Market Liquidity and Inventory Cycle

Ziqi Qiao *

October 15, 2020

Click here for the latest version

Abstract

Inventory behavior is an important part of business cycle. The structural VAR shows that inventory investment is pro-cyclical, the inventory-sales ratio is counter-cyclical and aggregate markup is pro-cyclical. However, the existing explanations for inventory holdings can’t adequately capture these empirical regularities. To fill the gap, this paper rationalizes inventory holding via search friction and explains its cyclical behavior by market liquidity – the trade-off between markup and selling speed. In the proposed model, sellers stock goods and post prices, while buyers choose which sellers to visit. Due to the lack of coordination, the sales are stochastic and there is leftover inventory. The frictional trade reveals a new incentive to hold inventory. Sellers compete in not only pricing but also buyers’ probability of consumption. As a result, carrying additional inventory allows sellers to post more profitable terms of trade. This new incentive has negative externality, such that sellers largely overstock in equilibrium. In contrast to the Walrasian framework, under this model, sellers can adjust prices in addition to quantities, so the quantity and price responses to shocks can fit the empirical findings.

*Department of Agricultural & Applied Economics, University of Wisconsin – Madison. Email: zqiao7@wisc.edu. I am grateful to Randall Wright for his invaluable guidance and Jean-Paul Chavas for his support. I would also like to thank Dean Corbae, Rasmus Lentz and Kenneth West for their helpful comments. All remaining errors are my own.
1 Introduction

Without understanding inventory behavior, we can’t understand business cycles. \(^1\) Inventory accounts for an average of 15% of GDP, representing significant resource allocation. Meanwhile, inventory investment highly correlates to business cycles and is as volatile as changes in output. Figure 1 plots log differences in four U.S. quarterly series: GDP, inventory, sales, and the inventory sales (I/S) ratio. As shown in the top panel, GDP and inventory comove closely, suggesting a strong correlation between the growth rate of the economy and inventory investment. On the other hand, inventory investment doesn’t keep up with sales growth: sales and the I/S ratio move in the opposite direction, as shown in the bottom panel. Given this long-term regularity, it’s hard to ignore inventory behaviors significance in business cycles. Nonetheless, few contemporary studies of business cycles incorporate inventory.

Figure 1: U.S. quarterly series, log difference

This paper documents inventory cycles via a structural VAR to demonstrate that (1) inventory investment is pro-cyclical, (2) the inventory sales ratio is counter-cyclical and persistent, and (3) the aggregate markup is pro-cyclical. It then proposes a model to rationalize inventory holdings via search friction. After calibrating the parameters to the long-term average moments, we see that the model produces similar impulse responses to the structural VAR responses.

Rationalizing inventory holdings isn’t trivial. The literature contains three classes of models, none of which can effectively predict the empirical behaviors of inventory. The first

\(^1\) The literature has long recognized inventory as important for business cycles. (Metzler, 1941; Blinder & Maccini, 1991; Ramey & West, 1999)
class models inventory as a consequence of production smoothing, due to volatile demand or cost. In the volatile demand case, the models predict that production will be less variable than sales, the opposite of what the data suggests. (Blinder, 1986). The volatile cost explanation (West, 1990; Bils & Kahn, 2000; Kryvtsov & Midrigan, 2012) requires markup to be counter-cyclical, but other empirical data suggests that the markup is pro-cyclical (Nekarda & Ramey, 2013). The second model class explains inventory by the stock-out avoidance motive (Kahn, 1987; Coen-Pirani, 2004). This explanation predicts that inventories fall when sales rise and rise when sales fall, which contradicts the observed pro-cyclical inventory investment. The third class considers a fixed restock cost, such that firms store materials for production (Arrow et al., 1951; Haltiwanger & Maccini, 1988; Khan & Thomas, 2007). In such models, inventory goes down gradually but jumps back up when it’s below a certain threshold. The persistence of the I/S ratio in the data doesn’t favor this explanation. Moreover, it only accounts for intermediate inventory and is silent on the final goods inventory.

Aside from the explicit explanations given for inventory holdings, the broader literature also considers inventory in reduced form. Kydland & Prescott (1982) and Christiano (1988) modeled inventory as a production factor, while Wen (2011) and Auernheimer & Trupkin (2014) put inventory in terms of utility. Nonetheless, the empirical performance of these models isn’t fully satisfactory: they often fail to predict correct signs and the great persistence of inventory-related responses.

