Notes on
Firm Dynamics

Pau Roldan*
New York University

Abstract
These are notes that I took in the class Advanced Topics in Macro (ECON-GB.3332) at NYU Stern, taught by Gian Luca Clementi in the Spring semester of 2016. Please be aware that typos and errors are possible, and are my sole responsibility and not that of the instructor. Please use them at your own discretion.

Contents
1 Introduction ........................................... 2
  1.1 Data sources ....................................... 2
  1.2 Stylized facts on production heterogeneities in firms 2
2 Productivity and Firm Dynamics ....................... 4
  2.1 Hopenhayn (ECMA, 1992) .......................... 4
  2.2 Clementi and Palazzo (AEJ: Macro, 2016) .......... 9
3 Financial Frictions and Firm Dynamics ............... 12
  3.1 Cooley and Quadrini (AER, 2001) ................. 13
  3.2 Clementi and Hopenhayn (QJE, 2006) ............. 14
  3.3 Albuquerque and Hopenhayn (REStud, 2004) ...... 21
4 Investment and Firm Dynamics ....................... 23
  4.1 Caballero, Engel and Haltiwanger (BPEA, 1995) ...... 23
  4.2 Caballero and Engel (ECMA, 1999) ............... 25
  4.3 Kahn and Thomas (ECMA, 2008) ........................ 31
  4.4 Extensions of Khan and Thomas (2008) .......... 36
5 Endogenous Growth and Firm Dynamics ............. 42
  5.1 Klette and Kortum (JPE, 2004) .................. 42
6 Other Topics ......................................... 51
  6.1 Uncertainty Shocks and Firm Dynamics ............. 52
  6.2 Credit Shocks and Firm Dynamics .................. 54
  6.3 Asset Pricing and Firm Dynamics .................. 60

*Department of Economics, New York University, 19 West 4th Street, 6th Floor, New York, NY 100012. E-mail: pau.roldan@nyu.edu
1 Introduction

We start by describing the main data sources for firm dynamics, and the main empirical challenges that the firm dynamics literature has traditionally faced, both at the macro and the micro level.

1.1 Data sources

- In the literature, the unit of analysis is establishments (or plants) and firms (or enterprises). This is an important distinction in IO but macro models are silent on it. In the data, however, they are different (in entry rates, etc).

- In the U.S., the main data source at the micro level if from the Census Bureau. Their main databases are the Longitudinal Research Database (LRD) and the Longitudinal Business Database (LBD). The unit for the Census is the establishment, not the firm. Both cover the universe of establishments. The LRD is very detailed, covering only manufacturing. The LBD is for non-firm establishments and the information included is more limited (e.g, unlike the LRD, which can’t track the ownership structure). The LBD includes only age, size and location.

- Accessing it is very difficult, but a somewhat more aggregated subsample called BDS is publicly available.

- Other datasets are the Business Employment Dynamics (BDM) at the US Bureau of Labor Statistics, including firms filing to the Unemployment Insurance program, at the quarterly frequency (relatively unused data); and Compustat, from Standard and Poor’s, coming from filings that firms make with the Security and Exchange Commission, thus including only publicly traded firms. Compustat is very detailed but it is accounting data, so it is subject to big measurement error (e.g, depreciation for accountants is not the same as what economists understand by depreciation). In this sense, the Census data is more reliable because it is confidential. Moreover, Compustat has a selection issue, as it includes only relatively big and old firms. All firms in Compustat are in the LRD, so it’s good to bridge data with LRD (e.g, through BDS).

- Another option is always to use foreign data: for instance, the Encuesta Nacional Industrial Anual for Chile; the Colombian Annual Manufacturing Survey for Colombia; the Quadros do Pessoal for Portugal; or USTAN for Germany.

1.2 Stylized facts on production heterogeneities in firms

- Size distribution (from Cabral and Mata, AER, 2004):
  - Distribution of log employment has remained constant over time. Distribution is right-skewed, so size is far from being log-normal. This invalidates early theories that argued that size is log-normal and Gibrat’s law holds.
  - The size distribution of a typical cohort becomes less and less right-skewed as it ages. One reason is selection out of exit: fast growing firms are the only ones that survive. The other is that growth rates are greater for small firms (heterogeneity in firm growth). Cabral and Mata argue that selection is not really the relevant channel, but most differences

\[^4\] In the US the entry rate went down but those that were born remained of the same size (firm cohorts have aged).
must be coming from growth heterogeneities across sizes. One hypothesis is the following: small firms might be small by choice or because there are financing constraints that prevent them from growing larger which might become less tightening as firms get older.

- In the GR, small young firms were hit the most. Typically they grow faster than large older ones. But this difference in growth rates between these two groups shrunk a lot during the GR.

• Employment hares by Size and Age:
  - Fort, Haltiwanger, Jarmin and Miranda (IMF, 2013): About 50% employment share held by firms that are old and large, although there are only a few large firms. There are a lot of small firms but hold a small share of employment.

• Entry and Exit:
  - Dunne, Roberts and Samuelson (RAND, 1988): Defined entry and exit rates empirically for establishments. Using LBD (using BDS), we know entry rate is on average greater than exit (the opposite in manufacturing), and both rates have drifted down, with quite a bit of variation and high cyclicality (e.g. see Great Recession). Data has some noise though (5-year seasonal component because data is compiled every 5 years).
  - Entry rates tend to be large (share of entrants is about 10-15% of total); they are quite similar across countries; have been declining in the US (Pugsley and Sahin, WP, 2014); entry (exit) is pro-(counter-) cyclical; entrants tend to be smaller than incumbents (in LBD about half the size). An unanswered question is why new firms enter small. One explanation is technical efficiency (as Hopenhayn models would imply), same reason why small firms grow faster. One hypothesis is that entering firms are not less efficient than incumbents, but they receive stronger idiosyncratic demand shocks (perhaps because they haven’t built a reputation).
  - Survival rates (probability of no exit conditional on having survived so far) have gone up due to the decrease in exit rates across firm age (measured by job destruction). Exit rates wrt job destruction have also declined in firm size, both conditionally and unconditionally.

• Firm growth:
  - Growth can be measured as employment growth at establishment, firm or aggregate level. Careful: growth rate accounts for so-called “reversion to the mean” bias, accounting for transitory shocks that momentarily make firms appear smaller or bigger than they actually are. Where i is establishment and s is firm,

\[
g_{it} = \frac{E_{it} - E_{it-1}}{X_{it}}
\]

where \(X_{it} = \frac{E_{it} + E_{it-1}}{2}\), and firm’s growth rate is \(g_{st} = \sum_{i \in s} \frac{X_{it}}{X_{it}} g_{it}\). [see slides to check formulas].

- Job reallocation is reshuffling in employment between small and big firms, equal to the sum of job creation (jobs added across growing establishments) and destruction (jobs lost across shrinking establishments). Excess reallocation is defined as reallocation net of employment growth.
- Data (LBD): growth is declining in size if we don’t take the reversion to the mean bias into account. Taking into account average size of firms this relationship disappears. Con-
trolling for age, in fact smaller firms grow slower. But the previously found decreasing relationship between size and growth was actually capturing mostly young firms. Conditional on survival, small firms do grow faster. However, unconditionally, small firms don’t necessarily grow faster. The unconditional positive correlation between growth and size found in the past was almost all driven by age.

– Young smaller firms’ growth rates plunged in the GR more than other firms.
– In the US, excess reallocation (creation plus destruction minus employment growth rates) is massive and mostly driven by younger firms. In general, continuing establishment create and destroy most jobs, following by new establishments of existing firms, and new firms. Creating and destruction rates are relatively very low for older firms.

• Investment rates:
  – Doms and Dunner (RED, 1998): A lot of lumpiness at the establishment level. About 50% of establishments have growth rates of zero or below. Weighted by investment, the fraction of plants that invest more than 25% is substantial. Thus, there is a big share of firms that are inactive, and a substantial number that increase their capital stock in a high degree.
  – Moreover, more than half of the firms had almost 50% growth rate or higher in the year in which they invested the most. So most firms concentrate most of their investment in a small period of time, and then go through long periods of time in which they do not invest. Therefore, there is a lot of lumpiness in investment. Models with convex adjustment costs are rejected by this evidence, and the theoretical literature has tended to using fixed adjustment costs (e.g, (s,S) models). However, for multi-plant firms there is little lumpiness. I.e, lumpiness occurs at the establishment level, not the firm level.

2 Productivity and Firm Dynamics

This section analyzes a first set of models linking firm growth to size and age. In these models, productivity is the proxy for firm size. We will start with the seminal paper by Hopenhayn (ECMA, 1992), and follow-up work by Clementi and Palazzo (AEJ: Macro, 2016) incorporating a more meaningful role for age.

2.1 Hopenhayn (ECMA, 1992)

This paper is the first technical attempt to formally model firm dynamics. The theory is simple, there is no capital accumulation, and only one state variable. There is no role for age, and the model is about size.

Environment

• Time is discrete and infinite.
• There is a continuum of firms (LLN applies), with common and time-invariant discount factor \( \beta \in (0, 1) \).
• Firms produce homogenous products (same product all firms) with one input only (e.g, labor or any input that can be adjusted instantaneously).
• $p$ denotes the product price, and $w$ denotes the input price.

• Firms are price-takers in the two markets in which they participate: product and labor market.

• Aggregate (inverse) demand is exogenously given, $D(Q)$; being $N$ the aggregate labor, $w(N)$ is the exogenous labor demand.

**Assumption 1** The following regularity conditions apply:

1. $D$ is continuous, strictly decreasing, and $\lim_{x \to +\infty} D(x) = 0$.
2. $w$ is continuous, non-decreasing, and strictly bounded.

A firm’s output is $q = f(\varphi, n)$, where $\varphi \in S \equiv [0, 1]$ is a productivity shock, and $n$ is labor input. Typically, $f(\varphi, n) = e^{\varphi n^\alpha}$ with $\alpha \in (0, 1)$. The state $\varphi$ is Markov, with $\varphi \sim F(\varphi'|\varphi)$.

• There is a fixed cost $c_f > 0$, constant in time and the same across all firms. This ingredient is needed to create exit in equilibrium, and it will play the role of firm leverage. The exit decision will come after production, not ahead of time, so $c_f$ will not affect entry.

• Firm profit is denoted $\pi(\varphi, p, w)$, while quantity supplied and labor demand are denoted $q(\varphi, p, w)$ and $n(\varphi, p, w)$, respectively.

• Timing: for an incumbent firm, within a given period,

1. Shock $\varphi$ is observed.
2. Taking state $(\varphi, p, w)$ as given, firm decides on labor demand.
3. Firm receives profits $\pi - c_f$.
4. Given price beliefs and the shock’s expected transition, firm decides whether or not to exit.

• To close the environment, a few more assumptions:

  – First, a few technical constraints:

    **Assumption 2** The following regularity conditions apply:

    1. $q$ and $n$ are single-valued functions (not correspondences).
2. $\pi$ is continuous and strictly increasing in $\varphi$.

  – Second, there exist no discontinuities in productivity, and the higher the productivity, the higher the chance of having a good shock next period. The latter ensures, together with Assumption 2.2, that the value function of the problem is increasing in current productivity.

    **Assumption 3** Distribution of the shock:

    1. $F$ is continuous in both $\varphi$ and $\varphi'$.
2. $F$ is strictly decreasing in $\varphi$.

  – Third, there are no absorbing states in productivity. The assumption will ensure that almost surely the productivity process visits regions in which there is exit for some firms.

    **Assumption 4** For every $\epsilon > 0$, $\exists n \in \mathbb{Z}$ such that $F^n(\epsilon|\varphi) > 0$, where $F^n(\epsilon|\varphi)$ is the distribution of $\varphi_{t+n}$, given $\varphi_t = \varphi$.

2The Melitz (ECMA, 2003) model is a generalization of Hopenhayn (ECMA, 1992) to a monopolistically competitive market with time-variation in markups, which is absent here.
• Entry and exit:
  – Values of firm that exit is zero for all firms and time periods (a normalization for the outside value).
  – Entrants pay fixed cost \( c_e > 0 \), which is sunk. Upon entry, and only in the first period of operation, they draw a shock \( \varphi \sim \nu \). From then on, conditional on survival, shocks are drawn from \( F \).
  – Having entrants and incumbents draw productivity from different distributions allows us to derive a non-trivial firm size distribution.

**Equilibrium**

• Let \( \mu_t(S) \) be the measure of total size of the industry (total mass of firms in the industry), where \( S = [0, 1] \) is the support of the productivity shock. For every Borel set \( A \subseteq S \), \( \mu_t(A) \) is the mass of firms drawing some \( \varphi \in A \).

• The aggregate supply is

\[
Q^s(\mu_t, p_t, w_t) \equiv \int q(\varphi, p_t, w_t)\mu_t(d\varphi)
\]

and the aggregate demand of labor is

\[
N^d(\mu_t, p_t, w_t) \equiv \int n(\varphi, p_t, w_t)\mu_t(d\varphi)
\]

• Given an initial distribution of firms \( \mu_0 \) and a perfectly foreseeable sequence of prices \( z_t = \{ (p_t, w_t, \mu_t) \}_{t \in \mathbb{Z}_+} \), value functions for incumbents \( \{ v_t(\varphi, z_t) \}_{t \in \mathbb{Z}_+} \), entrants \( \{ v^e_t(z_t) \}_{t \in \mathbb{Z}_+} \), a mass of entrants \( \{ M_t \}_{t \in \mathbb{Z}_+} \), and exit thresholds \( \{ x_t(\varphi, z_t) \}_{t \in \mathbb{Z}_+} \) are such that:

1. The value function for incumbents solves

\[
v_t(\varphi, z_t) = \pi(\varphi, z_t) + \beta \max \left\{ 0, \int v_{t+1}(\varphi', z_{t+1})F(d\varphi' | \varphi) \right\}
\]

2. The exit decision is given by exit iff \( \varphi \leq x_t(\varphi, z_t) \), where

\[
x_t \equiv \inf \left\{ \varphi \in S : \int v_{t+1}(\varphi', z_{t+1})F(d\varphi' | \varphi) \geq 0 \right\}
\]

and \( x_t = 1 \) if the set is empty, in which case the industry dies.

3. The value function for entrants solves

\[
v^e_t(z_t) = \int v_t(\varphi, z_t)\nu(d\varphi)
\]

4. The free-entry condition reads

\[
v^e_t(z_t) \leq c_e
\]

so that \( M_t = 0 \) if \( v^e_t(z_t) < c_e \), and \( M_t > 0 \) otherwise.

5. The law of motion for the total mass of firms (incumbents and entrants) is

\[
\mu_{t+1}([0, \varphi']) = \int_{\varphi \geq x_t(\varphi, z_t)} F(\varphi')\nu_t(d\varphi) + M_{t+1}G(\varphi')
\]

\underbrace{\text{Surviving incumbents}}_{x_t(\varphi, z_t)} \underbrace{\text{Entrants}}_{M_{t+1}}
for any \([0, \varphi'] \subseteq S\), where \(G\) is the cdf corresponding to the \(\nu\) measure.

- Alternatively, to define transitions for incumbents more compactly, we can define the operator \(\hat{P}\) such that

\[
\hat{P}_t(\varphi, A) = \begin{cases} 
\int_A F(d\varphi|\varphi) & \text{if } \varphi \geq x_t(\varphi, z_t) \\
0 & \text{otherwise}
\end{cases}
\]

and then define the law of motion of the measure of firms by

\[
\mu_{t+1} = \hat{P}_t \mu_t + M_{t+1} \nu
\]

where \(\hat{P}_t \mu_t(A) = \int \hat{P}_t(\varphi, A) \mu_t(d\varphi)\), for any Borel subset \(A \subseteq S\).

**Definition 1** A **stationary** equilibrium with positive entry is prices \((p^*, w^*)\), aggregate quantity \(Q^*\), labor demand \(N^*\), a mass of entrants \(M^*\) and a distribution \(\mu^*\), such that

1. Markets clears: \(p^* = D(Q^*)\) and \(w^* = W(N^*)\).
2. The entry decision is

\[
x^* = \inf \left\{ \varphi \in S : \int v(\varphi') F(d\varphi|\varphi) \geq 0 \right\}
\]

3. Because of positive entry, free-entry holds with equality: \(v^e = c_e\), where \(v^e = \int v(\varphi) \nu(d\varphi)\), meaning that \(\int v(\varphi') F(d\varphi|\varphi) = 0\), \(\forall \varphi' \in S\).
4. Entry and exit flows are equalized, and therefore \(\mu^*\) is defined by

\[
\mu^* = \hat{P} \mu^* + M^* \nu
\]

**Empirical performance**

**General comments:**

- Note that in the stationary equilibrium there is only one state, \(\varphi\).
- To bring the model to the data, labor \(n\) is the measure of size. Since \(n\) is monotonic in \(\varphi\), size can be indexed by productivity draws.
- Unconditionally, age matters because firms age as they survive in the market over time.
  - Assuming firms come in small (an assumption on \(\nu\)), as in the data, then firms survive only if they draw high productivities, in which case they become larger.
  - Thus, the model predicts that larger firms are old and survive because they are efficient in production. This is of course by construction, as firms stay in the market if they are lucky.
  - However, conditional on size, age is irrelevant. The model is thus ill-suited to study the empirical facts studied in Lecture 1. In particular, in the data, small firms are in fact not inefficient.
- In the data, however, young firms are seen to be small not because they are particularly technically more efficient than old incumbents, but mostly because they have more favorable
idiosyncratic demand conditions. Hopenhayn doesn’t distinguish between the two, as they are both encapsulated inside $\varphi$.

Parametrization:

- Suppose the production function $q(s, n) = e^s n^\alpha$, with $\alpha \in (0, 1)$. The productivity index evolves

$$s' = \rho s + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$ and $\rho \in (0, 1)$.

- Then, $\pi(s) = \max_n \{e^s n^\alpha - w n\}$, so that the labor demand is

$$n(s) = \left(\frac{\alpha e^s}{w}\right)^{1/\alpha}$$

and profits can be written as $\pi(s) = (1 - \alpha)e^{s/\alpha} \left(\frac{w}{2}\right)^{\alpha}$. The value function is $v(s) = \pi(s) - c_f + \max \{0, \int v(s') F(ds'|s)\}$

- We are interested in two applied questions:
  1. How firm growth (given by $n'/n$) evolves over size $(s)$
  2. How firm exit depends on size.

Assessing empirical performance:

- Some algebra shows

$$\frac{n'}{n} = e^{-\frac{\rho}{1-\alpha} s + \frac{\varepsilon}{1-\alpha}}$$

- Unconditionally, we want both the mean and variance of firm growth to decrease with size. Note

$$\mathbb{E} \left[ \frac{n'}{n} | s \right] = e^{-\frac{\rho}{1-\alpha} s + \mathbb{E} \left[ \frac{\varepsilon}{1-\alpha} \right]}$$

$$\mathbb{V} \left[ \frac{n'}{n} | s \right] = e^{-2\frac{\rho}{1-\alpha} s + \mathbb{V} \left[ \frac{\varepsilon}{1-\alpha} \right]}$$

so indeed they are both decreasing in $s$ because of the fact that $s$ is mean-reverting (high productivity firms have a higher chance of keep drawing high productivity draws than low productivity ones).

- For example, if $\rho = 1$, Gibrat’s law holds: average and variance of growth are independent of size.

- However, since $\rho < 1$, both 1st and 2nd moments of growth decrease with size.

- Simply to the extent that age is correlated with size (assuming $\nu$ so that entrants are smaller), then we would also get that younger firms grow faster if $\rho < 1$. Again, however, conditional on size, growth is independent of age.

---

3Haltiwanger in particular is pushing this idea. Syverson et al. (AER) has recently develop a methodology to disentangle empirically the two effects (price vs technical efficiency). Gourio and Rudanko (2014) improves on the theory front.

4Strictly, the measure of size is $n$. Since $n'(s) > 0$, we can index size with $s$, wlog.
• The prediction of exit relative to size is stark and trivial: there is an $x$ such that firms exit if $s < x$. Namely, only big enough firms survive. In the data, we also see big firms exiting with low probability.

Computation:
• To estimate the model, the key equilibrium variables of the stationary equilibrium allocation of this economy are the price $p$ and the total mass of firms $M$.
• For identification, what is key is the market clearing condition $D(p) = Q^*(p)$, where $Q^*(p) = \int e^s n(s, p)\, d\mu(s)$, and the free-entry condition $v^e(p) = c_e$. Note $M$ is given implicitly by the law of motion $\mu = P\mu + M\nu$.
• Algorithm:
  1. Conjecture a price, $p$.
  2. Obtain the value $v(p, s)$ for any $s \in S$, and the value of an entrant, $v^e(p) = \int v(p, s)\, d\mu(s)$.
  3. Obtain $p$ from free entry: $p = (v^e)^{-1}(c_e)$. Update $p$ appropriately and repeat 1-3 until converging for $p$ (convergence is ensured by the fact that $v^e$ is monotonic in $p$).
  4. Once we have $p$, obtain $M$ from the $\mu$ law of motion.

2.2 Clementi and Palazzo (AEJ: Macro, 2016)

The Hopenhayn model has no role for age once one controls for size. Now, we look at an extension of the model that incorporates capital, time-to-build and capital depreciation, as well as an adjustment cost for capital. These will generate a role for age.

Environment
• We look at a partial equilibrium, stationary version of the Clementi and Palazzo (AEJ: Macro, 2016) model.
• Firms have a common discount factor $1/R$ with $R > 0$, and produce the same homogenous, given by

$$y_t = s_t (K_t^\alpha L_t^{1-\alpha})^\theta$$

where $\alpha \in (0, 1)$, $\theta \in (0, 1)$ is the “span of control parameter” that controls the amount of decreasing returns to scale, and

$$\ln s_{t+1} = \rho \ln s_t + \sigma \varepsilon_{t+1}$$

where $\varepsilon_{t+1} \sim N(0, 1)$, $\rho \in (0, 1)$, and $\sigma > 0$.
• The labor supply is given by $L_t(w) = w^\gamma$, with $\gamma > 0$.
• Like in Hopenhayn, firms need to pay an operating cost $c_f > 0$, where $\ln c_f \sim N(\mu_{c_f}, \sigma_{c_f}^2)$.
  - This generates an exit pattern that is more empirically meaningful than the original Hopenhayn framework.

