Market Liquidity and Inventory Cycles

Ziqi Qiao *

January 8, 2026

Link to the latest version

Abstract

Inventories play a central role in business cycles, yet standard models struggle to jointly account for three key facts: inventory investment is procyclical, the inventory-sales ratio is countercyclical, and markups are procyclical. I address this puzzle with a directed search model where *ex ante* identical firms invest in inventories, post prices, and face stochastic sales. Beyond smoothing production and avoiding stockouts, firms may also hold extra inventory to enhance their pricing power, creating a crucial link between profitability and the speed of sales. This channel successfully reconciles the observed behavior of inventories and markups. In the calibrated model, firms collectively overstock by 1.6%, which amounts to approximately 0.2% of GDP. Meanwhile, large firms overstock while small firms understock, generating a welfare loss of 0.27% of GDP relative to the constrained optimum. In addition, this firm heterogeneity is central to cyclical dynamics.

*University of Wisconsin – Madison. Email: zqiao7@wisc.edu. I thank Randall Wright, Guanming Shi, Rasmus Lentz, Kenneth West, Hengjie Ai, Rishabh Kirpalani, José-Víctor Ríos-Rull, Russell Cooper and Chao He for their helpful comments. All remaining errors are my own.

1 Introduction

In U.S. data, inventories account for roughly 13% of GDP, making them a substantial component of resource allocation. But is this allocation efficient? To answer this question, we need first understand the role inventories play in the economy. A natural way to test our understanding is to compare theoretical predictions with data. Fortunately, inventories do exhibit several robust empirical regularities that provide a benchmark for evaluation.[1] Yet standard models cannot *jointly* account for these findings: matching some regularities often comes at the expense of missing others.

Inventories are highly volatile, moving closely with output and the business cycle. Three empirical regularities define this relationship: inventory investment is pro-cyclical, the inventory-sales (I/S) ratio is counter-cyclical, and aggregate markups are pro-cyclical. These patterns hold across multiple measures of cycles, providing a natural testbed for theory and highlighting the central role of inventories in macroeconomic fluctuations. As Alan Blinder famously puts it, *"Business cycles are, to a surprisingly large degree, inventory cycles."* Despite this, inventories remain underrepresented in modern macroeconomic models. This paper seeks to place them back at the center of macroeconomic analysis.

Three main explanations in the literature account for why firms hold inventories: production smoothing, stockout avoidance, and fixed restocking costs. While each provides useful insights, none fully explains the observed cyclical patterns. The production smoothing view emphasizes that firms hold inventories to buffer against volatile demand or markups, but this mechanism predicts either that production is less volatile than sales or that markups are counter-cyclical – predictions inconsistent with the data (Blinder, 1986; West, 1990; Bils & Kahn, 2000; Kryvtsov & Midrigan, 2012; Nekarda & Ramey, 2013). The stockout avoidance view stresses that firms maintain inventories to meet unpredictable demand, yet it implies inventories should fall when sales rise, contradicting the pro-cyclicality of inventory investment

[1] The importance of inventories in business cycles is well established. See Metzler (1941), Blinder & Maccini (1991), and Ramey & West (1999).

(Kahn, 1987; Coen-Pirani, 2004). The fixed restocking cost view highlights (s, S) policies, under which firms allow inventories to deplete gradually before replenishing them once they hit a threshold, but this mechanism cannot account for pro-cyclical inventory or the persistence of the inventory-sales ratio (Arrow *et al.*, 1951; Haltiwanger & Maccini, 1988; Khan & Thomas, 2007). To reconcile these findings, Luo *et al.* (2025) add ambiguity aversion to existing explanations with success, though they do not tackle markup patterns.

Beyond these explicit mechanisms, some macroeconomic models treat inventories in reduced form. Kydland & Prescott (1982) and Christiano (1988) model inventories as a factor of production, while Wen (2011) and Auernheimer & Trupkin (2014) embed them in utility. These approaches introduce inventories in tractable ways but fail to replicate the observed cyclical properties and the persistence of inventory dynamics. This disconnect motivates a framework that incorporates inventories in a more structural and empirically consistent manner.

This paper proposes a different microfoundation for inventory holdings by incorporating endogenous search frictions (Burdett *et al.*, 2001) into a heterogeneous-agent dynamic general equilibrium framework (Aiyagari, 1994). In the model, firms invest in inventories and post prices, while consumers decide which firms to purchase from. Because consumers cannot coordinate on where to shop, they adopt mixed strategies in equilibrium and randomly visit different firms, which generates endogenous search frictions. As a result, firms face stochastic sales: at times, excess demand leads to stockouts, while at other times insufficient demand leaves unsold inventories. This stochasticity is endogenous, as firms internalize that their inventory and pricing decisions influence the probability of attracting consumers. Firms therefore make joint decisions on quantities and prices – something not possible in a Walrasian framework – which proves crucial for understanding inventory cycles.

Firms make portfolio decisions between capital and inventories. During booms, more aggregate capital reduces interest rates, so firms allocate more resources to inventories, gen-

erating pro-cyclical inventory investment. At the same time, both larger inventory holdings and lower returns to capital incentivize firms to charge higher prices. As a result, firms set higher prices, generating pro-cyclical markups, while inventory adjustment slows, producing a counter-cyclical I/S ratio. The key mechanism is that firms can adjust prices as well as quantities, so their quantity response is less pronounced than if quantity were the only margin. This dual-margin adjustment within a directed search framework provides a unified explanation for the joint behavior of inventories, the I/S ratio, and markups, capturing both their cyclicality and persistence.

The model replicates salient features of firm behavior and market dynamics: firms choose both quantities and prices; sales are stochastic; markets do not clear; and fire sales – pricing below cost – can arise rationally. Beyond explaining aggregate patterns, the model's predictions also align with evidence from recent micro studies. For instance, Kim (2021) show that capital insufficiency can trigger price cuts and inventory liquidation; Kryvtsov & Vincent (2021) find that the frequency of price cuts is counter-cyclical; and Cavallo & Kryvtsov (2023) demonstrate that inventory stockouts exert inflationary effects.

By successfully explaining the empirical findings, the model provides a nuanced understanding of inventory behavior and enables a more confident assessment of inventory efficiency. The calibrated model suggests that, relative to the planner's solution under the constraint that consumers do not coordinate on which firms to buy from, the economy is overstocked by 1.6% – equivalent to 0.2% of annual GDP. The associated welfare loss is larger due to distributional effects: in equilibrium, large firms tend to overstock while small firms understock. Accounting for this dispersion, the welfare loss from inefficient inventory allocation amounts to 0.27% of GDP. This result points to the potential role of progressive taxation in restoring efficiency.

The paper also contributes to the broader search literature.[2] Classical search models

[2]For surveys, see Rogerson *et al.* (2005) on labor search, Lagos *et al.* (2017) on money search, and Wright *et al.* (2021) on directed and competitive search.

typically feature one-to-one matching,[3] limiting their scope for analyzing inventories. Since this paper focuses on inventory behavior, I model many-to-one matching. Specifically, I extend the static game of Burdett *et al.* (2001) to allow for an arbitrary inventory distribution and dynamic incentives. Watanabe (2010) studies a similar game to analyze the emergence of middlemen, but restricts attention to a two-point inventory distribution in a static setting. Geromichalos (2012, 2014) also model many-to-one matching with directed search, but considering contingent contracts that do not generate inventories. My model, by contrast, incorporates pre-committed inventory stocks, which induce greater matching frictions: some firms are left with unsold goods while others run out of stock. Another related paper is Shevchenko (2004), which studies inventories under random search in continuous time and shows that the holdup problem leads to underinvestment in capacity. In my model, the posting mechanism and firm heterogeneity mitigate the holdup problem, leading instead to aggregate overstocking. Li *et al.* (2024) study inventory management of intermediaries under directed search and apply the model to used-car dealers. They focus on intermediation and assume free entry on both the production and consumption sides, while this paper models dynamic production and consumption decisions to address macroeconomic issues.

Because many-to-one matching complicates analytical solutions, I solve the model numerically, adapting the methods from heterogeneous-agent models (Huggett, 1993; Aiyagari, 1994; Krusell & Smith, 1998), but with a key difference. Standard heterogeneous-agent models typically assume some *ad hoc* Markov process for idiosyncratic shocks, such as shocks to labor productivity. In my model, by contrast, the idiosyncratic sales shocks arise endogenously from the best responses and equilibrium conditions: inventory holdings determine the support, and prices determine the transition probabilities. Modeling endogenous shocks yields new insights into capital distribution. Firms with more capital exhibit greater risk tolerance, enabling them to post riskier terms of trade and pursue higher profits. This rich-

[3]Examples include Diamond (1982) for goods, Mortensen & Pissarides (1994) for labor, Kiyotaki & Wright (1989) for money, Cavalcanti & Wallace (1999) for banking, Duffie *et al.* (2005) for OTC markets, and Burdett & Coles (1997) for marriage.

get-richer mechanism can generate a heavy right tail in the capital distribution, offering broader insights into wealth inequality.

The rest of the paper proceeds as follows. Section 2 documents the empirical behavior of inventories and markups. Section 3 presents the model and its mechanisms. Section 4 calibrates the model and evaluates its empirical performance. Section 5 concludes.

2 Empirical regularities

One might expect inventory stocks to decline over time as technologies advance and firms adopt just-in-time inventory management. However, the aggregate data suggest otherwise. Figure 1 plots the aggregate inventory-to-GDP ratio using U.S. quarterly data from the Bureau of Economic Analysis. On average, inventories represent about 13% of annual GDP. The ratio rises during the 1970s, declines between 1980 and 2010, and then returns to its average level. At its lowest point, the ratio is still above 11%. Thus, there is no clear downward trend, and inventories remain a significant component of resource allocation. To put this in perspective, consuming half of the inventory stock would raise aggregate consumption by 10%.[4]

Beyond its substantial size, inventory is also closely linked to business cycles. As shown in Figure 1, the inventory-to-GDP ratio declines sharply during the 2008 financial crisis and the COVID-19 pandemic, reflecting urgent liquidation, reluctance to restock, or both. More broadly, inventory responds not only in crises but throughout the cycle. Its close comovement with the business cycle becomes evident when comparing the growth rates of inventory and GDP. Figure 2 plots these growth rates as log differences, showing that inventory and GDP move together not only in direction but also in magnitude. This strong comovement underscores the central role of inventory in business cycles, particularly during

[4]The data do not distinguish between intermediate and final goods. It is suggestive that final goods account for about 50% of total inventories.

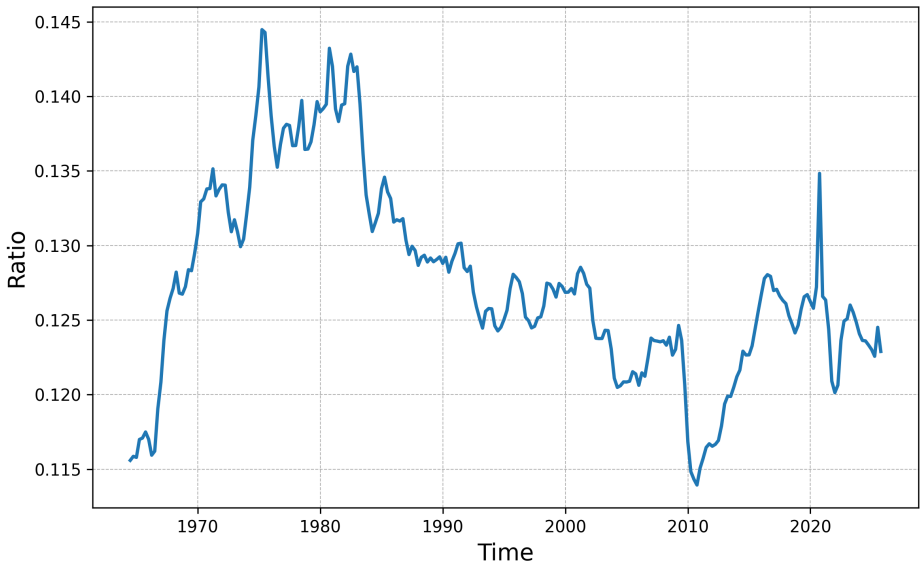

Figure 1: Inventory-GDP ratio, U.S. quarterly 1964Q1–2025Q1.

Inventory is the real non-farm private inventory. GDP is the real GDP. Data source: U.S. Bureau of Economic Analysis.

downturns.

Table 1 reports descriptive statistics on the volatility and correlations among inventory, sales, and markups. The data come from the U.S. Bureau of Economic Analysis and the U.S. Bureau of Labor Statistics, spanning 1964-2025 at a quarterly frequency. Total output y is measured by real GDP, inventory i by real private nonfarm inventories, and sales s by real final sales of domestic products. Because marginal cost is unobservable, I construct an aggregate markup m as the ratio of the Producer Price Index (PPI) to wages. Specifically, PPI refers to the Producer Price Index by Commodity for Final Demand (Finished Goods), weighted by the implicit GDP deflator to control for inflation. Wages are measured as Average Hourly Earnings of Production and Nonsupervisory Employees (Total Private).