To rationalize inventory holdings, I embed search friction (Burdett et al., 2001) in the heterogeneous-agent general equilibrium framework (Aiyagari, 1994). In the model, sellers post prices and inventories, while buyers visit sellers randomly due to the lack of coordination. It follows that the sales are stochastic to sellers, but in an endogenous way – sellers can choose to post a lower price to attract more buyers and increase the probability of sales. In line with the literature, sellers still have the production smoothing and stock-out avoidance motives. Additionally, holding more inventory allows sellers to charge higher prices or attract more consumers. But holding inventory is costly, so market liquidity is a concern – sellers face a trade-off between markup and selling speed. In order to sell the inventory fast, they have to lower their prices. Since sales are stochastic, the sales risk also affects the stocking and pricing decisions. It follows that the seller’s capital holdings matter as a risk buffer: given the same inventory size, sellers with more capital can charge a higher price.

The proposed model is parsimonious in parameterization and captures the inventory behaviors through economic mechanisms. In a good business cycle, high aggregate capital lowers the interest rate so the sellers invest more in inventory, meaning inventory investment is pro-cyclical. Unlike in a Walrasian market where price is a given, sellers also respond
through pricing in addition to inventory, so the quantity response of inventory isn’t as big. It follows that the inventory-sales ratio is counter-cyclical while markup is pro-cyclical. Given the concern of market liquidity, it’s optimal for sellers to charge a higher price and hold inventories longer. This leads to the high persistence of the responses compared to the standard models.

In addition to explaining the inventory cycle, the proposed model advances our understanding of economics in multiple ways. The paper adds to the literature on decentralized market models that study pricing and trading mechanisms. The sellers in the model face trade-offs between the posted price and speed of sales. This mechanism not only produces price dispersion but also permits fire sales, i.e. some posted prices are lower than the item’s production cost, which we observe in reality but which can’t happen in the standard models. Contributing to the search literature, this paper considers an environment of many-to-one matches and shows that the differentiation in the available quantities affects the equilibrium terms of trade. Finally, the stochastic process in the model is endogenous as the sellers choose the supports, e.g. how much inventory to hold, and the transition probabilities, e.g. what price to charge so a certain number of customers will show up. This provides a microfoundation for the idiosyncratic shocks in the heterogeneous-agent models (Huggett, 1993; Aiyagari, 1994) and produces different implications on agent mobility.

The rest of the paper is organized as follows. Section 2 documents the empirical behavior of aggregate inventory. Section 3 sets up the model and discusses the economic mechanisms. Section 4 calibrates the model and evaluates the model’s empirical performance. Section 5 concludes.

2 Empirical regularities

This section documents the empirical inventory behaviors in the sample. After giving some descriptive statistics, I compute a structural VAR model and plot the impulse responses.

The data are from the U.S. Bureau of Economic Analysis and U.S. Bureau of Labor Statistics. The data frequency is quarterly and the values are log-transformed. Total output $y$ is the real GDP, inventory $i$ is the real non-farm private inventory, and sales $s$ is the real final sales of domestic product. Aggregate markup $mk^*$ is calculated as the ratio of the

---


3Rogerson et al. (2005) surveyed labor search and Lagos et al. (2017) surveyed money searches.
Producer Price Index (PPI) to wages. I focus on the results using data from 1974 to 2007, and report the full sample results (1964 to 2020) along the side. I exclude data before 1974 because of the consensus on the structural break in 1974 (Ramey & West, 1999) and to improve comparability with previous studies. Along similar lines, the model introduced in the next section only considers a frictionless capital market without credit, so the financial crisis isn’t a good study object. Thus, I exclude data after the first quarter of 2007. Nonetheless, the results using the selected sample are very similar to the results from the full sample.

Table 1: Descriptive statistics, U.S. quarterly data

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>1974 - 2007</th>
<th>1964 - 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volatility</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma(\Delta y_t)$ GDP</td>
<td>0.008</td>
<td>0.010</td>
</tr>
<tr>
<td>$\sigma(\Delta s_t)/\sigma(\Delta y_t)$ sales to GDP</td>
<td>0.850</td>
<td>0.857</td>
</tr>
<tr>
<td>$\sigma(\Delta i_t)/\sigma(\Delta y_t)$ inventory to GDP</td>
<td>0.999</td>
<td>0.862</td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho(\Delta s_t, \Delta y_t)$ sales and GDP</td>
<td>0.750</td>
<td>0.883</td>
</tr>
<tr>
<td>$\rho(\Delta i_t, \Delta y_t)$ inventory and GDP</td>
<td>0.349</td>
<td>0.477</td>
</tr>
<tr>
<td>$\rho(\Delta i_t, \Delta s_t)$ inventory and sales</td>
<td>0.030</td>
<td>0.315</td>
</tr>
<tr>
<td>$\rho(i s_t, \Delta y_t)$ I/S ratio and GDP</td>
<td>-0.211</td>
<td>-0.106</td>
</tr>
<tr>
<td>$\rho(i s_t, is_{t-1})$ I/S autocorrelation</td>
<td>0.967</td>
<td>0.970</td>
</tr>
</tbody>
</table>