---

Note it is straightforward to obtain this endogenously with a representative household with CRRA preferences whose labor Frisch elasticity is $\gamma > 0$. 
− In particular, every firm in every time, even very large firms, has a non-zero probability of exiting.

• Firms face a capital adjustment cost:

\[ g(x, K) = \chi(x)c_0K + c_1 \left( \frac{x}{K} \right)^2 K \]

where \( c_0, c_1 > 0 \) are constants, \( \chi : \mathbb{R}_+ \to \{0, 1\} \) with \( \chi(0) = 0 \) and \( \chi(x) = 1, \forall x > 0, K \) is the capital stock and \( x \) is the level of gross investment. Therefore:

− If firms replace depreciated capital, they do not pay fixed costs.
− If firms invest beyond that level, they face a fixed cost \( c_1K \), plus a cost that increased convexly with gross investment.

• Entry and exit:

− Upon exit, firms recoup undepreciated capital net of adjustment costs of driving it to zero.
− On the other hand, there is a constant mass \( M \) of potential entrants.
  1. Upon entry, firms receive a signal \( q \sim Q \), where \( Q \) is a Pareto distribution

\[ Q(q) = \left( \frac{q}{\bar{q}} \right)^{\xi} \]

where \( \xi > 1 \) is the shape parameter, and \( \bar{q} > 0 \) is a number.
  2. This signal is then part of the initial productivity draw for entrants:

\[ \ln s_1 = \rho \ln q + \sigma \varepsilon_1 \]

3. When entering, firms pay \( c_e \geq 0 \).

• Timing within any given period:

− Incumbents:
  1. Observe the \( s \) shock.
  2. Hire labor and produce.
  3. Draw the \( c_f \) cost shock, and decide whether to invest \( x \) and adjust capital, or exit.
− Potential entrants:
  1. Receive signal \( q \sim Q \).
  2. Decide whether or not to enter. If entry, pay entry cost and invest \( x \).

• Comments:

− Note the \( q \) draw introduces heterogeneity: firms see a signal, and above a threshold firms will enter. Since the investment decision has to be made one period in advance, entering firms will make an investment decision based on that signal. The investment will turn out to be increasing in this signal, and therefore entrants will be heterogeneous in every period.
− Assuming that the operating cost is now random can give us that any firm has a chance of exiting, even very productive ones. This allows us to match attrition rates more freely.
− Another novelty is the capital adjustment cost. The way it has been set up, we will have lumpiness in investment in equilibrium, which is a key feature of the data.
– Note that both the fixed cost $c_1K$ (for those that want to grow) and explicit decreasing returns in production ($\theta < 1$) put a limit to how large firms can grow by introducing decreasing returns to scale on firms. Both of these ensure stationarity in the firm distribution. Still, both forces prevent very large firms to exist, and even though growth rates will be decreasing with size, the model is unable to explain why in the data one cannot reject that large enough firms, conditional on survival, live in a Gibrat regime.

**Equilibrium**

- **Objective functions:**
  - The profit function of a firm $s$ is
    $\pi(K,s) = \max_B \left\{ s(K^{\alpha}L^{1-\alpha})^\theta - wL \right\}$
  - The value of exit is
    $v_x(K) = K(1-\delta) - g(-K(1-\delta),K)$
    where $\delta \in (0,1)$ is the capital depreciation rate, and $-K(1-\delta) \equiv x$ is the gross investment rate.
  - The value of an incumbent is
    $v(K,s) = \pi(K,s) + \int \max \left\{ v_x(x), \bar{v}(x,s) - c_f \right\} dG(c_f)$
    where $\bar{v}$ is the value of continuing:
    $\bar{v}(x,s) = \max_x -x - g(x,K) + \frac{1}{R} \int v(K',s')dH(s'|s)$
    s.t. $K' = K(1-\delta) + x$
  - The value of an entrant that draws a signal $q$ is
    $v_e(q) = \max_{K'} -K + \frac{1}{R} \int v(K',s)dH(s'|q)$
    and the firm enters if $v_e(q) \geq c_e$.

**Definition 2** A recursive competitive stationary equilibrium are (i) value function $v$, $\bar{v}$, $v_e$; (ii) policy functions $x(K,s)$, $\ell(K,s)$ and $K'(q)$; (iii) a wage $w$ and an entrants measure $\epsilon$; such that:

1. $v$, $\bar{v}$, $x$ and $\ell$ solve the incumbent’s problem.
2. $v_e$, $K'(q)$ solve the entrant’s problem.
3. Labor market clears:
   $$\int \ell(K,s)d\Gamma(K,s) = L_s(w)$$
   where $\Gamma$ is the measure of incumbents.
4. Stationary measures:
   (a) For all Borel sets $s \times K \subseteq \mathbb{R}_+^2$ of the state space, the measure of entrants satisfies:
\[ \epsilon(s, K) = M \int \int_{B_e(K)} dQ(q) dH(s'|q) \]

where \( B_e(K) \equiv \{ q : K'(q) \subseteq K \text{ and } v_e(q) \geq c_e \} \) is the set of signals \( q \) such that the optimal entrant’s choice of capital is within \( K \), and the entrant prefers to enter.

(b) For all Borel sets \( s \times K \subseteq \mathbb{R}_+^2 \) of the state space, the measure of incumbents satisfies:

\[ \Gamma(s, K) = \int_s \int_{B(K)} d\Gamma(s, K) dG(cf) dH(s'|s) + \epsilon(s, K) \]

where \( B(K) \equiv \{(s, K) : \tilde{v}(s, K) - cf \geq v_x(x) \text{ and } K(1-\delta) + x(s, K) \in K \} \) is the set of states for which current incumbents prefer to stay in.

**Empirical performance**

- Empirical improvements with respect to Hopenhayn (ECMA, 1992):
  - As in the Hopenhayn model, age matters *unconditionally*: young firms have very low productivity on average, so they are more likely to exit.
  - However, age now also matters *conditional on productivity*, because firms are not only indexed by their random productivity draw, but also by the amount of capital they have.
  - In particular, age matters even when we condition on size, the latter now proxied by the level of capital. For a given level of \( K = K_0 > 0 \), firms differ in the cross-section due to \( s \). Therefore, the only reason why age would matter is if age is correlated to \( s \). If age is uniformly distributed in the \( s \) shock, given capital, then age would again be irrelevant.
  - So how can we make age correlate with this shock, and how to do it in a way that is meaningful empirically?
  - In this paper, the authors choose to calibrate \( Q(q) \) is such a way that entrants get a productivity shock that is lower than the average productivity level of incumbents, so that on average the size of entrants is lower than (in fact, about half of) the size of incumbents. This creates scope for small firms to grow, conditional on survival.
  - Given this restriction, for any given \( K = K_0 > 0 \), in the cross-section of firms with capital \( K_0 \), firms with low (high) \( s \) are typically older (younger). The reason is that for firms to have gotten large, they must have gotten lucky at first, and if they have managed to stay in with low productivity is because they must have gotten lucky in the past, and thus they must be old.
  - Moreover, old firms are shrinking and young firms are growing. Further, the higher \( K \) we look at, the lower the chance of observing young firms, because the exit threshold \( s^* \) (determining optimal size) is higher (thus the probability \( Pr(s \geq s^*) \), determining the mass of entrants, is constant).
  - To sum up, in order for age to matter as in the data, we need to impose: (i) small firms to enter with relatively low productivity; (ii) the \( s \) shock must be mean-reverting.

3 Financial Frictions and Firm Dynamics

The literature has focused on other ways to motivate firm dynamics. For instance, financing constraints might impede firms to jump to their efficient size right away. This can generate a
dependence of growth on size beyond the selection effect that we have investigated in the Section 2. We will look at two models on this. First, we will look at a simple model of firm dynamics in which firms face financing constraints. Secondly, we will micro-found this constraint with informational frictions in the form of moral hazard.

3.1 Cooley and Quadrini (AER, 2001)

This is the first paper to look at the importance of firm financing constraints on firm dynamics. Here, we present an extremely simplified version of it.

- Time is discrete and infinite.
- Suppose firms hold one-period debt $b'$ (paying an interest rate $r$). These contracts are one period, so that capital is chosen one period ahead of time.
- Firms produce according to

  $$ y = zF(k + b) $$

  where $k$ is owned capital, $b$ is borrowed capital and $k + b$ is total capital, $F$ is the production function and $z \sim i.i.d.$ (note we now do not need persistence in the shock). There is no capital depreciation for simplicity.

- The value of the firm is

  $$ v(k, b) = \max_{b', k'} \left( zF(k + b) - rb + \int_{New \text{ capital}} v(z', k')dF(z') \right) $$

  such that

  $$ b' \leq k' $$

  - The last equation is a financing constraint, which we take as given for now.
  - Firms can grow by (i) borrowing (as much as its capital in place); and (ii) reinvesting capital by distributing dividends.
  - In particular, in equilibrium the efficient size is obtained from the fact that firms borrow up to the point in which the marginal value of equity capital equals its marginal cost:

    $$ zF'(k + b) = r $$

    where $\bar{z} \equiv E[z]$.

  - Therefore, firms optimally borrow at capacity, i.e.

    $$ b' = k' $$

    and make zero profits, so that

    $$ k' = zF(2k) - rk + k $$

    as long as the firm is short of its efficient size.
• Therefore, the growth rate of the firm is

\[
\frac{k' - k}{k} = zF(2k) - r
\]

which is decreasing in \(k\), even if \(F\) is CRTS and \(z\) is not mean-reverting.

• In sum:
  - With a simple model of borrowing constraints, we can already obtain that growth rates are decreasing in size.
  - Importantly, note that this is true even if \(z\) is not mean-reverting, which was an assumption we could not go without in Section 2.
  - The constraint introduces a market incompleteness which impedes firms to jump directly to their efficient size.

### 3.2 Clementi and Hopenhayn (QJE, 2006)

• General environment:
  - This is a principal-agent (e.g., lender-borrower) model.
  - Let \(K\) be the amount of working capital invested into a project by the agent (called “borrower”). For simplicity, capital depreciates fully between periods.
  - The project is successful with probability \(p \in (0, 1)\), in which case the borrower collects revenues \(R(K)\), where \(R' > 0 > R''\) with \(R(0) = 0\).
  - The borrower needs the principal (called “lender”) to finance the project. In return, he commits to paying back \(\tau\), which is state-contingent.
  - Both borrower and lender are risk-neutral, and the contract’s outcomes are assumed to be fully enforceable (both lender and borrower are willing to commit to a long-term arrangement).

#### One-shot, complete information game

• First, we focus on the case in which there is symmetric information in the one-shot game.
  1. First, the lender advances capital \(K\).
  2. Then, the outcome is observed, i.e. the revenue realizes. With probability \(p \in (0, 1)\), the revenue outcome is \(R(K)\) and the borrower pays \(\tau_G\) back to the lender. Otherwise, the revenue outcome is zero, and the borrower pays back \(\tau_B\). If \(\tau_x < 0\), the borrower is getting money from the lender in state \(x \in \{B, G\}\).

• The surplus for the lender and the borrower are \(W^L_x(K) \equiv \tau_x - K\) and \(W^B_x(K) \equiv R_x(K) - \tau_x\), respectively, in each state \(x \in \{B, G\}\), where \(R_G(K) = R(K)\) and \(R_B(K) = 0\). The total surplus in this arrangement is, then

\[
W(K) \equiv \sum_{x \in \{B, G\}} p_x [W^L_x(K) + W^B_x(K)]
\]

\[
= p[(\tau_G - K) + (R(K) - \tau_G)] + (1 - p)[(\tau_B - K) + (0 - \tau_B)]
\]

\[
= pR(K) - K
\]
• A contract is a set of capital and payment schedules \( \{K, \tau_G, \tau_B\} \) that adheres to the above assumptions. An *optimal* contract is a contract that maximizes total surplus, or \( K^* = \arg\max K W(K) \). That is, \( K^* \) solves

\[
pR'(K^*) = 1
\]

Henceforth, the level \( K^* \) is referred to as the *efficient* level of capital.

One-shot, incomplete information game

• We now focus on what happens in equilibrium when information is not complete. In particular, we assume that revenue outcome is private information to the borrower, and focus on how such asymmetry in information can give rise to a moral hazard problem and financial frictions that may affect optimal firm decisions. Specifically, we will show that credit rationing, namely \( K < K^* \), can arise in equilibrium.

• In this case, the contract solves

\[
\begin{align*}
\max_{K, \tau_G, \tau_B} & \quad p(\tau_G - K) + (1 - p)(\tau_B - K) \\
\text{s.t.} & \quad R(K) - \tau_G \geq R(K) - \tau_B \quad [IC_G] \\
& \quad 0 - \tau_B \geq 0 - \tau_G \quad [IC_B] \\
& \quad \tau_G \geq R(K) \quad [PC_G] \\
& \quad \tau_B \leq 0 \quad [PC_B]
\end{align*}
\]

The constraints say, respectively:

1. The first two constraints impose that the reporting strategy is truthful. In both scenarios, the lender is required to report the true revenue.
2. The last two constraints require that the borrower is willing to participate (limited liability constraints).

• We then obtain that \( \tau_G = \tau_B = 0 \) and \( p\tau_G + (1 - p)\tau_B < 0 \), a contradiction. In sum, a contract does not exist, in a one-shot game, that is incentive compatible and delivers a positive value to the borrower. There does not exist a contract that forces the agent to report a good outcome.

Two-period, incomplete information game

• We show that this non-existence result hinges on the fact that this is a one-shot, one-period game. We thus move to a two-period model, with \( t = 0, 1 \). The timing is as follows:

1. In \( t = 0 \), the lender advances \( K \).
2. The borrower observes \( R(K) \) or zero and makes a report to the borrower, \( \tau_G \) and \( \tau_B \), respectively.
3. In \( t = 1 \), contracts can be made contingent on the history. If the report was \( \tau_x \), the lender will advance \( K_x \), with \( x \in \{G, B\} \).
4. For each case, the borrower will observe \( R(K_x) \) of zero, and make reports \( \tau_{xG} \) or \( \tau_{xB} \), respectively.
• We assume that the borrower has no way of saving other than by contracting with the lender. We also do not allow for inter-period storage.

• The optimal contract now solves:

\[
\max_{K, \tau} \left[ p\tau_G + (1-p)\tau_B - K + \delta p(\tau_{GG} + (1-p)\tau_{GB} - K_G) + \delta (1-p)(\tau_{BG} + (1-p)\tau_{BB} - K_B) \right]
\]

s.t. \[
\begin{align*}
\text{[IC}_{GB}] & : R(K_G) - \tau_{GG} \geq R(K_B) - \tau_B \\
\text{[IC}_{GG}] & : 0 - \tau_{GB} \geq 0 - \tau_{GG} \\
\text{[IC}_{BG}] & : R(K_B) - \tau_{BG} \geq R(K_B) - \tau_{BB} \\
\text{[IC}_{G}] & : R(K) - \tau_G + \delta (p[R(K_G) - \tau_{GG}] + (1-p)[-\tau_{GB}]) \geq R(K) - \tau_B + \delta (p[R(K_B) - \tau_{BG}] + (1-p)[-\tau_{GB}]) \\
\text{[IC}_{B}] & : -\tau_B + \delta (p[R(K_B) - \tau_{BG}] + (1-p)[-\tau_{GB}]) \geq -\tau_G + \delta (p[R(K_G) - \tau_{GG}] + (1-p)[-\tau_{GB}])
\end{align*}
\]

and (PC)'s where \( \delta < 1 \) is the discount factor. The four first constraints impose truthful revelation in the second period, for all contingencies (good and bad scenarios in first period). Likewise, the fifth and sixth impose incentive compatibility in the first period, for the good and bad scenarios, respectively. We should add to this the limited liability constraints for the borrower.

• We will solve this problem by considering its recursive counterpart formulation, and solving backwardly the problems at \( t = 1 \) and then at \( t = 0 \). Conditional on \( G \) in period \( t = 0 \), define

\[
\begin{align*}
V_G & \equiv p[R(K_G) - \tau_{GG}] + (1-p)[0 - \tau_{GB}] \\
V_B & \equiv p[R(K_B) - \tau_{BG}] + (1-p)[0 - \tau_{BB}]
\end{align*}
\]

to be the second-period values to the lender when the good scenario realized in the first period. Then, conditional on outcome \( G \), the problem can be written as follows:

\[
\max_{K, \tau_G, \tau_{GB}} \left[ p\tau_G - K_G + (1-p)\tau_{GB} - K_B \right]
\]

s.t. \[
\begin{align*}
V_G & = p[R(K_G) - \tau_{GG}] + (1-p)[-\tau_{GB}] \quad [PU_G] \\
-\tau_{GG} & \geq -\tau_{GB} \\
-\tau_{GB} & \geq -\tau_{GG}
\end{align*}
\]

and similarly for the bad outcome, where \( [PU_G] \) denotes “promised utility” constraint.

• Notice that we are claiming that the portion of the optimal contract \( \{ K, \tau_{GG}, \tau_{GB} \} \) can be obtained by solving a simpler problem, namely by maximizing the payoff to the lender only (the objective function here) subject to the borrower obtaining value \( V_G \). Similarly, one can
also solve the problem when the bad outcome realized in the first period. This is a typical trick in this literature that reduces the dimensionality of the problem. Once this problem has been solved for a given $V_G$ (or $V_B$) in the first period, one can go on to find the optimal $V_G$ (or $V_B$) in the first period.

- In sum, combining the above problem for any period-$(t = 0)$ outcome, in period $t = 1$ we solve

$$
\max_{K', \tau_G', \tau_B'} \quad p[\tau_G' - K'] + (1 - p)[\tau_B' - K']
$$

$$
s.t. \quad \begin{align*}
V &= p[R(K') - \tau_G'] + (1 - p)[-\tau_B'] \quad [PU] \\
-\tau_G' &\geq -\tau_B' \quad [IC_G] \\
-\tau_B' &\geq -\tau_G' \quad [IC_B] \\
\tau_G' &\leq R(K') \quad [PC_G] \\
\tau_B' &\leq 0 \quad [PC_B]
\end{align*}
$$

where $V$ is the value that is being promised to the borrower, taken as given. The optimal contract will be indexed by $V$, as will be shown shortly.

The limited liability (or participation) constraints and $[IC_B]$ imply $\tau_G' \leq \tau_B' \leq 0$, which means that the payment in the good state is nonpositive. Consequently, we obtain that the contract also includes the constraint $[IC_B]$, as written above. Then, we note again that $\tau_G' = \tau_B' = \tau' \leq 0$. Notice that the value that accrues to the lender at the end of the second period (when the payoff that is promised to the borrower is $V$) is non-negative, as $B(V) = p\tau_G' + (1 - p)\tau_B' - K \leq 0$. Moreover, the problem can be simply written as

$$
\max_{\tau', K'} \quad \begin{cases}
\tau' - K' \text{ s.t. } V = R(K') - \tau' \text{ and } \tau' \leq 0
\end{cases}
$$

which leads to $\tau = 0$ and thus

$$
K = R^{-1}(V/p)
$$

- Now, recall that the efficient capital was $K^*$ s.t. $R'(K^*) = 1/p$. Therefore, we get

$$
(K', \tau') = \begin{cases}
(R^{-1}(V/p), 0) & \text{if } R^{-1}(V/p) \leq K^* \\
(K^*, pR(K^*) - p) & \text{if } R^{-1}(V/p) > K^*
\end{cases}
$$

and the total surplus $W(V)$ when the promised payoff to the lender is $V$ is

$$
W(V) = pR(K') - K' = \begin{cases}
V - R^{-1}(V/p) & \text{if } V < pR(K^*) \\
pR(K^*) - K^* & \text{otherwise}
\end{cases}
$$

Thus, we achieve the constrained efficient capital for $V \geq \tilde{V} \equiv pR(K^*)$, but $K < K^*$ if we cannot promise a high-enough value to the lender. Correspondingly, the surplus is concave and increasing in $V$ for all $0 \leq V \leq \tilde{V}$ (with $W(0) = 0$), and flat at $W(K^*)$ otherwise.

- Now, by backward induction, we move to the first period. The problem is
\[
\operatorname{max}_{K, \tau_G, \tau_B} \quad [p \tau_G + (1 - p) \tau_B] - K + \delta [pB(V_G) + (1 - p)B(V_B)]
\]

\[
\text{s.t.} \quad V = p[R(K) - \tau_G] + (1 - p)[-\tau_B] + \delta [pV_G + (1 - p)V_B] \quad \text{[PU]}
\]

\[
R(K) - \tau_G + \delta V_G \geq R(K) - \tau_B + \delta V_B \quad \text{[IC_G]}
\]

\[
-\tau_B + \delta V_B \geq -\tau_G + \delta V_G \quad \text{[IC_B]}
\]

\[
\tau_G \leq R(K) \quad \text{[PC_G]}
\]

\[
\tau_B \leq 0 \quad \text{[PC_B]}
\]

where \( V \) is now the promised payoff to the agent (the borrower).\(^6\) Recall that \( B(V) \) is the value that accrues to the lender at the end of the second period, when payoff \( V \) is promised to the borrower. We can then note:

1. First, the two IC’s say \( \delta [V_G - V_B] \geq \tau_G - \tau_B \) and \( \delta [V_G - V_B] \leq \tau_G - \tau_B \), and therefore \( \tau_B - \tau_G = \delta (V_G - V_B) \).