The baseline measure of cycles is the log difference. For comparison with the literature, I also report statistics using the HP-filtered cyclical components. The two approaches yield results that are qualitatively similar and quantitatively close. Because the HP filter can generate spurious dynamic relations (Hamilton, 2018), I focus on log differences, although

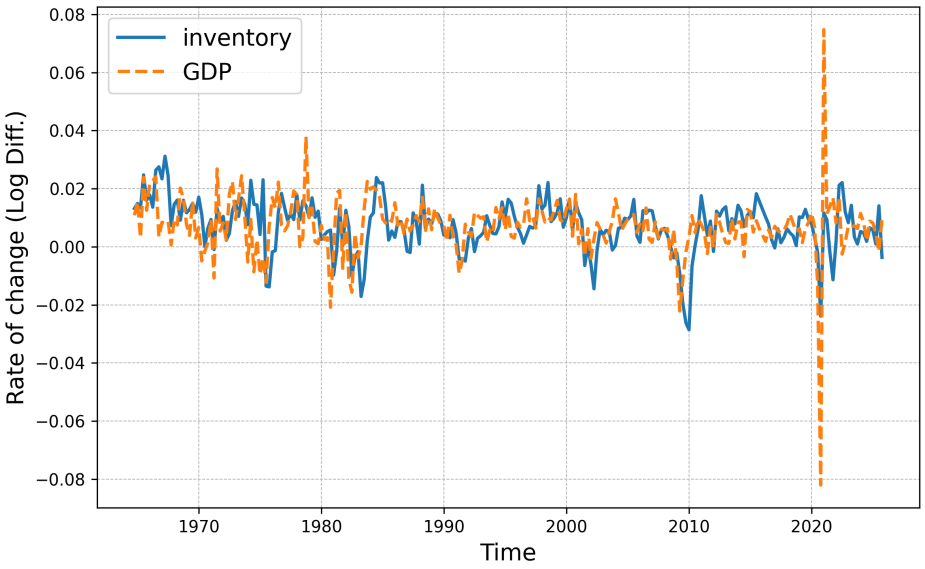

Figure 2: Inventory-GDP comovement, U.S. quarterly 1964Q1–2025Q1

Both variables are in log difference. Inventory is the real non-farm private inventory. GDP is the real GDP. Data source: U.S. Bureau of Economic Analysis.

the patterns are robust across cycle measures.

The top panel of Table 1 compares the volatility of output, sales, and inventory. Based on log differences, inventory is 84% as volatile as GDP, the same as sales. This finding rules out production smoothing as the sole explanation for inventory holding. If inventories were held only to buffer production against volatile sales, output would be less volatile than sales. Yet the data show the opposite.

The bottom panel of Table 1 reports various correlations. The correlation between inventory and output is 0.41, indicating that inventory is procyclical. This contradicts stockout avoidance as the only motive for holding inventory, since that mechanism predicts inventories should fall during expansions. By contrast, the inventory-to-sales (I/S) ratio has a negative correlation of -0.12 with GDP growth, implying that inventories rise with output but less than proportionately with sales. Moreover, the I/S ratio exhibits an autocorrelation of about 0.97, reflecting strong persistence. Such persistence is difficult to reconcile with fixed restocking cost models, which predict restocking only once inventories fall below a threshold

Table 1: Descriptive statistics, U.S. quarterly data 1964Q1-2025Q1

		Cycle measures	
Variable	Description	Log diff.	HP-filtered
	Volatility		
$\sigma(y)$	GDP	0.01	0.02
$\sigma(s)/\sigma(y)$	Sales over GDP	0.84	0.85
$\sigma(i)/\sigma(y)$	Inventory over GDP	0.84	1.18
	Correlation		
$\rho(s, y)$	Sales and GDP	0.88	0.96
$\rho(i, y)$	Inventory and GDP	0.41	0.55
$\rho(is, y)$	I/S and GDP	-0.12	-0.16
$\rho(is_t, is_{t-1})$	I/S autocorrelation	0.97	0.75
$\rho(m, y)$	Markup and GDP	0.12	0.31

Markup is defined as PPI/wage. The HP filter multiplier is 1600. Data sources: the U.S. Bureau of Economic Analysis and the U.S. Bureau of Labor Statistics.

and subsequent gradual depletion, implying much weaker persistence. Finally, markup is mildly procyclical, with a correlation of 0.12 with output. This again challenges production smoothing and stockout-avoidance motives, both of which imply or require countercyclical markups.

To further examine cyclical dynamics, I estimate a structural VAR model with three variables: sales, inventory, and aggregate markup. All variables are in logs and left unfiltered to account for potential unit root concerns. Given the quarterly frequency, I use eight lags and include both constant and trend terms. Structural identification follows a Cholesky decomposition. To study the impulse response to a sales shock, I order the variables as sales, inventory, and markup, assuming that a sales shock can immediately affect inventory, while prices and production costs remain fixed within the period. Reordering inventory and markup does not qualitatively change the results.

Figure 3 plots the impulse responses of sales, inventory, the I/S ratio, and aggregate

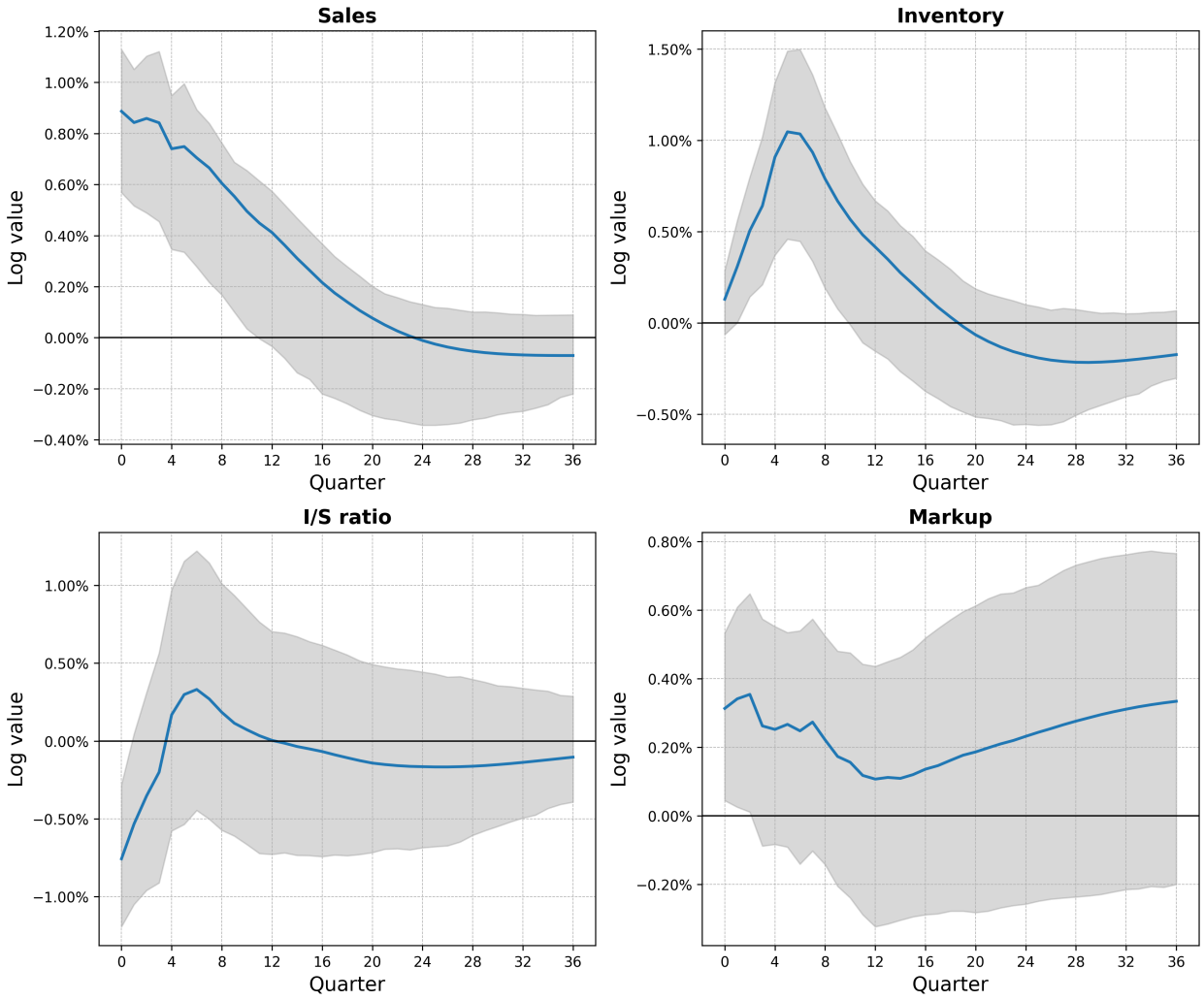

Figure 3: Impulse responses to sales shock, U.S. quarterly 1964Q1–2025Q1

Structural VAR using Cholesky decomposition. Vector order: (sales, inventory, markup). VAR specification: 8 lags with both constant and trend. All variables are in log form and unfiltered.

markup to a sales shock. The shaded areas represent 95% confidence intervals, [5] which highlight several key cyclical patterns: (1) inventory is procyclical, (2) the I/S ratio is countercyclical, [6] and (3) aggregate markup is procyclical. The impulse responses also display strong persistence. Their signs and overall patterns remain robust when variables are HP-filtered, as reported in Appendix A. The associated confidence intervals are narrower

[5]Confidence intervals are non-cumulative and constructed using 500 bootstrap replications.

[6]To address potential cointegration between inventory and sales, I conduct a Johansen test and re-estimate the VAR using the constructed inventory-sales relation. The counter-cyclicality of the I/S ratio persists, as shown in Appendix B.

in the HP-filtered case, since the filter removes the smooth trend.

As discussed above, existing models struggle to account for both the signs and the persistence of the observed movements. The next section introduces a directed search model designed to capture these patterns. The impulse responses in Figure 3 serve as the benchmark for evaluating the new framework.

3 The model

Time is discrete and continues forever. The economy features three types of agents: a continuum of sellers with measure 1, a continuum of buyers with measure μ, and a representative firm. Each period consists of two markets that operate sequentially: a Walrasian market (WM) followed by a search market (SM). The representative firm participates only in the WM, while sellers and buyers are active in both markets. This sequential structure resembles that in Lagos & Wright (2005), but serves a different purpose. Whereas Lagos & Wright (2005) use it to collapse the distribution of money holdings into a degenerate form, here the persistence of a nondegenerate distribution of state variables is essential for generating dynamic responses, as I will discuss later. The purpose of the sequential setup is to reduce computational complexity: joint decisions over inventories and prices are decomposed into separate stages. Specifically, quantities are chosen in the WM, while prices are determined in the SM. Let's first focus on stationary equilibrium, so the time subscript is omitted.

In the WM, sellers begin with a capital stock $k \geq 0$ and an inventory stock $x \in \{0, 1, 2, \ldots\}$, while buyers have no state variable, due to the linear preference introduced later.[7] The representative firm rents capital K from sellers, hires labor L from both sellers and buyers, and produces output Y according to a time-invariant technology $Y = f(K, L)$ with constant returns to scale. The markets for output, capital, and labor are competitive,

[7]It is not crucial that buyers only make static decisions. The results hold if buyers make intertemporal decisions with quasi-linear preference. The linear preference keeps the buyer homogeneous when they enter SM, which is explained below.

implying a capital return of $r = f_K(K, L)$ and a wage rate of $w = f_L(K, L)$. Given constant returns to scale, the representative firm earns zero profit in equilibrium, and sellers and buyers are compensated through their capital and labor incomes.

Given an inventory stock x, sellers may freely dispose of any amount and incur a unit storage cost δ on the remaining stock. They earn capital income rk and labor income $w\ell$.[8] Sellers can also transform WM output into additional inventory $\hat{x} \geq x$ at cost $C(\hat{x} - x)$, where $C'(\cdot) > 0$ and $C''(\cdot) \geq 0$. Although sellers do not directly consume x, they may sell it to buyers in the SM. Sellers' WM choices therefore include consumption c_t, labor supply ℓ, savings $\hat{k}$, and inventory holdings $\hat{x}$. They derive utility $u(c)$ from consumption and disutility $h(\ell)$ from labor, with standard assumptions: $u'(\cdot) > 0$, $u''(\cdot) < 0$, $h'(\cdot) > 0$, and $h''(\cdot) > 0$. On the other hand, buyers supply labor ℓ^* in WM with disutility $h(\ell^*)$ and carry their wage income $w\ell^*$ as payment instrument into the SM. Note that all buyers bring the same amount numeraire to SM, since they do not have state variable. Buyers being homogeneous in SM greatly simplifies the posting game.

At the start of the SM, each seller posts a price p. Buyers then observe both the posted price p and the available inventory $\hat{x}$ before choosing which seller to visit. Suppose n buyers arrive at a given seller. If $n \leq \hat{x}$, each buyer obtains the good; if $n > \hat{x}$, each buyer secures the good with probability $\hat{x}/n$. A successful purchase at price p yields utility $\eta - p$, subject to the budget constraint $p < w\ell^*$. Regardless of whether they purchase in the SM, buyers allocate their remaining income to outside consumption, which provides unit utility.

