+\rho+\delta)^{-1} + \lambda^{-1}\zeta\right)$ is the profit from selling a unit of output that adjusts for the effect of discounting and depreciation during production time ($\tau/(\tau+\rho+\delta)$), and it includes the maintenance capitalized maintenance cost ζv incurred over the time it takes to find a customer ($\lambda^{-1}\zeta$). The denominator comprises a discount factor and depreciation adjusted production time $(\tau+\rho+\delta)^{-1}$ and the matching time $\lambda(\tilde{p},\tilde{p}^*)^{-1}$. As expected, setting $\tau = \infty$ boils down to the standard monopoly

[19] The textbook formula of the user cost of capital assumes one period delay between investing and obtaining a unit of capital is: User Cost of Capital = Price of Capital Good * (Interest Rate + Depreciation Rate). The only difference is that the delay between investing and obtaining a unit of stock is τ^{-1} (in expectation) in our model.

problem of the form:

$$V_1 \to_{\tau \uparrow \infty} (\rho + \delta)^{-1} \left(\max_{\tilde{p} \geq 0} \left\{ \lambda \left(\tilde{p}, \tilde{p}^* \right) \left(\tilde{p} - v \right) \right\} - \zeta v + \dot{V}_1 \right),$$

and so the difference between these two formulas is key to understand how pricing in our model differs from the standard case.

To that end, consider the following prototypical representation of profit maximization implied by our model:

$$\max_{\tilde{p}} \frac{\tilde{p} - \hat{v}}{T \left(\tilde{p}, \tilde{p}^* \right) + T_0} \Rightarrow FOC : p - \hat{v} = \frac{T \left(\tilde{p}, \tilde{p}^* \right) + T_0}{T_{\tilde{p}} \left(\tilde{p}, \tilde{p}^* \right)} = \frac{\Lambda^{-1} f \left(\tilde{p}, \tilde{p}^* \right) + T_0}{\Lambda^{-1} f_{\tilde{p}} \left(\tilde{p}, \tilde{p}^* \right)}.$$

where $T \left(\tilde{p}, \tilde{p}^* \right) = \Lambda^{-1} f \left(\tilde{p}, \tilde{p}^* \right)$ captures how matching time depends on the aggregate demand Λ and the quoted price via some function $f(\cdot)$, T_0 is adjusted production time, and $\hat{v}$ is the adjusted marginal cost. Note that the standard case corresponds to $T_0 = 0$, and so what is of interest is the effect of $T_0 > 0$. It is clear from the first order condition on the right that the key effect has to do with the fact that when $T_0 > 0$, aggregate demand Λ affects the markup set by the producer over the effective marginal cost $\hat{v}$; in particular, the markups becomes procyclical in response to demand. In contrast, when $T_0 = 0$,which is the standard monopoly case, markups are positive but they do not vary with demand Λ.

Why is this the case? Intuitively, profit maximizing producers with $T_0 > 0$ take into account how matching time, which depends on the markup and aggregate demand, *relates to* production time. They adjust markups because matching time lagging production time reduces utilization of the fixed resources and that is costly. In particular, if selling time is longer, producers are willing to sacrifice markups to increase matching rate and the payoff is a higher utilization rate of their sales infrastructure.

3.5.2 General equilibrium

In general equilibrium, wholesale prices depend on retail prices by (21) and (9). Obtaining a closed form solution is not possible, and to analyze our model analytically we resort to a

log-linear approximation of the logarithmic term in (9) around the steady-state solution, which we established earlier. In particular, we approximate the term

$$\log\left(\frac{c_0}{\eta_0}\frac{v}{P}\right) - \log\left(\frac{c_0}{\eta_0}\frac{v}{P_{ss}}\right) \approx \left(\frac{P}{P_{ss}}\right)^{-1} - 1, \tag{24}$$

soak up the term "$\log\left(\frac{c_0}{\eta_0}\frac{v}{P_{ss}}\right)$" from the calculation of the steady state by defining

$$\tilde{\chi} := \chi - \eta_0 P_{ss} \log\left(\frac{c_0}{\eta_0}\frac{v}{P_{ss}}\right), \tag{25}$$

and plug in these results to (21) and (9) to obtain

$$\tilde{p}^* = \mathscr{X}\,(1+\eta_0) + \Theta v, \;\; P = \mathscr{X} + \Theta\frac{v}{\eta_0}, \tag{26}$$

where

$$\Theta := \eta_0 \frac{\eta_0\,(1+\phi\tau^{-1}\,(\delta+\rho)) + \tilde{\chi}}{1-\eta_0}. \tag{27}$$

These are the equilibrium prices that approximate the linear relationships around the deterministic steady state. We next use them to obtain a full characterization of local equilibrium dynamics.

3.5.3 Impulse response of markups and inventory-to-sales ratio to a demand shock

We now characterize how equilibrium markup and the inventory-to-sales ratio respond to a demand shock that moves Λ because Q falls and M is sticky due to the inaction region. The key friction, of course, is the fixed cost $\phi v > 0$ (amid $\tau^{-1} > 0$) and the assumption that δ is fairly small so that the active adjustment margin is minuscule. Below, we consider two cases for comparison: the *frictionless model* with $\phi = 0$ and the *baseline model* with $\phi > 0$. (Our quantitative model endogenizes this shock as MP shock and generates it from ρ.)

Frictionless model ($\phi = 0$): We consider this case to highlight the key result derived by BK to show that the markups and the inventory-to-sales ratio are intimately linked without

friction. Note that when $\phi = 0$, $\mathscr{X} \equiv v$ (all t) and hence $\dot{\mathscr{X}} \equiv 0$. Intuitively, this case arises because, effectively, the producer has full control over the mass of stocked access posts. At any instance of time, the producer can add access posts at an infinite rate, and then terminate any excess ones at no cost.