2. Second, wlog we can set \( \tau_B = 0 \).\(^7\) We denote \( \tau \equiv \tau_G \).

- Then, the problem now reads

\[
\operatorname{max}_{K, \tau} \quad pR(K) - K + \delta [pW(V_G) + (1 - p)W(V_B)]
\]

\[
\text{s.t.} \quad V = p[R(K) - \tau] + \delta [pV_G + (1 - p)V_B] \quad \text{[PU]}
\]

\[
-\tau \leq \delta [V_G - V_B] \quad \text{[IC]}
\]

\[
\tau \leq R(K) \quad \text{[PC]}
\]

Since \( S \) is weakly increasing, then note \(-pd\tau + \delta pdV_G = 0\), and so \( \tau = R(K) \) wlog and the problem further simplifies to

\[
\operatorname{max}_{K, \tau} \quad pR(K) - K + \delta [pW(V_G) + (1 - p)W(V_B)]
\]

\[
\text{s.t.} \quad V = \delta [pV_G + (1 - p)V_B] \quad \text{[PU]}
\]

\[
R(K) \leq \delta [V_G - V_B] \quad \text{[IC]}
\]

To solve this, note that any \( K < K^* \) (where recall \( K^* \) is such that \( pR'(K^*) = 1 \)) with \( R(K) < \delta [V_G - V_B] \) is not a solution, as \( K \) can always be increased optimally up to the point when \( \text{[IC]} \) binds. Thus, if \( K < K^* \), \( \text{[IC]} \) must bind. Since \( \text{[PU]} \) already holds with equality, then for any \( K < K^* \) we can solve the system of two equations to get \( R(K)/\delta + V_B = V_G \) and \( V = pR(K) + \delta [pV_B + (1 - p)V_G] \), and thus

\(\text{[PU]}\) this is actually equivalent to maximizing total surplus and maximizing \( pR(K) - K + \delta [pW(V_G) + (1 - p)W(V_B)] \) with respect to \( K, \tau_G, \tau_B, V_G \) and \( V_B \), subject to the same constraints.

\(\text{[IC]}\) To show this, we show that if \( \tau_B < 0 \), then surplus can always be increased by setting a higher \( \tau_B \). In the problem, if \( \tau_B < 0 \) then \( \text{[PC_B]} \) is slack. To ensure \( \text{[PU]} \) still holds with equality and \( \tau_B - \tau_G = \delta (V_G - V_B) \) is still true, we must ensure that \(-\delta(1 - p)d\tau_B + \delta (1 - p)dV_B = 0\), namely

\[
d\tau_B = \delta dV_B
\]

which shows what we wanted.
\[ V = pR(K) + \delta V_B \]

so

\[ V_G = \frac{V + (1 - p)R(K)}{\delta}, \quad V_B = \frac{V - pR(K)}{\delta} \quad (1) \]

and we can re-write the problem as

\[ K(V) \equiv \arg \max_K \left\{ pR(K) - K + \delta \left[ pW \left( \frac{V + (1 - p)R(K)}{\delta} \right) + (1 - p)W \left( \frac{V - pR(K)}{\delta} \right) \right] \right\} \]

whenever \( K < K^* \). Since we know the shape of the revenue function this delivers a policy function \( K(V) \), which will be the solution for all \( V \) such that \( K(V) < K^* \). Otherwise, for the remaining values of \( V, K = K^* \). In sum,

\[ K = \begin{cases} K(V) & \text{if } V \leq \tilde{V} \\ K^* & \text{o/w} \end{cases} \]

where recall \( \tilde{V} \equiv pR(K^*) \). Therefore, for \( V \geq \tilde{V} \), we have \( V_G = V_B \) and \( W(V) = (1 + \delta)[pR(K^*) - K^*] \). In contrast, for \( V < \tilde{V} \) we must have \( K < K^* \), on which the surplus \( S \) is concave, and \( V_G > V_B \), so \( K(V_G) \geq K(V_B) \). The result says that even though the shock is i.i.d., it creates persistence: if the bad shock is realized, the optimal level of capital may be sub-optimal, and even lower than the capital level that would have been advanced in the good shock. Critically, this works endogenously through incentive compatibility.

• Intuitively, in the constrained case \( K < K^* \) (credit rationing), in the first period it is optimal to leave nothing to the borrower (as \( \tau = R(K) \)): the borrower is paid in the second period through \( V \). If we want to increase \( K \), we must then increase the spread in the future between the lenders’ payoffs, i.e. increase \( V_G - V_B \). The tension is therefore that what prevents us from extending \( K \) to the efficient level is (i) the IC, according to which increasing capital would require to spread out the future lenders’ payoff; (ii) the curvature in the revenue function \( R \), which explains why the surplus function is concave for \( V \in [0, \tilde{V}] \).

**Determining \( V \)**

• So far we have indexed contracts with the promised utility for the borrower, \( V \).

• To determine \( V \) endogenously, we assume a sunk cost for the borrower \( I_0 \) (e.g. a fixed initial investment) with \( M < I_0 \) (where \( M \) is the cash that the borrower has on hand), and that the lending market is perfectly competitive.

• Then, we solve

\[ \max \{ V \text{ s.t. } W_0(V) - V - (I_0 - M) = 0 \} \]

i.e, we impose a break-even condition for the lender, where \( I_0 - M \) is the capital that needs to be advanced by the lender in order for the contract to take place. Here, \( W_0 \) is the first-period total surplus, which can be shown to look similar to the second-period surplus derived above (in particular, also increasing and concave up to a certain \( V \), and flat from then on.)
Then,
\[ W_0'(V) dV - dV + dM = 0 \]
and thus
\[ \frac{dV}{dM} = \frac{1}{1 - W_0'(V)} > 0 \]
where the inequality holds as long as \( W_0'(V) < 1 \), e.g. if the 45-degree line crosses \( W_0 \) in the increasing part. If so, we see that the value that must be promised to the borrower must be an increasing function of his initial cash at hand.

**Summing up: the 2-period problem**

- In sum, in a two-period contract, we have found:
  - Due to informational constraints there is credit rationing in equilibrium (\( K < K^* \)).
  - The agent (borrower) is getting value only in the final period (as \( \tau = R(K) \)).
- We now examine briefly how these insights extend to the infinite-horizon model.

**Infinite-horizon, incomplete information game**

- This is the original formulation in Clementi and Hopenhayn (QJE, 2006).
- Most insights from the two-period model extend to the infinite-horizon framework, but the model additionally allows for exit, so it will have implications for survival rates.

**Environment:**

- Time is discrete and infinite, \( t = 1, 2, \ldots \)
- Both entrepreneur and lender are risk-neutral, and both discount the future at rate \( \delta \in (0, 1) \).
- Both agents can commit to long-term contracts, for simplicity.
- The entrepreneur has net worth \( M \) and the project has an initial required investment of \( I_0 > M \) and a scrap value of \( S \).

**Timing:**

- The principal liquidates at the beginning of time with probability \( \alpha_t(h_t^{-1}) \), in which case entrepreneur receives \( Q_t(h_t^{-1}) \) and lender receives \( S - Q_t(h_t^{-1}) \).
- Otherwise, if there is no liquidation, nature draws a shock (high or low), the lender provides capital \( k_t(h_t^{-1}) \), and the entrepreneur promises payments \( \tau_t(h_t^{-1}) \), both of which are state-contingent.

- Conditional on not having liquidated, one can express the problem recursively, as we discussed above, so that

\[
\hat{W}(V) = \max_{K, \tau, V_G, V_B} \left[ pR(K) - K + \delta [pW(V_G) + (1 - p)W(V_B)] \right] \\
\text{s.t. } p(R(K) - \tau) + \delta [pV_G + (1 - p)V_B] = V, \ \tau \delta (V_G - V_B), \ \tau \leq R(K), \ V_G, V_B \geq 0.
\]

\(^8\)To understand the implications of relaxing this assumption, see Section 3.3.
Going one step back in the period, the liquidation problem reads

\[ W(V) = \max_{\alpha \in [0,1], Q_c} \alpha S + (1 - \alpha)Q(V_c) \]

s.t. promise-keeping and incentive-compatibility constraints. This step was missing in the previous model, because we assumed the project was never liquidated.

Equilibrium:

- Once again, we can show that the value contingent on no liquidation, \( \hat{W} \), is weakly increasing and weakly concave in \( V \). It intercept at \( \delta S \) and becomes flat for all \( V \geq \hat{V} \), for some \( \hat{V} \) below which there is credit rationing.
- Interestingly, as in the above we have that the value promised to the borrower is

\[ V_l = \delta [pV_{G,t+1} + (1 - p)V_{L,t+1}] \]

and therefore \( V_l \) is a sub-martingale, namely \( V_l \) grows positively and unboundedly in expectation. This means that \( V_l \rightarrow +\infty \) is an absorbing state. The paper also shows that having the liquidation risk makes it a second absorbing state, so that \( V_l \rightarrow 0 \) is also an equilibrium.

- Sub-martingaleness in \( V \) is inherited by \( K(V) \), though one cannot show analytically that \( K \) is monotonic in \( V \). The exit hazard rate (probability of exit conditional on not having exited up to a particular age) can be shown to be decreasing in \( V \), again as a consequence of \( V \) being a sub-martingale. However, simulations of the model show that as the relationship ages, the average size of the project (average \( K \)) is increasing and concave, the exit hazard rate is decreasing on average, and so is both average and variance of growth. The “age” dimension here has meaning because it is associated with size. Given an initial size, size grows as age advances. Thus, one again, conditional on size, age is once again irrelevant.

### 3.3 Albuquerque and Hopenhayn (REStud, 2004)

- In this paper, credit rationing will come not from an informational asymmetry as in the previous model, but from limited commitment from the borrower. In particular, the borrower can take the revenues from the contract and run away, so that she cannot commit to the written contract.

Environment:

- There are two periods. In the last period, the problem is

\[ \max \ p\tau_G + (1 - p)\tau_B - K \]

s.t. \( V = p[R(K) - \tau_G] + (1 - p)[-\tau_B] \) (promised utility), \( R(K) - \tau_G \geq R(K) \) (IC constraint) and \( -\tau_B \geq 0 \).

- The second constraint here is the commitment constraint. We assume that if the borrower runs away she can take the whole revenue \( R(K) \) with her (the LHS is the payoff if she pays back; the RHS is the payoff if she runs away).

---

\(^9\)For \( V < V_r \), the exit probability is of course \( \alpha_t \). For \( V > \hat{V} \), the exit probability is zero. For \( V \in [V_r, \hat{V}] \), the exit probability is zero today but expected to be positive in the future. Thus, hazard rates do not decrease smoothly with age.
We note $\tau_G \leq 0$ and $\tau_B \leq 0$. Again, wlog we set $\tau_B = 0$ (otherwise, if $\tau_B < 0$, we can always raise $\tau_B$ to increase the payoff without violating the constraints), and the problem reads:

$$\max_{K, \tau} \ p\tau - K$$

s.t. $p[R(K) - \tau] = V$ and $\tau \leq 0$.

- Note that this is the same problem as that of Clementi and Hopenhayn above, meaning that both stories are equivalent. In particular,
  1. If $V < pR(K^*)$, then $\tau = 0$ and $K = R^{-1}(V/p)$.
  2. Otherwise, if $V \geq pR(K^*)$, then $K = K^*$, where $K^*$ is the efficient $K$, solving $pR'(K) = 1$.

- Therefore, the value function is the same as in the asymmetric information model (increasing and concave up to $\tilde{V} = pR(K^*)$, a region in which $K < K^*$, and then flat, where $K = K^*$). This means that any difference must come from the first period.

- In the first period, the problem is

$$\max \ pR(K) - K + \delta[pW(V_G) + (1 - p)W(V_B)]$$

s.t. $p[R(K) - \tau_G] + (1 - p)[-\tau_B] + \delta[pV_G + (1 - p)V_B] = V; \ R(K) - \tau_G + \delta V_G \geq R(K)$, and participation constraints. The difference then is that we can write $\tau \leq \delta V_G$ and $\tau \leq R(K)$ (using again $\tau_B = 0$ wlog). Thus, $\exists \tilde{V}_0$ such that $V \geq \tilde{V}_0$ implies $K = K^*$ in both periods and $V < \tilde{V}_0$ implies $\tilde{V}_0 = pR(K^*)[1 + \delta(1 - p)]$. Note that the only difference is therefore the term $(1 - p)$ in $\tilde{V}_0$, which was not there in the CH model. Indeed, while now we have $R(K) \leq \delta V_G$, before we required $R(K) \leq \delta[V_G - V_B]$. Thus,

$$V_G = \frac{R(K)}{\delta} \quad V_B = \frac{V - pR(K)}{\delta(1 - p)}$$

which we can compare to [1].

- In sum, we have very similar implications: in spite of i.i.d. shocks, we have persistence in the level of capital, as state-dependent promised utilities are different. This is once again because incentive compatibility requires a sub-optimal level of capital in some states of the world.

- Notes on the dynamics of the infinite-horizon version (the original version in the paper):
  1. Once can show using the recursive formulation that $R(K) = \frac{V(1 - \delta(1 - p))}{p}$, $V_G = \frac{V[1 - \delta(1 - p)]}{\delta p}$, and $V_B = V$.
  2. So note that $V_G > V_B = V$, meaning that in the bad shock the borrower stays put, and in the good shock she upgrades her value. That is, the value for the borrower only stays constant or goes up, but can never go down (unlike in the asymmetric information case).
  3. This feature is of course counter-factual. To solve this problem, Albuquerque and Hopenhayn simply introduced a mean-reverting shock in the $R$ function (s.t. $R(K) = zK^\alpha$, with $z \sim AR(1)$), which gives us back a conditional monotonicity result. The problem is then that there are confounding sources of growth, hard to disentangle in the data: average capital size can grow both exogenously through $z$ or endogenously through the limited commitment constraints.
4 Investment and Firm Dynamics

In this section, we study models of firm dynamics that try to reconcile empirical evidence on investment both at the micro and aggregate level. Firstly, at the intensive margin, in the data, firm investment is lumpy: firms invest just enough to maintain their undepreciated capital, and then invest intensely at peak seasons. Secondly, at the extensive margin, investment is procyclical at the aggregate level: when output is above trend, the fraction of firms that invest increases. Finally, investment in booms reacts more than proportionally to aggregate shocks, relative to recessions, i.e. the elasticity of investment is time-varying at the aggregate level.

An early approach was to assume quadratic costs of capital adjustment. However, these models do not deliver the aforementioned key features of the data. With a CRS production function of the form \( e^z K \), with \( z' = \rho z + \sigma \epsilon, \; \epsilon \sim \mathcal{N}(0,1) \), and capital adjustment costs of the form \( \phi(K', K) = \frac{1}{2} \left( \frac{K'}{K} - (1 - \delta) \right)^2 K \), then the problem is

\[
V(K, z) = e^z K - K' - \phi(K, K') + \beta \int V(K', \rho z + \sigma \epsilon) f(\epsilon) d\epsilon
\]

so that the FOC is

\[
-x - \delta + \beta e^{\sigma z} \int e^{\sigma \epsilon} f(\epsilon) d\epsilon = 0
\]

where \( x \equiv \frac{K'}{K} \) is the investment rate. Therefore, the optimal investment rate is a linear function of productivity shocks, and therefore innovations on productivity always have the same impact on investment across states. This linearity is problematic of course, because shocks in the data have a non-linear impact on investment.

The literature then introduced models with fixed (as opposed to quadratic) costs of adjustment, which give rise to \((s, S)\) investment rules. Some of these theories are also counter-factual because firms that invest have stronger adjustment than the adjustment of those firms that are disinvesting, while \((s, S)\) models assume symmetry in the cost of adjustment.

Caballero and Engel (ECMA, 1999) then attempted to reconcile the facts with asymmetric \((s, S)\) rules. They present a tractable model with capital adjustment costs which allows for time-varying reaction of investment to aggregate shocks. We analyze this model next. We will later look at a GE version of the model due to Khan and Thomas (JPE, 2013), who examine the feedback responses on investment through the effect that shocks have on prices, and to Winberry (WP, 2015), who points out that Khan and Thomas rely on a counter-factual pro-cyclicality of the interest rate, and fixes this issue by departing from CARA utility and introducing habit formation. Both of these papers will have mean-reverting shocks and relax the assumption of Gibrat’s law which bought tractability to Caballero and Engel (1999).

We start with an empirical paper before jumping to the theory.

4.1 Caballero, Engel and Haltiwanger (BPEA, 1995)

This is an empirical paper looking at LRD micro data on investment.

- The authors first define the concept of mandated investment, \( x \):
  - \( x \) is defined to be the investment rate that would be mandated by our theory of firm-level investment, absent an adjustment cost.
If \( x > 0 \) (\(< 0 \)), then there are shortages (excess) of capital. If \( x = 0 \), the firm is at its efficient size.

Looking at data, one can then back out the density of firms over the r.v. \( x \) at a certain period in time, \( f(x, t) \).

• We also define the **adjustment rate function**, \( A(X, t) \):
  
  - \( A(x, t) \) is thus the fraction of the mandated investment rate which is in fact undertaken.
  - The average investment rate for a firm with mandated investment level \( x \) is \( \kappa A(x, t) \), where \( \kappa \) is the initial size.

• Then, the **average investment rate for the population** is:
  
  \[
  I(t) = \int \kappa A(x, t)f(x, t)dx 
  \]

Note that the average investment rate for the population of firms coincides with the aggregate investment rate of the economy if the latter is independent from the initial size \( \kappa \) (i.e. if Gibrat’s law holds).

- The authors characterize the shape of
- Note that if adjustment cost were quadratic, then \( A \) would be independent of \( x \), so \( A(t) = A(x, t) \), and thus
  
  \[
  I(t) = A(t) \int \kappa f(x, t)dx 
  \]

i.e, aggregate investment is only a function of the mean of the \( f \) distribution, \( \int \kappa f(x, t)dx \).

• What the authors show is that:
  
  - In the data, in fact, the function \( A \) is far from being constant in \( x \). It is low for \( x < 0 \) (excess capital), and then it increases sharply around the switching point to \( x > 0 \). Conditional on \( x > 0 \) (capital shortage), as \( x \) increases, \( A(x, \cdot) \) increases concavely.
  - What they show is that the average investment rate depends not only on the mean of the distribution, but also on higher moments of the distribution. This seems to point out that adjustment costs are not quadratic.
  - The distribution of firms across mandated investment levels seems to be bell-shaped around zero (the efficient capital level), so an \( (s, S) \) rule could not really capture the data either.
  - They find also that the average adjustment rate function is non-linear in mandated investment: firms that disinvest undertake small adjustments and the intensity of investment adjustments is even smaller the closer mandated investment is to zero (from below), but among firms that invest positively the adjustment is much larger and increases rapidly when mandated investment is away from zero (from above).
  - Finally, the histograms across investment rates (investment-capital ratios) among firms with high positive mandated investment and among firms with low positive mandated investment both spike at zero, but the spike is much higher in the low group. I.e. conditional on investment, a large fraction of firms actually undertake the investment when mandated investment is high, while the fraction of firms that undertake the investment is much lower when the mandated investment level is low.
In sum, the asymmetry in the average adjustment rate function seems to say that firm investment has a non-linear response when there exist shocks that create a deviation from the efficient size \( x = 0 \). This reconciles the aggregate-level evidence saying that the elasticity of investment responses is time-varying. Moreover, since the distribution of firms is shifting left and right due to shocks, the share of firms that are investing is very much cyclical.

4.2 Caballero and Engel (ECMA, 1999)

Environment

- Firms produce a horizontally differentiated set of products and thus face a downward-sloping demand for goods

\[ Q(p) = Bp^{-\eta} \]

with \( \eta > 1 \). Here, \( B \) is a stochastic demand shifter.
- Firms operate with a CRTS technology:

\[ F(K, L) = AK^\alpha L^{1-\alpha} \]

with \( \alpha \in (0, 1) \). Here, \( A \) is a stochastic TFP shock.
- Denote the discount rate by \( r \), the depreciation rate by \( \delta \), and the wage rate by \( w \), the latter taken to be a stochastic exogenous process. There are spot labor markets which allow firms to hire and fire at wage \( w \).
- Once again, the aggregate state is \((A, B, w)\), which is exogenous and taken as given.
- Define the cost of producing a certain quantity \( q \) by

\[ C(q) = \min \{ wL \text{ s.t. } AK^\alpha L^{1-\alpha} = q \} \]

Equilibrium (Static Model)

- Solving for the cost function, we get \( C(q) = wq^{\frac{1}{1-\alpha}} A^{-\frac{1}{1-\alpha}} K^{\frac{\eta(1-\alpha)}{\eta(1-\alpha) + 1}} \).
- Profits are

\[ \Pi(L, K; A, B, w) = \max \{ pQ(p) - C(Q(p)) \} \]

where \( Q(p) \) is the demand function. Plugging in the cost and the demand functions gives price

\[ p(K; A, B, w) = \left( \frac{\frac{\eta - 1}{\eta} - \frac{\alpha - 1}{\alpha}}{AB^{-\alpha} K^\alpha} \right)^{\frac{1-\alpha}{\eta(1-\alpha) + 1}} \]

- Profits are then

\[ \Pi(A, B, w) = \max_K \{ \theta K^\beta - (r + \delta)K \} \]

where \( \theta \) is a function of \((A, B, w)\) and \( \beta \equiv \frac{\alpha(q-1)}{1+\alpha(q-1)} \).
• Note \( \beta < 1 \), and thus we have a decreasing returns to scale function. On the other hand, \( \theta \) is a convoluted shock that includes demand, supply and wage shocks. This is an illustration that when incorporating shocks other than productivity TFP shocks, one can micro-found reduced-form specifications on DRTS production functions.

• The optimal choice of capital then solves \( \theta \beta K^{\beta-1} = r + \delta \), and thus the convolution of shocks has to satisfy

\[
\theta = \xi (K^*)^{1-\beta}
\]

where \( \xi \equiv \frac{r+\delta}{\beta} \).