Buyers do not coordinate on which seller to visit, and this lack of coordination generates endogenous search frictions. The posting-and-search game admits a unique equilibrium in which buyers visit sellers probabilistically (Burdett *et al.*, 2001). Moreover, since all buyers bring the same income $w\ell^*$ into the SM, the equilibrium satisfies the market utility property: in equilibrium, the expected payoff from visiting any seller must be equal. Intuitively, if

[8]Labor income prevents sellers from hitting corner solutions in the absence of credit markets.

one seller offers a higher expected payoff, buyers would deviate toward that seller, thereby equalizing payoffs across sellers. Buyer homogeneity thus greatly simplifies the structure of the posting-and-search equilibrium.

Mathematically, the game can be described as follows. Each seller posts a tuple $(\hat{x}, p, n)$, where $\hat{x}$ denotes the available inventory for sale, p the posted price, and n the buyer-seller ratio, subject to the market utility condition that each buyer's expected payoff equals J. Here, J is an equilibrium object. Given this formulation, the next step is to derive the meeting and consumption probabilities needed to set up the dynamic programming problem.

The matching is many-to-one: each buyer only meets a single seller, while each seller can be visited by multiple buyers. In each submarket $(\hat{x}, p, n)$, each buyer visits each seller with equal probability. [9] It follows that the expected number of customers is n for all sellers in submarket $(\hat{x}, p, n)$. Therefore, the probability π_s of making sales s follows a truncated Poisson distribution.

$$\pi_s(\hat{x}, n) = \begin{cases} \frac{n^s e^{-n}}{s!} & \text{if } s < \hat{x} \\ 1 - \sum_{i=0}^{\hat{x}-1} \frac{n^i e^{-n}}{i!} & \text{if } s = \hat{x} \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

Given the market tightness, the price does not affect the sales probability. Nonetheless, sellers internalize that different prices will lead to different tightness. Note that stockout avoidance plays a role here, as the sales are constrained by the available inventories. Recall the buyers have equal probabilities of consumption if the number of buyers exceeds the inventory holdings. The probability α of a buyer consuming the SM goods at a seller with

[9] This holds true in the unique non-coordinate equilibrium (Watanabe, 2010; Galenianos & Kircher, 2012).

$(\hat{x}, p, n)$ is then

$$\begin{aligned}\alpha(\hat{x}, n) &= \sum_{i=0}^{\hat{x}-1} \frac{n^i e^{-n}}{i!} + \sum_{i=\hat{x}}^{\infty} \frac{n^i e^{-n}}{i!} \frac{\hat{x}}{i+1} \\ &= \sum_{i=0}^{\hat{x}-1} \frac{n^i e^{-n}}{i!} + \frac{\hat{x}}{n} \left(1 - \sum_{i=0}^{\hat{x}} \frac{n^i e^{-n}}{i!}\right) \end{aligned} \tag{2}$$

which is again independent of p. Note that a single buyer's consumption probability depends on the probabilities of how many other buyers show up, so the threshold is at $\hat{x} - 1$. Knowing these key probabilities, we now turn to the preferences and technology before formally defining the equilibrium.

The representative firm has a Cobb-Douglas production function $Y = AK^{\gamma}L^{1-\gamma}$, where A is the time-invariant TFP and γ is the capital share. Let the disutility from labor be $\zeta \ell^{\epsilon}/\epsilon$, where ϵ measures the labor supply elasticity and ζ converts the disutility to util. Then, in WM, buyers make a labor decision by solving the following optimization problem:

$$\max_{l \in [0,1]} \; wl - \zeta \frac{l^{\epsilon}}{\epsilon} \tag{3}$$

where w is the wage rate and ℓ is the labor supply. Note that the benefit $w\ell$ is linear, as buyers' preference is linear in SM. It follows that the buyers' optimal labor supply is $l^* = \left(\frac{w}{\zeta}\right)^{1/(\epsilon-1)}$ and their WM income is wl^*. In SM, buyers derive utility η from consuming x and gain linear utility from consuming the remaining numeraire. Due to the linear preference, borrowers do not save. Again, the linear assumption simplifies the analysis but is not necessary. In the posting-and-search game, buyers decide which submarket $(\hat{x}, p, n)$ to search, and within the submarket, they visit each seller with equal probability. The equilibrium satisfies the market utility condition – all submarkets yield the same expected payoff

J for buyers. In particular, the expected payoff is

$$J = \max_{(\hat{x},p,n)} \ \alpha(\hat{x},n)(\eta + w\ell^* - p) + [1 - \alpha(\hat{x},n)]w\ell^*. \tag{4}$$

The sellers' problem is dynamic, with two state variables: capital k and inventory x. In WM, they supply capital and earn rk. At the same time, each unsold unit of inventory incurs a storage cost of δ. To ensure that sellers have enough resources to cover the storage costs, I assume that sellers also supply labor ℓ in WM and earn $w\ell$ income. The disutility from labor for sellers is the same as that for buyer's. The total available resources for a seller at the beginning of WM is $(1+r)k + wl - \delta x$. In terms of spending, they choose their consumption c, investments in capital $\hat{k}$ and inventory $\hat{x}$ that can be sold in the subsequent SM. In SM, they post the price p and the buyer-seller ratio n, along with the inventory size $\hat{x}$. Assume free disposal and let the cost of producing $C(\hat{x} - x) = \max\{\hat{x} - x, 0\}^\kappa / a$. Note that for $\kappa > 1$, the production cost of SM goods is convex, giving sellers an incentive to smooth production.

Taking into account the buyers' expected utility J, we can write the sellers' SM value $V(\cdot)$ function as

$$V(\hat{k},\hat{x};w,J) = \max_{p,n} \ \sum_{s=0}^{\hat{x}} \pi_s(\hat{x},n) \cdot W(\hat{k} + sp, \hat{x} - s; r', w', J') \tag{5}$$

$$\text{s.t.} \quad J = \alpha(\hat{x},n)(\eta - p) + wl^* \tag{6}$$

$$p \leq \min\{w\ell^*, \eta\} \tag{7}$$

where $W(\cdot)$ is the WM value function and (r', w', J') represent the market conditions in the next period. Seller's SM payoff is the sum of the probability of sales times the corresponding WM payoff, as seller's flow payoff only happens in the WM. For each unit sold, the inventory goes down by 1 unit and the capital goes up by p. Constraint (6) determines how p affect n

given $\hat{x}$, while (7) captures buyers' payment and participation constraints. Note that sales are stochastic, in an endogenous way. The possible events depend on $\hat{x}$, and the transition probabilities π_s depend on $(\hat{x}, p, n)$ and market condition J.

The WM payoff is

$$W(k, x; r, w, J) = \max_{c, \hat{k}, \hat{x}} \frac{c^{1-\sigma}}{1-\sigma} - \zeta \frac{\ell^{\epsilon}}{\epsilon} + \beta V(\hat{k}, \hat{x}; w, J) \tag{8}$$

$$\text{s.t.} \quad c + \hat{k} + C(\hat{x} - x) = (1 + r)k + w\ell - \delta x \tag{9}$$

Sellers gain utility from consuming c and disutility from supplying labor ℓ. They carry $\hat{k}$ and $\hat{x}$ to the next SM, subject to the budget constraint (9). They discount future payoff by β.

Definition 1. *A stationary equilibrium is the value functions W and V, market values (r, w, J), aggregate quantities (K, X), policy functions $(\hat{k}, \hat{x}, p^*, n^*)$, and measure $F_{K,X}$, such that*

1. *Optimality: given (r, w, J), $(p^*, n^*, \hat{k}, \hat{x})$, W and V solve (5) - (9).*
2. *Clearing: $r = f_K(K, \mu l^* + \bar{l})$, $w = f_L(K, \mu l^* + \bar{l})$, $K = \int k \, \mathrm{d}F_{K,X}$ and $\mu = \int n^* \, \mathrm{d}F_{K,X}$.*
3. *Stationarity: $F_{K,X}(k, x) = \int \sum_{s=0}^{\hat{x}} \pi_s(\hat{x}, n^*) \mathbb{1}\{\hat{k} + sp^* \leq k\} \cdot \mathbb{1}\{\hat{x} - s \leq x\} \, \mathrm{d}F_{K,X}$*

Note that when SM production is costly, the market may shut down, resulting in no inventory. To study inventory, I focus on the parameter range in which SM remains active. Meanwhile, the stochastic process here is endogenous, unlike in the standard heterogeneous-agent framework, and thus it is not necessarily ergodic. Intuitively, wealthy sellers may accumulate capital and dominate the SM indefinitely, without facing the downturn risk of low sales. The precise properties of the model depend on parameter values. Consequently, the standard proofs guaranteeing uniqueness of a stationary distribution do not apply in the absence of ergodicity. Rather than pursuing an analytical solution, I follow the approach

common in heterogeneous-agent models and solve the model numerically. The calibrated model does converge to a stationary equilibrium, which is reported in the next section. Before turning to the numerical results, it is useful to highlight the key economic forces at play without fully solving the equilibrium.

The endogenous stochastic process reflects the trade-off between the probability of sales and markup. Hence, sellers face market liquidity when making decisions; to sell faster, they must lower their prices. This relationship can be formally captured by the tightness elasticity of price derived from (6):

$$\lambda(\hat{x}, p, n) \equiv \frac{\mathrm{d}\log p}{\mathrm{d}\log n}|_{(J,w)} = \frac{\alpha_n(\hat{x}, n)n}{\alpha(\hat{x}, n)}\left(\frac{\eta}{p} - 1\right) < 0 \tag{10}$$

Market liquidity λ represents the price cut needed to improve market tightness. It can be grouped into two parts. The first part, $\frac{\alpha_n(\hat{x},n)n}{\alpha(\hat{x},n)}$, represents the elasticity of the consumption probability, while the second part, $\left(\frac{\eta}{p} - 1\right)$, reflects the consumer surplus. When the chance of consumption is very elastic or when the surplus is high, sellers must reduce prices drastically to attract more consumers and liquidate inventories quickly. This creates a new incentive to hold inventory, as the elasticity of the consumption probability depends on the inventory holding.

Proposition 1. *In the case of overstock, $\hat{x} \geq n$,*

$$\left|\frac{\alpha_n(\hat{x}+1, n)n}{\alpha(\hat{x}+1, n)}\right| < \left|\frac{\alpha_n(\hat{x}, n)n}{\alpha(\hat{x}, n)}\right| \tag{11}$$

It follows immediately that

$$|\lambda(\hat{x}+1, p, n)| < |\lambda(\hat{x}, p, n)| \tag{12}$$

Holding more inventory improves market liquidity.

Proof. See Appendix C. □

Proposition 1 shows that holding more inventory makes the consumption probability less elastic. Hence, to attract more buyers, the magnitude of the price cut is smaller when inventory holdings are greater. In other words, sellers have an additional incentive to hold inventory, which allows them to post more profitable terms of trade. Note that this incentive arises in the overstock case $\hat{x} \geq n$, where inventory exceeds the expected number of consumers. In the understock case $\hat{x} < n$, the sign can be ambiguous. Nonetheless, the calibrated model shows that sellers overstock in equilibrium, which aligns with our daily observations.

On the other hand, we can immediately see that $\frac{\partial|\lambda(\hat{x},p,n)|}{\partial p} < 0$. The concern for market liquidity diminishes when prices are higher, so sellers who can charge higher prices worry less about market liquidity at the margin. Such complementarity has equilibrium consequences. In the model, sellers make portfolio decisions between capital and inventory. Sellers with more capital can tolerate a higher risk of low sales and thus tend to post higher prices. A higher price alleviates the market liquidity concern, which encourages them to post even higher prices. Therefore, the capital level affects individual profitability; wealthier agents are also more profitable. This mechanism can generate a heavy tail in the wealth distribution, which poses an empirical puzzle for the Walrasian framework. In the Walrasian framework, the law of one price governs profitability for all agents. Given the uniform profitability across all wealth levels, wealth accumulation solely depends on the diminishing marginal benefit of consumption, and agents have no incentive to maintain a high wealth level. In my model, agents' profitability is wealth-dependent: the richer have better profitability. Even though the agents are *ex ante* homogeneous, in the stationary equilibrium, some agents may cluster at a high-wealth level in the distribution, as seen in the numerical solution.

To illustrate the market liquidity channel does help to explain inventory cycles, I later compare the model impulse responses to the ones from structural VAR. Before turn to the

numerical exercise, I show analytically the economic mechanisms.

Proposition 2. *Given the distribution of posting strategies $G(x,p)$ that delivers market utility J, a change in buyer utility $\tilde{\eta} > \eta$ leads to $\tilde{J} > J$.*

Moreover, the new equilibrium tightness satisfies $\frac{\alpha(x_1,\tilde{n}_1)}{\alpha(x_1,n_1)} < \frac{\alpha(x_2,\tilde{n}_2)}{\alpha(x_2,n_2)}$ for any $p_1 > p_2$ for all x_1, x_2 in the support.

It follows that $\int p\mathbb{E}_s[s\pi_s(x,\tilde{n}_{x,p})]\,\mathrm{d}G(x,p) > \int p\mathbb{E}_s[s\pi_s(x,n_{x,p})]\,\mathrm{d}G(x,p)$.

Proof. See Appendix D. □

Proposition 2 examines an aggregate sales shock generated by a preference shock to η, which occurs after sellers post terms but before buyers search. When buyers experience a positive utility shock, they adjust their visiting behavior to satisfy the new market utility condition, i.e., all submarkets must deliver the same expected utility. Since inventories and prices are pre-committed by sellers, the only margin that adjusts is market tightness. As buyers derive more utility from consumption, they tolerate higher prices and shift toward submarkets that offer a higher consumption probability. Consequently, aggregate sales increase.