By (23), it is immediate that the markup is given by

$$\mu^* = -\mathcal{E}\left(\tilde{p}^*\right)^{-1} \frac{\tilde{p}^*}{v} = \eta_0 \frac{P}{v}, \tag{28}$$

which yields the standard monopoly pricing formula, and hence it is constant by (26). As noted before, the key property of this formula is that the level of demand Λ is irrelevant, and that is why the markups are constant. The retail price P also does not change after the shock.

As for the the inventory-to-sales ratio, we plug in $\mathscr{X} = v, \dot{\mathscr{X}} = 0$ to the HJB equation in (20) to calculate Λ^{-1}, which gives

$$\Lambda^{-1} \equiv \frac{M}{Q} = \frac{\Theta + \eta_0}{\delta + \zeta + \rho} \mu^*, \tag{29}$$

where, recall, Λ^{-1} corresponds in the model to the inventory-to-sales ratio. What this formula shows is that the assumed path for Λ is simply infeasible unless the markup moves. This result is the analog of the result obtained by Bils and Kahn (2000), since we assumed here a constant marginal cost.

Baseline model ($\phi > 0$): Consider now the baseline setup assuming $\phi > 0$ and suppose the shock is large enough so that the economy enters the inaction zone (M then slowly depreciates at rate δ). Suppose the shock is transient and

$$\dot{\Lambda} \equiv \left(\Lambda^{ss} - \Lambda\right) \omega \Rightarrow \begin{cases} \dot{\Lambda} > 0 & \Lambda < \Lambda^{ss} \\ \dot{\Lambda} < 0 & \Lambda > \Lambda^{ss} \end{cases}, \tag{30}$$

where $\omega > 0$ is some assumed mean-reversion rate of the shock. By definition, the inventory-to-sales ratio Λ^{-1} in this case rises and steadily declines towards the steady state.

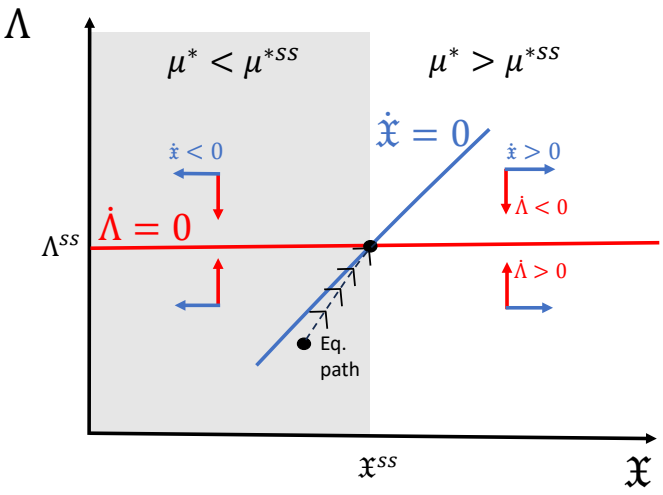

Figure 5: Phase Diagram for Proposition 1.

Plugging (26) to the combined HJB equation in (20), it is clear that the value of inventory follows the differential equation of the form

$$\dot{\mathscr{X}} = \mathscr{X}\left(\delta + \rho + \tau - \eta_0 \Lambda\right) - \Lambda v \Theta - v\left(\tau - \zeta\right), \tag{31}$$

where $\delta + \rho + \tau - \eta_0 \Lambda > 0$, as long as $\tau > \zeta$ (which we assumed) and $\eta_0 \Lambda$ is not too large—or else in steady state we could not have $\dot{\mathscr{X}} = 0$. Accordingly, in the neighborhood of the steady state, we have

$$\begin{cases} \dot{\mathscr{X}} > 0 & \mathscr{X} > \mathscr{X}^{ss} \\ \dot{\mathscr{X}} < 0 & \mathscr{X} < \mathscr{X}^{ss} \end{cases}. \tag{32}$$

Plugging in to the retail price P from (26) to (23), the formula for the markup is exactly identical,

$$\mu = \underbrace{\frac{\mathscr{X} - v}{v}}_{=v^m/v} + \eta_0 \frac{\tau + (\delta + \rho)\,\phi}{\tau}. \tag{33}$$

but since $\mathscr{X}$ is no longer constant, markup is no longer constant. As explained earlier, in this case, we are in the inaction zone and producers see a long selling time. This moves both the markup and $\mathscr{X}$ and the fixed point of that process is captured by the equation above.

To show negative correlation between μ^* and Λ^{-1}, consider the phase diagram in the space

of $(\Lambda, \mathscr{X})$, as locally implied by (31) and (30). The signs of the vector field (arrows in the figure) are given by (32) and (30), which implies that, along the stable manifold, $\mathscr{X}$ drops together with Λ and converges back to the steady state as the shock mean reverts. Accordingly, on impact we have $\mathscr{X} < v + \phi v$ and thereafter $\dot{\mathscr{X}} > 0$, which means that and $\mathscr{X}$ slowly converges towards its steady state value $v + \phi v$. Since the inventory-to-sales ratio corresponds to Λ^{-1}, the inventory-to-sales ratio goes up as the markup falls. Accordingly, we have established that the markup and the inventory ratio are *negatively* correlated.

Intuitively, the friction changes how markups move vis-à-vis the inventory-so-sales ratio is because access posts are firm's long-lived asset and scrapping them early in response to transient shocks is not optimal. But if that's the case, the producer is better off utilizing these assets. This raises the inventory-to-sales ratio and the fact that selling time is longer results in lower markups. In equilibrium, these affects are supported by a decline in the user cost v^m in (33). Of course, in equilibrium, lower markups cannot change the aggregate arrival rate, which is still Λ. Put differently, the market becomes more competitive and for this reason, markups decline.

Formally, the HJB equations show that what sustains the negative correlation between the markup and the inventory-to-sales ratio is the fact that the value of inventory change. This is clear from the formula

$$\Lambda^{-1} = \frac{M}{Q} = \frac{\eta_0 \mathscr{X} + v\Theta}{\mathscr{X}\left(\delta + \zeta + \rho\right) - \dot{\mathscr{X}} - v\left(\tau - \zeta\right)} \mu^*, \tag{34}$$

and the fact that the first term is no longer constant since it involves $\mathscr{X}$.