• Defining \( z \equiv \ln \left( \frac{K}{K^*} \right) \) (reminiscent of the “mandated level of capital” described above), then \( K = K^* e^z \), and profits are

\[
\Pi(z, K^*) = K^\beta \xi (K^*)^{1-\beta} - (r + \delta)K
\]

\[
= \left( \frac{K}{K^*} \right)^{\beta} \xi K^* - (r + \delta)K
\]

\[
= \xi \left( e^{z\beta} - \beta e^z \right) K^*
\]

\[
= \pi(z) K^*
\]

Note that profits are linear in the efficient level of capital, i.e. there is homogeneity. Of course, profits are maximized at \( z = 0 \). Otherwise, they are low the farther \( z \) is away from zero, both from below or above.

**Equilibrium (Dynamic Model)**

• In the dynamic version of the model, \( \theta \) is importantly assumed to be a random walk, possibly with drift. That is,

\[
\theta_{t+1} = \theta_t e^{s_t}
\]

where \( s_t \sim \text{i.i.d. } N(\mu, \sigma^2) \) is the convolution of possibly idiosyncratic and aggregate shocks (recall that \( \theta \) is a shock including \( A, B, \) and \( w \)).

• Firms are assumed to face a capital adjustment cost given by

\[
\Psi(K, \theta) = \omega K^\beta \theta
\]

where \( \omega \sim \Gamma(\varepsilon) \) and \( \Gamma \) is the Gamma distribution. Here the interpretation is that when capital is adjusted from some capital in place \( K \) to some \( K' \), then a fraction \( \omega \) of sales \( K^\beta \theta \) is lost.\(^{10}\) This is an \((s, S)\) rule: the cost does not depend on where the firm adjusts to \( K' \), only where it adjusts from \( K \).

• Using the optimal level of capital \( K \), we get

\(^{10}\)Indeed, note \( K^\beta \theta = \Pi(K, \theta) + (r + \delta)K \), where the RHS are sales.
\[ \Psi(K, \theta) = \omega K^\beta \xi (K^*)^{1-\beta} \]
\[ = \omega \left( \frac{K}{K^*} \right)^\beta \xi K^* \]
\[ = \omega e^{\beta z} \xi K^* \]
\[ = \psi(z) K^* \]

and once again we get linearity in the efficient level of capital, \( K^* \).

- **Timing:**
  1. Each period starts with the state \((z_t, K^*_t, \omega_t; \theta_t)\). Recall \( z_t \) is a measure of distance from the efficient level \( K^*_t \).
  2. Then, the firm decides whether or not to adjust.
     (a) If no adjustment, the firm earns profits \( \pi(z_t) K^*_t \).
     (b) If adjustment, the firm pays the adjustment cost \( \psi(z) K^*_t \) and jumps immediately to the new level of capital, described by a new distance-to-the-efficient-level, call it \( c_t \). Earned profits are \( \pi(c_t) K^*_t \).
  3. Finally, capital depreciation realizes. A new period starts with state \((z_{t+1}, K^*_{t+1}, \omega_{t+1}; \theta_{t+1})\), with \( z_{t+1} \in \{(1-\delta)c, (1-\delta)z_t\} \).

- Since \( \theta_t = \xi (K^*_t)^{1-\beta} \), then
  \[ \frac{K^*_t}{K^*_{t-1}} = \left( \frac{\theta_t}{\theta_{t-1}} \right)^{\frac{1}{1-\beta}} \]
  and thus since \( \theta_t \) is assumed to be a random walk, so will the efficient level of capital. Using the process for \( \theta \) above,
  \[ \frac{K^*_t}{K^*_{t-1}} = e^{\mu t} \]
  and so, in particular, there is no persistence in the efficient capital level. This assumption is of course critical.

- The value function of the firm is
  \[ V^*(z_t, K^*_t, \omega_t) = \max \{ V(z_t, K^*_t), V(c, K^*_t; \omega_t) \} \]
  where
  1. The value of not adjusting is
  \[ V(z_t, K^*_t) = \pi(z_t) K^*_t + \frac{1}{1+r} E_t [V^*(z_{t+1}, K^*_{t+1}, \omega_{t+1})] \]
  2. Similarly, the value of adjusting is
  \[ V(c, K^*_t; \omega_t) = \max_{c'} V(c', K^*_t) - \psi(z_t, \omega_t) K^*_t \]
  where once again \( V(c', K^*_t) = \pi(c') K^*_t + \frac{1}{1+r} E_t [V^*(z_{t+1}, K^*_{t+1}, \omega_{t+1})] \).
• Thanks to the homogeneity results, we can divide both sides by $K_t^*$ in the value of no adjustment to get:

$$v(z_t) \equiv V(z_t, K_t^*) \frac{K_t^*}{K_t^*} = \pi(z_t) + \frac{1}{1 + \tau} \mathbb{E}_t \left[ V^*(z_{t+1}, K_{t+1}^*, \omega_{t+1}) \frac{K_{t+1}^*}{K_t^*} \right]$$

$$= \pi(z_t) + \frac{1 - \delta}{1 + \tau} \mathbb{E}_t \left[ v(z_{t+1}, \omega_{t+1}) e^{-\Delta z_{t+1}} \right]$$

where we have used that $e^{zt} = \frac{K_t^*}{K_t^*}$, so $\frac{K_{t+1}^*}{K_t^*} = e^{-\Delta z_{t+1}} \frac{K_{t+1}^*}{K_t^*} = (1 - \delta)e^{-\Delta z_{t+1}}$ (for the last equality, note the timing assumption that depreciation occurs inter-period while capital adjustment occurs intra-period). Normalizing the value of adjusting by $K_t^*$ as well, then we can write the whole value function as

$$v^*(z_t, \omega_t) = \max \{ v(z_t), v(c; \omega_t) \}$$

where $v(c; \omega_t) = \max c', v(c') - \psi(z_t, \omega_t)$ and $v(c') = \pi(c') + \frac{1 - \delta}{1 + \tau} \mathbb{E}_t \left[ v(z_{t+1}, \omega_{t+1}) e^{-(c'-c)} \right]$.

• Figure 2 in the paper plots the two value functions:
  - The inaction band (region of $z_t$ for which $v(z_t) \geq v(c; \omega_t)$) is centered around zero, so that adjustment occurs for some positive values of $z_t$ and some negative values of $z_t$.
  - Moreover, in the action region, firms always jump to $c$, which is the level of $z$ for which $v^*$ is maximized, namely the value of adjusting is highest (because $v(z_t) \leq v(c; \omega_t)$ always at $z_t = c$).
  - Finally, the width of the inaction region is increasing in $\omega$, as one would expect.

• Adjustment hazard rates:
  - Define the function $\Omega(z)$ as the largest adjustment cost $\omega$ such that a firm with a capital discrepancy $z$ find it optimal to adjust.
  - By definition, the indifference condition

$$v(z_t) = \psi(z_t, \Omega(z_t))$$

must hold. Using the $\psi$ function, then $\Omega(z_t) = \frac{v(c) - v(z_t)}{\xi e^{2z_t}}$.

- Intuitively, $\Omega(z)$ is the $\omega$ that makes a firm $z$ indifferent between adjusting and not adjusting. Of course, we obtain $\Omega(c) = 0$ (a firm that does not find it optimal to adjust cannot face a positive cost any more), and

$$\Omega'(c) = \frac{1}{\xi} \frac{-v'(z_t) e^{\beta z_t} \beta e^{\beta z_t} [v(c) - v(z_t)]}{\xi e^{2\beta z_t}}$$

so $\Omega'(c) = \frac{1}{\xi} \frac{-e^{\beta z_t} \beta e^{\beta z_t} [v(c) - v(z_t)]}{\xi e^{2\beta z_t}}$. Noting $v'(c) = 0$ by the envelope condition, we obtain $\Omega'(c) = 0$. The function is plotted in Figure 3A in the paper.

- For any $\omega \sim \Gamma$, define also $L(\omega)$ (or $U(\omega)$) as the maximum shortage (or excess) of capital tolerated by a firm with draw $\omega$. The function is plotted in Figure 3B in the paper. We see that higher $\omega$ leads to higher discrepancy and, thus, to a larger inaction region.

- Finally, we ask what is the probability that a firm with discrepancy $z$ will invest. For this, define $x = z - c$, where recall that $c$ is the optimal discrepancy. For a given draw $\omega \sim \Gamma$, then a firm adjusts if $\omega < \Omega(x + c)$, and thus the adjustment hazard is given by
\[ \Lambda(x) \equiv \Gamma\left(\Omega(x + c)\right) \]

where \( \Gamma(\cdot) \) is the cdf of the Gamma distribution.

– Figure 4A in the paper shows \( \Lambda(x) \) for different shapes of the \( \Gamma \) distribution.

1. The dotted line plots \( \Lambda(x) \) for a low-mean, low-variance \( \Gamma \) distribution. The dashed line plots \( \Lambda(x) \) for a high-mean, high-variance \( \Gamma \) distribution.

2. Hazard rates are higher for the latter case when there is medium or small discrepancy, and vice versa when discrepancy is very large. In the low-mean, low-variance case, the hazard rates are similar to the standard \((s, S)\) rules.

3. What is more, for very large discrepancies, the hazard rate is flat for the high-mean, high-variance case, while the low-mean, low-variance \( \omega \)-draw case depends more on the current discrepancy when the latter is far away from zero.

4. In contrast, if \( \omega \) is drawn from a high-mean, high-variance distribution, the amount of adjustment does not depend too much on how much discrepancy there is, which is clearly counter-factual. The low-mean, low-variance case will be better suited to replicate the data, as it exhibits stronger dependence with the state.

• Aggregate moments:

– We can also compute the expected investment conditional on adjusting, and the aggregate investment rate. When a firm of discrepancy \( x = z - c \) adjusts, it goes from \( K' = e^x \) to \( K' = e^c \), and thus investment for such a firm will be

\[ I(x) \equiv e^c K^* - e^x K^* = \frac{e^c - e^x}{e^x} e^x K^* = (e^{-x} - 1)K(x) \]

where \( K(x) \equiv e^x K^* \). Then, the average investment rate of firms with discrepancy \( x \) is

\[ \mathbb{E}_\omega[I(x)|x] = \Lambda(x)(e^{-x} - 1)K(x) \]

This function is plotted in Figure 4B in the paper, again for a high-mean, high-variance \( \Gamma \) and a low-mean, low-variance \( \Gamma \). We see that the latter specification has a better chance of delivering an aggregate investment rate that is sufficiently cyclical as in the data. In particular, the aggregate investment in the model is given by

\[ I \equiv \int (e^{-x} - 1)\Lambda(x)\bar{K}(x)\hat{f}(x)dx \]

where \( \bar{K}(x) \) is the average capital across firms that are at distance \( x \) from the target, and \( \hat{f} \) is the distribution of firms in the \( x \) space. The aggregate investment rate is, then

\[ \frac{I}{\bar{K}} = \int (e^{-x} - 1)\frac{(e^{-x} - 1)\Lambda(x)}{K} \bar{K}(x)\hat{f}(x)dx \]

where, in an abuse of notation, here \( K \) stands for the aggregate capital stock. Note that adding-and-subtracting \( \int (e^{-x} - 1)\Lambda(x)\hat{f}(x)dx \) on the RHS gives
\[
\frac{I}{K} = \int (e^{-x} - 1)\Lambda(x)\tilde{f}(x)dx + \int \frac{(e^{-x} - 1)\Lambda(x)}{(A)}\left(\frac{\hat{K}(x)}{K} - 1\right)\tilde{f}(x)dx
\]
\[
= \int (e^{-x} - 1)\Lambda(x)\tilde{f}(x)dx + \int (e^{-x} - 1)\Lambda(x)\tilde{f}(x)dx \int \frac{\hat{K}(x)}{K} - 1 \tilde{f}(x)dx
\]
\[
= \int (e^{-x} - 1)\Lambda(x)\tilde{f}(x)dx
\]

Assuming that the objects \((A)\) and \((B)\) are uncorrelated, then we can split up the integral and write the second line above. The third line follows from the fact that \(\int \left(\frac{\hat{K}(x)}{K} - 1\right)\tilde{f}(x)dx = 0\). Thus, the aggregate investment rate is computed as a weighted average hazard rate. Even though the uncorrelated assumption seems odd (as the hazard rate need not be constant in \(x\)), they show numerically that this gives a good approximation typically. The reason why this is is because, as we argued, they need a low-mean, low-variance \(\Gamma\) distribution in order to match the aggregate data, but for this distribution the vast majority of firms lie in the inaction region, in which the hazard rate is flat on \(x\) (recall Figure 4A and note that \(\Lambda(x)\) is flat exactly around zero, where most firms are bunched together).

- In the latter case in which most firms are inactive, we can use that \(x = 0, e^{-x} - 1 \approx -x\) (via a Taylor expansion), the investment rate is

\[
\frac{I}{K} \approx -\Omega_0 \int x\tilde{f}(x)dx
\]

assuming most firms are located around \(x = 0\), where \(\Omega_0\) is approximately \(\Omega(0)\). Thus the investment rate is a function only of the mean of the distribution (an aggregation result). Consequently, there is no state dependence: aggregate shocks shift the investment rate up and down, but symmetrically so. This is a reason why, in this model, we would like to stay away from constant hazard.

In words, in order to generate implications for the aggregate investment rate, we need a non-trivial distribution of firms over the \(x\) space. Since the \(\Lambda\) hazard function will be shifted back and forth by shocks, it will be critical to understand where the mass of firms is located. For a positive aggregate shock, \(\Lambda\) will shift up and. If a large mass of firms were around \(x = 0\), then the shock will be cause a relatively large mass of them to be now located in the lower end, and so a lot of them will adjust, causing the aggregate investment rate to increase as well. However, if the mass of firms were to be further away from the inaction region to begin with, such procyclicality of the investment rate would not be achieved. This is a point that Kahn and Thomas (2008) and Winberry (2015) will improve upon.

- Recall that the random-walkness of the shock was critical in the theory. Assuming

\[
\mu_t = v_t + u_t
\]

where \(v_t\) is an aggregate shock and \(u_t\) is an idiosyncratic one, then recall \(K_t^* = K_{t-1}^*e^{v_t+u_t}\) and thus
\[ \Delta x_t = z_t - c - (z_{t-1} - c) = \Delta z_t \]
\[ = \ln \left( \frac{K_t}{K_{t-1}} \right) - \ln \left( \frac{K^*_t}{K^*_{t-1}} \right) \]
\[ = \ln(1 - \delta) - (v_t + u_t) \]
\[ \approx -\delta - v_t - u_t \]

Therefore, the distance \( x \) might change because of three reasons: (i) depreciation, (ii) shocks hitting all firms alike, and (iii) shocks hitting at the firm level. Adding a drift in technology (in \( v_t \), for example imposing that \( v_t \) is a permanent shock) is then one simple way of generating the state dependence of aggregate shocks on the aggregate investment level that motivated this section.

Using the timing that, at the end of period \((t - 1)\) when there is a mass \( f(x, t - 1) \) of firms with distance \( x \), depreciation occurs and aggregate shocks occur (creating the \( f(x, t) \) distribution) and idiosyncratic shocks hit (creating the \( f(x, t) \) distribution), then we get

\[ \tilde{f}(x, t) = f(x + \delta + v_t, t - 1) \]

where

\[ f(x, t) \equiv \int \Omega(y) \tilde{f}(y, t) dy \quad g_u(-x) + \int [1 - \Lambda(x + u)] \tilde{f}(x + u, t) g_u(-u) du \]

where \( g_u \) is the pdf of the idiosyncratic shock.

### 4.3 Kahn and Thomas (ECMA, 2008)

Caballero and Engel’s paper is in partial equilibrium, in the sense that prices are given and, in particular, the interest rate is constant. Next, we look at a similar framework that looks at a GE version of the model and micro-founded the household side. Given the micro advances made by the previous paper, their focus is now to quantitatively match features of both the macro and micro data with time-varying prices. This work has become a workhorse model in the literature.

**Environment**

- Firms:
  - Firms produce with a DRTS production function

\[ y_t = z_t \varepsilon_t F(k_t, \gamma^t n_t) \]

where \( z \in \{z_1, \ldots, z_N\} \) is an aggregate shock and \( \varepsilon \in \{\varepsilon_1, \ldots, \varepsilon_N\} \) is idiosyncratic, both following a Markov process with \( Pr[z' = z_i | z = z_i] = \pi_{ij} > 0 \) and \( Pr[\varepsilon' = \varepsilon_m | \varepsilon = \varepsilon_e] = \pi_{\varepsilon m} > 0 \). Moreover, \( \gamma^t n_t \) is efficient units of labor, where \( \gamma > 0 \) is a parameter.
  - The firm has an adjustment cost in units of labor of
\( \psi(w) = w\xi \)

where \( w \) is the wage rate (determined in equilibrium) and \( \xi \in [0, \xi] \), with \( \xi > 0 \), is an i.i.d idiosyncratic shock drawn from some distribution \( G \). In particular, if investment \( i \) is such that \( i \notin [a_k, b_k] \) where \( a < 0 \) and \( b > 0 \), then investment is constraint and requires paying the adjustment cost.

- The LOM of capital is

\[ \gamma k' = (1 - \delta)k + i \]

where the LHS is labor-augmented capital.

- The aggregate states are \( z \) and \( \mu \), where \( z \) is the aggregate shock and \( \mu \) is the aggregate distribution, with \( \mu' = \Gamma(\mu, z) \) (where \( \Gamma \) is endogenous). The value function for firm with individual states \((\varepsilon, k, \xi)\) is

\[ V^0(\varepsilon, k; z, \mu) = \int_0^1 V^1(\varepsilon, k, \xi; z, \mu)G(d\xi) \]

where

\[ V^1(\varepsilon, k, \xi; z, \mu) = \max_{n, k^*} \left\{ \begin{array}{c} n w(z, \mu) + (1 - \delta)k + \max \{-\xi w(z, \mu) + R(\varepsilon, k^*; z, \mu'), R(\varepsilon, k^c; z, \mu')\} \end{array} \right\} \]

\( k^* \) is the efficient level of capital, \( k^c \neq k^* \) is a different choice of capital, and \( R(\cdot) \) is the continuation value conditional on a choice of capital, defined by

\[ R(\varepsilon, k'; z, \mu') = -\gamma k' + \sum_{j=1}^{N_i} \pi_{ij} d_j(z; \mu) \sum_{m=1}^{N_c} \pi_{em} V^0(\varepsilon, k'; z, \mu') \]

where \( d_j(z; \mu) \) is the stochastic discount factor (equal to the marginal utility of households since markets are complete), and \( \mu' = \Gamma(z, \mu) \) is taken as given.

- Households:

- There is a representative household holding a portfolio \( \lambda \) of shares of the firm and solving

\[ W(\lambda; z, \mu) = \max_{c, n^h, \lambda'} u(c, 1 - n^h) + \beta \sum_{j=1}^{N_i} \pi_{ij} W(\lambda'; z, \mu') \]

subject to \( \mu' = \Gamma(z, \mu) \) and the budget constraint

\[ c + \int \rho_1(\varepsilon, k; z, \mu)\lambda'[d_i(\varepsilon \times k)] \leq w(z, \mu)n^h + \int \rho_0(\varepsilon, k; z, \mu)\lambda[d_i(\varepsilon \times k)] \]

where \( c \) is consumption, \( n^h \) is supply of labor, \( \rho_1 \) is the end-of-period price of the portfolio (i.e, the value of the firm without the dividend pay-out, or “ex-dividend”) and \( \rho_0 \) is the price at the beginning of the period (i.e, the value of the firm last period, including the dividend it paid out, or “cum-dividend”).

**Equilibrium**

**Definition 3** An equilibrium is a tuple
\[
\left\{ w, (d_j)_{j=1}^{N^*_x}, \rho_0, \rho_1, V^1, N, K, W, C, N^h, \Omega^h \right\}
\]

such that

1. \( V^1 \) solves the value function of the firm, with associated policies \((N, K)\).
2. \( W \) solves the value function of the household, with associated policies \((N^h, \Omega^h)\).
3. The stock market clears, meaning that the household’s portfolio choice is consistent with the measure of stocks paid out by firms

\[
\Omega^h(\varepsilon_m, k'; \mu, z) = \mu'(\varepsilon_m, k')
\]

for all pairs \((\varepsilon_m, k') \in \{\varepsilon_1, \ldots, \varepsilon_{N^*_e}\} \times \mathbb{R}_+\).
4. The wage rate clears the labor market, or

\[
N^h(\mu; z_i, \mu) = \mu \cdot \left[ \int N(\varepsilon, k; z_i, \mu) + \int \xi \cdot \iota \left( \frac{2\gamma K(\varepsilon, k, \xi; z_i, \mu) - (1 - \delta)k}{k} \right) G(d\xi) \right]
\]

for every \(i \in \{1, \ldots, N^*_z\}\), where \(\iota\) is an indicator function defined by \(\iota(x) = I[ x \in [a, b] ]\).
5. The consumption policy function is

\[
C(\mu; z_i, \mu) = \int \left[ z_i \varepsilon F(k, N(\varepsilon, k; z)i, \mu) - \int 0^\xi \left( \gamma K(\varepsilon, k, \xi; z_i, \mu) - (1 - \delta)k \right) G(d\xi) \right] \mu(d\varepsilon \times dk)
\]

for every \(i \in \{1, \ldots, N^*_z\}\).
6. The law of motion of the measure of firms solves

\[
\mu'(\varepsilon_m, B) = \int_{\{\varepsilon, k, \xi; K(\varepsilon, k, \xi; z_i, \mu) \in B\}} \pi^\varepsilon_m G(d\xi) \mu(d\varepsilon \times dk)
\]

for every Borel set \(B\) of the state space, and every \(m \in \{1, \ldots, N^*_e\}\).

\begin{itemize}
  \item Note here that \(\iota\) captures the investment rate levels that are subject to the labor-denominated adjustment costs. Note that these costs are simply modeled as pure transfers to the household, so they are not subject to any deadweight loss. Moreover, consumption is computed as total output net of total investment.
  \item To characterize the equilibrium, this literature uses the following trick. First, we use the intra-temporal condition to note that the wage rate is

\[
w(z_i, \mu) = \frac{u_1(c, 1 - n^h)}{u_2(c, 1 - n^h)}
\]

and, by complete markets, the SDF is

\[
d(z_i, \mu) = \beta \frac{u_1(c', 1 - n^{h'})}{u_1(c, 1 - n^h)}
\]

Using further that \(p(z, \mu) = u_1(c, 1 - n^h)\) is the price of the consumption good, then we can express the SDF as a function of this price and normalize the value function to be written as follows:
\[
v^1(\varepsilon, k, \xi; z_i, \mu) = \max_n \left[ z_i F(k, n) - wn + (1 - \delta)k \right] p + \max \left\{ -\xi wp + \max_{k' \in R^+} r(\varepsilon, k'; z_i, \mu'), r(\varepsilon, k^c; z_i, \mu') \right\}
\]

where \( r \) is the normalized continuation value \( R \) from above, defined by

\[
r(\varepsilon, k'; z_i, \mu') = -\gamma k' p + \beta \sum_{j=1}^{N_e} \sum_{e=1}^{N_e} \pi_{ij} \pi_{em} v_0(\varepsilon, k'; \nu, \mu')
\]

and

\[
v_0(\varepsilon, k'; z_j, \mu') = \int_0^\xi v^1(\varepsilon, k; \xi, \mu) G(d\xi)
\]

- What we have done is simply to multiply both sides of the value function \( V^1 \) by the marginal utility of consumption (i.e., \( v^1 \equiv pV^1 \)). Conveniently, the marginal utility of consumption shows up in fewer places now. This gives more computational tractability.