This sales shock effectively raises the total capital stock in the next period: buyers pay more for SM goods and leave less income unspent on other consumption. In other words, higher sales transform what would otherwise have been consumed output into sellers' capital accumulation. With more capital available, the interest rate falls, making capital investment less attractive. When sellers make their portfolio decisions, they allocate more toward inventory, rendering inventory pro-cyclical.

At the same time, sellers weigh the price margin. As shown in Proposition 1, higher inventory improves market liquidity. This allows sellers to charge higher prices, which reduces buyers' visits and limits the incentive to hold excessive inventory. Two implications

follow immediately. First, although inventory rises, it does not increase as quickly as sales, producing a counter-cyclical inventory-sales ratio. Second, higher posted prices imply that markups are pro-cyclical.

In short, portfolio choices drive pro-cyclical inventory, while pricing behavior generates a counter-cyclical inventory-sales ratio and pro-cyclical markups. In the numerical exercise below, I show how the model's responses reconcile those from the VAR.

4 Calibration and results

To solve the model, the numerical algorithm departs slightly from the standard approach because the stochastic process is endogenous. In equation (5), there are three aggregate states: the interest rate r, the wage rate w, and market utility J. Since r and w are bijective, we only need to track (r, J) as the relevant aggregate states. Because J governs how different prices translate into market tightness, the stochastic processes must be defined relative to J. Furthermore, the idiosyncratic shocks depend on agents' optimal choices of inventory $\hat{x}$ and price p. The sequential structure of the problem is helpful: the quantity decisions $\hat{k}$ and $\hat{x}$ are separable from the pricing decisions p and n, which substantially reduces computational complexity.

The numerical algorithm proceeds as follows. First, we define a grid over the state space (k, x), where k is continuous and x is discrete by construction of the model. The computation loop then follows the steps outlined below.

(I) Guess a pair (r^0, J^0). Use the CRS property of $f(K, L)$ to obtain w^0 and K^0.

(II) Guess an initial value function V^0 on the grid of (k, x)

  (i) For each $(\hat{k}, \hat{x})$, grid search for optimal (p, n). Note that given J, p and n are bijective, so we only search over n. This provides us the optimal value $\hat{V}^0$ over $(\hat{k}, \hat{x})$, along with the policy functions for p and n.

(ii) For each (k, x), grid search for optimal $(\hat{k}, \hat{x})$ using $\hat{V}^0$. This gives a new value function V^1 and the policy functions for $(\hat{k}, \hat{x})$.

(iii) If V^0 and V^1 are distant, replace V^0 with V^1 and iterate to (i). If V^0 and V^1 are close, exit the inner loop.

(III) State an initial distribution F^0 over grids (k, x).

– Use the policy functions from (II) to iterate the distribution until the convergence reaches some F^1.

(IV) Use F^1 to compute the average market tightness $\hat{\mu} = \int n^* \, \mathrm{d}F$. If $\hat{\mu} > \mu$ ($\hat{\mu} < \mu$), increase (decrease) J^0 and return to (II). If $\hat{\mu}$ and μ are close, go to the next step.

(V) Use F^1 to compute the aggregate capital K. If $K > K^0$ ($K < K^0$), decrease (increase) r^0 and go to (II), unless K and K^0 are close.

A potential concern in solving the model numerically is that different initial distributions may converge to different steady states. However, since the set of steady states is not dense, small perturbations in the initial distribution do not alter the outcome. That said, the non-uniqueness introduces some degree of ad hocness into the calibration, even though the choice of initial distribution does not qualitatively affect the impulse responses. In what follows, the iteration begins with a uniform distribution.

4.1 Calibration

To calibrate the model, I match model-implied moments to their long-run empirical counterparts, as reported in Table 2. The parameters in the top panel are set without solving for the equilibrium. Total factor productivity A is normalized to 0.03, and the risk-aversion parameter σ is fixed at 2, a standard value in the real-business-cycle (RBC) literature. The parameters γ and ϵ are chosen to directly match the observed capital share and the elasticity of labor supply, respectively.

Table 2: Model calibration

Parameter		Description	Target	Data	Model
		Non-equilibrium object			
A	0.03	TFP	normalization	-	-
σ	2.0	CRRA param.	risk aversion	-	-
γ	0.33	capital share	capital share	0.3 - 0.4	0.33
ϵ	1.53	labor preference	labor supply elasticity	1.90	1.90
		Equilibrium object			
β	0.992	time preference	annual interest rate	0.03	0.03
ζ	0.033	labor preference	labor hour	0.51	0.51
κ	1.6	SM scale elasticity	inventory/GDP	0.14	0.16
δ	0.03	maintenance cost	sales/GDP	0.99	0.99
η	0.04	SM utility	consumption/GDP	0.65	0.68
μ	2.6	agg. buyer-seller ratio	wage/GDP	0.46	0.47
a	220	SM technology	markup	1.2 - 2.3	1.47

The parameters in the bottom panel are calibrated to match equilibrium objects and therefore require a joint search. The time preference parameter β is set to target an annual interest rate of 3%, while the leisure preference parameter ζ is chosen to match an average labor supply of 0.51. The inventory-to-GDP ratio pins down the SM scale elasticity κ, estimated at 1.6, which implies convex costs and thus an incentive to smooth production. SM utility η is calibrated to the consumption-to-GDP ratio, and the aggregate buyer-seller ratio μ is chosen to match the wage-to-GDP ratio. Finally, SM productivity a is set to target the markup. Overall, the model matches the targeted equilibrium objects to their data counterparts quite well.

It is worth noting that the calibration of β differs from the standard approach in macroeconomics. Typically, β alone determines the steady-state interest rate through the Euler equation $1 = \beta(1 + r)$, since saving in capital is the only way to transfer resources across time. In this model, however, sellers can also invest in SM goods. The optimal portfolio choice therefore equates the marginal return on capital with the marginal return on SM goods, adjusted for sales risk. Consequently, $1 \neq \beta(1 + r)$ in general. Moreover, the more advanced the SM technology, the higher the steady-state interest rate. In the calibrated

steady state, we obtain $1 > \beta(1 + r)$. This contrasts with standard RBC models, or most heterogeneous-agent models, where long-run productivity does not affect the return to capital.

Another important target moment is the markup. The magnitude of the aggregate markup remains understudied, with existing estimates ranging from 1.2 to 2.3. The underlying source of the markup in this model, however, differs from the standard interpretation. Traditional markup estimation attributes it to market power (Berry *et al.*, 1995; De Loecker *et al.*, 2020), where a higher markup reflects stronger firm-level market power. By contrast, the present model operates in a competitive environment in which search frictions generate the markup. Both market tightness and inventory holdings influence its level. Consequently, the appropriate markup target is not clear-cut. In calibration, the model delivers a markup that falls within the mid-range of estimates in the literature. Importantly, choosing a different target markup does not qualitatively alter the impulse responses.

4.2 The steady state

I report the steady state for the calibrated parameters below. Figure 4 shows that the value functions are increasing in both capital and inventory. Figure 5 presents the WM policy functions. In the left panel, the inventory policy functions are increasing in capital holdings, indicating that larger sellers produce more SM goods. In the right panel, the capital policy functions are not monotonically increasing, as sellers begin to allocate resources to inventory once they have accumulated sufficient capital.

A novel feature of this paper is the characterization of price-posting policy functions, shown in Figure 6. The left panel plots potential best responses as dots, while the right panel displays the actual price-tightness pairs that carry positive mass in equilibrium, with larger dots indicating greater mass. In both panels, the curves represent isoquants of market utility: each point along a curve delivers the same expected payoff to buyers. Different curves

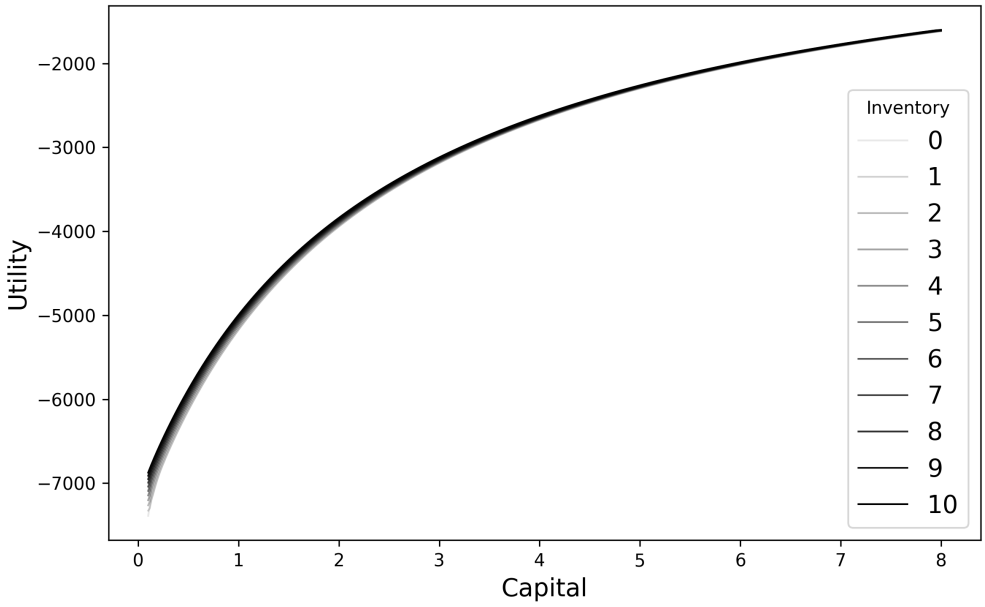

Figure 4: Value function

The value functions across different inventories are close. Lower inventory yields lower value function.

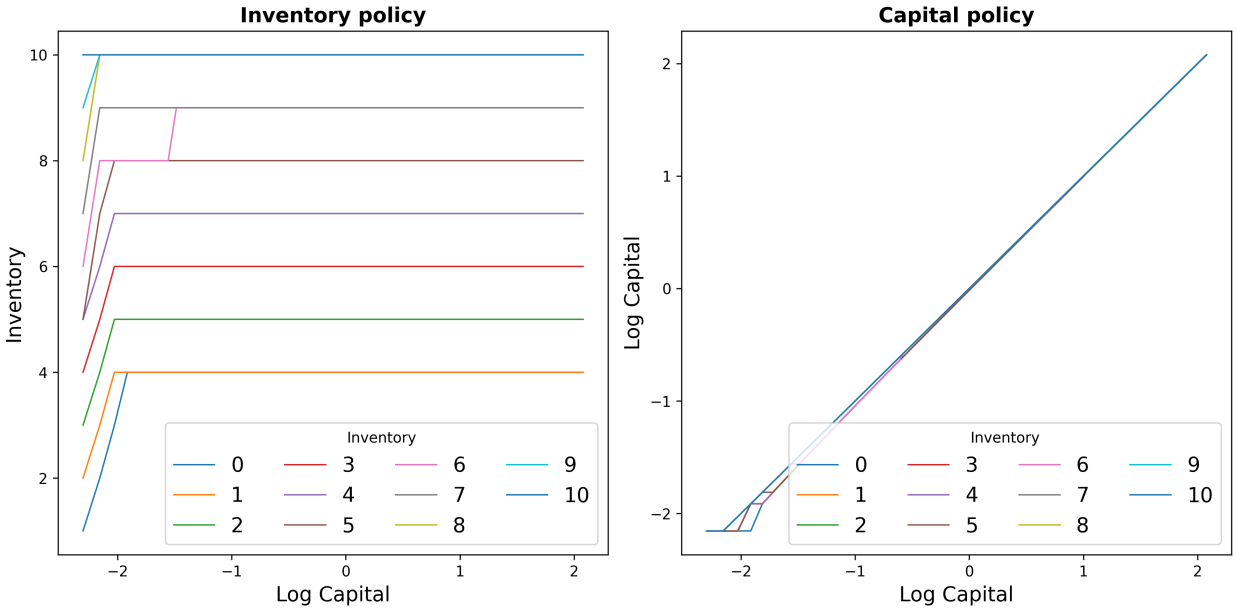

Figure 5: WM policies: inventory and capital

The differences across different inventory levels concentrate at the low capital level. Therefore, the x-axis is log transformed.

correspond to different inventory levels, yet all satisfy the market utility condition. Holding the buyer-seller ratio fixed, higher inventory raises the probability of consumption, which

in turn supports a higher price. Moreover, consistent with Proposition 1, greater inventory improves market liquidity, illustrated by flatter isoquants at higher inventory levels. Finally, note that different dots correspond to different stochastic processes.

