RIETI Discussion Paper Series 23-E-087

The U-shaped Law of High-growth Firms

Yoshiyuki ARATA
RIETI

MIYAKAWA, Daisuke
Waseda University

MORI, Katsuki
National Tax College

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 23-E-087
December 2023

The U-shaped Law of High-growth Firms[1]

Yoshiyuki Arata[2]
Research Institute of Economy, Trade and Industry
Daisuke Miyakawa
Waseda University
Katsuki Mori
National Tax College

Abstract

This paper investigates firm growth dynamics by using the theory of stochastic processes and data on corporate tax records covering almost all firms in Japan. We show that the growth path of high-growth firms (HGFs) is characterized by a single large jump rather than a gradual increase. Specifically, before the jump occurs, the growth path of a HGF is similar to that of non-HGFs, but then it experiences a rapid increase in size. This growth pattern with a jump is typical (i.e., most likely) for HGFs. To provide further empirical evidence, we consider the ratio of the growth rate in the first period to the entire growth rate over two periods. The histogram of this ratio exhibits a U-shaped curve for HGFs, indicating that high growth over the two periods is explained by high growth either in the first or second period (but not both). This U-shaped curve is consistent with the idea that a single large jump determines the growth path of HGFs.

Keywords: High-growth firms; Random walk; Subexponential distributions
JEL classification: D21; D22; L10

[1]This study is a part of "Joint Statistical Research Program" conducted in collaboration with the National Tax College (NTC), under the approval of the National Tax Agency (NTA) (in March 2022), in accordance with "Guideline on the Utilization of National Tax Data in the Joint Statistical Research Program." This study is also a part of the project "Firm Dynamics, Industry, and Macroeconomy" conducted at the Research Institute of Economy, Trade and Industry (RIETI). The views expressed herein are those of the authors and do not necessarily reflect the views of NTC, NTA, or RIETI. Arata appreciates the financial support by the Japan Society for the Promotion of Science (JSPS) (KAKENHI Grant Numbers: 21K13265). Miyakawa appreciates the financial support from JSPS (KAKENHI Grant Numbers: 21K01438). Conflict of interest: none.
[2]Corresponding author: y.arata0325@gmail.com

1 Introduction

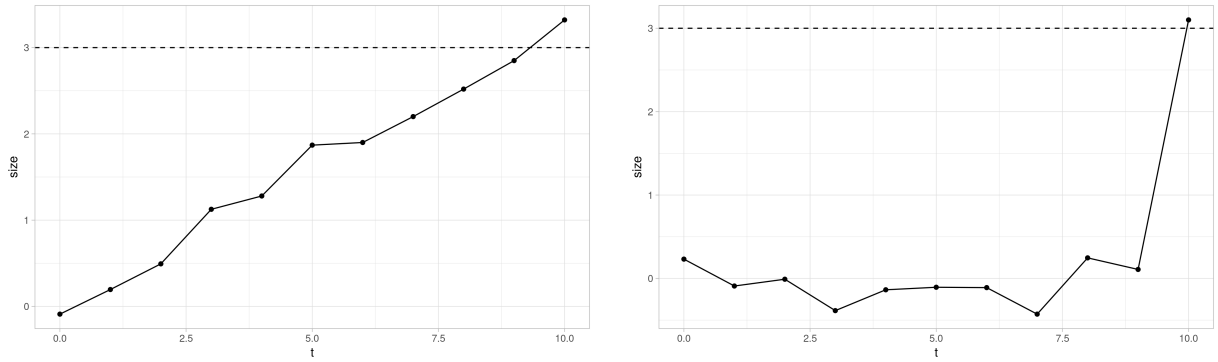
What drives firms' growth? How do firms grow over time? These questions about firm growth dynamics are one of the classical and most important themes in economics. Especially in the last decade, it has been recognized that high-growth firms (HGFs) are a driving force of job creation, the emergence of new markets, and economic growth. Thus, a better understanding of firm growth dynamics is needed not only for researchers but also for policymakers.

However, many empirical studies investigating firm growth dynamics have reached an unpleasant conclusion: we are unable to identify firms that will be HGFs in the future. Although some variables relevant to firm growth dynamics are identified (e.g., firm age and size), it is known that the explanatory power of models is quite weak and cannot be used for prediction. For example, after surveying empirical studies in the literature, Geroski (2000) concludes that "[t]he most elementary 'fact' about corporate growth thrown up by econometric work on both large and small firms is that firm size follows a random walk." Does it mean that the firm growth is completely random and that we have nothing to say about it? Is there any way to improve our understanding of firm growth dynamics?