(a) All variables are in log values  
(b) $\Delta x_t = x_t - x_{t-1}$

Table 1 reports the volatilities and correlations of the core variables. The top panel compares the volatility of output, sales, and inventory. We see that inventory is about as volatile as output, while sales is 85% as volatile as output. This stylized fact rejects the production smoothing motive as the sole explanation for holding inventory. If agents held inventory only to smooth production against volatile sales, inventory investment would have to be less volatile than sales. The bottom panel of Table 1 presents the correlations. The correlation between inventory and output is 0.35 and the correlation between inventory and sales is 0.03. These positive correlations don’t support stock-out avoidance as the sole explanation for inventory holding. If the purpose of holding inventory is to avoid a stock-out situation, inventory should go down when sales go up. Moreover, the I/S ratio has a negative correlation of -0.21 to GDP growth. Although inventory investment is pro-cyclical,

\footnote{Specifically, PPI is the Producer Price Index by Commodity for Final Demand (Finished Goods). I weigh it by the Implicit Price Deflator for GDP to control for inflation. Wages data come from the Average Hourly Earnings of Production and Nonsupervisory Employees (Total Private).}

\footnote{The signs of the impulse responses are the same for both samples, but the magnitude and persistence of the impulse responses are greater when using the full sample. Appendix A reports the results.}
it doesn’t keep up with sales, which is puzzling in the standard models where agents only adjust quantities. Lastly, the autocorrelation of the I/S ratio is about 0.97, very persistent. Hence, it’s unlikely for the fixed restock cost to solely explain inventory holding. When facing a fixed cost to restock, agents restock inventory when it’s below a certain threshold and use it up over time, so the I/S ratio won’t be persistent. These raw moments provide various bivariate relations that don’t support the existing explanations of inventory holdings. However, these variables also influence each other beyond the bivariate relations. To see how they move together, I consider a structural VAR model below.

Figure 2: VAR, impulse responses to sales shock, 1974 – 2007

The three variables in the VAR are sales, inventory, and aggregate markup, as defined above. All the variables are unfiltered and in log values, given the concern of unit root. Since the data is quarterly, I use 8 lags and include both constant and trend. Using 12 lags produces similar results. The structural...
identification uses Cholesky decomposition. To investigate the impulse response to a sales shock, I order the variables by sales, inventory, and markup. The logic is that a sales shock immediately affects inventory size while the posted price and production cost are fixed for the period. Altering the order of inventory and markup doesn’t qualitatively change the impulse responses.

Figure 2 plots the impulse response of sales, inventory, I/S ratio, and aggregate markup to the sales shock for the sample of 1974 – 2007. The shaded areas depict the 95% confidence intervals. The signs of the responses present the cyclicality to be explained: (1) Inventory investment is pro-cyclical, (2) I/S ratio is counter-cyclical, and (3) aggregate markup is pro-cyclical. Moreover, the impulse responses are very persistent as they return to zero in around 9 years. The existing models have difficulty capturing either the signs of the movements or their persistence. The next section proposes a new model to rationalize inventory holding. I then compare the impulse responses from the model to the ones in Figure 2, to evaluate the model’s performance.

3 Model setup

This section introduces the dynamic general equilibrium model with heterogeneous agents and endogenous idiosyncratic risks.

In this model, time is discrete and continues forever. There are three types of agents – measure 1 of continuum capitalists, measure $\mu$ of continuum workers, and a representative firm. In each period, two markets operate sequentially: a Walrasian market (WM) and a search market (SM). The representative firm only plays an role in WM, while the capitalists and workers also make decisions in SM. Theoretically these two markets can operate simultaneously, as we don’t degenerate the distributions of capital and inventory. However, sequential operation greatly reduces the computational burden. To further simplify the problem, let the workers have linear utility so they don’t save and make only static decisions. This simplification eliminates the heterogeneity among workers and permits the application of market utility, a great help to the many-to-one match problem I introduce later.