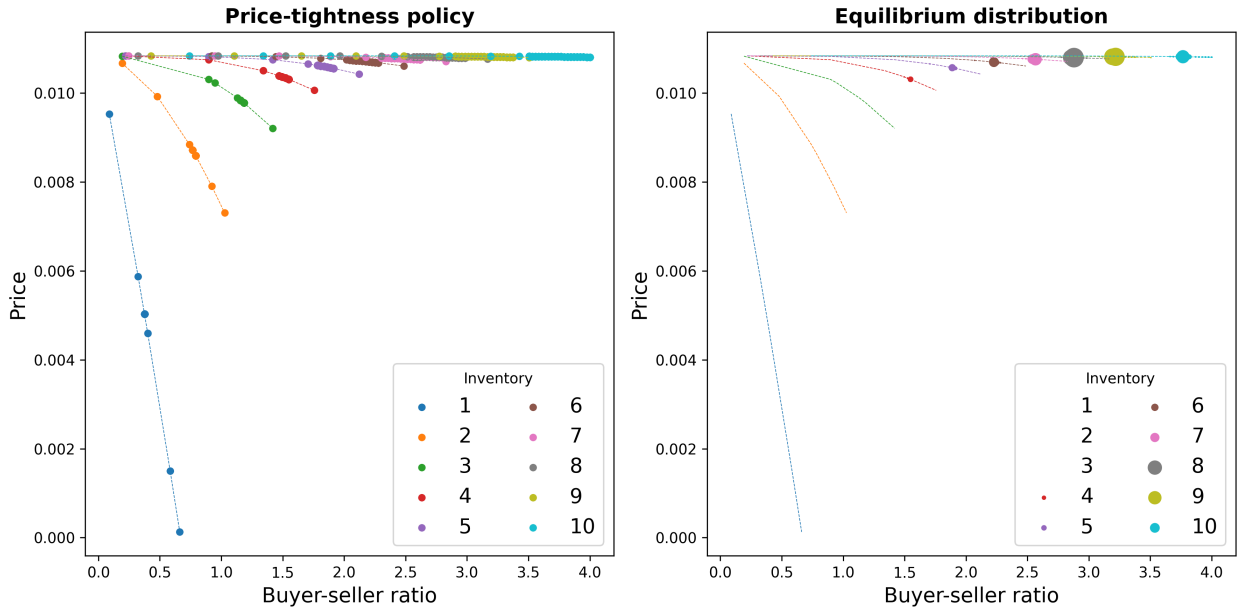

Figure 6: SM policies: price and tightness

The dots present the optimal responses. The dashed lines are the isoquants for the market utility condition. The left panel plots the policies for all states. The right panel plots the policies with positive mass in equilibrium. Bigger dots represent greater equilibrium mass.

Figure 7 presents the stationary distributions. The left panel plots inventory distributions, while the right panel plots capital distributions. Appendix F provides the joint distributions. In both panels, the pre-trade distribution captures state heterogeneity at the beginning of the SM. After trades occur, the inventory distribution shifts left and the capital distribution shifts right, producing the post-trade distributions. The post-trade inventory distribution appears roughly symmetric, whereas the pre-trade distribution is right-skewed, implying that average inventory holdings exceed the median. Capital distributions exhibit similar skewness, suggesting a relatively larger share of big firms compared to small ones. The pre- and post-trade capital distributions differ only slightly, as the capital stock is roughly twenty times larger than total inventories.

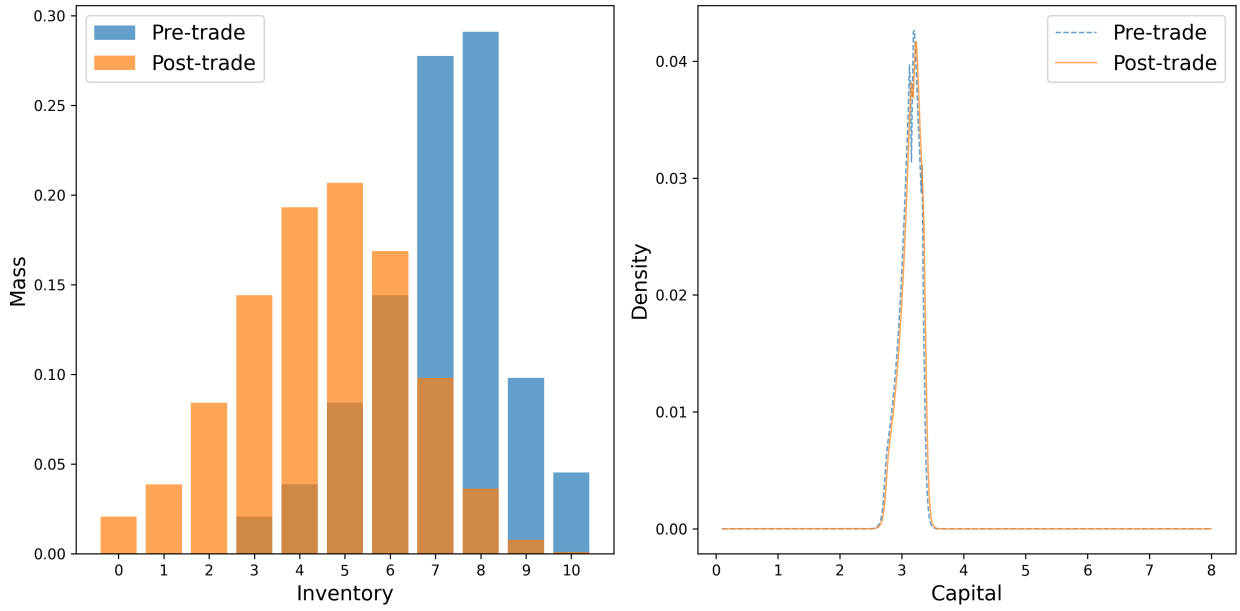

Figure 7: Marginal distributions

The distributions are from the stationary equilibrium. The left panel plots the marginal distribution for inventory. The right panel plots the marginal distribution for capital. The pre-trade state is after posting before trading. Appendix F reports the joint distributions.

With the *ex post* heterogeneity in inventory and capital, the model naturally generates price dispersion. Figure 8 shows the price distribution, normalized so that the highest price equals 1. The distribution is right-skewed, indicating that fewer sellers charge lower prices. In magnitude, the lowest price is about 90% of the highest price. While this degree of price dispersion is smaller than what retail data suggest (Kaplan & Menzio, 2015), reliable estimates of aggregate price dispersion are limited. Another implication is that smaller shops exhibit higher price variance, as shown in Figure 6, consistent with general observations.

Greater price dispersion could be generated by increasing variance in capital—for example, by using a different initial capital distribution instead of a uniform one. Although the model has the potential to better match observed heterogeneity, the focus of this paper is not on capturing heterogeneity per se. Given the stationary equilibrium, we can verify whether the model's responses to sales shocks align with empirical observations.

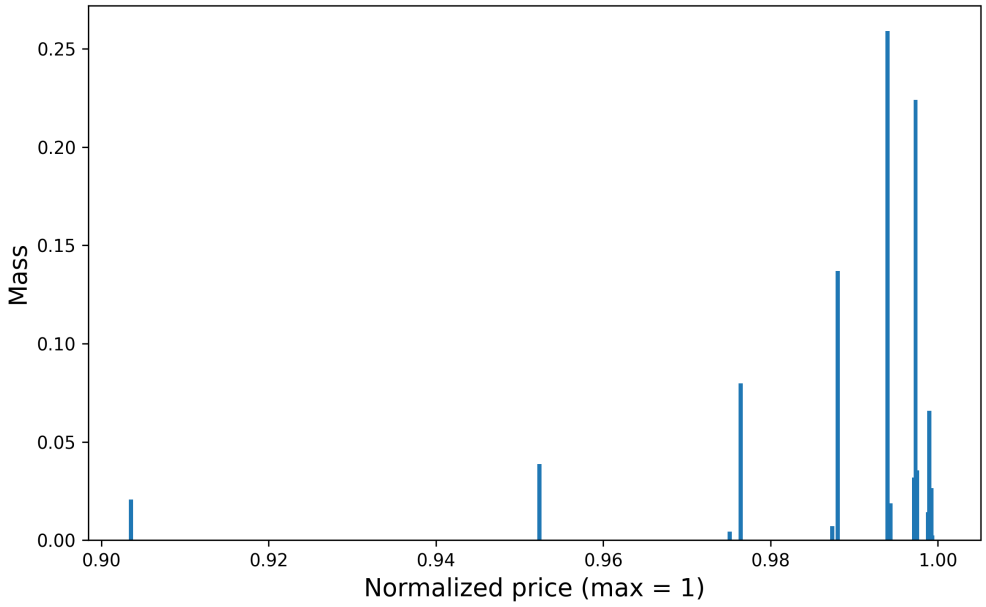

Figure 8: Stationary distribution: price

The graph depicts the price distribution. The normalization applies a constant multiplier to all prices, with the highest price being normalized to 1.

4.3 Impulse responses

To evaluate model performance, I compare the model's impulse responses with those from the structural VAR (Figure 3). To introduce a sales shock in the model, I consider a one-time increase in buyers' SM utility η, occurring after sellers post terms but before buyers search. The choices of η and timing ensure precise identification of the shock. In the VAR analysis, a Cholesky decomposition is used with the vector ordering (sales, inventory, markup), which implies that a sales shock should trigger an immediate response in inventory while leaving markup initially unchanged. Because the search and matching processes are microfounded through the game, the only parameter capable of generating such a shock is the SM utility η. Placing the shock after sellers post ensures that prices and thus markups do not respond immediately.

Specifically, the new utility $\tilde{\eta} = 0.42$ is chosen to match the magnitude of the initial sales jump in the VAR impulse responses. Given the timing, posted inventories and prices remain fixed, and buyers adjust their search strategies to achieve a new market utility $\tilde{J}$ with

a corresponding market-tightness distribution $\tilde{G}$, as described in Proposition 2. To obtain the initial distribution F_0 after the shock, I iterate on $\tilde{J}$ until the average buyer-seller ratio equals the observed ratio μ.

For the impulse responses, I first guess sequences of interest rates r_t and market utility J_t for $t = 0, 1, ..., T$, where T denotes the final period. Using these sequences, I backward-compute all optimal responses $(\hat{k}_t, \hat{x}_t, n_t, p_t)$, and then forward-compute the evolution of the joint distributions F_t. This distribution path generates updated sequences r'_t and J'_t. I iterate this procedure until J_t and r_t converge, increasing T as necessary.

Figure 9 compares the model's impulse responses with those from the VAR. The directions of the initial movements are consistent, and the responses are highly persistent, although the magnitudes are not exact: sales decline slightly too quickly, while inventory and markup respond less than observed. Nonetheless, the model predictions generally remain within the 95% confidence intervals of the VAR results.

Given its parsimonious parameterization, the model performs well. More importantly, all movements are driven by economic mechanisms, underscoring the role of price adjustment. By incorporating this additional decision margin, agents' quantity responses align more closely with the data. The model allows agents to choose both quantities and prices by replacing the traditional market-clearing condition with the market utility condition. Figure 14 in Appendix E reports impulse responses for other major variables, showing aggregate patterns similar to those in a standard model.

It is worth noting that *ex post* heterogeneity is essential for generating cycles. If all sellers were homogeneous, the SM would consist of a single submarket in which all sellers hold the same inventory and post the same price. In that case, an SM preference shock would have little effect, since buyers would not adjust their search behavior when all sellers offer identical terms. Thus, *ex post* heterogeneity plays a crucial role in the dynamics, unlike in many other heterogeneous-agent models, where heterogeneity typically has only quantitative

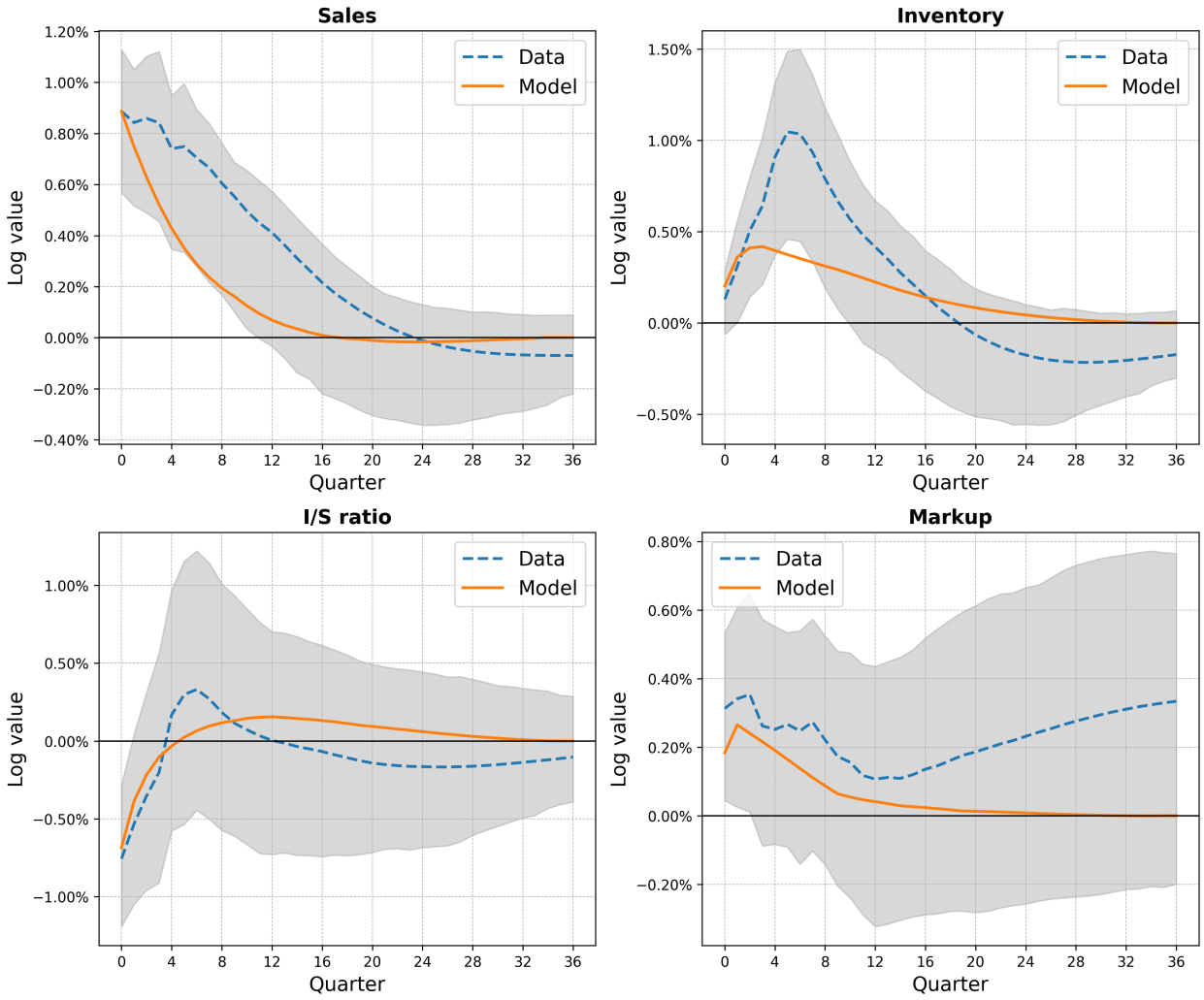

Figure 9: Model responses vs. VAR responses

A comparison between the equilibrium model and the VAR model. Both models construct a stationary relation among the variables. The impulse responses are induced by some one-time unexpected sales shocks. The VAR shock is a one standard deviation shock. The shock to the equilibrium is chosen to match the VAR shock. All variables are in log form.

effects.