This paper shows that even though we are unable to identify which firms will be HGFs, we can still obtain a meaningful implication about how firms grow over time. Rather than relying on a specific optimization model, we examine the typical growth path of HGFs by analyzing statistical regularities found in empirical data. The crucial assumption in our analysis is the distribution of firms' growth rates. It is well known that the distribution of empirical growth rates has a heavier tail than a Gaussian and is close to a Laplace distribution. Upon further examination, we find that the growth rate distribution has a strictly heavier tail than an exponential. Based on this empirical fact (and the random walk assumption), we show that firm growth is characterized by a sudden, large jump rather than a gradual increase. Prior to this jump, the growth path of a HGF is indistinguishable from that of other non-HGFs, but then it experiences a rapid increase in size (see **Figure 1**). We show that this growth pattern is the norm rather than the exception for HGFs.

Our analysis consists of two parts: theoretical analysis using probability theory and empirical analysis using comprehensive administrative data in Japan. For the theoretical analysis, we consider the two distribution classes: the light-tailed and heavy-tailed distributions. The light-tailed distributions are those with an exponentially bounded tail. Its tail probability decays rapidly, meaning that the probability of observing extreme values is low. Examples of this distribution class include Gaussian and Laplace distributions. On the other hand, the heavy-tailed distributions have a tail that decays slower than an exponential, leading to a higher probability of observing extreme values. This class includes distributions with a heavy tail, such as log-normal and Pareto distributions. Our theoretical analysis shows that the properties of firm growth dynamics can vary significantly depending on whether the growth rate distribution is light-tailed or heavy-tailed.

Assuming that the logarithm of firm size follows a random walk, we focus on the sample path of firms



(a) Gradual increase

(b) Increase by a sudden jump

Figure 1: The image of sample paths for HGFs. In Panel (a), firm size increases gradually and reaches a point as the consequence of many small successes. In Panel (b), the growth path is characterized by a sudden, large jump.

that grow rapidly over n periods (i.e., firms whose growth rate over n periods exceeds a large threshold value u). In particular, since the growth rate over n periods consists of n individual growth rates, we analyze how each of these growth rates contributes to the overall growth rate and reaches u. Our analysis (especially using the ruin theory, e.g., Asmussen and Albrecher (2010)) shows that if the growth rate distribution is light-tailed, a high growth rate over n periods is primarily determined by the cumulative effect of individual growth rates. Each growth rate contributes approximately equally to the overall high growth rate over the n periods, and thus, the sample path exhibits a gradual increase, as depicted in **Figure 1**(a). In other words, when the growth rate distribution is light-tailed, the most probable path (or event) leading to high growth over the n periods is a gradual increase.

In contrast, when the growth rate distribution is heavy-tailed, a high growth rate over the n periods is driven by the presence of a few large individual growth rates, or what we call a jump. More precisely, the probability that the sum exceeds u is asymptotically equivalent to the probability that the maximum among n individual growth rates exceeds u. A high growth rate over the n periods is dominated by a single burst in a period, while the contribution of growth rates in other periods is negligible. Thus, the heavy-tailedness of the growth rate distribution means that an infrequent, large jump characterizes the growth path of HGFs, as depicted in the **Figure 1**(b).

Given these theoretical backgrounds, our remaining task is to empirically test (1) the random walk assumption and (2) the distribution class that the growth rate distribution belongs to. For our empirical analysis, we use data on corporate tax records provided by the National Tax College, which includes sales revenues, profits, and the amount of corporate tax paid by firms. This is the population data in Japan and covers more than almost all firms in Japan from 2014 to 2020. Constructing a panel data tracking the growth path of these firms, we empirically test the two assumptions above.

For the random walk assumption, we focus on the autocorrelation of growth rates. For the analysis of dependence between successive growth rates, we avoid using Pearson's correlation coefficient, as it is not

appropriate for cases where extremes or outliers are of importance. Instead, we use dependence measures based on the copula theory, which is robust to such extremes. We find that the dependence between successive growth rates is–if any–weak, and furthermore, this dependence weakens as we consider the tail regions of the distribution. These results support the random walk assumption, particularly when focusing on HGFs. For the growth rate distribution, we consider one-year and three-year growth rates and use the density estimation and mean excess function to analyze the heaviness of the distribution tail. We find the growth rate distributions for both cases have a strictly heavier tail than an exponential; that is, the growth rate distribution is subexponential. These empirical findings imply that the growth path of HGFs is driven by a few large jumps rather than a gradual increase.

To provide another empirical support, we consider the following ratio:

$$r := \frac{X_1}{X_1 + X_2}$$

where X_1 and X_2 denote growth rates over the first- and second periods, respectively (e.g., X_1 and X_2 are growth rates in 2015 and 2016, respectively, and $X_1 + X_2$ is the growth rate for the two years). Ratio r represents the relative contribution of the growth rate in the first period to the entire growth rate over the two periods. We examine the distribution of r across a range of different values of $X_1 + X_2$. We find that the histogram of r exhibits a U-shaped curve with peaks at 0 and 1 when the growth rate over the entire period is high (i.e., when $X_1 + X_2$ is large). This suggests that when HGFs are considered, it is more likely that the high growth is caused by either high growth in the first period or high growth in the second period, but not both. This empirical finding is consistent with our theoretical analysis, which suggests that the sample path of HGFs is characterized by sudden, large jumps. The U-shaped curve in the histogram of r reflects the fact that these jumps tend to occur in one period or the other, rather than being spread evenly across both periods.