3.5.4 Response asymmetry

Our model is a representative agent model and it implies an asymmetric response to MP shocks. This is because the inaction region applies to negative shocks but it does not apply to positive shocks. However, the stark nature of this asymmetry should taken with a grain of salt because it is driven by the fact that our model is a representative agent setup. A simple extension that incorporates idiosyncratic demand shocks on a sectoral level would break this result because some firms/sectors would be in the inaction zone in the stationary steady state. Note also that

some asymmetry is consistent with the data because the largest spikes in the inventory-to-sales ratio do occur during recessions.

4 Quantitative analysis

We next calibrate our model and compare the implied impulse responses to the ones derived from the SVAR in Section 2. We begin by describing how we calibrate its parameters, and how we solve it numerically.

4.1 Parameterization

Period length in the model (integer values of t) is one month to match the frequency of the SVAR from Section 2. To calibrate c_0, η_0 and $\tilde{\chi}$, we set three joint targets: 1) The wholesale markup $\mathscr{T}_1 := (p - v)/v$ of 50 percent, which is roughly in line with the gross margin in the S&P Compustat for the more recent years (De Loecker et al., 2020),[20] 2) the distribution content in final consumption of 50 percent—as measured by the ratio $\mathscr{T}_2 := (P - \tilde{p}^*)/P$—which is roughly in line with the estimates of the share of distribution content of retail goods (Burstein et al., 2000), and 3) the average search cost $\mathscr{T}_3 := (c(\pi^*)v - \tilde{p}^*)/\tilde{p}^*$ of 2%. The mapping between these targets and the parameters is as follows:

$$c_0 = \mathscr{T}_3 \frac{(\mathscr{T}_2 + 1)(\tau\mathscr{T}_2 - \phi(\delta + \rho))}{\tau(\mathscr{T}_2 + \mathscr{T}_3 + \mathscr{T}_2\mathscr{T}_3) - \phi(\delta + \rho)}, \quad \eta_0 = \frac{(\mathscr{T}_1 - 1)(\phi(\delta + \rho) - \tau\mathscr{T}_2)}{\tau(1 + \mathscr{T}_2)}, \quad \tilde{\chi} = \frac{\mathscr{T}_1(1 + \mathscr{T}_2)}{1 - \mathscr{T}_1}.$$

To set τ, we match the delivery delay (time) of materials in the US manufacturing sector, which is estimated at 60 days.[21] According to Blinder et al. (1998) (page 96), 86 percent of sales by manufacturing firms are to other firms, so most sales are materials (intermediate goods). We set $\phi = 1/2$. This value generates a consistent SVAR decline in labor productivity after 10 quarters. We calibrate δ to match the decline in industrial production implied by our SVAR

[20] In the recent data gross margin is above 50 percent, but earlier in the sample is it slightly below.

[21] These estimates come from Deloitte, https://www2.deloitte.com/us/en/insights/industry/manufacturing/manufacturing-industry-outlook.html. The original source is the Institute of Supply Management.

after 10 quarters.[22] W set the steady state interest rate $\rho = \rho^*$ equal to 10 percent, which is within the range of estimates of the weighted average cost of capital (WACC), a measure broadly used in the valuation of US companies.[23] Finally, we set an arbitrary value of ζ equal to 10 percent. Parameter values are summarized in Table 1.[24]

Parameter	η_0	c_0	δ	ϕ	χ	τ	ζ	ρ
Value	.16	.065	10^{-4}	1/2	1.91	.5	.1	8×10^{-3}

Table 1: Assumed Parameter Values in the Calibrated Model.

4.2 Solution Method

To solve for impulse responses, we solve the differential equation in (31), which after plugging in the parameters becomes

$$\dot{\mathscr{X}}_t - \mathscr{X}_t(0.5 + \rho_t - 0.161\Lambda_t) + (0.32 - 0.031\rho_t)\Lambda_t + 0.4 = 0.$$

As long as $\mathscr{X}$ remains in the inaction zone (and stays positive), access posts depreciate at a constant rate δ, and this equation describes the evolution of $\mathscr{X}$ towards the steady state. We verify that it remains in the inaction zone. Our terminal condition is that at some distant horizon, $\mathscr{X}_t$ returns to the steady state, and we solve the above differential equation using the finite element method with that terminal condition.

The inputs to solve the above differential equation are the assumed paths of Λ_t and ρ_t. To that end, we assume that the path of ρ_t is such that Λ_t^{-1} follows the SVAR-implied impulse response function in Figure 3 (leftmost panel)—which we approximate using a smooth function that closely fits the data.[25] In particular, we back out the interest rate from the Euler equation

[22] The calibrated value is small relative to typical measures of capital depreciation rate. However, note that the amortization cost of access posts can be arbitrarily large if there are maintenance costs. Such costs in our model are part of ζv in (18) and (16).

[23] See the range of estimates at https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/wacc.html.

[24] As we show in the robustness analysis in the Online Appendix, may of these other parameters have a significant impact on the results. What is needed for our results to go through is that ϕ is not too low and δ is not too high. In particular, search costs could be significantly lower, or higher.