The WM operates like a standard Walrasian general equilibrium. Specifically, the representative firm rents capital $K_t$ from the capitalists, hires labor $L_t$ from the workers, and

---

7The confidence intervals are noncumulative and calculated by bootstraps with 500 runs.
8To address the concern of cointegration between inventory and sales, I do a Johansen test and rerun the VAR with the constructed inventory-sales relation. The counter-cyclical persist, as shown in Appendix B.
produces outputs $Y_t$ with a time-invariant production technology $Y_t = f(K_t, L_t)$ that has a constant return to scale. The markets for $Y$, $K$, and $L$ are competitive, so the return to capital is $r_t = f_K(K_t, L_t)$ and the wage rate is $w_t = f_L(K_t, L_t)$. Since $f(\cdot)$ has a constant return to scale, the representative firm makes 0 profit in equilibrium. The capitalists and workers then purchase the WM outputs as the numeraire. The capitalists have a technology to convert the WM output $Y$ into the SM goods $x$. They don’t directly consume $x$ but can sell them to workers for WM output. Before entering SM, the capitalists make all the quantity decisions – consumption $c_t$, direct saving $\hat{k}$, and inventory, holding $\hat{x}$ for SM.

After all the quantity decisions in WM, the capitalists make the pricing decisions in SM. Due to a lack of coordination, the meetings between capitalists and workers are subject to search friction. The capitalists decide what price to charge, while taking into account the workers’ visiting responses. In the price posting problem, each capitalist posts a tuple $(\hat{x}, p, n)$, where $\hat{x}$ is the available inventory for sale, $p$ is the price, and $n$ is the buyer-seller ratio. After observing all the posts, each worker chooses which capitalist to visit. Note that at this point, the workers are still homogeneous such that market utility applies. Specifically, the ex ante utility of visiting each capitalist should be same across all workers. Denote market utility as $J$, an equilibrium object. Knowing the processes, we just need to work out the meeting and consumption probabilities to establish the dynamic programming problem.

An advancement of this model is the consideration of many-to-one matching – 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 goes to each seller with equal probability. It follows that the expected number of visits is $n$ for all sellers in submarket $(\hat{x}, p, n)$. Therefore, the probability $\pi_s$ of making sales $s$ follows a truncated Poisson distribution.

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

Note that the stock-out avoidance plays a role here, as the probability of sales is truncated beyond the inventory level. Assume the buyers have equal probability of consumption if the number of buyers exceeds the inventory holding. The probability $\alpha$ of a buyer getting to

---

9This has to be true in the unique non-coordinate equilibrium (Galenianos & Kircher, 2012).
consume the SM goods is then

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

$$= \sum_{i=0}^{\hat{x}-1} \frac{n^i e^{-n}}{i!} + \frac{\hat{x}}{n} \left( 1 - \sum_{i=0}^{\hat{x}} \frac{n^i e^{-n}}{i!} \right) \tag{2}$$

Note that a single buyer’s consumption probability calculates 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 preference and technology so we can define the equilibrium formally.

The representative firm has a Cobb-Douglas production function $Y = AK^\gamma L^{1-\gamma}$, where $A$ is the time-invariant TFP and $\gamma$ is the capital share. In WM, workers make a labor decision by

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

where $w$ is the wage rate, $l$ is the labor supply, $\epsilon$ measures the labor supply elasticity, and $\zeta$ alters the level of disutility from working. It’s straightforward that the workers’ optimal labor supply is $l^* = \left( \frac{w}{\zeta} \right)^{1/(\epsilon-1)}$ and their WM income is $wl^*$. In SM, workers get utility $\eta$ if they consume $x$ and gain linear utility from consuming $Y$. Their optimization problem entails deciding which submarket $(\hat{x}, p, n)$ to search. Workers consume all remaining income since they don’t save. Their market utility can be then written as

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

The capitalists’ problem is dynamic with two state variables – capital $k$ and inventory $x$. In WM, they supply capital and earn $rk$. They then choose their consumption $c$ and investment in capital $\hat{k}$ and inventory $\hat{x}$ that can be sold immediately. In SM, they post the price $p$ and buyer-seller ratio $n$, along with the inventory size $\hat{x}$. Assume free disposal and let the cost of producing $\hat{x}$ given $x$ be $\max\{\hat{x} - x, 0\}^{\kappa}/a$. Note that for $\kappa > 1$, the production cost of SM goods is convex and the capitalists have motives to smooth production. Furthermore, each unit of inventory costs $\delta/a$. Taking into account the workers’ expected utility $J$, we
can write the capitalists’ SM value $V(\cdot)$ function as

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

s.t. $J = \alpha(\hat{x}, n)(\eta - p) + w^*$

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

where $W(\cdot)$ is the WM value function calculated below.

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

s.t. $c + \hat{k} + \frac{\max\{\hat{x} - x, 0\}^+ + \delta x}{a} = (1 + r)k$

Note that the posted price has two potential upper bounds – workers’ utility $\eta$ from consuming SM goods and their budget constraint $w^*$. With the dynamic programming problem established, we are ready to define the equilibrium formally.