4.4 Efficiency

Having established that the model can account for the puzzling inventory cycles, we can use it to evaluate the efficiency of inventory allocation. Returning to the stationary equilibrium, we first compare sellers' inventory stocks with the expected number of customers. Figure

10 plots the buyer-seller ratio for each inventory level. The dashed line, with a slope of 0.5, represents the threshold where inventory is sufficient to serve twice the expected number of buyers.

The first observation is that all sellers overstock: equilibrium buyer-seller ratios lie well below inventory levels. Most sellers hold enough inventory to serve more than twice their expected buyers, and, according to the bottom-right dot, some stock enough to serve up to ten times their expected customers. Despite this apparent overstocking, the welfare implication requires additional analysis, as holding more inventory increases buyers' probabilities of consumption.

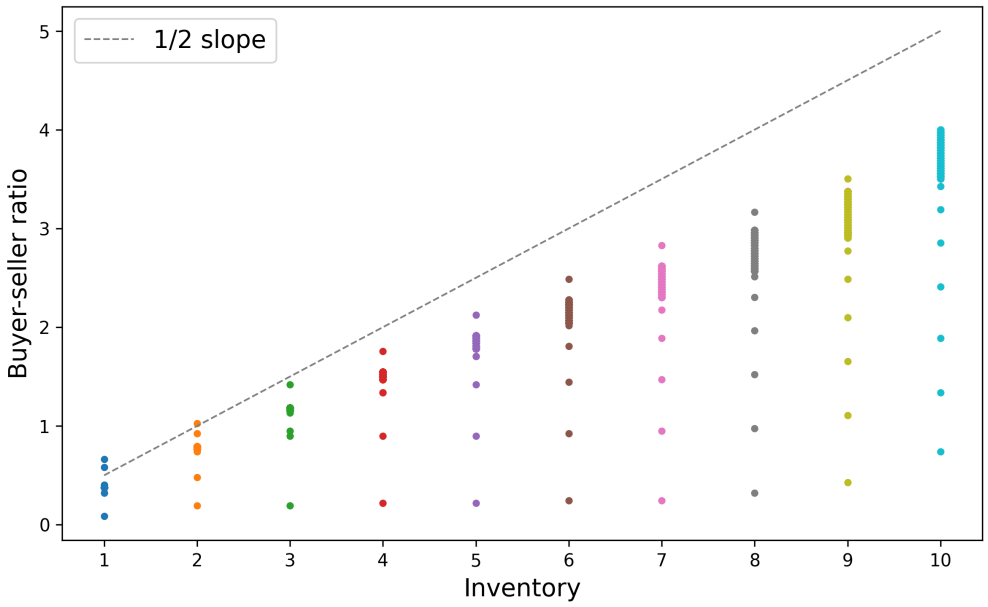

Figure 10: Overstock: tightness vs. inventories

The dots represent the equilibrium buyer-seller ratio across different inventory levels. For any given inventory level, the variation of buyer-seller ratio comes from the capital dispersion. More capital is associated with higher buyer-seller ratio. The dashed line has a slope of 0.5.

Consider the planner's problem under the constraint that buyers cannot coordinate which sellers to purchase from. From the game, the urn-ball matching mechanism implies that the optimal market structure consists of a single submarket, in which all sellers hold the same amount of inventory and market tightness is determined by the aggregate buyer-seller ratio

μ. The planner's problem is

$$x^* = \max_x \ \mu\eta\alpha(x,\mu) - a^{-1}\sum_{s=0}^{x}\pi_s(x,\mu)\left[s^\kappa + \delta(x-s)\right]. \tag{13}$$

Holding more inventory increases buyers' probability of consumption but comes at the cost of higher maintenance expenses. The actual level of the constrained optimum depends on the model parameters. In the calibrated economy, the optimal inventory level is $x^* = 7$. However, as shown in the left panel of Figure 7, only 28% of sellers hold exactly 7 units. Of the remaining sellers, 44% stock 8 or more units, while 28% hold 6 or fewer units. Overall, the economy overstocks by 1.6%, equivalent to roughly 0.2% of annual GDP.

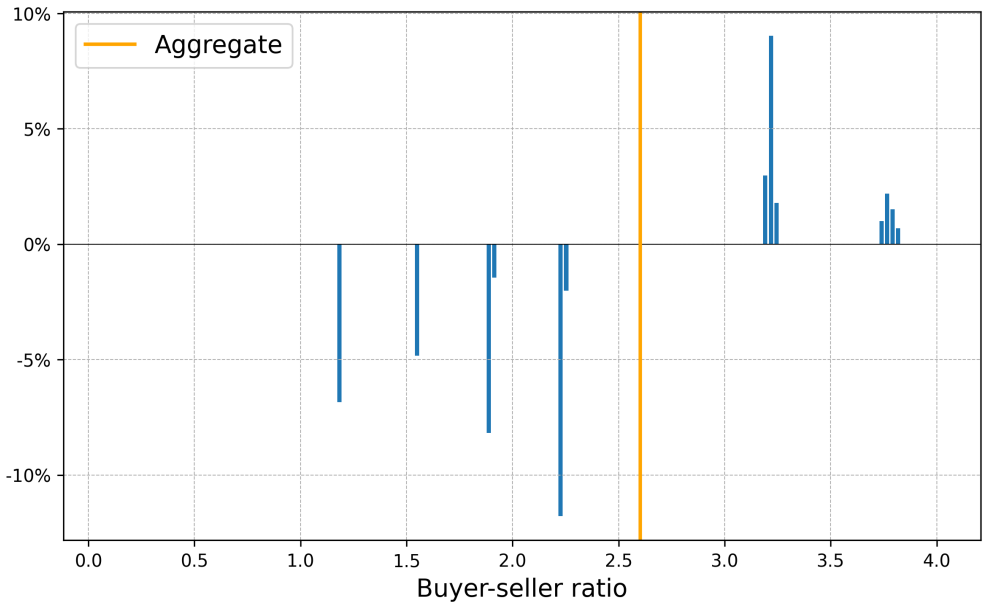

Figure 11: Over/under-stock heterogeneity

The graphs depicts the percentage of over/under-stock given the equilibrium tightness, weighted by the distribution density.

The actual welfare loss must account for distributional effects, as illustrated in Figure 11. Consequently, the efficiency evaluation has two components. The first is over- or under-stocking relative to market tightness, which generates a 0.06% welfare loss compared with the constrained optimum. The second is the dispersion in market tightness caused by capital heterogeneity, contributing a 0.49% welfare loss – about eight times larger than the first

component. Together, these effects produce a total welfare loss of 0.55%, equivalent to roughly 0.27% of annual GDP.

Note that larger sellers tend to overstock, while smaller sellers understock. To restore efficiency, a natural policy implication is the implementation of a progressive tax on either capital or inventory. Such a tax could have two beneficial effects: it would improve inventory efficiency across sellers of different sizes and simultaneously encourage a more balanced capital distribution. Because the welfare impact of a progressive tax depends heavily on the precise capital distribution, I leave the detailed policy design – which would target agent heterogeneity more precisely – for future work.

5 Conclusion

This paper brings inventories back to the center of macroeconomic analysis. Despite their substantial size and well-documented importance for business cycles, inventories have remained peripheral in most macroeconomic models. I begin by documenting three robust empirical regularities: inventory investment is pro-cyclical, the inventory-sales ratio is counter-cyclical, and markups are pro-cyclical. Existing explanations cannot jointly account for these patterns.

To address this gap, I develop a dynamic general equilibrium model with endogenous search frictions, in which firms jointly choose prices and inventories. This framework generates stochastic sales and captures firms' dual adjustments along both price and quantity margins. The resulting mechanism rationalizes the observed empirical regularities, as the model successfully produces impulse responses that closely match those from the structural VAR.

In the calibrated economy, inventories are overstocked relative to the planner's allocation by 1.6%, resulting in an annual welfare loss of 0.27% of GDP. Most of this inefficiency stems

from firm heterogeneity, with large firms overstocking and small firms understocking. This pattern suggests that a progressive tax on either capital or inventory could improve welfare.

The model provides new insights into the mechanisms driving inventory cycles. Price adjustments interact with inventory holdings to produce the observed co-movements: higher sales increase both inventories and markups while moderating the response of quantities. The complementarity between high inventories and high prices also generates persistence, extending the effects of shocks without relying on exogenous shock persistence. These features make the framework a promising foundation for studying business cycle dynamics more broadly.

On the other hand, the model has limitations that point to directions for future research. The current framework underpredicts the degree of wealth inequality and price dispersion observed in the data. Incorporating credit markets and richer forms of heterogeneity could help bridge this gap. Moreover, although this paper focuses on inventories, the many-to-one matching structure provides a foundation for studying other markets – such as financial assets and labor – where liquidity, price formation, and firm-level capacity decisions play a central role.

In sum, by incorporating microfoundations of inventories into a dynamic general equilibrium framework, this paper offers a unified explanation of their cyclical behavior, quantifies the welfare costs of inefficient inventory allocation, and provides a framework with broader applicability for macroeconomic analysis. This perspective also resonates with the enduring relevance of the Lucas critique.

References

Aiyagari, S Rao. 1994. Uninsured idiosyncratic risk and aggregate saving. *The Quarterly Journal of Economics*, **109**(3), 659–684.

Arrow, Kenneth J, Harris, Theodore, & Marschak, Jacob. 1951. Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, 250–272.

Auernheimer, Leonardo, & Trupkin, Danilo R. 2014. The role of inventories and capacity utilization as shock absorbers. *Review of Economic Dynamics*, **17**(1), 70–85.

Berry, Steven, Levinsohn, James, & Pakes, Ariel. 1995. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Bils, Mark, & Kahn, James A. 2000. What inventory behavior tells us about business cycles. *American Economic Review*, **90**(3), 458–481.

Blinder, Alan S. 1986. *More on the speed of adjustment in inventory models*.

Blinder, Alan S, & Maccini, Louis J. 1991. Taking stock: a critical assessment of recent research on inventories. *Journal of Economic perspectives*, **5**(1), 73–96.

Burdett, Ken, & Coles, Melvyn G. 1997. Marriage and class. *The Quarterly Journal of Economics*, **112**(1), 141–168.

Burdett, Kenneth, Shi, Shouyong, & Wright, Randall. 2001. Pricing and matching with frictions. *Journal of Political Economy*, **109**(5), 1060–1085.

Cavalcanti, Ricardo de O, & Wallace, Neil. 1999. Inside and outside money as alternative media of exchange. *Journal of Money, Credit and Banking*, 443–457.

Cavallo, Alberto, & Kryvtsov, Oleksiy. 2023. What can stockouts tell us about inflation? Evidence from online micro data. *Journal of International Economics*, **146**, 103769.

Christiano, Lawrence J. 1988. Why does inventory investment fluctuate so much? *Journal of Monetary Economics*, **21**(2-3), 247–280.

Coen-Pirani, Daniele. 2004. Markups, aggregation, and inventory adjustment. *American Economic Review*, **94**(5), 1328–1353.

De Loecker, Jan, Eeckhout, Jan, & Unger, Gabriel. 2020. The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, **135**(2), 561–644.

Diamond, Peter A. 1982. Aggregate demand management in search equilibrium. *Journal of political Economy*, **90**(5), 881–894.

Duffie, Darrell, Gârleanu, Nicolae, & Pedersen, Lasse Heje. 2005. Over-the-counter markets. *Econometrica*, **73**(6), 1815–1847.