[25] We fit a function of the form: $at^c e^{-bt}$, where a, b, c are coefficients to minimize the distance to the empirical

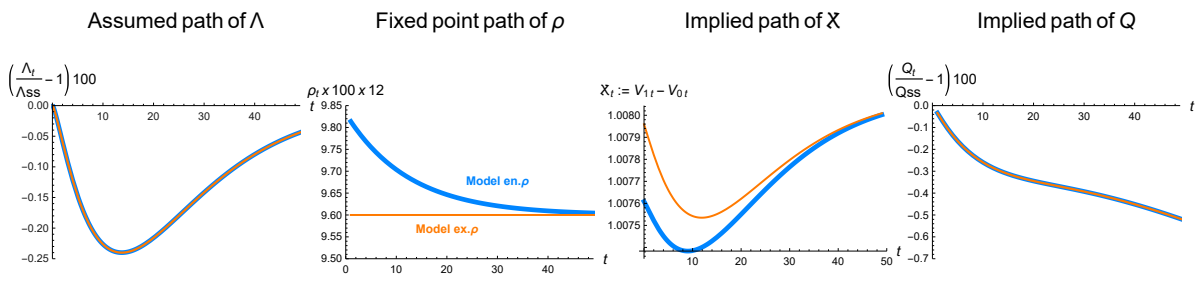

Figure 6: Model: Impulse responses to demand shock (first panel on the left).

in (22), plug in the new path for interest rates, and solve for the fixed point. After solving for this fixed point, we the solve for the paths of N and M from $\dot{M}_t = -\Lambda_t M_t + \tau N_t - \delta M_t$ and $\dot{N}_t = \Lambda_t M_t - \tau N_t - \delta N_t$—since we know $x^* = 0$ in the inaction zone. (When $\mathscr{X}_t$ hits the steady state, the laws of motion change to $\dot{M}_t = -\Lambda_t M_t + \tau N_t - \delta M_t$ and $\dot{N}_t = \Lambda_t M_t - \tau N_t + (x_t^* - \delta)(N_t + M_t)$, for some large x^*, but this occurs beyond the horizon of the plotted impulse responses.)

Figure 6 shows the obtained paths of the key variables. This figure provides a decomposition of the impact of endogenous interest rate—which corresponds to the difference between the blue and the orange line, labeled "Model ex.ρ" and "Model en.ρ," respectively.

4.3 Mapping between Model and the Data

We associate real sales in the data with Q in the model and industrial production with the number of switches from an unstocked to a stocked state, τN. Accordingly, we associate capacity utilization in the data with the ratio of total labor usage in the economy (numerator) to the maximum labor usage had all units been producing output (denominator), which gives:

$$\text{Capacity Utilization} = \frac{\tau N + \zeta M}{\tau (M + N)}.$$

Mapping labor productivity presents some ambiguity because it is not clear how one should handle the entry cost $\phi x^*(M + N)$. If the entry cost involves investment in R&D, that part

impulse response function. The estimated function is $0.232697 - 6.88643 \times 10^{-5} e^{-0.0944353t} t^{1.29557}$. To ensure it returns to steady state in finite time, we assume that after $T = 100$ we add an infinitesimal drift $-10^{-5}t$ to this function. This drift term is set small enough to ensure it has no impact on the resulting path.

should not be included because R&D investment is capitalized in national accounting statistics. We opt for the worst-case scenario of treating the entry cost as an expensed cost and define labor productivity as:[26]

$$\text{Labor Productivity} = \frac{\text{Industrial Production}}{\zeta M + \tau N + \phi x^*(M+N)}.$$

We map firm-level measured markups onto the wholesale markup $\log(\tilde{p}^*/v) \approx (\tilde{p}^* - v)/v$. Finally, we map P onto the CPI price level. The idea here is that retail prices include all markups and all distribution costs, which is represented by P in the model.[27]

4.4 Quantitative results

Figure 7 shows our quantitative results. For comparison, we report both a constant and an endogenous path for ρ, labeled "Model ex.ρ" and "Model en.ρ," respectively. As we have shown analytically in the previous section, the model generates procyclical markups and countercyclical inventory-to-sales ratio. The other impulse responses are also in line with the data. In particular, gross output-to-sales ratio moves little, capacity utilization and labor productivity both declines and consumer prices move little. Labor productivity initially rises because firms no longer bear the cost of replacing depreciating access posts—as we discussed in the previous section—but it declines afterward.

The interest rate path corresponds to the path of ρ converted from a monthly to an annual rate using the expectations theory to be comparable to the one-year T-bill rate. The model matches the path from SVAR well until monetary policy changes course, which the model cannot match the given assumed path for the inventory-to-sales ratio that we match by construction.

[26] This is the worst-case scenario because our calibration targets the decline in labor productivity. This assumption reduces the size of the inaction zone. The presence of a sufficiently wide inaction zone is central to our mechanism.

[27] This mapping of P onto retail prices presents some ambiguity because distributors in our model are an abstract concept that in part capture market activities of retailers and in part stand in for distribution activities done by households. In the latter case, the CPI would comprise some retail prices and some wholesale prices, depending on how much distribution is done within households. This source of mismeasurement is fairly small given that 86 percent of transactions are B2B transactions in inventory-holding industries, and about 70 percent in the economy as a whole (Blinder et al., 1998).