- Defining

\[
E(\varepsilon, z, \mu) \equiv \max_{k' \in R^+} r(\varepsilon, k'; z, \mu')
\]

as the value of not adjusting, and

\[
E^c(\varepsilon, k, z, \mu) \equiv \max_{k' \in R^+} r(\varepsilon, k^c; z, \mu')
\]

as the value of adjusting, then we know that the firm adjusts if \( E(\varepsilon, z, \mu) - \xi wp \geq E^c(\varepsilon, k, z, \mu) \), which defines a threshold \( \hat{\xi}(\varepsilon, k, z, \mu) \) defined by

\[
\hat{\xi}(\varepsilon, k, z, \mu) = \frac{E(\varepsilon, z, \mu) - E^c(\varepsilon, k, z, \mu)}{wp}
\]

Using \( \xi^T(\varepsilon, k, z, \mu) \equiv \min\{\hat{\xi}(\varepsilon, k, z, \mu)\} \), then we can conclude the following policy rule for capital next period:

\[
k' = K(\varepsilon, k; z_i, \mu) = \begin{cases} k^*(\varepsilon, z_i, \mu) & \text{if } \xi \leq \xi^T(\varepsilon, k, z, \mu) \\ k^c(\varepsilon, k, z_i, \mu) & \text{o/w} \end{cases}
\]

for any \( i \in \{1, \ldots, N_z\} \).

- Knowing that the optimal capital investment follows a threshold rule, the consumption policy is then

\[
C = \int z \varepsilon F(k, N(\varepsilon, k; z, \mu)) - \underbrace{G(\xi^T(\varepsilon, k, z, \mu))}\_{\text{Unconstrained firms}} [\gamma k^*(\varepsilon, z, \mu) - (1 - \delta)k] \\
- \underbrace{(1 - G(\xi^T(\varepsilon, k, z, \mu)))}\_{\text{Constrained firms}} [\gamma k^c(\varepsilon, k, z, \mu) - (1 - \delta)k]
\]

and labor is

34
\[ N = \int [N(\varepsilon, k; z, \mu) + \int_0^{\xi^T(\varepsilon, k, x, \mu)} \xi G(d\xi)] \mu(d\varepsilon \times dk) \]

Finally, the LOM for the distribution of firms across idiosyncratic states can be simplified to

\[ \mu'(\varepsilon_m, \hat{k}) = \sum_{\varepsilon=1}^{N_{\varepsilon}} \pi_{\varepsilon m} \left[ \int G(\xi^T(\varepsilon, k; z, \mu)) \mu(d\varepsilon \times dk) \right. \\
\left. + \int \left( \hat{k} - k^\ast(\varepsilon, k; z, \mu) \right) \left( 1 - G(\xi^T(\varepsilon, k; z, \mu)) \right) \mu(d\varepsilon \times dk) \right] \]

and \( i(x) \equiv 1[x = 0] \).

**Empirical performance**

- **In the computation of the equilibrium, firms with a given idiosyncratic exogenous state \( \varepsilon_e \) today will choose different levels of capital and will lie in different \( \varepsilon_m \)'s tomorrow according to the Markov transition of shocks. In the end, we will find a distribution of firms of discrete support (a pmf) in the \((\varepsilon_e, \varepsilon_m)\) space with \((N_{\varepsilon})^2\) mass points, which can be easily computed in the computer.

- **Calibration:**
  - Knowing that the constraint investment is in the region \( i \in [ak, bk] \) then, with respect to the frictionless economy, the calibration of the constrained economy requires pinning down \( b \) (as they assume \(|a| = b\)) and moments (critically the standard deviation) of the \( \xi \) process, which they assume to be an AR(1) process:
    \[ \xi' = \rho_z \xi + \sigma_z \varepsilon^\xi \]
    with \( \varepsilon^\xi \sim \mathcal{N}(0, 1) \), \( \rho_z = 0.95 \) and \( \sigma_z = 0.007 \).
  - They calibrate these few parameters by matching the key moments on micro lumpiness in investment that we mentioned at the beginning of the section, which they take from Cooper and Haltiwanger (RES, 2006)[1] In particular, they match the inaction region, the positive and negative spikes and the share of firms doing positive and negative investment (see performance in Table II of their paper).

- **Results:**
  - A key result (see Table II) is that the traditional model (with \( b = 0 \)) requires a lot of inaction in order to generate the factual spikes. However, there does not seem to be this much inaction in the data, so Kahn and Thomas effectively shrink the inaction region by assuming \( b > 0 \). By doing this, they get the right frequency of positive spikes with respect to negative spikes
  - Another comparison they make (see Table III) is comparing their model with the PE version of it (similar to Caballero and Engel). On the aggregate investment rate, the PE model does not create enough persistence and generates too much skewness and excess kurtosis. The PE model assuming away quadratic adjustment costs (called “PE

---

[1] This paper uses LRD data and provides a comprehensive set of moments on firm investment in the U.S from 1972 to 2006.
frictionless” in the table) does a slightly better job because, in contrast to that other model, it does not strongly inherit the properties of the aggregate shocks.

- The data does not show skewness nor kurtosis, which does not give support to using \((s, S)\) rules (the “PE lumpy investment” model in the table) either. In other words, the data says that there is state-dependence in first and second moments of the aggregate investment distribution, but not on third or fourth moments. That is, shocks do not seem to generate asymmetries in the distribution emerging from nonlinearities, nor they make its tails thicker or thinner.

- The PE models also create a lot of skewness and kurtosis. Intuitively, the hazard function shifts when hit by shocks and, without an endogenous readjustment in prices, this increases the mass of firms investing at the same price without a counteracting adjustment in the intensive margin. At the extensive margin, firms react more to a positive shock than to a negative shock without a change in prices, which creates skewness.

- In GE, the full-blown version of their model, these effects are muted out due to the price reaction. When there is a positive shock, the hazard shifts to the right but the interest rate also goes up, depressing the SDF and counteracting the increase in investment at the extensive margin. Getting this countercyclicality in interest rates is challenging, however. A paper by Beaudry and Guay (JEDC, 1996) already pointed out that in a plain-vanilla investment model with normal shocks the interest rate is procyclical: a positive productivity shock increases the demand for investment goods because the marginal product of capital goes up, but the persistence of the shock also creates a wealth effect that makes households want to consume more. In order for households to give up consumption to provide the capital that firms desire, the interest rate then must go up. This does not square well with the data, as interest rates are typically countercyclical. They then show that augmenting the standard model with habit and quadratic adjustment costs can go some way into fixing this problem.

4.4 Extensions of Khan and Thomas (2008)

The reason why Khan and Thomas (2008) has become a seminal work is that they provided what has become a prototypical model of firm investment dynamics in the business cycle. A limitation of their model, however, is that it does not feature entry and exit, which is an important part of the data. Moreover, there is no role for age, while recently the data has unveiled, as we saw in Section 1, that most growth is coming from young, and not only small, firms. Generally, there is no role for the firm’s life cycle. Papers such as Winberry (WP, 2015), Khan and Thomas (JPE, 2013) and Clementi and Palazzo (AEJ: Macro, 2016) thus augment Khan and Thomas’ model to incorporate some of these important dimensions. Moreover, these models return to the above discussion on interest rates.

Winberry (WP, 2015): The countercyclicality of interest rates

Taking Beaudry and Guay’s insight on the necessity of having countercyclicality in the interest rate, Winberry introduces internal habit and quadratic adjustment costs into the Khan and Thomas’ model.

\[^{12}\text{The intuition is that quadratic shocks feature linearly in the optimal allocation, and particularly linear in the shocks.}\]
• Environment:
  – The model is very similar to Khan and Thomas (ECMA, 2008). The differences are in preferences and capital adjustment costs.
  – Household preferences are GHH with habit, given by:
    \[ u(C_t, M_t, N_t) = \left( \frac{C_t - M_t - \chi^{\frac{1+\alpha}{1+\alpha}}}{1 - \sigma} \right)^{1-\sigma} - 1 \]
  – The habit process is as in Campbell and Cochrane (1999):
    \[ \log S_{t+1} = (1 - \rho_s) \log S_t + \rho_s \log S_t + \lambda \log \left( \frac{C_{t+1}}{C_t} \right) \]
    where \( S_t \equiv (C_t - M_t)/C_t \).
  – The other main difference with respect to Khan and Thomas is that he assumes quadratic adjustment costs \( \psi(K, K') = \frac{1}{2} \left( \frac{K'}{K} - (1 - \delta) \right)^2 K \).

• Results:
  – The GHH specification kills wealth effects on the labor supply, and \( M_t \) is the habit term. Habit generates the right response in consumption as described above, but a productivity shock might create a counterfactual shift in labor supply that would counteract the desired reaction in interest rates. The author shuts down this response simply by resorting to GHH preferences.
  – Table 2 in his paper shows that the calibrated model does a good job aggregate investment rates in the data. Critically, the combination of habit and quadratic adjustment costs delivers the correct countercyclicality of interest rates.
  – Winberry also computes variances of the implied distribution of investment rates \( I/K \) in different percentiles of the distribution to argue that there is significant conditional heteroskedasticity in empirical investment rates and shows that his model can replicate these numbers, unlike the Khan and Thomas framework with no adjustment costs.

Clementi and Palazzo (2016) revisited: The role for entry and exit

• One recent feature of the data that has received a lot of attention is the permanent output trend deviation that the U.S. has experienced after the Great Recession. Clementi and Palazzo (AEJ: Macro, 2016) show the evidence for the BDS data (see, e.g. their Figure 15). For instance, there has been a dramatic drop in the number of establishments, coming from a stark decline in entry. The share of employment accounted for by young and newly born firms has been steadily going down. More recently born firms have entered smaller and smaller (the new entering cohorts were small) and they have accounted for a smaller and smaller share of total employment. Interestingly, conditional on being born, these firms were not pathologically small with respect to other periods or incumbents within the same period. That is, entry has depressed at the extensive margin, but not at the intensive one. Moreover, conditional on survival, young firms have continued growing relative to incumbents, so age was still very relevant during the Great Recession. In short, exit rates haven’t particularly gone down in the recent years.

• Accordingly, we now augment the Khan and Thomas model with entry and exit, and a role for the firm’s life cycle.
Environment

- We studied this model in the context of productivity-driven firm dynamics in Section 2.2. We now highlight the model as an extension of Khan and Thomas (2008), and incorporate aggregate shocks to the version presented in Section 2.2. The model is in PE (for a GE version, see Khan and Thomas (JPE, 2013)).
- We now provide a brief reminder of the environment, fully outlined in Section 2.2. The production function is DRTS:
  \[ y_t = z_t s_t [K^\alpha_t L_t^{1-\alpha}]^\theta \]
  where \( \alpha, \theta \in (0, 1) \), and
  \[
  \begin{align*}
  \log z_{t+1} &= \rho_s \log z_t + \sigma_z \varepsilon_{z,t+1} \\
  \log s_{t+1} &= \rho_s \log s_t + \sigma_s \varepsilon_{s,t+1}
  \end{align*}
  \]
  are aggregate and idiosyncratic shocks, respectively.
- The discount factor is \( 1/R \), where \( R > 1 \) is fixed. Assuming an acyclical discount factor does not crucially affect the main results.
- The labor supply is given by
  \[ L^S(w) = w^\gamma \]
  where \( \gamma > 1 \) is the Frisch elasticity.
- Capital depreciates at rate \( \delta \in (0, 1) \), and capital adjustment costs are quadratic and given by
  \[ g(x, K) = \chi(x) c_0 K + c_1 \left( \frac{x}{K} \right)^2 K \]
  where \( c_0, c_1 > 0 \) are constants, and there is an operating cost \( c_f \sim \log N(\mu_{c_f}, \sigma_{c_f}) \).
- There is a constant mass of potentially entrant firms \( M > 0 \) and, in the Hopenhayn tradition, firms receive a signal \( q \sim Q \) and, upon entry, draw \( s \) from \( H(s'|q) \), with \( \log s' = \rho_s \log q + \sigma_s \varepsilon_s \). If a firm enters, it must pay a cost \( c_e > 0 \).
- Timing:
  1. An incumbent firm first observes \( s_t \) and \( z_t \). Would-be entrants, at the same time, draw a signal \( q \) and observe \( z_t \).
  2. Then, incumbents hire labor, produce and draw the cost of operation \( c_f \). Would-be entrants decide whether to stay out, or enter by paying \( c_e \).
  3. After this, incumbents decide whether to stay and invest, or exit. Entrants, in turn, invest.

Equilibrium

- Characterization:
  - The aggregate state is \( \lambda_t = \{ z_t, \Gamma_t \} \), where \( \Gamma_t \) is the joint cdf of idiosyncratic states.
Period profits for an incumbent firm \((K,s)\) are given by
\[
\pi(\lambda; K,s) = \max_L sz[K^\alpha L^{1-\alpha}]^\theta - wL
\]
and the value function of this firm is
\[
V(\lambda; K,s) = \pi(\lambda; K,s) + \int \max \left\{ V_x(K), \tilde{V}(\lambda; K,s) - c_f \right\} dG(c_f)
\]
where \(V_x\) is the value of exit, and the value of continuing is
\[
\tilde{V}(\lambda; K,s) = \max_x -x - g(x,K) + \frac{1}{R} \int \int V(\lambda'; K'; s) dM(s'|s) dJ(\lambda'|\lambda)
\]
with \(K' = K(1-\delta) + x\).

For would-be entrants, the value of entering is
\[
V_e(\lambda,q) = \max_{K'} -K' + \frac{1}{R} \int \int V(\lambda'; K', s) dH(s'|q) dJ(\lambda'|\lambda)
\]
so that a firm enters if, and only if, \(V_e(\lambda,q) - c_e \geq 0\).

**Computation:**
- The implementation of the equilibrium is made through a Krusell-Smith type of technique. The reason is that agents in the model have to made forecasts of the time evolution of wages, which in turn depend on the entire distribution of the idiosyncratic states.
- To see this formally, note that the labor market equilibrium reads
\[
L^S(w) = \int L(\lambda; s,K) d\Gamma(s,K)
\]
where the RHS is labor demand. Since \(L^S(w) = w^\gamma\), then
\[
\log w_t = \frac{\log ((1-\alpha)\theta z_t)}{1 + \gamma(1 - (1-\alpha)\theta)} + \frac{1 - (1-\alpha)\theta}{1 + \gamma(1 - (1-\alpha)\theta)} \Omega_t
\]
where
\[
\Omega_t \equiv \log \left( \int (s_t K_t^\alpha)^{1 - \frac{1}{1 - \gamma}} d\Gamma_t(s_t, K_t) \right)
\]
Therefore, the wage today depends on the current realization of the aggregate state \(z_t\), as well as the distribution of states \(\Gamma_t\). In order to forecast wages going forward, agents must then be able to forecast the evolution of such distribution.

- Using a Krusell-Smith type of insight, we guess-and-verify that \(\Omega_t\) can be approximated by an affine function of itself and the aggregate shock:
\[
\Omega_{t+1} = \eta_0 + \eta_1 \Omega_t + \eta_2 \log z_t
\]
Now, conveniently, plugging the approximation into the market-clearing condition, \(\log w_{t+1}\) is a function of \(w_t, z_{t+1}\) and \(z_t\), but crucially not \(\Omega_t\):
\[
\log w_{t+1} = \beta_0 + \beta_1 \log w_t + \beta_2 \log z_{t+1} + \beta_3 \log z_t
\]
where \(\beta_0, \beta_1, \beta_2, \beta_3\) are some convolution of parameters \(\alpha, \theta, \eta_0, \eta_1, \eta_2\).
Conveniently, the aggregate state $\lambda$ has now reduced to simply $\lambda_t = \{z_t, w_t\}$ (i.e., the distribution is no longer an aggregate state), and the model can now be simulated by looping over the guess and getting new $\hat{\beta}$ estimates until convergence (namely, until old and new guesses for the $\beta$'s are close enough to each other). To obtain new estimates, we plug the guess into the market clearing condition in order to obtain the market-clearing wage with which the old guess can be compared. The algorithm converges because we have assumed log-linear and Gaussian shocks that have no skewness or kurtosis, which are consistent with an affine guess on a log-linear transformation of the distribution.

**Performance**

- **Allocation:**
  - As in the version of Section 2.2, there is a threshold strategy in the $q$ space describing entry. There is a $\hat{q}$ such that a firm enters if it draws $q \geq \hat{q}$.
  - The novelty with respect to Section 2.2 is that now there are aggregate shocks, so the threshold moves with aggregate shocks, too. A higher aggregate shock moves the threshold down, so worse firms will have a better chance to enter. In bad times, however, only the very best firms will enter. In other words, there is an important role for selection at entry: the mean idiosyncratic productivity goes up in recessions. This type of recession is often called a “cleansing recession”, because it prevents unproductive firms from entering.

- Accordingly, in the model, a one-time mean-reverting positive productivity shock creates, in an average across many different simulations, creates a spike in entry next period, after which it comes back down, and a fall in the exit rate, after which it goes back up. Because of the selection-at-entry effect, after such a shock the average productivity of entrants declines, and so does the one for exiters. Because of high entry and low exit, the number of establishments grows for some periods. Conditional on surviving, their productivity increases because they tend to enter unproductive and have scope for productivity growth. Thus, labor demand grows too, which explains most of the increase in wages. In fact, the wage remains high for a long period (it reverts slowly), which means that the entry rate converges from below (i.e. increases slightly for a number of periods until it converges back to its initial level).

- **Entry and exit as a source of output persistence:**
  - Entry and exit are therefore critical at both extensive and intensive margins in terms of the impact of shocks on output. With entry and exit, the effect on output is stronger in all periods and more persistent. This is because more firms entering drive output up. On the other hand, upon entering, surviving firms generate more output. On impact, the IRF of output with and without entry/exit are similar, but in the longer term entry/exit are critical for generating a persistence, and only by adding this new dimension can the model generate a slow recovery similar to the one seen after the Great Recession.
  - Intuitively, the reason why the model with entry/exit generates more output persistence can be seen clearly in a version of the model without capital accumulation. Suppose production is

\[\text{For this selection mechanism to work, however, it is crucial for wages to not move too much (as they would respond positively to the shock and would push the threshold in the opposite direction), so the Frisch elasticity has to be sufficiently low.}\]
The FOC wrt labor is \( s_t z_t t_{it}^{\alpha - 1} = w_t \), so

\[
l_t = s_t \left( \frac{\alpha z_t}{w_t} \right)^{\frac{1}{1-\alpha}} \]

Where \( \Gamma_t(s) \) is the distribution over aggregate shocks, then market clearing reads

\[
L_t = \left( \frac{\alpha z_t}{w_t} \right)^{\frac{1}{1-\alpha}} \int s^{\frac{1}{1-\alpha}} \Gamma_t(s) ds
\]

implying that the firm’s output is \( y_t = z_t s_t^{\frac{1}{1-\alpha}} \left( \frac{\alpha z_t}{w_t} \right)^{\frac{\alpha}{1-\alpha}} \), and total output is

\[
Y_t = z_t \left( \frac{\alpha z_t}{w_t} \right)^{\frac{1}{1-\alpha}} \int s^{\frac{1}{1-\alpha}} \Gamma_t(s) ds
\]

Combining total output \( Y_t \) with total labor \( L_t \) above, we get \( L_t^{\alpha} = \left( \frac{\alpha z_t}{w_t} \right)^{\frac{\alpha}{1-\alpha}} \left[ \int s^{\frac{1}{1-\alpha}} \Gamma_t(s) ds \right]^{\frac{\alpha}{1-\alpha}} \), which back into total output gives \( Y_t = z_t L_t^{\alpha} \left[ \int s^{\frac{1}{1-\alpha}} \Gamma_t(s) ds \right]^{1-\alpha} \). Equivalently, where \( N_t \) is the mass of firms in the economy at time \( t \), we can write

\[
Y_t = z_t \left[ \int s^{\frac{1}{1-\alpha}} \hat{\Gamma}_t(s) ds \right]^{1-\alpha} \left[ \int s^{\frac{1}{1-\alpha}} \Gamma_t(s) ds \right]^{\frac{\alpha}{1-\alpha}} \]

Distr. of productivity over firms (entr. and incamb.)