**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) and (7).
2. **Clearing:** $r = f_K(K, \mu^*)$, $w = f_L(K, \mu^*)$, $K = \int k \, dF_{K,X}$ and $\mu = \int n^* \, dF_{K,X}$.
3. **Stationarity:** $F_{K,X}(k, x) = \int \sum_{s=0}^{\hat{s}} \pi_s(\hat{x}, n^*) 1\{\hat{k} + sp^* \leq k\} \cdot 1\{\hat{x} - s \leq x\} \, dF_{K,X}$

Note that when the SM production is costly, the SM can shut down so there’s no inventory. For our purposes, we only consider the parameter range when SM opens. Different from the standard heterogeneous-agent framework, the stochastic process here is endogenous, i.e., the sellers choose the supports – how much inventory to hold – and the transition probability – what price to charge and the mean sales. This process isn’t necessarily ergodic, so the standard proofs for the uniqueness of stationary distribution don’t apply here. The numerical exercise does show convergence, while the uniqueness isn’t guaranteed as different initial distributions of capital can lead to a different stationary equilibrium. I impose additional parametric restrictions on the initial distribution and discipline the calibration by wealth distribution. Details are in the next section.

Before calibrating the model, we should review the new economics in the model. The en-
dogenous stochastic process marks the trade-off between the probability of sales and markup. Hence, the capitalists face market liquidity when making decisions – to sell fast, they have to lower prices. This can be formally captured by the tightness elasticity of price derived from (6):

$$\frac{d \log p}{d \log n}(J,w) = \frac{\alpha_n(\hat{x}, n)n}{\alpha(\hat{x}, n)} \left( \frac{\eta}{p} - 1 \right) < 0$$

(8)

This elasticity can be grouped into two parts. The first part $\frac{\alpha_n(\hat{x}, n)n}{\alpha(\hat{x}, n)}$ is the elasticity of the consumption probability, while the second part $\left( \frac{\eta}{p} - 1 \right)$ is the consumer surplus. When the chance of consumption is very elastic or when the surplus is high, capitalists have to reduce prices drastically to attract more consumers and liquidate inventories fast. A new incentive to hold inventory appears, as the elasticity of the consumption probability depends on the inventory holding.

**Proposition 1.** In the case of overstock, $\hat{x} \geq n$, holding more inventory improves market liquidity.

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

(9)

**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 additional incentive to hold inventory which allows them to post more profitable terms of trade. Note that this incentive happens in the overstock case $\hat{x} \geq n$, where the inventory is greater than the expected number of consumers. In the understock case $\hat{x} < n$, the sign can be ambiguous. Nonetheless, the calibrated model has the sellers overstocking in equilibrium, which also fits our daily observation.

On the other hand, the concern of market liquidity diminishes when prices are higher, so the capitalists who can charge high prices don’t worry much about market liquidity at the margin. It follows that the capitalists holding more capital can tolerate a higher risk of low sales and post a higher price, which reinforces their profitability through the market liquidity mechanism.

Lastly, this mechanism implies the cyclical behavior of inventory. When a positive...
sales shock happens, the optimal responses involve charging higher prices, instead of just producing more units to drive down the price. Therefore, the I/S ratio is counter-cyclical while the markup is pro-cyclical. Moreover, given the margin of price adjustment, the quantity adjustments take longer to return to the steady state level. The persistence of the impulse responses emerges as a consequence. The standard Walrasian framework can’t imply these results since quantity adjustments are the only choices. With these intuitions in mind, we now turn to model calibration.

4 Empirical performance

I first describe the numerical algorithm used to find a stationary equilibrium. After establishing parametric restrictions on the initial capital distribution, I then calibrate the model parameters to match the long-run average moments. Reports on steady state and dynamic responses follow.

4.1 Algorithm

The numerical algorithm benefits from the sequential solutions of (9) and (10). We define a grid on the space of state variables \((k, x)\), where \(k\) is continuous and \(x\) is discrete by nature. The computation loop is as follows.

(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 $\hat{\mu} = \int n^* dF$. If $\hat{\mu} > \mu$ ($\hat{\mu} < \mu$), increase (decrease) $J^0$ and return to (II). If $\hat{\mu}$ and $\mu$ 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.