Galenianos, Manolis, & Kircher, Philipp. 2012. On The Game-Theoretic Foundations Of Competitive Search Equilibrium. *International economic review*, **53**(1), 1–21.

Geromichalos, Athanasios. 2012. Directed search and optimal production. *Journal of Economic Theory*, **147**(6), 2303–2331.

Geromichalos, Athanasios. 2014. Directed search and the Bertrand paradox. *International Economic Review*, **55**(4), 1043–1065.

Haltiwanger, John C, & Maccini, Louis J. 1988. A model of inventory and layoff behaviour under uncertainty. *The Economic Journal*, **98**(392), 731–745.

Hamilton, James D. 2018. Why you should never use the Hodrick-Prescott filter. *Review of Economics and Statistics*, **100**(5), 831–843.

Huggett, Mark. 1993. The risk-free rate in heterogeneous-agent incomplete-insurance economies. *Journal of economic Dynamics and Control*, **17**(5-6), 953–969.

Kahn, James A. 1987. Inventories and the volatility of production. *The American Economic Review*, 667–679.

Kaplan, Greg, & Menzio, Guido. 2015. The morphology of price dispersion. *International Economic Review*, **56**(4), 1165–1206.

Khan, Aubhik, & Thomas, Julia K. 2007. Inventories and the business cycle: An equilibrium analysis of (S, s) policies. *American Economic Review*, **97**(4), 1165–1188.

Kim, Ryan. 2021. The effect of the credit crunch on output price dynamics: The corporate inventory and liquidity management channel. *The Quarterly Journal of Economics*, **136**(1), 563–619.

Kiyotaki, Nobuhiro, & Wright, Randall. 1989. On money as a medium of exchange. *Journal of political Economy*, **97**(4), 927–954.

Krusell, Per, & Smith, Jr, Anthony A. 1998. Income and wealth heterogeneity in the macroeconomy. *Journal of political Economy*, **106**(5), 867–896.

Kryvtsov, Oleksiy, & Midrigan, Virgiliu. 2012. Inventories, markups, and real rigidities in menu cost models. *Review of Economic Studies*, **80**(1), 249–276.

Kryvtsov, Oleksiy, & Vincent, Nicolas. 2021. The cyclicality of sales and aggregate price flexibility. *The Review of Economic Studies*, **88**(1), 334–377.

Kydland, Finn E, & Prescott, Edward C. 1982. Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, 1345–1370.

Lagos, Ricardo, & Wright, Randall. 2005. A unified framework for monetary theory and policy analysis. *Journal of political Economy*, **113**(3), 463–484.

Lagos, Ricardo, Rocheteau, Guillaume, & Wright, Randall. 2017. Liquidity: A new monetarist perspective. *Journal of Economic Literature*, **55**(2), 371–440.

Li, Fei, Murry, Charles, Tian, Can, & Zhou, Yiyi. 2024. Inventory management in decentralized markets. *International Economic Review*, **65**(1), 431–470.

Luo, Yulei, Nie, Jun, Wang, Xiaowen, & Young, Eric R. 2025. Production and inventory dynamics under ambiguity aversion. *Journal of Monetary Economics*, 103767.

Metzler, Lloyd A. 1941. The nature and stability of inventory cycles. *The Review of Economics and Statistics*, **23**(3), 113–129.

Mortensen, Dale T, & Pissarides, Christopher A. 1994. Job creation and job destruction in the theory of unemployment. *The review of economic studies*, **61**(3), 397–415.

Nekarda, Christopher J, & Ramey, Valerie A. 2013. *The cyclical behavior of the price-cost markup*. Tech. rept. National Bureau of Economic Research.

Ramey, Valerie A, & West, Kenneth D. 1999. Inventories. *Handbook of macroeconomics*, **1**, 863–923.

Rogerson, Richard, Shimer, Robert, & Wright, Randall. 2005. Search-theoretic models of the labor market: A survey. *Journal of economic literature*, **43**(4), 959–988.

Shevchenko, Andrei. 2004. Middlemen. *International Economic Review*, **45**(1), 1–24.

Watanabe, Makoto. 2010. A model of merchants. *Journal of Economic Theory*, **145**(5), 1865–1889.

Wen, Yi. 2011. Input and output inventory dynamics. *American Economic Journal: Macroeconomics*, **3**(4), 181–212.

West, Kenneth D. 1990. The sources of fluctuations in aggregate inventories and GNP. *The Quarterly Journal of Economics*, **105**(4), 939–971.

Wright, Randall, Kircher, Philipp, Julien, Benoît, & Guerrieri, Veronica. 2021. Directed search and competitive search equilibrium: A guided tour. *Journal of Economic Literature*, **59**(1), 90–148.

Appendix A VAR using HP-filtered variables

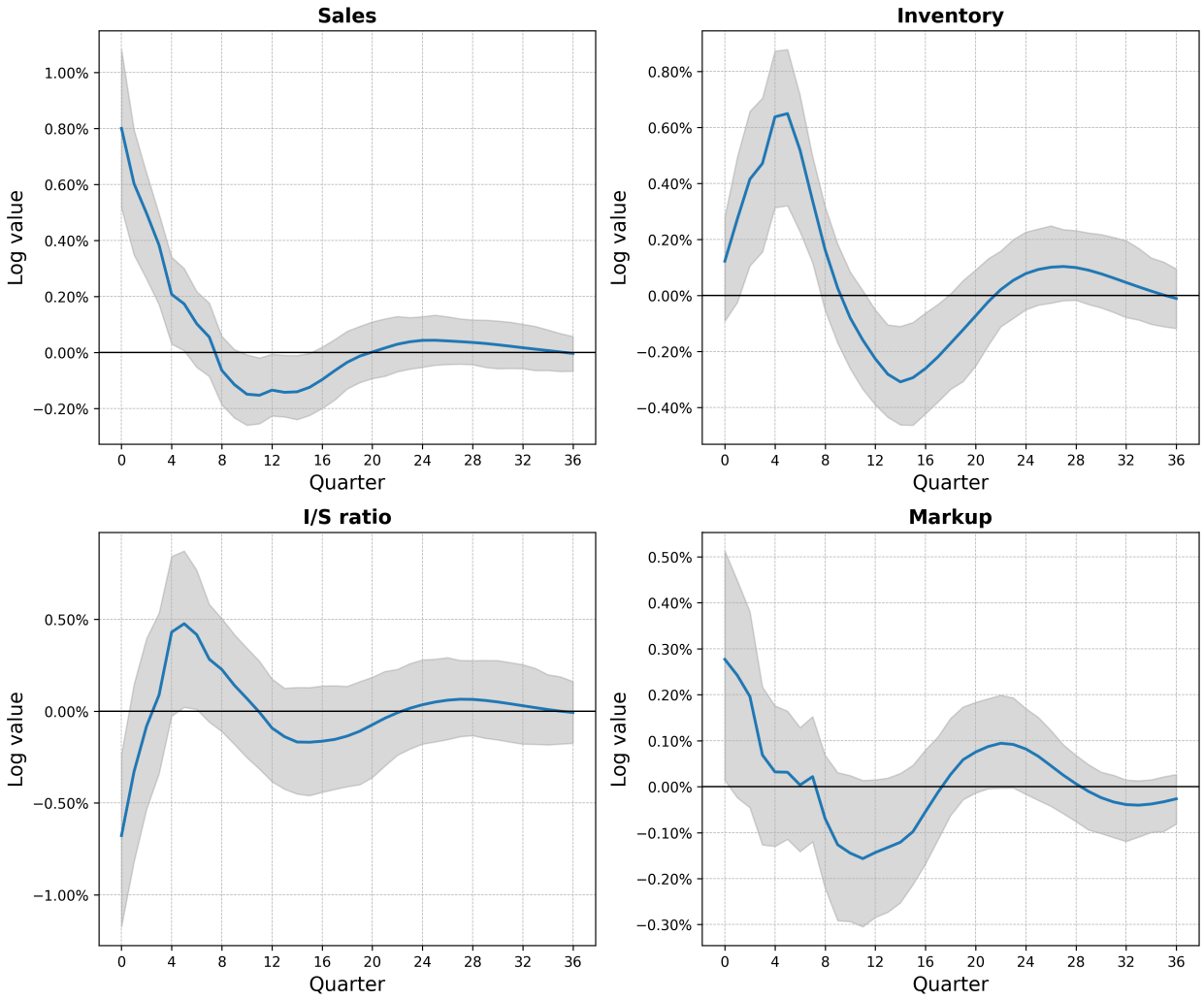

Figure 12: Impulse responses to sales shock, 1964 – 2023

Structural VAR using Cholesky decomposition. Vector order: (sales, inventory, markup). VAR specification: 8 lags with both constant and trend. All variables are in log form and HP-filtered.

This section reports the VAR results from the HP-filtered variables. Figure 12 plots the results where the shaded areas represent the 95% confidence interval. The directions of movements are same as using the selected sample. Inventory is pro-cyclical, I/S ratio is counter cyclical and aggregate markup is pro-cyclical. The confidence intervals cover origin in long-run as using the selected sample.

Appendix B Cointegration test

For the concern of cointegration, I consider a stationary relation between log inventory and log sales $i_t - \theta s_t$. To estimate θ and test the cointegration, I use Johansen procedure with eigenvalue method and long-run VECM for error correction. Table 3 reports the results.

Table 3: Johansen test

	Unfiltered		HP-filtered		Critical value		
Lag	$\hat{\theta}$	Test stat.	$\hat{\theta}$	Test stat.	0.10	0.05	0.01
2	0.89	20.31	1.45	74.67	12.91	14.90	19.19
4	0.90	16.88	1.40	82.45	12.91	14.90	19.19
8	0.91	16.13	2.88	42.13	12.91	14.90	19.19

[a] All variables are in log form. [b] Test with eigenvalue
[c] Error correction: long-run VECM

$\hat{\theta}$ is around 0.9 for the unfiltered cycle, robust to the lag lengths from 2 to 8. The HP-filtered cycle has $\hat{\theta}$ ranges from 1.4 to 2.7. In either setting, we reject the null of no cointegration at the 0.10 level. Most cases also reject the null at the 0.05 level, except the case with log difference and lag length 8. The test statistic 14.46 is very close to the critical value 14.90. Overall, these results suggest that inventory and sales are cointegrated.

To deal with the cointegration, I construct the inventory-sales (I/S) relations using the estimated $\hat{\theta}$ for the two cycle measurements. Figure 13 plots the corresponding impulse responses. The inventory-sales relation is also counter-cyclical. Replacing inventory-sales ratio by the cointegrated relation doesn't change the empirical regulation.

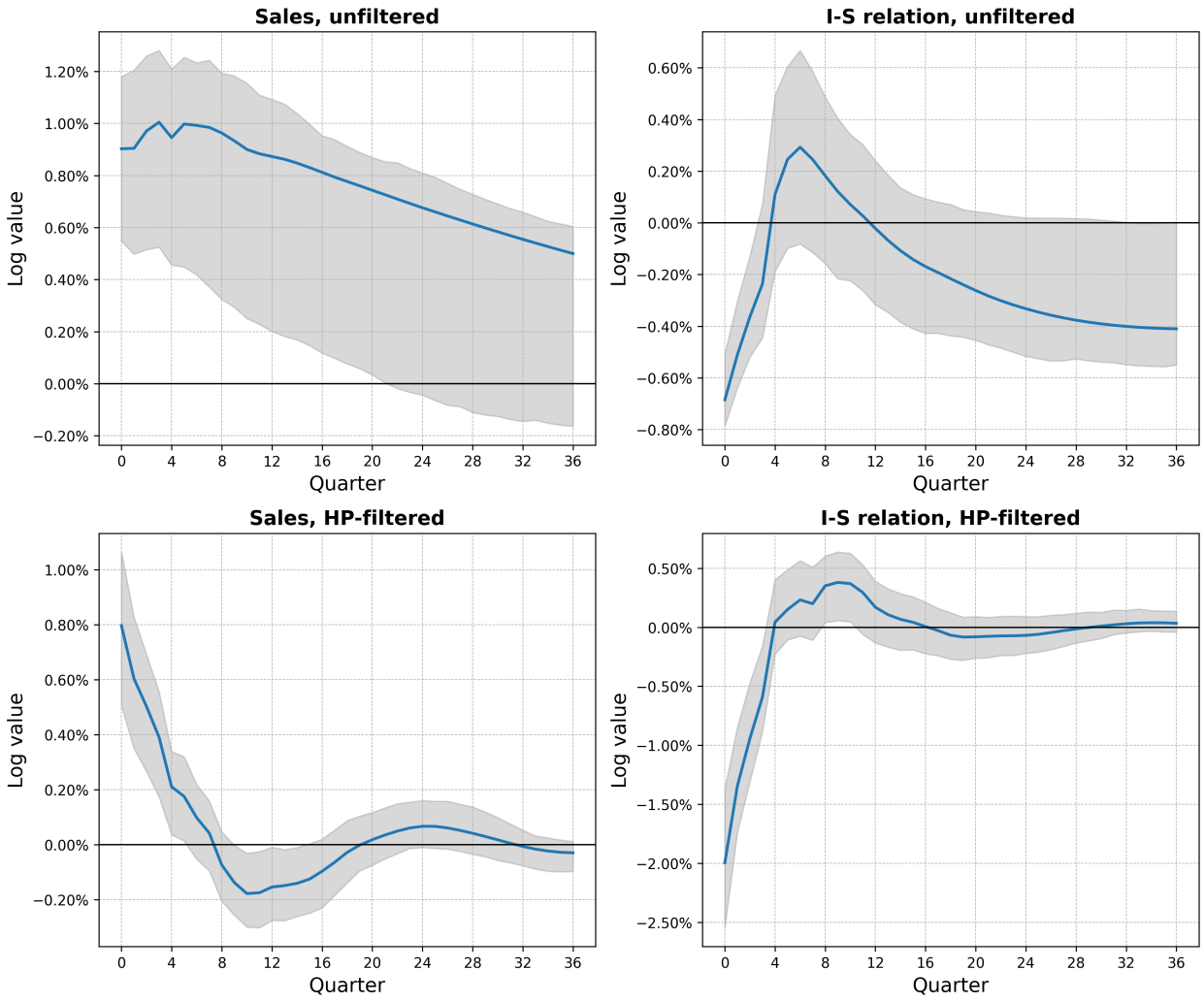

Figure 13: Impulse responses to sales shock, 1964 – 2023

Structural VAR using Cholesky decomposition. Vector order: (sales, inventory-sales relation). VAR specification: 8 lags with both constant and trend. All variables are in log form. The variables in the top panels are unfiltered. The variables in the bottom panels are HP-filtered.