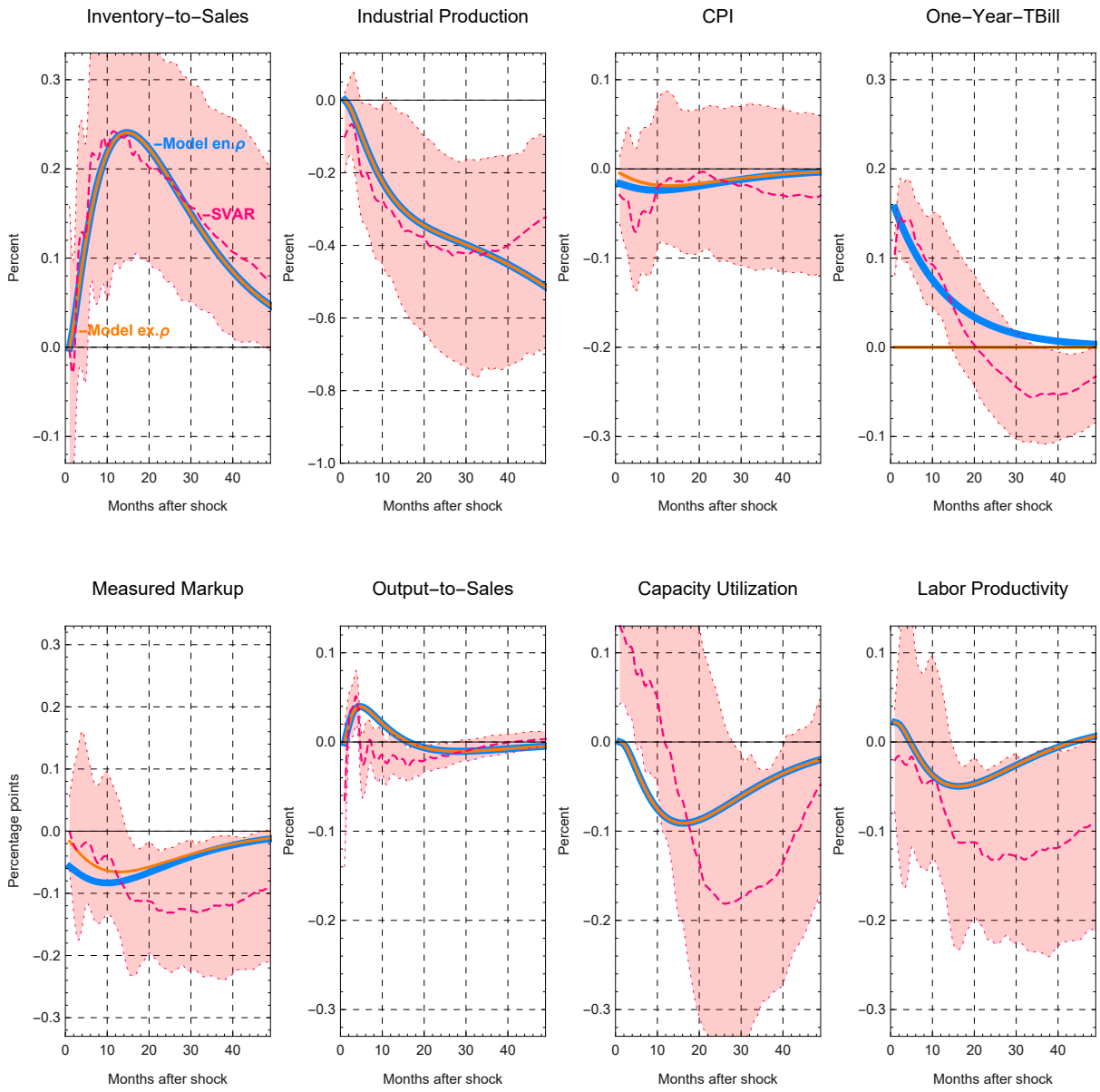

Figure 7: Calibrated Model versus SVAR

In particular, our model predicts that the interest rate must be above the trend until the inventory-to-sales ratio returns to the steady state and that the recovery starts only after the inventory-to-sales ratio returns to the steady state. Hence, the model overpredicts the persistence in other variables vis-à-vis the persistence of the inventory-to-sale ratio. It should be noted, however, a faster decline in the inventory-to-sales ratio in the data is still consistent with the depicted confidence bands, and the empirical results are not conclusive as to whether this represents our model's failure.

Lastly, while in our baseline model real wage (v/P) rises, it does not have to be the case. In particular, suppose that $v = wL^{-(1-\alpha)}$, as would be the case in the presence of capital as a fixed factor. In that case, as discussed in the Online Appendix, after MP contraction, even if w declines, v can be constant as in the baseline model because labor supply L declines and hence $L^{-(1-\alpha)}$ rises. We conclude that, qualitatively, given scant change in the price level P, our model is broadly consistent with real wages being acyclical—as, arguably, is the case in the data (Christiano et al., 1997).

4.5 Steep marginal cost: A Discussion

We return to the issue of steep marginal costs and demonstrate this explicitly using our model. As argued in Section 2.4, mismeasurement of variable costs implies that marginal cost curves are fairly steep. However, if this is the case, the point that inventories imply countercyclical markups becomes moot. Violations of identification assumptions 1 and 2 produce a similar effect, and the same argument extends to these assumptions (Nekarda and Ramey, 2013).

Consider the frictionless calibrated model with $\phi = 0$ and assume that the marginal cost curve on the aggregate level is given by $v(L_t) = v_0 + 2\left(\frac{L_t - L_{ss}}{L_{ss}}\right)$, where L is aggregate labor. As in the baseline setup, the producer takes the marginal cost as given, and so the assumed curvature is an aggregate phenomenon. What is important is that the producer understands how the marginal cost evolves with employment after the shock hits. The value of the coefficient ("2" in front of the linear term) is the worst-case scenario for mismeasurement, as discussed in Section 2.4.

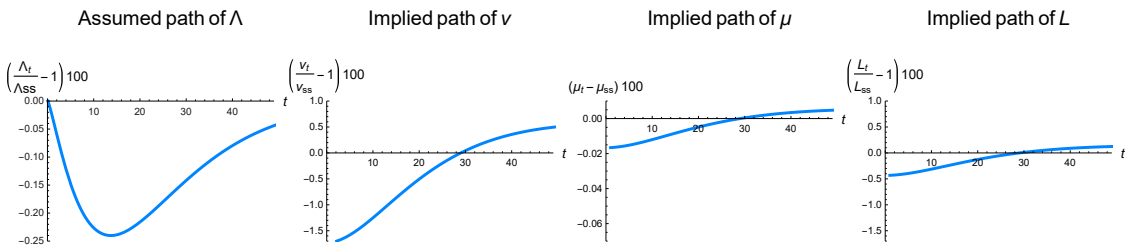

Figure 8: Frictionless Model ($\phi = 0$) with Increasing Marginal Cost Schedule.

To make markups move over the cycle, assume $\tilde{\chi}(L_t) = \tilde{\chi} + .2\left(\frac{L_t - L_{ss}}{L_{ss}}\right)$. This crude assumption suffices to examine the correlation when markups do fall. One interpretation of this specification is that the costs of distribution simply fall as employment falls in the economy, and so does the match surplus y^*. The source of markup movements is not important to establish their consistency with the dynamics of the inventory-to-sales ratio.