\( \# \) firms

\( N_t^{1-\alpha} L_t^\alpha \)

where \( \hat{\Gamma} \equiv \Gamma / N \). In the model without entry and exit, both \( N_t \) and \( \hat{\Gamma}_t \) converge to a constant, and thus all output dynamics are governed by the evolution of \( z_t \). That is, \( Y_t \) inherits the properties of \( z_t \), and the model converges to a simple neoclassical economy\(^{14}\). With entry and exit, however, there is a selection effect: entering firms enter small, meaning that \( N_t \) decays slowly following a positive productivity shock (as discussed earlier). Similarly, since most firms enter small and conditional on survival they grow fast, \( \hat{\Gamma} \) also persists high after the shock. Thus, both \( N_t \) and \( \hat{\Gamma} \) keep the marginal product of labor persistently high in the half-life of the shock, generating persistence in output after the shock over an above that which is directly inherited from \( z_t \).

– The model can in fact quantitatively account for a fair share of the slow recovery following the Great Recession. The model can replicate well that, following a negative productivity shock, the adjustment comes mostly from the extensive margin (the drop in the number of establishments) rather than the intensive margin (the selection margin, i.e. the drop in the employment from the firms that are already producing)\(^{15}\).

\(^{14}\) Adding capital to this merely amplifies the impact of shocks as \( K \) is a slow-moving variable, as it is standard in neoclassical models.

\(^{15}\) Recall Figure 15 in their paper for the empirical evidence, and see Figure 16 for how the model can deliver a persistent drop in employment when the model is shocked with a negative productivity shock to entrants on top of the negative TFP shock. For the model’s performance in terms of the extensive and intensive margin of employment, see Figure 17. In the latter figure, we see that the version with only a TFP shock versus the model with entry-specific shocks on top of the TFP shock deliver an identical evolution of the intensive margin, meaning that the calibrated model has little role for selection.
5 Endogenous Growth and Firm Dynamics

This literature is interested in the nexus between long-term growth, i.e. low-frequency fluctuations in output, and firms conducting R&D to create technological innovation. From the theoretical point of view, firm innovation is the engine of industry dynamics. From the empirical point of view, these theories try to match stylized facts on R&D and firm growth. Klette and Kortum (JPE, 2004) is the pioneering work in this literature.

5.1 Klette and Kortum (JPE, 2004)

Empirical facts This paper is motivated by a few stylized facts relating to R&D expenditures:

1. Productivity and R&D expenditures are positively correlated, but productivity growth is not correlated with R&D.
2. Patents are positively correlated with R&D expenditures.
3. R&D intensity (namely, R&D expenditures scaled by some measure of size, e.g. sales) is independent of size. Namely, R&D expenditures are proportional to size.
4. The distribution of R&D intensity is right-skewed, or many large firms do little R&D and few do a lot. In fact, there are lots of firms that do not do R&D, even after conditioning for size.
5. Differences in R&D across firms are persistent.
6. R&D at the firm level is a geometric random walk. That is, growth rates of R&D expenditure are uncorrelated with size.

More generally, the authors document the following facts regarding industry dynamics:

- The size distribution is right-skewed. Namely, there are few large firms and many small ones.
- The variance of growth rates is declining with size.
- Smaller firms have higher exit rates. Younger firms have higher exit rates as well.

The last point has been cast into doubt recently with Census data (recall Section 1), a set of facts that this paper predates. In this respect, the model will have no particular prediction, as age is irrelevant once controlling for size.

The Model Klette and Kortum then write a model of firm dynamics with R&D expenditures to account for these facts.

- Basics:
  - Time is continuous and infinite.
  - There is a $[0, 1]$ continuum of goods, and a firm is defined as a portfolio of goods.
  - By assumption, every single good generates a flow profit $\tilde{\pi} \in [0, 1]$ for the firm. A firm producing $n \geq 1$ goods then has flow profits of $\tilde{\pi}n$.
- R&D:
  - Firms gain and lose products on the basis of a Schumpeterian process of creative destruction. R&D is undirected and, if the firm is successful, it results into business-stealing: the firm acquires the good produced by someone else and starts producing it at a higher productivity, thereby allowing the new innovator to out-price the old producer.
When choosing R&D, the firm undertakes an expenditure to raise a probability \( \lambda > 0 \) of gaining a product. Similarly, let \( \mu > 0 \) be the probability of losing a product.

Formally, to raise a Poisson arrival of innovation \( I \), a firm of size \( n \) has to undertake an expenditure \( R \), where

\[
R = C(I, n) = nc \left( \frac{I}{n} \right)
\]

where \( c \) is increasing an strictly convex in \( I \), and \( n \), the number of products in the firm’s portfolio, can be thought of a proxy for the firm’s stock of knowledge (as it is correlated to the number of times that the firm has been successful in innovation). The assumption of strict convexity means that there exist decreasing returns to research effort.

An implication of the decreasing returns assumption is that R&D cost per good is

\[
\frac{C(I, n)}{n} = c \left( \frac{I}{n} \right)
\]

and thus

\[
\frac{\partial}{\partial n} c \left( \frac{I}{n} \right) = c' \left( \frac{I}{n} \right) \left( -\frac{1}{n^2} \right) < 0
\]

Namely, average cost is decreasing in size. Yet, note that

\[
C(\kappa \cdot I, \kappa \cdot n) = \kappa \cdot nc \left( \frac{\kappa \cdot I}{\kappa \cdot n} \right) = \kappa \cdot C(I, n)
\]

for any constant \( \kappa > 0 \), namely the cost function is homogeneous of degree one in \((I, n)\).

Since R&D is the sole engine of firm growth, and because flow profits are linear in size, this will immediately yield Gibrat’s law at the firm level (unconditional on survival): firm value is linear in size, and thus the growth rate of all firms will be independent of size. Conditional on survival, growth rates will be higher for small firms, and smaller for large firms.

**Optimal R&D choices:**

- The value function of a firm satisfies the following HJB equation:

\[
rV(n) - \dot{V}(n) = \max_i \left\{ \tilde{\pi}n - C(I, n) + I \left( V(n + 1) - V(n) \right) - \mu n \left( V(n) - V(n - 1) \right) \right\}
\]

Notice that we are assuming partial-equilibrium for now by taking \( \tilde{\pi} \) as given by the firm. Later on we will solve the GE version in which \( \tilde{\pi} \) is endogenous and, in particular, grows over time. Since we are assuming for now that there is no growth in the economy (as \( \tilde{\pi} \) is fixed), we impose stationarity so that \( \dot{V} = 0 \). In the GE version, of course, \( \dot{V} \neq 0 \).

- Guessing that \( V(n) = vn \), the FOC reads

\[
c' \left( \frac{I}{n} \right) = v
\]

meaning that all firm have the same R&D intensity, \( \lambda \equiv \frac{I}{n} \), regardless of size. Using the method of undetermined coefficients to determine \( v \), we can plug the optimal \( \lambda \) back into the objective to obtain
\[ \text{run} = \tilde{\pi}n - nc(\lambda) + \lambda n v - \mu n v \]

which yields

\[ (t + \mu - \lambda)v = \tilde{\pi} - c(\lambda) \]

- We therefore obtain that innovation intensity \((\lambda)\) is increasing in \(\tilde{\pi}\), decreasing in \(r\), and decreasing in \(\mu\) (the creative-destruction rate). Similarly, research intensity is \(C(\lambda, 1) / n = C(\lambda, 1) = c(\lambda)\), which is increasing in \(\lambda\). Therefore, research intensity is also increasing in \(\tilde{\pi}\) and decreasing in both \(r\) and \(\mu\).

- Firm’s life cycle:

  - Let \(P_n(t; n_0)\) be the probability of having \(n\) goods for a firm at time \(t\) given it had \(n_0\) goods at time \(t = 0\). This is described by the following ODE:

    \[
    \dot{P}_n(t; n_0) = \left( (n - 1)\lambda P_{n-1}(t; n_0) + (n + 1)\mu P_{n+1}(t; n_0) - n(\lambda + \mu)P_n(t; n_0) \right)\
    \text{Gaining the } n^{th} \text{ good through R&D}\
    \text{Losing the } (n+1)^{th} \text{ good due to creative destruction}\
    \text{n-sized flowing out due to gain or loss}\
    \text{Inflow (into } n)\
    \text{Outflow (out of } n)\
    \]

    for any \(t > 0\) and \(n \geq 1\), with boundary condition

    \[
    \dot{P}_0(t; n_0) = \mu P_1(t; n_0)\
    \]

    Gaining or losing two or more products does not feature into the accounting equation because these events are \(o(n)\), thanks to the fact that innovation is modeled as Poisson events. Moreover, note that this ODE takes advantage of the fact that R&D intensity is constant in \(n\), which allows us to write the evolution of the firm size distribution in a simple closed-form. This is one of the main technical contributions of the paper.

  - In all of the following, we will assume a typical firm is born with size \(n = 1\) without loss of generality. The solution to the ODE is

    \[
    P_0(t; 1) = \frac{\mu(1 - e^{(\mu - \lambda)t})}{\mu - \lambda e^{-(\mu - \lambda)t}} = \frac{\lambda}{\mu} \gamma(t)\
    P_1(t; 1) = [1 - P_0(t; 1)][1 - \gamma(t)]\
    P_n(t; 1) = \gamma(t)P_{n-1}(t); \ n = 2, 3, \ldots
    \]

    where

    \[
    \gamma(t) = \frac{\lambda(1 - e^{(\mu - \lambda)t})}{\mu - \lambda e^{-(\mu - \lambda)t}}\
    \]

    Since we will show (in the GE version) that \(\lambda < \mu\), then we obtain that \(\lim_{t \to +\infty} P_0(t; 1) = 1\). In words, firms exit the market almost surely\(^{16}\). The probability of being size \(n\) at time \(t\), conditional on survival, is

    \[
    \frac{P_n(t; 1)}{1 - P_0(t; 1)} = (1 - \gamma(t))^{n-1}\
    \]

\(^{16}\)Of course, there is a counter-balancing force coming from endogenous entry, which we are not specifying yet.
Conditional on survival, firms keep growing as their probability of exiting decays down to zero. Indeed, we have

$$\Pr_t[n \leq n^*|\text{Survival}] = (1 - \gamma(t)) \sum_{j=1}^{n^*} (\gamma(t))^{j-1}$$

and so \(\lim_{t \to +\infty} \Pr_t[n \leq n^*|\text{Survival}] = (1 - \lambda \mu) \sum_{j=1}^{n^*} (\lambda \mu)^{j-1}\), meaning that conditional on having survived firms keep growing until they eventually exit with probability one.

- The probability that a firm of size \(n\) exits within \(n\) periods is
  
  $$P_0(t; n) = [P_0(t; 1)]^n$$

  That is, the probability that a large firm of size \(n\) loses all of its \(n\) products is the product of the probability that \(n\) firms of one product exit. This is, there is independence across firm sizes, allowing us to write the joint probability as a product of the marginals. This, again, is a by-product of the linearities that are built into the value of the firm.

- The hazard rate of exit is
  
  $$\frac{\dot{P}_0(a; 1)}{1 - P_0(a; 1)} = \frac{\mu(\mu - \lambda)}{\mu - \lambda e^{-(\mu - \lambda)a}} = \mu(1 - \gamma(a))$$

  Intuitively, the hazard rate is the product of the probability of having reached size \(n = 1\) by time \(a\) (the term \(1 - \gamma(a)\)) and the probability of losing that one product (the term \(\mu\)). Note that the hazard rate is decreasing in \(a\), namely the probability of exiting (conditional on survival) in decreasing in the age of the firm. In the limit,

  $$\lim_{a \to +\infty} \frac{\dot{P}_0(a; 1)}{1 - P_0(a; 1)} = \mu - \lambda > 0$$

  namely firm exit almost surely, even conditional on survival.

- Conditional on survival, the expected size of a firm of size \(a\)

  $$\sum_{n=1}^{+\infty} n P_0(a; 1) = \frac{1}{1 - \gamma(a)}$$

  which is increasing in \(a\). Therefore, conditional on survival, a firm grows in expectation as it ages.

- The average size by age \(a\) of a cohort of \(m\) firms starting with \(n = 1\) goods is

  $$m \sum_{n=0}^{+\infty} n P_0(a; 1) = m \frac{1 - P_0(a; 1)}{1 - \frac{2}{\mu} P_0(a; 1)}$$

  Thus, the size of a cohort shrinks monotonically as it ages, and in the limit is shrinks down to zero (consistently with the fact that all firms exit almost surely, both conditionally and unconditionally). Both of these facts are consistent with the data.

- We can also compute firm growth. Let \(N_t\) be the size of a firm as of time \(t\), and \(G_t \equiv \frac{N_t - N_{t-1}}{N_{t-1}}\) the growth of a firm from time zero to \(t\). Then,

  $$\mathbb{E}[G_t|N_0 = n_0] = e^{-(\mu - \lambda)t} - 1$$
Therefore, the growth rate of a firm from zero to $t$ is independent of the firm’s initial size ($N_0$), an unconditional version of Gibrat’s law. Conditionally on survival

$$E[G_t | N_t > 0; N_0 = n_0] = \frac{e^{-(\mu - \lambda)t}}{1 - P_0(t; 1)^n} - 1$$

which is now decreasing in the initial size. Thus, the model delivers a deviation from the conditional Gibrat’s law: conditional on survival, small firms grow faster. Both of these predictions are, once again, consistent with the stylized facts documented by the authors.

- The variance of growth rates is

$$\text{Var}[G_t | N_0 = n_0] = \frac{\lambda + \mu}{n_0(\mu - \lambda)} e^{-(\mu - \lambda)t} [1 - e^{-(\mu - \lambda)t}]$$

which is also decreasing with $N_0$, again consistent with the data. Unconditional on survival, smaller firms have more volatile growth rates.

- Heterogeneity in research intensities:

• Now, we suppose that there is heterogeneity in flow profits per good. Let $\pi \in [0, 1]$ now be the flow product per good, with $\tilde{\pi} = E[\pi]$. The goal is to have differences in research intensity without having differences in growth, so we can still obtain the unconditional Gibrat’s law and the conditional deviation predictions. In order to obtain this result, we can simply restate the problem by assuming that firms are heterogeneous in their R&D cost functions.

- In particular, suppose the cost function is indexed by $\pi$, such that

$$C_\pi(I, n) = \frac{\pi}{\bar{\pi}} C(I, n)$$

where $C(I, n)$ is the cost function from above. Then, the value of firm $\pi$ is

$$rV_\pi(n) - V_\pi(n) = \max I \left\{ \frac{\pi n - \pi nc}{n} \left( \frac{I}{n} \right) + I \left( V_\pi(n+1) - V_\pi(n) \right) + \mu n \left( V_\pi(n) - V_\pi(n-1) \right) \right\}$$

so that, using again that $V_\pi(n) = v_\pi n$, the FOC gives

$$-\frac{\pi}{\bar{\pi}} nc' \left( \frac{I}{n} \right) \frac{I}{n} + v_\pi = 0$$

and thus $c' \left( \frac{I}{n} \right) = v$, where $v = \frac{\pi}{\bar{\pi}} v_\pi$. Therefore, $I/n$ is the same across all $\pi$’s, as sought, so that we have that all firms have the same $\lambda$. Since all firms have the same innovation intensity, they will all grow at the same rate and the Gibrat’s law predictions are still holding. However, there is now heterogeneity in research intensity, as

$$\frac{\pi}{\bar{\pi}} nc(\lambda) = \frac{\pi}{\bar{\pi}} c(\lambda)$$

which says that smaller firms (now measured as firms with productivity lower than the mean productivity $\bar{\pi}$) spend relatively more in research per good, in accordance to the stylized facts listed above.

\textsuperscript{17}Seen differently, above we have solved the problem for the firm whose productivity per good is exactly equal to the mean productivity in the economy. Now, we study the problem of any other firm.
• Equilibrium innovation rates:
  – Let \( M_n(t) \) be the measure of firms with \( n \) products at time \( t \). Since there is a mass-one of products, then we must have that
    \[
    \sum_{n=1}^{+\infty} nM_n(t) = 1
    \]
  for all \( t \). The industry-wide innovation rate is, then:
    \[
    \sum_{n=1}^{+\infty} M_n(t)I(n) = \sum_{n=1}^{+\infty} M_n(t)\lambda n = \lambda
    \]
  Since there is endogenous entry, there is an entry rate \( \eta \) equal to the probability that any incumbent loses a good to a size-zero firm, where of course
    \[
    \eta + \lambda = \mu
    \]
  where recall that \( \mu \) is the probability of losing a good to either an entrant or an incumbent (the creative-destruction rate). In particular, we assume there is an entry cost of \( F > 0 \) (a parameter) and, upon entry, the firm draws \( \pi \) from some cdf \( \phi(\pi) \). Thus, \( F = \mathbb{E}[\nu] \), meaning \( v = F^{18} \) The mass of potential entrants is unity at any instant.

  – Using that \((r + \mu - \lambda)\) \( v = \tilde{\pi} - c(\lambda) \), \( v = F \) and \( \mu = \lambda + \eta \), we obtain the following entry rate in equilibrium:
    \[
    \eta = \frac{\tilde{\pi} - c(\lambda)}{F} - r
    \]
  so in order to have positive entry we need \( \frac{\tilde{\pi} - c(\lambda)}{F} > r \).

• Firm size distribution:
  – We are now ready to compute the firm size distribution. For any \( n \geq 2 \), we have
    \[
    \dot{M}_n(t) = (n-1)\lambda M_{n-1}(t) + (n+1)\mu M_{n+1}(t) - \frac{n(\lambda + \eta)M_n(t)}{n \text{-sized firms either losing or gaining}}
    \]
  while for \( n = 1 \) firms,
    \[
    \dot{M}_1(t) = \eta + 2\mu M_2(t) - \mu M_2(t) + \eta M_2(t)
    \]
  where we have used that the individual innovation rate by incumbents (\( \lambda \)) is also the industry-wide innovation rate.

  – In steady state, \( \dot{M}(t) = \eta - \mu M_1(t) \), where \( M_1 = \frac{n}{\pi} \), so using the flow equations we get the steady state distribution:
    \[
    M_n = \frac{\lambda^{n-1}\eta}{n\mu} = \frac{\theta}{n} \left( \frac{1}{1 + \theta} \right)^n
    \]
  where \( \theta \equiv \eta/\lambda \). This allows us to compute the steady-state unconditional probability of having \( n \) goods at any time, \( P_n \), which is given by

\[ ^{18} \text{Recall } v_r = \frac{\pi}{\bar{\pi}} \text{ and } \mathbb{E}[\pi] = \tilde{\pi}. \]
\[ P_n = \frac{M_n}{M} = \frac{(1 + \theta)^{-n}}{n \ln ((1 + \theta)/\theta)} \]

- GE version:

  - Suppose that the above economy is endowed with a total mass \( L \) labor. There are three types of labor, all receiving a wage \( w \): manufacturing labor, used to produce the final good \( (L_x) \), labor used to innovate at start-ups \( (L_s) \), and labor used to innovate at incumbents \( (L_R) \).
  
  - The entry cost is now in labor units, given by \( F = wh \), where \( h \) is a parameter equal to the number of researchers needed for entrants to create an innovation.
  
  - The cost of innovation for an incumbent is equal to \( wC_\pi(x) \), where
    \[
    C_\pi(x) = \frac{\pi}{n} n \ell_R \left( \frac{nx}{n} \right)
    \]
    is the number of skilled workers needed to generate an innovation rate of \( nx \) for a firm \( \pi \), where \( \ell_R \) is monotone and convex.
  
  - Now we make explicit the Schumpeterian process of innovation: when a firm innovates upon someone else’s product, the new producer not only acquires the good, but is also able to produce it at the same cost but a higher quality/efficiency. This allows the most recent innovator to price the good in such a way that the follower (last innovator, who faces the same cost but can produce the good at a lower productivity) is indifferent between producing it or not. In this way, a monopolistically competitive structure arises endogenously: there is a continuum of vintages, as measured by the highest quality with which each good is produced, where each vintage is produced by only one firm, who acts as a monopolist of that variety and whose productivity advantage enables it to out-price any of its competitors. There is Bertrand competition between the new and the old producer, and we will assume a tie-breaking rule such that, given equal prices, the consumer will always prefer the latest vintages.
  
  - In particular, every cohort of entrants makes a draw from \( \Psi(q) \), with support \([1, +\infty)\), which it keeps for the duration of its firm. The draw \( q \) indexes the multiple by which this firm improves the quality of the goods that it acquires from incumbents through creative destruction.
  
  - Consumers maximize the following preferences:
    \[
    \int_0^{+\infty} e^{-\rho t} \ln C_t dt
    \]
    subject to
    \[
    \int_0^1 \sum_{k=1}^{J_t(j)} p_t(j; k)x_t(j; k) dj = 1
    \]
    where
    \[
    C_t \equiv \exp \left( \int_0^1 \ln \left( \sum_{k=1}^{J_t(j)} x_t(j; k)z_t(j; k) \right) dj \right)
    \]
    where \( J_t(j) \) is the both number of vintages of product \( j \) available as of time \( t \) and the latest vintage of good \( j \), \( z_t(j; k) \) is the quality of product \( j \) of vintage \( k \), and \( x_t(j; k) \) is
the consumer’s demand.

- The FOC with respect to vintage \( k \) of good \( j \) is

\[
\frac{z_t(j; k)}{\sum_{k=1}^{J_t(j)} x_t(j; k) z_t(j; k)} = \lambda p_t(j; k)
\]

where \( \lambda \geq 0 \) is the Lagrange multiplier. This implies that

\[
\frac{p_t(j; k)}{p_t(j; k-1)} = \frac{z_t(j; k)}{z_t(j; k-1)} = q
\]

where the last equality is by definition: the ratio of the newest to the second-to-newest quality of good \( j \) is the step size \( q \) that was drawn from \( \Psi \) by the firm that made such latest innovation when it first entered into the market. Thus, each good \( j \) has an associated distribution of step sizes, which is precisely \( \Psi \).