4.2 Calibration

One concern of the model solution is that different initial distributions might converge to different stationary equilibria. Though different initial distributions don’t affect the impulse responses qualitatively, the non-uniqueness adds ad-hocness to the calibration. To deal with this concern, I impose some parametric restrictions on the initial distribution of capital. Inspired by the shape of U.S. wealth distribution, I use a Pareto-like distribution with the scale parameter $k_m$ and shape parameter $\theta$. The exact distribution cannot be Pareto, because the support of capital has to be finite for computation. I truncate the distribution at the upper limit of the capital grid. For a large enough upper bound, the results don’t change with the upper bound. In the calibration, the two parameters are calibrated to match the moments of wealth inequality.

Table 2 reports the calibrated parameters and compares the model moments to the target moments. The targets in the top panel don’t involve any equilibrium objects. I normalize the TFP $A$ to 0.05 and choose a preference parameter $\sigma$ of 2.0, a typical value in the business cycle literature. $\gamma$ and $\epsilon$ are chosen to directly match the capital share and labor supply elasticity explicitly. The parameters in the middle panel target equilibrium objects and require a joint search. The model can match the equilibrium objects to data moments quite well. The bottom panel compares the wealth inequalities from the model and the data (Quadrini & Rios-Rull, 1997). Unsurprisingly, the model produces a much more equal distribution than the data suggests. In the data, the mode of wealth distribution is close to zero. However, in the model, the agents can’t borrow and have to save for economic activities, so their wealth levels move away from zero in the steady state. Without a credit market, it’s hard to match the inequality we observe in the data. Nonetheless, the calibrated model preserves the general shape of wealth distribution. We discuss the steady-state results next.

---

10The equilibria are not dense, so a marginal change in the initial distribution doesn’t affect the outcome.

11The magnitude of aggregate markup is still understudied. Traditionally, the estimation relies on market power (Berry et al., 1995; De Loecker et al., 2020). My model has a competitive environment, so the search friction generates the markup. I use the traditional estimates as the data targets for now.
Table 2: Model calibration

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target</th>
<th>Data</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Non-equilibrium object</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>$0.05$</td>
<td>normalization</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>$2.0$</td>
<td>risk aversion</td>
<td>-</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$0.33$</td>
<td>capital share</td>
<td>$0.3 - 0.4$</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>$1.53$</td>
<td>labor supply elasticity</td>
<td>$1.90$</td>
</tr>
<tr>
<td></td>
<td><strong>Equilibrium object</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>$0.992$</td>
<td>annual interest rate</td>
<td>$0.03$</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>$0.07$</td>
<td>labor hour</td>
<td>$0.51$</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>$1.6$</td>
<td>inventory/GDP</td>
<td>$0.14$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>$0.1$</td>
<td>sales/GDP</td>
<td>$0.99$</td>
</tr>
<tr>
<td>$\eta$</td>
<td>$3.2$</td>
<td>consumption/GDP</td>
<td>$0.65$</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$2.8$</td>
<td>wage/GDP</td>
<td>$0.46$</td>
</tr>
<tr>
<td>$a$</td>
<td>$269$</td>
<td>markup</td>
<td>$1.2 - 2.3$</td>
</tr>
<tr>
<td></td>
<td><strong>Wealth distribution</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_m$</td>
<td>$4.8$</td>
<td>top 20% share</td>
<td>$0.80$</td>
</tr>
<tr>
<td>$\theta$</td>
<td>$4.0$</td>
<td>Gini Index</td>
<td>$0.78$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>coefficient of variation</td>
<td>$6.09$</td>
</tr>
</tbody>
</table>

4.3 Steady-state results

For the calibrated parameters, we can visualize the steady state. As shown in Figure 3, the value functions are increasing in both capital and inventory, given the free disposal. Figure 4 plots the CM policy functions. In the left panel, the policy functions of inventory are increasing the capital holdings, so larger capitalists produce more SM goods. In the right panel, the policy functions of capital aren’t monotonically increasing, since the capitalists invest in inventory once they accumulate enough capital.

What’s interesting is the policy functions of price posting, displayed in Figure 5. The price dispersion naturally appears. At different inventory levels the customers have different probabilities of consumption, so the posted prices have to vary. On the other hand, given the same inventory, the more capital one holds, the higher price one can post. These predictions are empirically relevant but can’t be generated by the standard Walrasian model with uniform pricing.

The right panel of Figure 5 presents the corresponding buyer-seller ratio. The dash line has a slope of 0.5, depicting the threshold where the inventory is enough to serve twice
the expected level of customers. The first thing to notice is that all sellers overstock— the chosen buyer-seller ratios are well below the inventory level. Moreover, most of the sellers hold enough inventory to serve more than double the expected buyers. We see two layers of inefficiency here. First, due to the lack of coordination on the buyer side, the sellers face
stochastic sales and have to hold inventory. In the meantime, when a seller stocks more inventory she improves her buyers’ shopping experience, which has a negative externality on other sellers through the channel of market utility. As a result, we not only observe inventory but also a large degree of overstock.