Since $\mathscr{X}_t \equiv v(L_t)$ and $\dot{\mathscr{X}} = 2\dot{L}_t$, the HJB equation in (31) boils down to a differential equation for employment L_t of the form

$$0.73 + L_t\left(\Lambda_t\left(7.95 + L_t\right) - 2.03\right) + 13.57\dot{L}_t - 3\Lambda_t = 0.$$

To solve for the impulse response of markups, we plug in the fitted function to the data for Λ_t (as stated in footnote 25), and assume an exogenous path for ρ—which, as we have shown, has a scant impact on the results. The terminal condition is that at some distant horizon after the shock dies out L_t returns to the steady state. Note that this model qualitatively succeeds if L_t falls after fitting Λ_t, in which case we can say that the inventory-to-sales ratio is countercyclical. As Figure 8 shows, that the model succeeds: markups fall and the inventory-to-sales ratio rises after the shock (the inverse of Λ).

5 Conclusions

We develop a new theory of inventory dynamics that integrates other productive assets alongside inventories, reconciling the procyclical response of markups to monetary policy shocks with the

countercyclical response of the inventory-to-sales ratio. Emphasizing the putty-clay nature of many such assets, our model explains the observed data patterns effectively. In particular, our empirical analysis of the effects of monetary policy shows that the model aligns well with the observed impulse responses. Furthermore, we argue that any mismeasurement of markups due to deviations from our identification assumptions would suggest steep marginal cost curves, thereby rendering the point that inventories imply countercyclical markups moot.

References

Anderson, S. P. and R. Renault (1999): "Pricing, product diversity, and search costs: A Bertrand-Chamberlin-Diamond model," *The RAND Journal of Economics*, 719–735. 1, 14

Arias, J. E., J. F. Rubio-Ramírez, and D. Waggoner (2021): "Inference in Bayesian Proxy-SVARs," *Journal of Econometrics*, 225, 88–106. 9, 11

Bils, M. (1987): "The Cyclical Behavior of Marginal Cost and Price," *The American Economic Review*, 77, 838–855. 1, 2.1.1

Bils, M. and J. A. Kahn (2000): "What Inventory Behavior Tells Us about Business Cycles," *American Economic Review*, 90, 458–481. 1, 2.2, 3.5.3

Blanchard, O. J. and B. S. Bernanke (2023): "What Caused the US Pandemic-Era Inflation?" Working Paper 31417, National Bureau of Economic Research. 1

Blinder, A., E. Canetti, D. Lebow, and J. Rudd (1998): *Asking About Prices*, New York: Russell Sage Found. 1, 5, 4.1, 27

Blinder, A. S. and L. J. Maccini (1991): "Taking Stock: A Critical Assessment of Recent Research on Inventories," *Journal of Economic Perspectives*, 5, 73–96. 1, 2.2

Broer, T., N.-J. Harbo Hansen, P. Krusell, and E. Öberg (2019): "The New Key-

nesian Transmission Mechanism: A Heterogeneous-Agent Perspective," *The Review of Economic Studies*, 87, 77–101. 1

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. 4.1

Cheremukhin, A. and P. Restrepo-Echavarria (2020): "Wage Setting Under Targeted Search," *St. Louis Fed Working Paper 2020-041D*. 1, 14

Chow, G. and A. Lin (1971): "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series. Review of Economics and Statistics," . 11

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1997): "Sticky Price and Limited Participation Models of Money: A Comparison," *European Economic Review*, 41. 1, 4.4

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. 1, 2, 2.1.2, 4.1

Diamond, P. A. (1971): "A model of price adjustment," *Journal of economic theory*, 3, 156–168. 1

Drozd, L. A. and J. B. Nosal (2012): "Understanding International Prices: Customers as Capital," *American Economic Review*, 102, 364–95. 1

Fitzgerald, T. (1997): "Inventories and the business cycle: an overview," *Economic Review*, 11–22. 1, 2.2

Gertler, M. and P. Karadi (2015): "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics*, 7, 44–76. 1, 2.1.2, 9, 2.2, (document)

Gilchrist, S. and E. Zakrajsek (2012): "Credit Spreads and Business Cycle Fluctuations," *American Economic Review*, 102, 1692–1720. 2.1.2

Hall, R. E. (1988): "The Relation between Price and Marginal Cost in US Industry," *Journal of Political Economy*, 96, 921–947. 1

Khan, A. and J. K. Thomas (2007): "Inventories and the Business Cycle: An Equilibrium Analysis of (S, s) Policies," *American Economic Review*, 97, 1165–1188. 2

Kryvtsov, O. and V. Midrigan (2012): "Inventories, Markups, and Real Rigidities in Menu Cost Models," *The Review of Economic Studies*, 80, 249–276. 1

Nekarda, C. J. and V. A. Ramey (2013): "The Cyclical Behavior of the Price-Cost Markup," Working Paper 19099, National Bureau of Economic Research. 1, 2.1.1, 2.3, 4.5

Ortiz, J. (2022): "Spread Too Thin: The Impact of Lean Inventories," *Washington: Board of Governors of the Federal Reserve System*. 1

Quilis, E. M. (2018): "Temporal disaggregation of economic time series: The view from the trenches," *Statistica Neerlandica*, 72, 447–470. 11

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. 1, 2.1.1

Stahl, D. O. (1989): "Oligopolistic Pricing with Sequential Consumer Search," *The American Economic Review*, 79, 700–712. 16

Storesletten, K., J.-V. R. Rull, and Y. Bai (2011): "Demand Shocks that Look Like Productivity Shocks," Tech. rep. 1

Weber, I. M. and E. Wasner (2023): "Sellers' Inflation, Profits and Conflict: Why can Large Firms Hike Prices in an Emergency?" *Economics Department Working Paper Series*. 1

Wolinsky, A. (1986): "True Monopolistic Competition as a Result of Imperfect Information," *The Quarterly Journal of Economics*, 101, 493–511. 1, 14

Appendix

This Appendix contains omitted proofs, supplementary derivations, and the list of data sources. An additional Online Appendix and *Mathematica* notebooks are available online. They contain detailed derivations, extensions, and additional robustness exercises referenced in text.