- Given our tie-breaking rule and no price discrimination, the lowest price that the newest innovator can charge in order to out-price the latest innovator is precisely \( p = qw \), where \( w \) (the wage) is the marginal cost, and \( q \) (quality step) acts as the mark-up. Incidentally, this also implies that each firm is produced by at most one firm, giving rise to a monopolistically competitive structure, and that in equilibrium only the very latest vintages, those of highest quality, are produced.

- Thanks to log-preferences, the consumer spends the same amount of resources on each good-vintage pair, or

\[
p_t(j; k) x_t(j; k_j) = p_t(j'; k_j') x_t(j; k_j')
\]

for any \( (j, j') \in [0, 1] \times [0, 1] \). This means total expenditure, which can be written as

\[
\int_0^1 p_t(j; k_j) x_t(j; k_j) dj = 1
\]

where the RHS is a normalization, implies that

\[
x_t(j; k_j) = \frac{1}{p_t(j; k_j)} = \frac{1}{qw}
\]

where \( q \) is the latest upgrade on good \( j \), i.e., the quality step associated to the latest innovator of the good. Note flow profits for the producer of good \( j \) at time \( t \) is

\[
\pi_t(j) = (p_t(j) - w) x_t(j) = (qw - w) \frac{1}{qw} = 1 - \frac{1}{q}
\]

and therefore the distribution of \( \pi \), which we assumed exogenously in the PE version, is not pinned down by the distribution of \( q, \Psi \). In other words, the firms are heterogeneous in their ability of improving upon goods, \( q \), which stays fixed throughout the firm’s life. This translates into firm heterogeneity in flow profits, which justifies what we assumed initially.

**Definition 4 (Equilibrium)** A competitive equilibrium of this economy is an interest rate \( r \), a wage rate \( w \), firm value \( v \) (with \( v = \frac{V(n)}{n} \), where \( V(n) = n v \)), an innovation rate for incumbents \( \lambda \) and one for entrants \( \eta \), such that (i) entrants break even; (ii) incumbents maximize and their value is given by \( v \); (iii) consumers maximize; (iv) the wage rate clears the labor market.

- Noting that \( \mathbb{E}[v_n] = \mathbb{E}[\frac{\pi}{v}] = v \), then the free-entry condition reads
\[ v = wh \]

The incumbent’s problem is

\[ \max I \quad -wn \pi \ell_R \left( \frac{I}{n} \right) + I \pi v \]

meaning

\[ v \ell_R(\lambda) = v \]

Thus, similarly to what we found in the PE version of the model, a firm’s demand of skilled labor is invariant to \( n \) (the number of goods in its portfolio). The demand for researchers by all incumbents with \( n \) goods in their portfolios is

\[ L_R(n) = n \ell_R(\lambda) \]

as \( \ell_R(\lambda) \) is common across firms of the same \( n \), and thus the total demand for skilled labor for incumbents in the economy is

\[ L_R = \sum_{n=1}^{+\infty} n M_n \ell_R(\lambda) = \ell_R(\lambda) \sum_{n=1}^{+\infty} n M_n = \ell_R(\lambda) \]

The number of researchers demanded by entrants is simply \( L_S = \eta h \). Lastly, the amount of labor demanded by the final good sector is

\[ L_X = \int_0^1 \frac{1}{w q(j)} dj = \frac{1}{w} \int_0^1 \frac{1}{q(j)} = \frac{1 - \bar{\pi}}{w} \]

where we have used that \( \bar{\pi} = \int_0^1 \pi_j dj = \int_0^1 \left( 1 - \frac{1}{q(j)} \right) dj = 1 - \int_0^1 \frac{1}{q(j)} dj \). The wage then solves

\[ L = L_X(w) + \ell_R(\lambda) + \eta h \]

- The budget constraint for the consumer can be written as

\[ Y = wL + rv + \frac{v}{h} L_S - wL_S \]

In other words, the consumer’s expenditure is her labor income \((wL)\), her capital income from owning incumbent firms \((rv)\) and entrants \((\frac{v}{h} L_S)\), and the resources needed to finance entrants \((wL_S)\). Note the last two terms are equal to each other by free entry, and therefore

\[ Y = wL + rv \]

In a stationary equilibrium with no output growth, the transversality condition then implies that \( r = \rho^{19} \). Normalizing \( Y = 1 \) we get

\[ w = \frac{1}{L + rh} \]

Note that the economy is being solved for a stationary equilibrium with no aggregate economic growth \((g \equiv \frac{\dot{Y}}{Y} = 0)\), implying that total expenditures and firm value are also not growing. This does not imply, however, that product quality is not growing. In particular, the object \( \ln \int_0^1 x \ell(j) z(j) \) is growing, because so is \( z \).
and market clearing reads

\[ L_R + L_S = L - \frac{1 - \bar{\pi}}{w} = \bar{\pi}L - \rho h(1 - \bar{\pi}) \]

or \( L_S = \bar{\pi}L - \rho h(1 - \bar{\pi}) - \ell_R(\lambda) \). Moreover, recall that the creative destruction rate is \( \mu = \lambda + \eta \), so

\[ \mu = (\ell'_R)^{-1}(h) + \frac{L_S}{h} \]

Finally, we can get the growth rate of qualities. Note we can write consumption in good \( j \) as

\[
\ln \left( \sum_{k=1}^{J_t(j)} x_t(j; k) z_t(j; k) \right) = \ln \left( \frac{1}{w q_t(j; J_t(j))} \prod_{k=1}^{J_t(j)-1} q(j; k) \right) = \ln \left( \prod_{k=1}^{J_t(j)-1} q(j; k) \right) - \ln w = \sum_{k=1}^{J_t(j)-1} \ln (q(j; k)) - \ln w
\]

where, in the first line, \( \frac{1}{w q_t(j; J_t(j))} = x_t(j) \) is the demand for good \( j \), coming only from the latest quality (that is, formally \( x_t(j; k) = 0, \forall j \in \{1, \ldots, J_t(j) - 1\} \) and \( x_t(j; J_t(j)) = \frac{1}{w q_t(j; J_t(j))} \), and \( \prod_{k=1}^{J_t(j)-1} q(j; k) \) is a measure of the quality of the good.\(^{20}\) Thus, total consumption is

\[
\ln C_t = \int_0^1 \ln \left( \sum_{k=1}^{J_t(j)} x_t(j; k) z_t(j; k) \right) \, dj = \int_0^1 \left( \sum_{k=1}^{J_t(j)-1} \ln (q(j; k)) - \ln w \right) \, dj = \mu t \ln \bar{q} - \ln w
\]

meaning \( C_t = \bar{q}^{\mu t} w \), where we have defined \( \bar{q} \) as the geometric mean

\[ \bar{q} \equiv \exp \left( \int_1^{+\infty} \ln(q) \Psi(q) dq \right) \]

Therefore, notice that the growth rate of consumption is positive, and driven by growth in the aggregate quality of the economy, \( \bar{q} \).

6 Other Topics

In this final section, we provide a brief overview of other leading theories on firm dynamics.

We will analyze three recent trends of work: uncertainty shocks, credit shocks and asset pricing.

\(^{20}\)This is because initial quality is unity by normalization, so the product of all step sizes up to time \( t \) is also the total quality of the good by time \( t \).
6.1 Uncertainty Shocks and Firm Dynamics

In this section, we discuss two seminal papers on a recent literature examining the impact of uncertainty shocks on firm dynamics. Specifically, these papers study whether time-varying uncertainty can have an impact on investment. Bloom (ECMA, 2009) pioneered this literature with a PE model to assess the qualitative effect of uncertainty shocks on investment, and Bachman and Bayer (JME, 2013) provided a GE version of the model that they could calibrate to the U.S. economy. The main idea goes back to Keynes (1936), who already pointed out that expectations play a major role in the determination of capital, and that expectations fluctuate greatly over the cycle due to uncertainty. Bernanke (1983) was an early attempt to write a model in which irreversible investment decisions (e.g. investment subject to a fixed cost) create an option value of waiting to acquire more information that reduces uncertainty. This wait-and-see effect is the critical ingredient in this literature, and is typically generated by the increase in variance in investment returns and the irreversibility in investment (the presence of a fixed cost).

Bloom (ECMA, 2009)

Bloom (2009) departs from the observation that the implied volatility (the variance in expectations) in output is highly countercyclical (see Figure 1). Moreover, this measure of uncertainty is highly correlated with stock market volatility, the cross-sectional standard deviation of firm stock returns, firm profit growth, industry TFP growth, GDP forecasts and other idiosyncratic measures of uncertainty in the economy (see Table I in the paper). Moreover, he provides VAR evidence with innovations in these volatility indices to show that, at high frequencies, the shock causes very short-lived recessions followed by a rebound and a slow recovery from above (see Figures 2 and 3). Next, Bloom writes a PE model that shows how uncertainty shocks can generate drops in investment.

- Environment:
  - The output of a firm is
    \[ s(A, K, L, H) = A^{1-a-b} K^a (LH)^b \]
    where \( K \) is physical capital, \( H \) is hours, \( L \) is labor, \( a, b < 0 \) and \( a + b < 1 \), with \( A \) being TFP (say, productivity of demand shocks) given by
    \[ A = A^M \cdot A^F \]
    where \( A^M \) are aggregate shocks given by
    \[ A^M_t = A^M_{t-1} (1 + \sigma_{t-1} W^M_t) \]
    with \( W^M_t \sim \mathcal{N}(0,1) \), and \( A^F \) are firm-level shocks given by
    \[ A^F_{it} = A^F_{it-1} (1 + \mu_{it} + \sigma_{t-1} W^F_{it}) \]
    with \( W^F_{it} \sim \mathcal{N}(0,1) \), and \( \sigma_t \in \{\sigma_L, \sigma_H\} \) is a Markov process with transition matrix whose typical element is \( \pi_{kj} \equiv P(\sigma_t = j | \sigma_{t-1} = k) \).

\[ ^{21} \text{In particular, he uses indicator regressors that switch on for particularly stark peaks in implied volatility –about 18 so-called “tail” events in his 1960-2009 sample.} \]
The novelty of the paper is the time-varying nature of the volatility of the shock. Moreover, the shocks are common in both aggregate and idiosyncratic shocks.

There are adjustment costs in capital \((K)\) and labor \((L)\), and wages are a function of hours \((H)\), such that \(w(H) = w_1(1 + w_2H^\gamma)\), where \(w_1, w_2, \gamma\) are parameters that are pinned down empirically. Otherwise, prices are given.

- **Impulse responses:**

The exercise is then to shock \(\sigma\) and see how the economy reacts. The shock is to assume \(\sigma_t = \sigma_L\) for all \(t < T\), then \(\sigma_T = \sigma_H\) when the shock hits, and then let \(\sigma_t\) recover through the Markov process for \(t > T\). His calibration delivers a very fast reversion in the \(\sigma\) shock. The spirit of the exercise is to assume a large but short-lived spike in volatility that reverts quickly, like in the data.

The shock creates an initial decline in labor, followed by a recovery that overshoots and reverts back from above. The reason is that when the shock hits \((at t = T)\), the inaction region increases as expectations deteriorate and firms stop hiring momentarily. This is followed by a recovery, which is the result of the fast mean-reversion of the shock and the recovery in expectations: the labor-hiring inaction region shrinks quickly but firms were now far from their frictionless level, so hiring spikes and labor recovers persistently. Moreover, the shock generates higher volatility in outputs (the distribution of observables increases in variance as firms disperse in the distribution), so a bigger number of firms invest intensively while firms in the lower tails stay idle, which generates the subsequent overshooting.

The response can thus be decomposed in two parts (see Figure 9 in the paper). First, after the \(\sigma\) shock, firms revise their expectations and are more pessimistic about the future (the “uncertainty effect”). This, on its own, does not create a recession but only an expansion due to a precautionary motive. Second, in the opposite direction, once \(\sigma\) reverts quickly back to \(\sigma_L\) expectations are quickly re-adjusted (the “volatility effect”), so the reaction in investment is the opposite and, absent the uncertainty effect, the shock actually does not cause a boom. The combination of the two gives the observed IRFs: an initial short-lived recession, a rebounding expansion that overshoots and a mean-reversion from above.

Similarly, TFP first drops due to a large drop in reallocation from low- to high-productive firms. When the shock recovers and inaction regions shrink, the productivity distribution spreads out and creates both highly productive firms and highly unproductive ones, creating large reallocation from the latter to the former and a recovery in TFP that rebounds and overshoots.

**Bachman and Bayer (JME, 2013)**

Bachman and Bayer (2013) proposed a GE version of the same idea. Essentially, they introduce time-varying volatility into a Khan and Thomas (2008) framework and ask the same question as Bloom (2009): what are the effects of spikes in uncertainty when investment is irreversible?

- **Environment:**

Output is given by

\[
y_t = z_t\varepsilon_t K_t^\theta N_t^\nu
\]
where
\[
\log \varepsilon_{t+1} = \rho \log \varepsilon_{t-1} + \sigma_{\varepsilon} \eta_{\varepsilon,t+1}
\]
and
\[
\log z_{t+1} = \rho z_{t-1} + \sigma_{\varepsilon} \eta_{z,t+1}
\]
with \(\sigma_{\varepsilon} - \bar{\sigma}_{\varepsilon}\) and \(\sigma_{\varepsilon} - \bar{\sigma}_{\varepsilon}\) being also AR(1) processes. Barred variables denote unconditional means. The innovations to first (\(\eta\)) and second (\(\sigma\)) moments are orthogonal to each other.

- The structure is otherwise as in Khan and Thomas (2008). The difference now that \(\sigma_{\varepsilon} \) and \(\sigma_{\varepsilon} \) are state variables in the decision of the firm.

- **Solution:**
  - The computation follows Krusell and Smith’s approach. In particular, they keep track of aggregate capital with a log-linear rule that incorporates the standard deviation of the shocks (see equation (9a) in the paper).
  - They then do a similar experiment to Bloom by looking at the impulse responses after a one-time shock in the volatility of shocks, \(\sigma_{\varepsilon}\), which recovers quickly. They show (in Figure 1) the responses of a PE model in which prices are given, and a GE model in which interest rates clear markets.
  - The PE response looks like Bloom’s: a short-lived recession on impact, a rebound and an overshoot. The GE version shows a deeper recession but a lot smaller rebound. The reason is that the GE incorporates a wealth effect on the consumer. Investment goes down on impact and consumption goes up and a drop in interest rates, but the wealth effect on labor supply (preferences are CRRA, not GHH) causes households to work less and wages to go up. The reason why there is less overshooting is because the consumer wants to smooth consumption, so the increase in investment is a lot more limited since the response in the interest rates mutes out the rebound.
  - Table 5 then shows that firm-level risk fluctuations added to first moment productivity shocks lead to similar business cycle dynamics as in RBC models, unless firm-level risk is (counterfactually) highly volatile.

- **Extension:**
  - One critical aspect is that \(\sigma_{\varepsilon}\) (idiosyncratic) and \(\sigma_{\varepsilon}\) (aggregate) shocks are uncorrelated. However, if they were not orthogonal, a bad shock to idiosyncratic volatility would now signal a drop in aggregate productivity in the future as well.
  - Assuming the latter structure, they show (in Figure 3) that the strong recession-rebound-overshooting structure re-emerges. Moreover, the response is very similar to one coming from a model that has no correlation between shocks but allows for news shocks (i.e. direct signals) on the evolution of \(\sigma_{\varepsilon}\) (what they call the “forecasting model”), conveying that the response in the GE model with non-orthogonal shock is mostly due to the signal-induced forecasting channel.

### 6.2 Credit Shocks and Firm Dynamics

Recently, the Great Recession and its aftermath has motivated new theories on how financial frictions can impact firms. A recent literature, led by Mian and Sufi, tries to understand how
disturbances in the labor market led to erosion in households’ balance sheets, household deleveraging and a subsequent drop in aggregate demand. In a parallel body of work, however, some papers show that the same disturbances that forced households to deleverage also contributed to generating a credit crunch on the firm (supply) side.

Siemer (WP, 2014) shows using the LBD data that small (and especially young) firms tend to be more financially constrained than large firms, so in response to the lending shock in the cross-section the former firms had to reduce job creation by more in sectors that were more intensively being financed through the financial sector. Since small and young firms are the ones who generate most jobs (by the Haltiwanger facts), the financial shock created a large recession through a credit crunch.

In this section we review one of such theories.

Kahn and Thomas (JPE, 2013)

This paper imposes a borrowing constraint into the prototypical heterogeneous-firm business cycle model. In a complete-market environment, we know by the Modigliani-Miller theorem that the debt structure of firms should not matter. To break this, Khan and Thomas will introduce a deadweight loss for bankruptcy and assume that firms cannot issue equity and must finance themselves through debt. For this, firms are going to keep on borrowing, subject to a borrowing constraint, until their capital level reaches the optimal unconstrained level. One implication of the model is that large firms will not hold debt, which is counterfactual. However, this is a simple parsimonious way to introduce debt, and the paper’s results on business cycles do not hinge critically on the assumption about debt.

Environment

- The model builds on Kahn and Thomas (ECMA, 2008). The production function is

\[ y = z\varepsilon F(k, n) \]

where \( z \) and \( \varepsilon \) are aggregate and idiosyncratic Markov shocks, respectively, both of which have a discrete support. \( z \) is as usual reflecting both technical efficiency (“revenue TFP”) and idiosyncratic demand conditions.

- There is an exogenous exit rate \( \pi_d \). Namely, exit rates are constant over the cycle, and are the same across size and age classes. This is of course counterfactual, but not critical, and it merely allows us to achieve a constant mass of firms in equilibrium. Each firm that exits is replaced by an entrant, all entering with initial size \( k_0 \), no debt \( (b = 0) \), and the \( z \) shocks is drawn from the invariant distribution. Note that entrants (young by definition) enter small, so there is no significant role for age after controlling for size.

- The borrowing constraint for a firm with capital \( k \) is

\[ b' \leq \zeta \theta_k k \]

and \( \theta_k \in [0, 1] \) is a constant. The interpretation is that the firm can resell a fraction \( \theta_k \) of its capital if it default on its debt. Moreover, \( \zeta \in \{\zeta_o, \zeta_f\} \) is the financial shock, with \( \zeta_o > \zeta_f \), which is a Markov chain with transition matrix
A financial crisis is defined as a tightening of credit standards, as defined by a $\zeta_o$-to-$\zeta_\ell$ transition.

- The parameters $(p_o, p_\ell)$ are calibrated with the Reinhart and Rogoff (AER, 2009) facts, who reported the average duration a developed country spends on a typical financial crisis state. In particular, they will get $(p_o, p_\ell) = (0.9765, 0.3125)$ meaning that crisis times are very infrequent, and the expected duration of a crisis is only about 3 years. Moreover, $\zeta_o$ is calibrated to match that, in the data, leverage (the ratio of debt to assets) is on average about 37% in normal times.

- On the investment side, if there is positive investment the law of motion is as usual:
  \[ k' = k(1 - \delta) + i \]
  where $i$ is investment expenditure in consumption units to expand capital by $[k' - k(1 - \delta)]$ units of capital. If there is negative investment (i.e., if $k' < l(1 - \delta)k$), then the number of units of consumption recovered is $i = \theta_k[k' - (1 - \delta)k]$. In other words, the price of capital is unity if the firm buys it, and it is $\theta_k < 1$ if the firm sells it. This generates that if the marginal product of capital is within the interval $[\theta_k, 1]$, then the firm stays put and $i = 0$ (an inaction region).

**Equilibrium**

- The beginning-of-period problem of a firm with state $(k, b, \varepsilon)$ facing aggregate state $s \equiv (z, \zeta)$ and a distribution over states $\mu \equiv \Delta(k, b, \varepsilon)$ is as follows:

  \[
  v_0(k, b, \varepsilon; s, \mu) = \pi_d \max_n \left\{ \zeta \varepsilon_i F(k, n) - w(s, \mu)n + \theta_k(1 - \delta)k - b \right\} + (1 - \pi_d)v(k, b, \varepsilon; s, \mu)
  \]

  where $v(k, b, \varepsilon; s, \mu)$ is the value of no liquidation, given by

  \[
  v(k, b, \varepsilon; s, \mu) = \max \left\{ v^u(k, b, \varepsilon; s, \mu), v^d(k, b, \varepsilon; s, \mu) \right\}
  \]

  Here, the value of a firm buying capital (i.e., when $k' > k(1 - \delta)$) is

  \[
  v^u(k, b, \varepsilon; s, \mu) = \max_{n, k', b', D_{j\varepsilon}} \left\{ D_{j\varepsilon} + \sum_{m=1}^{N_x} \sum_{i=1}^{N_x} \pi_{ij} \pi_{sm} d_m(s, \mu) \sum_{j=1}^{N_x} \pi_{ij} v_0(k', b', \varepsilon_j; s_m, \mu) \right\}
  \]

  subject to

  \[
  b' \leq \zeta \theta_k k
  \]

  where $\pi^*$ is the combined Markov chain of the two $z$ and $\zeta$ aggregate Markov shocks, $d_m$ is the SDF in state $m$, and
are dividends, where \(q\) is the price of debt. The law of motion of the distribution of states is

\[ \mu' = \Gamma(\mu, s_e) \]

On the other hand, if \(k' \leq (1 - \delta)k\), i.e. the firm is selling (and not buying) capital, then it receives similarly the value

\[ v^d(k, b, \varepsilon; s_e, \mu) = \max_{n, k', b'} \left\{ D_{ei} + \sum_{m=1}^{N_e} \pi^a_{em} d_m(s_e, \mu) \sum_{j=1}^{N_e} \pi^a_{ij} v_0(k', b', \varepsilon; s_m, \mu) \right\} \]

again subject to

\[ b' \leq \zeta \theta k \]

but now

\[ 0 \leq D_{ei} \leq z \varepsilon F(k, n) - w(s_e, \mu)n + q(s_e, \mu)b' - b - [k' - (1 - \delta)k] \]

because the selling price is \(\theta k\) times lower than the buying price.