Figure 6: Stationary distribution

![Image](image_url)

Figure 6 plots the stationary distributions for inventory and capital. Despite missing the inequality measures, the capital distribution still preserves a Pareto-like shape, especially the long tail. In a typical heterogeneous-agent environment, the capitalists don’t have an incentive to hold too much capital, given the interest rate and uniform risk exposure. By contrast, in the proposed model, risk exposure is endogenous and hence heterogeneous across the distribution. Holding large capital serves as a risk buffer, so the capitalists can post higher prices. As discussed in the model section, a higher price reduces consumers’ surplus so the market liquidity becomes better, which dampens the pricing effect on sales reduction and improves profitability. The marginal benefit of holding more capital diminishes much slower than the case with uniform risk exposure. As a result, the equilibrium can have a long tail of wealth distribution.

4.4 Impulse responses

The goal here is to compare the impulse responses from the model to those from the data (Figure 2). To induce a sales shock, I inject a one-time shock to $\eta$, the workers’ preference for the SM goods. The timing of the shock is after the capitalists’ price posting but before the workers’ search. Thus, the posted prices are fixed, but we need to solve for the new market utility to shock $J_s$ so the average buyer-seller ratio equals the worker-capitalist ratio.
μ. To solve for the impulse responses, we first guess two paths of interest rate $r_t$ and market utility $J_t$, then backward-compute all the optimal responses ($\hat{k}_t, \hat{x}_t, n^*_t, p^*_t$). We can then use the after-shock distribution of the buyer-seller ratio to forward shoot capital-inventory distribution, and look for convergence in $J_t$ and $r_t$.

Figure 7: Model responses

Figure 7 compares the results. The impulse responses from the models match the ones from the VAR reasonably well. The directions of the initial movements are the same and the responses are very persistent. The magnitudes of the changes are not exact. For the first two periods, the sales drop more than the data suggests, while inventory accumulates faster. Hence the I/S ratio bounces back more quickly. On the other hand, instead of growing gradually, the markup jumps at the initial periods and returns to a steady level quickly. All in all, given the parsimonious parameterization, the model can predict the directions of quantity and price simultaneously. This highlights the importance of price adjustment, an
additional margin to choose so the quantity doesn’t respond as much.

5 Conclusive remarks

Despite the known importance of inventory in the business cycle, the literature typically overlooks inventory in its analysis. This makes it difficult to address two key questions: why do sellers hold inventory, and what motives can explain their inventory behavior? This paper rationalizes inventory holding via search friction and explains the empirical regularities of inventory through market liquidity, the trade-off between markup and the speed of sales. In the calibrated equilibrium, the sellers hold enough inventory for more than twice the number of expected buyers, because holding more inventory increases buyers’ shopping experience, allowing the sellers to charge higher prices. This new motive for holding inventory helps explain the inventory cycle. When sales increase, sellers not only stock more inventory but also charge higher prices. It follows that the quantity responses aren’t as large as a case in which sellers can’t adjust prices. Therefore, the inventory-sales ratio is counter-cyclical, markup is pro-cyclical, and the quantity responses are persistent.

Getting the correct signs for quantity and price simultaneously is not trivial. For example, the New Keynesian framework typically relies on nominal rigidity as a key transmission mechanism, which often implies counter-cyclical markup. However, this paper also accomplishes the sign predictions parsimoniously. The proposed model doesn’t rely on any shock sequences to alter the directions of variable movements, in contrast to published models that rely on additional degrees of freedom. Being correct and parsimonious builds confidence in the model mechanism, which highlights the importance of price margins. The insight helps us understand market activities better. It seems promising to extend the study scope to general business cycles. Beside this direct extension, the model has two other aspects that are worth exploring.

The calibration strongly calls for the inclusion of a credit market. The predicted wealth inequality is highly insignificant compared to the data. In the model, the capitalists have to accumulate capital themselves, so the wealth distribution moves away from 0. In the data, wealth distribution spikes at 0 as people rely on credit for activities. To match the wealth inequality in the data, adding a credit market is necessary. Meanwhile, the endogenous idiosyncratic risk can be useful for inequality studies. In the heterogeneous-agent literature, a long tail in wealth distribution is hard to obtain, because the idiosyncratic risks are uniform across the wealth levels. In this paper, however, the agents adjust their risk exposure
by choosing prices and inventories. As a result, the marginal benefit of holding capital di-
minishes slower and the equilibrium can sustain a long tail. Including a credit market is
not only an improvement to the calibration but also provides a micro-founded approach to
understanding wealth inequality.