Appendix C Proof of Proposition 1

The calculation below shows how inventory holding affects the elasticity of the consumption probability. From

$$\alpha(\hat{x}, n) = \sum_{i=0}^{\hat{x}-1} \frac{n^i e^{-n}}{i!} + \sum_{i=\hat{x}}^{\infty} \frac{n^i e^{-n}}{i!} \frac{\hat{x}}{i+1} \tag{14}$$

we can calculate the derivative

$$\begin{aligned}
\alpha_n(\hat{x}, n) &= \sum_{i=0}^{\hat{x}-1} \frac{in^{i-1}e^{-n} - n^i e^{-n}}{i!} + \sum_{i=\hat{x}}^{\infty} \frac{in^{i-1}e^{-n} - n^i e^{-n}}{i!} \frac{\hat{x}}{i+1} \\
&= \sum_{i=0}^{\hat{x}-1} \frac{in^{i-1}e^{-n}}{i!} + \sum_{i=\hat{x}}^{\infty} \frac{in^{i-1}e^{-n}}{i!} \frac{\hat{x}}{i+1} - \alpha(\hat{x}, n) \\
&\equiv g(\hat{x}, n) - \alpha(\hat{x}, n) < 0
\end{aligned} \tag{15}$$

Note that

$$\alpha(\hat{x}+1, n) - \alpha(\hat{x}, n) = \frac{n^{\hat{x}} e^{-n}}{\hat{x}!} - \frac{n^{\hat{x}} e^{-n}}{\hat{x}!} \frac{\hat{x}}{\hat{x}+1} > 0 \tag{16}$$

$$g(\hat{x}+1, n) - g(\hat{x}, n) = \frac{\hat{x}}{n} \left[\alpha(\hat{x}+1, n) - \alpha(\hat{x}, n)\right] \tag{17}$$

It follows that

$$|\alpha_n(\hat{x}+1, n)| - |\alpha_n(\hat{x}, n)| = \left[\alpha(\hat{x}+1, n) - \alpha(\hat{x}, n)\right] \left(1 - \frac{\hat{x}}{n}\right) \begin{cases} > 0 & \text{if } \hat{x} < n \\ = 0 & \text{if } \hat{x} = n \\ < 0 & \text{if } \hat{x} > n \end{cases}$$

Therefore, when $\hat{x} \geq n$, the case of overstock

$$\left| \frac{\alpha_n(\hat{x}+1, n) n}{\alpha(\hat{x}+1, n)} \right| < \left| \frac{\alpha_n(\hat{x}, n) n}{\alpha(\hat{x}, n)} \right| \tag{18}$$

Appendix D Proof of Proposition 2

The steady state market utility satisifes

$$J = \alpha(\hat{x}, n)(\eta - p) + wl^* \tag{19}$$

for all $\hat{x}$ and p in the equilibrium distribution $G(x, p)$. Note that

$$\frac{\partial J}{\partial \eta} = \alpha(\hat{x}, n) > 0 \tag{20}$$

for all sellers. By the Envelope theorem, a change in buyer utility $\tilde{\eta} > \eta$ leads to a higher market utility $\tilde{J} > J$. Now evaluate the new market equilibrium. Since $\hat{x}$ and p are pre-committed, a change from η to $e\tilde{t}a$ only affects n. Consider two sellers (x_1, p_1) and (x_2, p_2). In the old and new equilibrium, we have

$$\alpha(x_1, n_1)(\eta - p_1) = \alpha(x_2, n_2)(\eta - p_2) \tag{21}$$

$$\alpha(x_1, \tilde{n}_1)(\tilde{\eta} - p_1) = \alpha(x_2, \tilde{n}_2)(\tilde{\eta} - p_2) \tag{22}$$

Take a ratio of the two equations.

$$\frac{\alpha(x_1, \tilde{n}_1)}{\alpha(x_1, n_1)} \cdot \frac{\tilde{\eta} - p_1}{\eta - p_1} = \frac{\alpha(x_2, \tilde{n}_2)}{\alpha(x_2, n_2)} \cdot \frac{\tilde{\eta} - p_2}{\eta - p_2} \tag{23}$$

If $p_1 > p_2$, $\frac{\alpha(x_1, \tilde{n}_1)}{\alpha(x_1, n_1)} < \frac{\alpha(x_2, \tilde{n}_2)}{\alpha(x_2, n_2)}$ because

$$\frac{\tilde{\eta} - p_1}{\eta - p_1} = 1 + \frac{\tilde{\eta} - \eta}{\eta - p_1} > 1 + \frac{\tilde{\eta} - \eta}{\eta - p_2} = \frac{\tilde{\eta} - p_2}{\eta - p_2} \tag{24}$$

It follows that the expected sales increases with the posted price.

$$\frac{\sum_{s=0}^{x_1} s\pi(x_1, \tilde{n}_1)}{\sum_{s=0}^{x_1} s\pi(x_1, n_1)} > \frac{\sum_{s=0}^{x_2} s\pi(x_2, \tilde{n}_2)}{\sum_{s=0}^{x_2} s\pi(x_2, n_2)} \tag{25}$$

Since this is true for any (p_1, x_1) and (p_2, x_2) with $p_1 > p_2$, the total expected sales increases.

$$\int p \sum_{s=0}^{x} s\pi_s(x, \tilde{n}_{x,p}) \, \mathrm{d}G(x, p) > \int p \sum_{s=0}^{x} s\pi_s(x, n_{x,p}) \, \mathrm{d}G(x, p) \tag{26}$$

Appendix E Other impulse responses

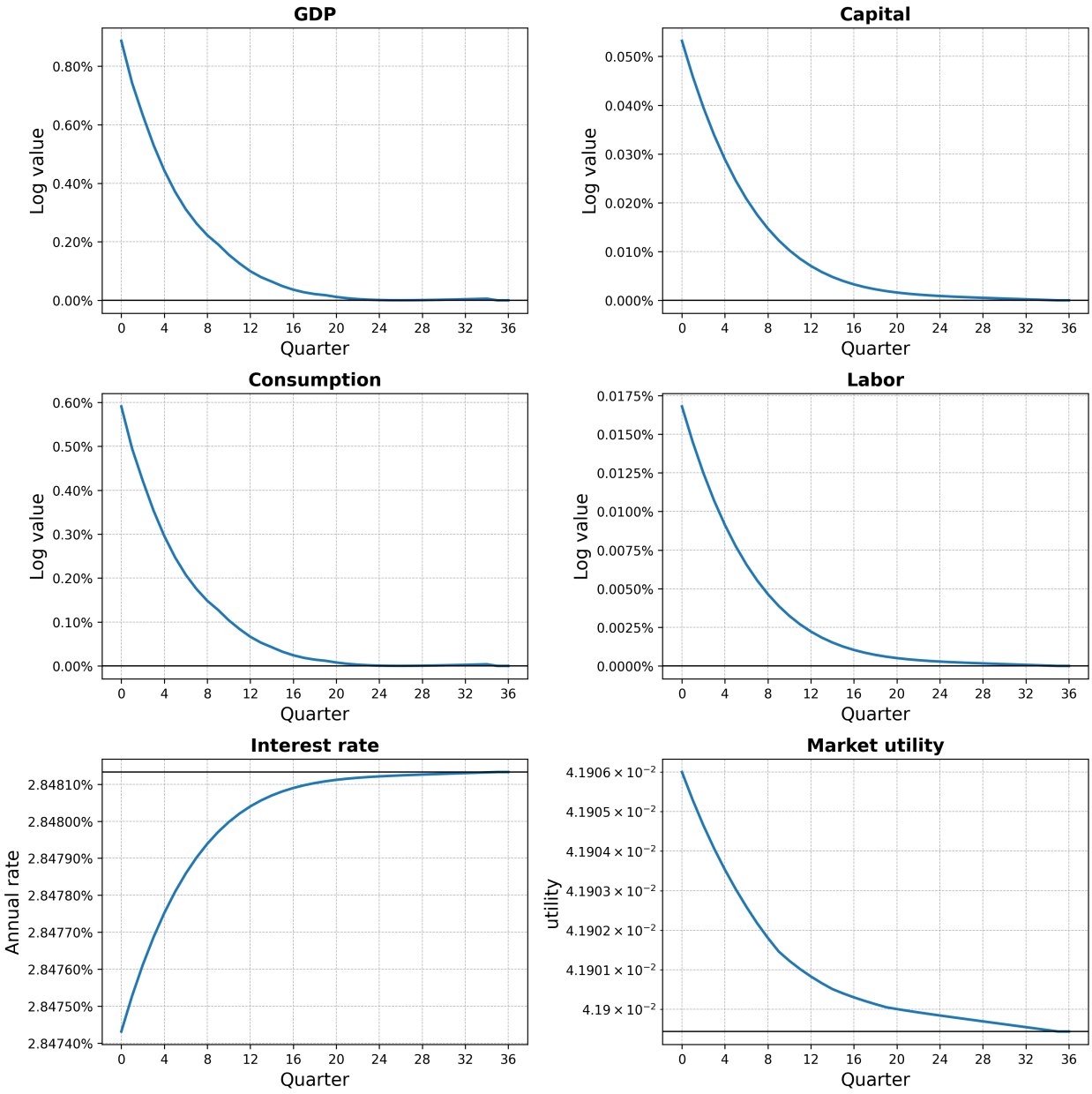

Figure 14: Model responses, other variables

The model responses to the sales shock. In the last graph, the market utility J is buyers' expected utility in the search market.

Appendix F Joint distributions

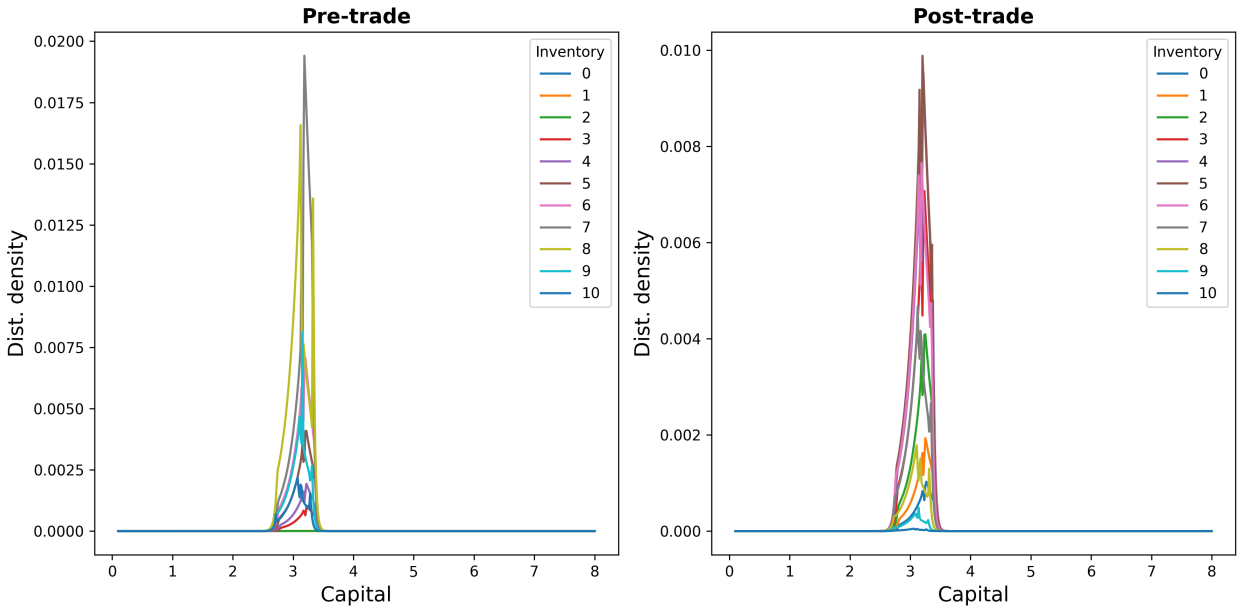

Figure 15: Joint distributions: steady state