- Note that firms are using the SDF to discount future flows because markets are complete (there is a complete set of Arrow securities for all realizations of the state).
- There is a representative household trading only with the firm. Household value is

\[ v^h(\lambda, \phi; s_e, \mu) = \max_{c, n^h, \lambda^h, \phi^h} \left[ u(c, 1 - n^h) + \beta \sum_{m=1}^{N_e} \pi^a_{em} v^h(\lambda', \phi'; s_m, \mu') \right] \]

subject to

\[ c + q(s_e, \mu) \phi' + \int_S \rho_1(k, b, \varepsilon; s_e, \mu) \lambda'(d[k \times b \times \varepsilon]) \leq w(s_e, \mu)n^h + \phi + \int_S \rho_0(k, b, \varepsilon; s_e, \mu) \lambda'(d[k \times b \times \varepsilon]) \]

and

\[ \mu' = \Gamma(\mu, s_e) \]

where \(c\) is consumption, \(n^h\) is supply of labor, and recall from Khan and Thomas (ECMA, 2008) that \(\rho_1\) is the end-of-period price of the portfolio (i.e, the value of the firm without the dividend pay-out, or “ex-dividend”) and \(\rho_0\) is the price at the beginning of the period (i.e, the value of the firm last period, including the dividend it paid out, or “cum-dividend”). Here, \(\phi\) are bond holdings (i.e, lending to the firm). In sum, all firms borrow at the same risk-free rate \((1/q)\), plus premium that, counterfactually, is the same for all firms.

**Definition 5** A recursive equilibrium is a tuple \(\{w, q, \rho_0, \rho_1, v_0, N, K, v^h, C^h, N^h, \Phi^h, \Lambda^h\}\) such that

1. \(v_0\) solves the firm’s problem and the associated policy functions are \(N\) (for labor demand), \(K\) (for capital demand), and \(B\) (for demand for borrowing).
2. $v^h$ solves the household’s problem, with associated policy functions $C^h$ (for consumption), $N^h$ (for labor), $\Phi^h$ (for bond holdings) and $\Lambda^h$ (for debt holdings, i.e. lending to firms).

3. Markets clear:
   
   (a) Stock market clearing:
   
   $$\Lambda^h(k, b, \varepsilon; s, \mu) = \mu'(k, b, \varepsilon; s, \mu)$$

   for all $(k, b, \varepsilon)$.  
   
   (b) Labor market clearing:
   
   $$N^h(\mu, \phi; s, \mu) = \int N(k, b, \varepsilon; s, \mu) \mu(d(k \times b \times \varepsilon))$$

   (c) Goods market clearing (resource constraint):
   
   $$C^h(\mu, \phi; s, \mu) = \int S\left\{\varepsilon F(k, N(\varepsilon, k; s, \mu)) - (1 - \pi_d)\varepsilon\left(K(k, b, \varepsilon; s, \mu) - (1 - \delta)k\right)\right\}$$

   where $\iota(x) = I[x > 0](1 - \theta_k)(\i.e., \iota = 1$ if the firm is buying capital $(k' > (1 - \delta)k)$

   and $\iota = \theta_k$ if the firm is selling capital $(k' \leq (1 - \delta)k)$.

4. The law of motion of the distribution is:

   $$\mu' = \Gamma(\mu, s)$$

   The bond market clears (i.e, $\Phi^h(\mu, \phi; s, \mu) = \int B(k, b, \varepsilon; s, \mu) \mu(d(k \times b \times \varepsilon))$) by Walras' law.

   - The solution algorithm is as in Kahn and Thomas (2008), namely we can renormalize values by the marginal utility of consumption and define

   $$p(s, \mu) = u_1(c, 1 - n)$$

   and

   $$w(s, \mu) = \frac{u_2(c, 1 - n)}{p(s, \mu)}$$

   so that the risk-free price of the bond is

   $$q(s, \mu) = \beta \sum_{m=1}^{N_s} \pi_m^s \frac{p(s_m, \mu')}{p(s, \mu)}$$

   The value of the firm can then be written as follows:

   $$v_0(k, b, \varepsilon; s, \mu) = \pi_d \max_n \left\{p(s, \mu) \left[z \varepsilon F(k, n) - w(s, \mu) n + \theta_k (1 - \delta) k - b\right]\right\} + (1 - \pi_d) v(k, b, \varepsilon; s, \mu)$$
• In terms of the Krusell-Smith type of numerical implementation, the aggregate state is going
to be $m$ (the mean of the capital distribution), $z$ (the mean of the aggregate shock), and the
($\zeta, \zeta_{-1}, \zeta_{-2}$) vector (the contemporaneous, plus two lags, of the binary financial conditions).
With respect to Khan and Thomas (2008), the computation also additionally keeps track of
debt $b$.

• In equilibrium, there are two types of firms. “Unconstrained firms” are firms that, no matter
their level of capital and financial conditions, they can always reach their unconstrained
efficient level of capital. “Constrained firms” are firms that are unable to in certain states of
the world. In particular, large firms have less debt and tend to be unconstrained, whereas
smaller firms are more constrained and usually cannot reach their efficient level because they
hold a lot of debt. Additionally, in order to generate the positive correlation between size
and leverage that we see in the data, the authors assume that there is a group of so-called
“no-constrained” firms, which are firms for which the borrowing constraint $b^i \leq \zeta_\theta k$ does
not apply. This third group of firms, unconstrained by definition, is added only for empirical
purposes.

**Empirical performance**

• In the estimated model, 9% of firms are unconstrained, 62% are constrained and 29% are
no-constraint firms. Entrants’ capital stock is on average 21% the size of the average capital
stock in the distribution.

• Figure 2 in their paper shows the life-cycle of debt and capital for a typical cohort. Firms
accumulate debt quickly to build up their capital stock, and once capital reaches its average
long-run level these firms start deleveraging. Thus, small firms grow faster through debt
accumulation and heavy leveraging.

• Tables 2 and 3 then report business cycle moments with and without the credit shocks. The
credit shock generates volatility with respect to output in all aggregate series, and makes
them slightly less procyclical. In an economy without productivity shocks and only credit
shocks (Table 4) we see that investment becomes very volatile, and consumption loses most
of its procyclicality. This is consistent with the fact the recession was followed by a slow
recovery and the deep drop in business investment compared to the drop in measured TFP
(see Figure 5).

• With IRFs, the authors show:
  – Figure 4: The productivity shocks (that replicate the drop in TFP in the data) alone
cannot account for the dramatic drop in investment and consumption and the slow
subsequent recovery. The reason is that the productivity shock affects all firms the
same.
  – Figure 6: Instead, when the shock is financial and there is a credit tightening, we get
different responses because the shock affects small and constrained firms more than
large unconstrained firms. The experiment is to assume that $\zeta = \zeta_\ell$ for four consecutive
periods, and $\zeta$ reverts back through the Markov chain from then on (on those periods,
the reported IRF is the average of hundreds of IRFs obtained by simulation). When
credit tightens, households supply less labor and small firms demand less labor and they
disinvest because the interest rate drops on impact (which incidentally is also responsible
for consumption going up on impact). Investment takes a few periods to reach the trough.
because the credit tightening acts as a financial accelerator that imposes a time-to-rebuild behavior on small firms. Moreover, measure TFP drops (even though exogenous TFP $z$ is constant) because the credit crunch generates capital misallocation: smaller growing firms that are short of their efficient size lose capital, which makes them even more inefficient. Yet, the drop is not as high as in the data. The drop in debt is matched because it is targeted in the calibration, but interestingly this generates a large response in investment. The overshooting in investment and employment is, however, counterfactual.

6.3 Asset Pricing and Firm Dynamics

Real business cycles models, while interested in the cyclical properties of economic aggregates, tend to provide counterfactual implications for asset pricing (for example, the equity premium puzzle). In 1992, Fama and French outlined the cross-sectional facts of asset returns and excess returns (i.e., return on equity in excess of the risk-free asset) and related them to firm-level characteristics.

1. Firstly, small firms (in the sense of low market capitalization, i.e. low stock market returns) typically earn higher returns. This observation is sometimes called the “size premium”.

2. Secondly, firms with low Tobin’s $q$’s (the ratio of stock market value to the book value of the firm, or the inverse of the book-to-market ratio) have also higher returns (the so-called “value premium”).

3. Finally, low investment rates also predict higher returns (the “investment rate premium”).

These empirical observations hold unconditionally, but also conditionally on one another, namely market cap, $q$ and the investment rates are all, separately and jointly, strong predictors of firm returns. Moreover, the three returns predictors are also strong predictors of risk: low market capitalization, low $q$’s and low investment rates are all typically observed in firms with riskier returns.

Based on these observations, theorists have attempted to write models of equity returns that yield these predictions. Unfortunately, finance and macroeconomics have been disconnected on this topics until fairly recently, when a new firm dynamics literature has emerged to explain these finance facts. An early example is Berk, Green and Nask (1999). In this section, we present a partial equilibrium model of firm dynamics, based on Clementi and Palazzo (NBER WP, 2015), that will speak to these facts, even though the pricing kernel (the stochastic discount factor) will be taken to be exogenous.

Clementi and Palazzo (NBER WP, 2015)

• There are three periods, $t = 0, 1, 2$.

• A firm is indexed by $(s_t, z_t, k_t)$, where $s$ (z) are idiosyncratic (aggregate) productivity levels, following $\text{AR}(1)$’s $z_t = \rho z_{t-1} + \varepsilon_{z,t}$ and $s_t = \rho s_{t-1} + \varepsilon_{s,t}$ for $t = 1, 2$. Shocks are orthogonal to each other, and $k_t$ is capital.

• In period $t = 0$, the firm may borrow and invest in capital. In $t = 1$, the dividend is $d_1 = e^{s_t + z_t} k_1 + k_1 (1 - \delta) - k_2$, where $k_2$ is the investment for the second period. Similarly, last period’s dividends are the liquidation value $d_2 = e^{s_2 + z_2} k_2 + k_2 (1 - \delta)$, after which the world ends.
Our goal is pricing equity, i.e. the cash-flows $d = (d_1, d_2)$. The value of the firm at time $t = 1$ is

$$v_1(k_1, s_1, z_1) = \max_{k_2} e^{s_1+k_1} k_1^\alpha + k_1(1 - \delta) - k_2 + E_1 \left[M_2 \left(e^{s_2+k_2} k_2^\alpha + k_2(1 - \delta)\right)\right]$$

where $M_2 \equiv M(z_1, z_2)$ is the SDF (the household’s price of risk), taken as given. Using backward induction, at time $t = 0$ the value is

$$v_0(k_0, s_0, z_0) = \max_{k_1} e^{s_0+k_0} k_0^\alpha + k_0(1 - \delta) - k_1 + E_0 \left[M_1 v_1(k_1, s_1, z_1)\right]$$

where $M_1 \equiv M(z_0, z_1)$. Notice we assume that the SDF is only affected by aggregate shocks, whereas idiosyncratic shocks are not being priced (the reason being that the household can always diversify idiosyncratic risk away by reshuffling her portfolio).

We define the expected equity return, $E_0 [R_e]$, by

$$E_0 [R_e] \equiv \frac{E_0 \left[v_1(k_1, s_1, z_1)\right]}{E_0 \left[M_1 v_1(k_1, s_1, z_1)\right]}$$

where the denominator is the $x$-dividend price as of time $t = 0$ (the value of cash-flows as evaluated by the risk-averse household), while the numerator is the expected value of the cum-dividend price of the firm (the actual value of the cash flow). In other words,

$$E_0 [R_e] = \frac{E_0 [d_1 + P_t]}{P_0}$$

where $P_t \equiv E_t [M_{t+1} v_{t+1}(k_{t+1}, s_{t+1}, z_{t+1})]$ is the expected price of the stock as of time $t = 0, 1$.

As of time $t = 0$, the only dimension of heterogeneity across firms is the idiosyncratic shock $s_0$. In order to understand the cross-sectional properties of equity returns, we then ask how expected equity returns $E_0 [R_e]$ are affected by $s_0$. Expanding the definition of equity returns, note:

$$E_0 [R_e] = \frac{E_0 \left[v_1(k_1, s_1, z_1)\right]}{E_0 \left[M_1 v_1(k_1, s_1, z_1)\right]} = \frac{E_0 [\Gamma_{cu}] + E_0 [\Gamma_{co}]}{E_0 [M_1 \Gamma_{cu}] + E_0 [M_1 \Gamma_{co}]}$$

where $\Gamma_{cu}$ is the payoff of the current asset and $\Gamma_{co}$ is the payoff of the continuation asset. Intuitively, we can think of equity as being a portfolio of two assets, current and continuation. The current asset pays $e^{s_1+k_1} k_1^\alpha + k_1(1 - \delta)$ today and nothing tomorrow, while the continuation value pays $-k_2$ today and $e^{s_2+k_2} k_2^\alpha + k_2(1 - \delta)$ tomorrow. Now, note we can expand

$$\frac{E_0 [\Gamma_{cu}] + E_0 [\Gamma_{co}]}{E_0 [M_1 \Gamma_{cu}] + E_0 [M_1 \Gamma_{co}]} = \frac{E_0 [M_1 \Gamma_{cu}] + E_0 [M_1 \Gamma_{co}]}{E_0 [M_1 \Gamma_{cu}] + E_0 [M_1 \Gamma_{co}]} + \frac{E_0 [\Gamma_{cu}] + E_0 [\Gamma_{co}]}{E_0 [M_1 \Gamma_{cu}] + E_0 [M_1 \Gamma_{co}]}$$

$$= \beta(s_0, z_0) + (1 - \beta(s_0, z_0))$$
where $\beta(s_0, z_0)$ is the loading of the portfolio on the current asset, and $1 - \beta(s_0, z_0)$ is the loading of the portfolio on the continuation asset. That is, $\beta$ is the fraction of firm value coming from current assets. Importantly, note that this decomposition shows that the equity value is a function of the states $(s_0, z_0)$ only through these loadings, while $E_0[\Gamma_{cu}]$ and $E_0[\Gamma_{co}]$ are the expected returns of the current and continuation values of the firm. What one can show is that $\beta$ is increasing in the shocks, so that the loading in the asset will be higher if the productivity of the firm is higher on the period at which the asset pays. In line with this, we assume

$$M_t = \exp\{-\gamma \epsilon_{z,t}\}$$

where $\gamma > 0$. Thus, if the firm pays a lot when the aggregate state is high, the expected return is high and the firm is not highly valued. In other words, the firm is value when it pays more in bad times, or the SDF is counter-cyclical. A GE model would of course deliver this by showing that the SDF is simply the ratio of marginal utilities. This reduced-form specification is enough for our purposes, however. For instance, the risk-free rate is

$$R_{ft} = \frac{1}{E_t[M_{t+1}]} = \frac{1}{E_t[\exp\{-\gamma \epsilon_{z,t+1}\}]} = \frac{1}{e^{-\frac{1}{2} \gamma^2 \sigma^2}}$$

where we have used log-normality in the last step. Note that the risk-free rate does not depend on the current state of nature. That is, the conditional risk-free rate is (albeit counterfactually) constant.

- The current asset’s risk is coming only through $(s_1, z_1)$. The expected return of this asset is, then

$$\frac{E_0[\epsilon^{s_1+z_1}k_1^0]}{E_0[M_1e^{s_1+z_1}k_1^0]} = \frac{E_0[e^{\rho s_0 + \epsilon_{s_1} + \rho z_0 + \epsilon_{z_1}}k_1^0]}{E_0[M_1e^{s_0 + \epsilon_{s_1} + \rho z_0 + \epsilon_{z_1}}k_1^0]} = \frac{E_0[e^{\epsilon_{s_1} + \epsilon_{z_1}}]}{E_0[e^{\epsilon_{s_1} + \epsilon_{z_1}}]} = \frac{E_0[e^{\epsilon_{s_1} + \epsilon_{z_1}}]}{E_0[e^{1 - \gamma \epsilon_{s_1} + \epsilon_{z_1}}]} = e^{\frac{1}{2} \gamma^2 \sigma^2}$$

where we have used the orthogonality of $z$ and $s$, and the last line uses log-normality. This means

$$\frac{E_0[\Gamma_{cu}]}{E_0[M_1\Gamma_{cu}]} = \chi e^{\gamma^2 \sigma^2} R_f + (1 - \chi) R_f$$

for some constant $\chi$ that we are not computing here. The current asset therefore has a risky piece (associated to the realization of the shocks tomorrow) and a riskless piece (the return on $k_1$, which is a given choice). The risky piece, calculated above, is only a function of the covariance between the SDF and the aggregate shock, magnified by the $\gamma$ parameter (which, in the GE version of the model, is the risk aversion coefficient). Similarly, the continuation asset has average return

62
\[
\frac{E_0[M_{co}]}{E_0[M_1M_{co}]} = \frac{E_0[e^{\frac{z_{t+1}}{1-\alpha}}]}{E_0[M_1e^{\frac{z_{t+1}}{1-\alpha}}]} = \cdots = e^{\frac{\gamma}{1-\alpha} \sigma_2^2 R_f}
\]

where we have skipped the straightforward algebra steps. Note that, now, \(\rho_z\) features into the return. If \(\rho_z\) is high, the covariance between the SDF and the payoff of the continuation asset tomorrow is high, because the \(z_1\) shock propagates into the returns tomorrow. Similarly, \(\alpha\) also shows up: if \(\alpha\) is high, then the payoff tomorrow is also highly correlated with the innovation of the \(z_1\) shock, because the more sensitive is the choice of \(k_2\) to the shock.

- **Facts:**
  - What we see therefore is that if \(\rho_z > 1 - \alpha\), then the continuation asset has a higher expected return than the current asset, because if so the covariance of the SDF and the innovations to the shocks is higher for the current asset.
  - Moreover, a low \(s_0\) (high \(s_0\)) will deliver higher (lower) returns. Since higher \(s_0\) means a higher capitalization for the firm (because here \(k_0\) is given and production, and thus capital, is monotonically increasing in \(k\)), the prediction is in accordance with the stylized facts that motivated this section.
  - A high book-to-value firm will also have higher returns, as in the data. For this, we can turn to the quantitative results of the infinite-horizon version of this model. In this model, the only changes are capital adjustment costs

\[
g(x, k) = \chi(x) c_0 k + c_1 \left(\frac{x}{k}\right)^2
\]

where \(c_0, c_1 \geq 1\) and \(x\) is the investment rate, and the SDF now reads

\[
\log M_{t+1} = \log \beta + [\gamma_0 z_t + \gamma_1 z_{t+1}]
\]

which now generates a counter-cyclical risk-free rate, and a constant price of risk\(^{22}\)

The goal is now to talk about market valuation, and Tobin’s \(q\), while being consistent with the data on the predictions in investment rate seen in Section 4.

- The optimization problem of the firm is

\[
v(\lambda, k, s) = \max_x \left\{ e^{x + k^\alpha} - x - g(x, k) + \int_{\lambda} M(z, z') v(\lambda', k', s') dH(s'|s) dJ(\lambda'|\lambda) \right\}
\]

such that \(k' = x + (1 - \delta)k\), and \(\lambda\) is the distribution of the aggregate state.

- The approach is then to simulate this economy and, for firms in different deciles of the distribution, compute equity returns. The model delivers the right prediction when sorting firms according to market capitalization (i.e., ranking by \(k\)): high market capitalization firms have low returns. However, the predictions of the model when sorting firms by book-to-values (i.e., Tobin’s \(q\)’s) are highly counterfactual: the firms that have high Tobin’s \(q\) are firms that have low idiosyncratic productivity, or

\(^{22}\)To be specific, we have \(M(z_t, z_{t+1}) = \beta E_t \left[ \frac{e^{(z_{t+1}) - z_t}}{(e^{z_t})^{-\sigma}} \right]\). Assuming \(\log M_{t+1} = \log \beta + \gamma_0 z_t + \gamma_1 z_{t+1}\) with \(\gamma_0 = -\gamma_1 = \gamma\) would give \(M_{t+1} = \beta e^{\gamma(z_{t+1} - z_t)}\), and a constant risk-free rate as before. Letting \(\gamma_0 \neq -\gamma_1\) is a general way of having the countercyclicality in the risk-free rate.
\[ \frac{\partial q}{\partial s} < 0 \]

where \( q \equiv \frac{v(\lambda, k, s)}{k} \). This is because even though a positive shock increases the productivity of capital, the average value of capital goes down due to the fact that there are decreasing returns to scale in production. This is symptomatic of models with capital adjustment costs, an observation that was made already by Hayashi (1982) in the context of capital adjustment costs and CRTS in production. An easy fix is to introduce an operating cost \( c_f > 0 \), i.e. a period-by-period fixed cost of production. This does not fundamentally change the policy functions of the firm, but now

\[ q = \frac{v(\lambda, k, s)}{k} - \frac{c_f/r_f}{k} \]

where \( v \) is as defined above, i.e. without including \( c_f \), and \( c_f/r_f \) is the PDV of the cost. Of course, this can overturn the prediction because \( \frac{\partial^2 q}{\partial c_f \partial r_f} > 0 \). In this way, firms with high book-to-market (low \( q \)) have higher returns than firms with low book-to-market (high \( q \)), as in the data.

The simulations of the model also show that low book-to-market firms as of time zero, which have low capital stock but relatively high value, have had a sequence of bad shocks in the past, and by mean-reversion they have a higher chance of having good shocks so they invest more. To do so, they have negative dividends, i.e. issue equity, in order to finance their investment. High book-to-market (i.e. low \( q \)) firms, on the other hand, will desinvest and have positive dividends (pay cash to shareholders). Since dividends are risk-less because they are distributed regardless of the shock, high book-to-market firms are less risky. This is problematic, because in the data firms with high book-to-market do not typically have positive cash payouts, but rather tend to finance themselves through equity (negative dividend flows).

Zhang (JF, 2005), for instance, improves upon these dimensions by building a very similar model that, however, delivers that high book-to-market firms generate higher returns, like in the data. The problem is that this is at the cost of not matching moments of the investment rate, which was an important focus of the model above. The reason is that he assumes very high adjustment costs, which implies a counterfactually low dispersion in investment rates.