Lastly, this paper provides a way to model many-to-one matching. Many-to-one match-
ing is commonly studied in the network literature, but in a different style that is not directly
applicable to the general equilibrium setting. In the labor search literature the matching is
largely one-to-one, where vacancies are independent and there is no definition of the firm.
Modeling many-to-one matching can make a meaningful difference. Firms often post multi-
ple vacancies that can be substitutes or complements. The inventory model is a special case
of perfect substitutes. In this case, the capacity not only affects the quantity of meetings but
also the terms of trade. Hence, the implication for the labor market is to build connections
between the matching rate and wage rate. Understanding price and quantity altogether
requires future research.
References


Chang, Briana, & Zhang, Shengxing. 2018. Endogenous market making and network formation. *Available at SSRN 2600242*.


Appendix A  VAR using full sample

Figure 8: VAR, impulse responses to sales shock, 1964 – 2020

This section reports the VAR results from the full sample. Figure 8 plots the results where the shaded areas represent the 98% 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. However, the responses have greater magnitude and persistence. 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 \( \theta \) and test the cointegration, I use Johansen procedure with eigenvalue method and long-run VECM for error correction. Table 3 reports the results.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\theta} )</td>
<td>Test stat.</td>
<td>( \theta )</td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
<td>17.04</td>
<td>0.88</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
<td>11.14</td>
<td>0.89</td>
</tr>
<tr>
<td>8</td>
<td>0.85</td>
<td>4.99</td>
<td>0.90</td>
</tr>
</tbody>
</table>

(a) Test with eigenvalue  (b) Error correction: long-run VECM

\( \hat{\theta} \) is around 0.86 for sample 1974 – 2007 and 0.89 for the full sample, robust to the lag lengths from 2 to 8. Using the selected sample, we reject the null of no cointegration at 0.05 level for in the lag 2 case. For the full sample, we reject the null at 0.05 level in all the lag cases. These results suggest that inventory and sales are cointegrated. We then construct the inventory-sales (I/S) relations using the estimated \( \hat{\theta} \) for the two samples and investigate the corresponding impulse responses.

Figure 9 and 10 plot the impulse responses for the two samples. After teasing out the cointegration, the stylized fact still holds that the inventory-sales ratio is counter-cyclical.
Figure 9: VAR, impulse responses to sales shock, 1974 – 2007

Figure 10: VAR, impulse responses to sales shock, 1964 – 2020
Appendix C  Proof of Proposition 1

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

\[
\alpha(\hat{x}, n) = \sum_{i=0}^{\hat{x}-1} \frac{n^i e^{-n}}{i!} + \sum_{i=\hat{x}}^\infty \frac{n^i e^{-n}}{i!} \frac{\hat{x}}{i+1} \tag{10}
\]

we can calculate the derivative

\[
\alpha_n(\hat{x}, n) = \sum_{i=0}^{\hat{x}-1} \frac{i^n e^{-n} - n^i e^{-n}}{i!} + \sum_{i=\hat{x}}^\infty \frac{i^n e^{-n} - n^i e^{-n}}{i!} \frac{\hat{x}}{i+1} - \alpha(\hat{x}, n) \\
\equiv g(\hat{x}, n) - \alpha(\hat{x}, n) < 0 \tag{11}
\]

Note that

\[
\alpha(\hat{x} + 1, n) - \alpha(\hat{x}, n) = \frac{n^\hat{x} e^{-n}}{\hat{x}!} - \frac{n^\hat{x} e^{-n}}{\hat{x}!} \frac{\hat{x}}{\hat{x} + 1} > 0 \tag{12}
\]

\[
g(\hat{x} + 1, n) - g(\hat{x}, n) = \frac{\hat{x}}{n} [\alpha(\hat{x} + 1, n) - \alpha(\hat{x}, n)] \tag{13}
\]

It follows that

\[
|\alpha_n(\hat{x} + 1, n)| - |\alpha_n(\hat{x}, n)| = [\alpha(\hat{x} + 1, n) - \alpha(\hat{x}, n)] \left(1 - \frac{\hat{x}}{n}\right) \begin{cases} 
> 0 & \text{if } \hat{x} < n \\
= 0 & \text{if } \hat{x} = n \\
< 0 & \text{if } \hat{x} > n
\end{cases} \tag{14}
\]

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

\[
\left|\frac{\alpha_n(\hat{x} + 1, n) n}{\alpha(\hat{x} + 1, n)}\right| < \left|\frac{\alpha_n(\hat{x}, n) n}{\alpha(\hat{x}, n)}\right| \tag{14}
\]