Derivation of $d(\pi)$ and $s(\pi)$ for Section 3.1.2: By definition of precision π and random search, an observation meets the criterion after exactly a single search with probability $1-\pi$, after two searches with probability $(1-\pi)\pi$, three searches with probability $(1-\pi)\pi^2$ and so on and so forth. This process implies a geometric distribution but the presence of discounting requires some care to derive the mean. Let the discount factor be $\beta \equiv e^{-\rho d\tau}$ and note that, as with standard geometric distribution, the expected cost until termination is

$$Sum := c_0(1-\pi)\beta + \pi c_0\beta + \pi\beta\left(\underbrace{c_0(1-\pi)\beta + \pi c_0\beta + \pi\beta(c_0(1-\pi)\beta + ...)}_{=Sum}\right),$$

which boils down to $c_0(1-\pi)\beta + \pi\beta c_0 + \pi\beta Sum = Sum$, and hence gives $Sum = c_0\beta(1-\pi\beta)^{-1} \to_{\beta\uparrow 1} c_0(1-\pi)^{-1}$. Applying an analogous reasoning to the expected surplus from a match—while noting that dt is small so that the expected change of static surplus is of second order in dt (see discussion at the end)—analogously yields the summation

$$Sum := s(\pi)(1-\pi)\beta + \pi 0\beta + \pi\beta\left(\underbrace{s(\pi)(1-\pi)\beta + \pi 0\beta + \pi\beta(s(\pi)(1-\pi)\beta + ...)}_{=Sum}\right),$$

and hence $\beta(1-\pi)s(\pi) + \pi\beta Sum = Sum$, which gives $Sum = s(\pi)(1-\pi)(1-\pi\beta)^{-1} \to_{\beta\uparrow 1} s(\pi)$ stated in text. If the static surplus contains a diffusion process, the expected value is still the static value as $dt \to 0$. The Wiener shock itself, while its variance is of order $\sqrt{dt}$, does not affect the expected value because of symmetry. The drift term is of order dt, and hence its effect vanishes in the limit. (See footnote in text for an explicit argument.)

Proof of Lemma 1: Note that in steady state equilibrium $\lambda(\tilde{p}^*, \tilde{p}^*) = \Lambda > 0$. By (20) in steady state we must have

$$\mathscr{X} := \frac{\Lambda(\tilde{p}^* - \kappa v) + \tau(1-\kappa)v}{\Lambda + \rho + \tau + \delta} = (1-\kappa)v + \phi v.$$

This defines the steady state value of Λ. Consider now the first order condition for $\tilde{p}^*$ in (21)

$$(*) : \tilde{p}^* - (1-\kappa)v - \phi v = \kappa v + \eta_0 \frac{P}{\tilde{p}^*}.$$

Observe that on the left-hand side we have a linear function that is strictly increasing in $\tilde{p}^*$ and has a negative vertical intercept, and that on the right-hand side ($\kappa v + \eta_0 \frac{P}{\tilde{p}^*}$) we have a downward sloping hyperbola for all $\tilde{p}^* \geq 0$. The range of the hyperbola is from ∞ at $\tilde{p}^* \to 0^+$ to $\kappa v > 0$ at $\tilde{p}^* \to \infty$. Therefore, (*) has a unique solution for any fixed value of $P > 0$. Denote that solution as $\tilde{p}^*(P)$ and note that it is parameterized by η_0. Note also that $\tilde{p}^*(P)$ is a strictly increasing and strictly concave function of P, which we establish in the following supplementary lemma (proof is at the end):

Lemma 1.1: $\tilde{p}^*(P)$ is a strictly increasing and strictly concave function.

Next, we take into account that steady state P must satisfy (9), and hence $P = P(\tilde{p}^*)$ is such that

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

The fixed point $P, \tilde{p}^*$ is defined by the system $P = P(\tilde{p}^*), \tilde{p}^* = \tilde{p}^*(P)$, and it must be such that $\infty > P > \tilde{p}^* > v$. We will construct a sequence $\{P_k\}$ which, if converges, converges to that fixed point. To establish the convergence of that sequence, we will show that the sequence is monotone increasing and bounded from above by another convergent sequence; that is, given $\{P_k\}$ we construct $\{\bar{P}_k\}$ such that $P_k < \bar{P}_k$ for all $k = 0, 1, 2, 3$, and show $\bar{P} := \lim_{k\to\infty} \bar{P}_k > 0$. We start from the definition of the sequence of interest: $\{P_k\}$.

Define the sequence P_0, P_1, P_2.... such that, given $P_0 = v$, i) $\tilde{p}_0^*$ solves (*) given P_0 (which we established is unique), ii) $P_1 = \tilde{p}_0^* + \chi v - \eta_0 P_0 \log\left(\frac{c_0}{\eta_0} \frac{v}{P_0}\right)$, and so on with P_1 replacing P_0 in i to define $\tilde{p}_1^*$ and $\tilde{p}_1^*$ replacing $\tilde{p}_0^*$ in ii to define P_2 etc... Note that this sequence is a monotone

increasing sequence if we pick a sufficiently small η_0. To see this, note that we can ensure that $\eta_0 P \log\left(\frac{c_0}{\eta_0}\frac{v}{P}\right)$ is a positive and arbitrarily small number for any $P \in [P_0, \bar{P}]$, since it can be tightly bounded from above by choosing η_0 small enough. This follows from: i) the properties of function $f(\eta_0) := \eta_0 \log\left(a\eta_0^{-1}\right)$ for any $a > 0$ (note: $\lim_{\eta_0 \downarrow 0} f(\eta_0) = 0$ and $f(\eta_0)$ strictly decreasing and positive-valued for sufficiently small $\eta_0 > 0$), and ii) the fact that we can then pick η_0 small enough to ensure $\log\frac{c_0}{\eta_0}\frac{v}{P} > 0$ for any $P \in [P_0, \bar{P}]$ and by i bound it as follows $0 < \eta_0 P \log\left(\frac{c_0}{\eta_0}\frac{v}{P}\right) < \eta_0 \bar{P} \log\left(\frac{c_0}{\eta_0}\frac{v}{v}\right) < \chi v$. Accordingly, $P_1 > P_0$, and by Lemma *1.1,* we know that $\tilde{p}_1^* > \tilde{p}_0^*$. Hence, as long as $P_k \leq \bar{P}$, this is a monotone increasing sequence. We next construct the bounding sequence $\{\bar{P}_k\}$ and show it converges.

Consider the sequence $\bar{P}_0, \bar{P}_1, \bar{P}_2$.... such that $\bar{P}_0 = v + \eta_0 v$ and i) $\tilde{p}_0^*$ solves given $\bar{P}_0$ (*), ii) $\bar{P}_1 = \tilde{p}_0^* + \chi v$, and so on and so forth with $\bar{P}_1$ replacing $\bar{P}_0$ in i to define $\tilde{p}_1^*$ and $\tilde{p}_1^*$ replacing $\tilde{p}_0^*$ in ii to define $\bar{P}_2$ etc... Note that $\bar{P}_1 > P_1$, since

$$P_1 := \tilde{p}_0^* + \chi v - \eta_0 P_0 \log\left(\frac{c_0}{\eta_0}\frac{v}{P_0}\right) < \tilde{p}_0^* + \chi v := \bar{P}_1,$$

given our choice of η_0. It is clear that this argument extends to $k = 2, 3...$ because an analogous inequality applies to the subsequent terms by noting that the corresponding $\tilde{p}_k^{*\prime}s$ of the bounding sequence $\left\{\bar{P}_k\right\}$ must be larger under the original sequence $\{P_k\}$ by Lemma *1.1.* This bounding sequence is also monotone increasing and bounded, and hence it converges. To see this, consider (*) again and plug in for $\bar{P}_k$, and note that the following evaluation applies by the fact that the underlying $\tilde{p}_k^{*\prime}s$ of this sequence form an increasing sequence such that:

$$\tilde{p}_k^* - \mathscr{X}(\tilde{p}_k^*) = \kappa v + \eta_0 \frac{\left(\tilde{p}_{k-1}^* + \chi v\right)}{\tilde{p}_k^*} < \kappa v + \eta_0 + \frac{\chi v}{\tilde{p}_k^*}.$$

If $\bar{P}_k \to_{k\uparrow\infty} \infty$, it is clear that the underlying $\tilde{p}_k^* \to \infty$. But this contradicts the above inequality because the left-hand side is strictly increasing in $\tilde{p}_k^*$ and the right hand side is bounded. It is straightforward albeit tedious to see that the rest of the equilibrium definition can be satisfied by a set of constant values consistent with the above. We omit the details. Q.E.D.

Proof of Lemma 1.1: Let $\tilde{p}^* - (1-\kappa)v - \phi v = a\tilde{p}^* - b$, where $0 < a \leq 1$, $b > 0$, as established in the main proof. Multiply both sides of (*) in main proof by $\tilde{p}^*$ and observe that the solution

of (*) is a positive solution of (**): $a\left(\tilde{p}^{*}\right)^{2}=(\kappa v+b)\tilde{p}^{*}+\eta_{0}P$. The left-hand side is a canonical quadratic function of $\tilde{p}^{*}$ and the the right-hand side is a linear function with the intercept that is linear in P. Accordingly, the positive solution $\tilde{p}^{*}$ of (*), which also solves (**), implies smaller and smaller increments as P is changed by an equal step ΔP, with the $sign\left(\Delta\tilde{p}^{*}\right)$ of the resulting increments being the same as $sign\left(\Delta P\right)$. This is the definition of concavity of an increasing function.

Steady state for Section 3.5 and Section 4.2: Steady state of the model is: $\mathscr{X}^{ss}=v+v\tau^{-1}\phi\left(\delta+\rho\right)$, $p^{ss}=v\frac{\mathscr{X}^{ss}/v+\eta_{0}\tilde{\chi}}{1-\eta_{0}}$, $P^{ss}=v\frac{\mathscr{X}^{ss}/v+\tilde{\chi}}{1+\eta_{0}}$, $\Lambda^{ss}=\frac{1-\eta_{0}}{\eta_{0}}\frac{(\delta+\zeta+\rho)+(\delta+\rho+\tau)\frac{\mathscr{X}^{ss}-v}{v}}{\mathscr{X}^{ss}/v+\tilde{\chi}}=\frac{1-\eta_{0}}{\eta_{0}}\frac{\zeta-\tau+(\delta+\rho+\tau)\mathscr{X}^{ss}/v}{\mathscr{X}^{ss}/v+\tilde{\chi}}$. The numerical values for the calibrated parameters are $v=1,p=1.5$, $P=3$. $\pi=0.865784$, $\Lambda=0.240453$, $X=1.0162$. For more details see the online *Mathematica* notebook.

Data sources: CPI (1979-2012), Industrial production (1979-2012), EBP (1979-2012), One-Year-Tbill rate (1979-2012), monetary policy shock instrument: Gertler and Karadi (2015). Data on markups, sales and cogs comes from S&P Compustat Quarterly Fundamentals, as detailed in text. Additional aggregate series include: capacity utilization (1979-2012) from the Board of Governors of the Federal Reserve System (BOG); real sales and inventories in manufacturing and trade industries from Bureau of Economic Analysis (BEA)—which are used to calculate the IS ratio and the gross output (as sales + change in inventories) in the SVAR; and labor productivity (1979-2012) in the business cycle from Bureau of Economic Analysis. Most of these series have been retrieved from FRED, Federal Reserve Bank of St. Louis between January 1st, 2024 and March 30th of 2024.