It is worth emphasizing that our finding is based on statistical regularities (i.e., the random walk and the growth rate distribution) and does not specify an optimization model of firms' behavior. As in Geroski (2000), we assume that a firm's growth is highly unpredictable (i.e., random), and we do not know which firm will be a HGF in the future. However, even when a firm's growth is random (or because of this randomness), there exists a robust feature characterizing firm growth dynamics because it is governed by the logic of probability theory. The U-shape curve for the ratio r, which we call the U-shape law, is one of the examples that randomness gives rise to an empirical law for economic phenomena.

Related literature

This paper belongs to the literature on firm growth dynamics, which aims to understand observed empirical regularities (see Coad (2009), Coad et al. (2014), Dosi et al. (2017), and Coad et al. (2022) for a survey). In particular, a series of empirical studies discuss our two assumptions: the heavy-tailedness of the growth rate distribution and the random walk assumption.

For the former, since the seminal work by Stanley et al. (1996), it has been recognized in the literature

that the growth rate distribution deviates from a Gaussian and is close to a Laplace distribution (see, e.g., Bottazzi et al. (2001), Bottazzi and Secchi (2006), and Arata (2019)). This is one of the most robust empirical regularities in the sense that this distribution shape is observed across different countries, times, and sectors. Furthermore, several recent papers (e.g., Buldyrev et al. (2007);Bottazzi et al. (2011);Dosi et al. (2020)) empirically show that the tail of the growth rate distribution is strictly heavier than that of the Laplace distribution (i.e., an exponential tail). In particular, Bottazzi et al. (2011) propose the Subbotin family, which includes a Laplace distribution as a special case, rejecting the null hypothesis that growth rates follow a Laplace distribution. In our paper, consistent with these empirical studies, we find that the tail of the growth rate distribution is strictly heavier than an exponential. However, we do not specify the functional form of the growth rate distribution. Only the fact that the tail is heavier than an exponential matters for our finding.

Regarding the random walk assumption, there is a strand of empirical studies discussing the persistence of growth rates (e.g., Coad (2007); Coad and Hölzl (2009); Frankish et al. (2013); Dosi et al. (2020)). Their results are mixed; for example, Coad (2007) shows that while growth rates are negatively autocorrelated for small firms, large firms exhibit positive autocorrelation. However, if any, this autocorrelation is generally weak, and in most cases, "lagged growth is a poor signal of future growth" (Coad et al. (2013), p.617). Furthermore, in recent years, many researchers have focused on the persistence of high growth in the context of HGFs (e.g., Delmar et al. (2003); Daunfeldt and Halvarsson (2015); Guarascio and Tamagni (2019); Esteve-Pérez et al. (2022)). They show that HGFs are "one-hit wonders" (Daunfeldt and Halvarsson (2015)); that is, firms that experience a high-growth period does not repeat another high-growth period again. In addition, there is no typical path unique to high-growth periods, meaning that the sample path of HGFs is erratic. These studies suggest that the random walk assumption is a reasonable one to describe firm growth dynamics. [1] [2]

The closest paper to ours is Coad et al. (2013), which assume a simple random walk with increment ± 1. They compare the frequency of patterns in growth dynamics (such as four successive growth $++++$ and alternating pattern $+-+-$) with that of the simple random walk and show that the simple random walk provides a good approximation for growth dynamics. Following the spirit of Coad et al. (2013), we assume that firm growth dynamics follow a random walk but extend this idea by considering the distribution of its increments. Our analysis shows that depending of the heaviness of the distribution tail of increments, the sample path properties of the random walk qualitatively change. By this method, our analysis provides a unified explanation entailing the heaviness of the distribution tail, (non-)persistence, and the sample path

[1] Another empirical regularity which is consistent with the random walk assumption is Gibrat's law, which states that a growth rate is independent of its firm size. See, e.g., Lotti et al. (2009).

[2] Another important empirical finding that makes our stochastic approach more appealing is the lack of firm attributes characterizing firm growth dynamics. For example, Bianchini et al. (2017) and Moschella et al. (2019) show that the persistence of high growth is not related to any firm characteristics. These results suggest that it is necessary to use stochastic modeling without relying on micro-founded modeling.

properties of firm growth dynamics.

Outline

This paper is organized as follows. Section 2 considers a random walk model with increments following a subexponential distribution. Section 3 provides empirical results using data on corporate tax records in Japan. Section 4 concludes. In the Appendix, we check the robustness of our finding by considering firms' age.

2 Probabilistic Method

This section provides probabilistic methods to analyze firm growth dynamics. Section 2.1 introduces a random walk and the two distribution classes. Section 2.2 examines the relation between the summation and maximum of iid shocks. Section 2.3 discusses the sample path properties of a random walk.

2.1 Random walk

Let S_k be the size of a firm at time k. We analyze the evolution of its logarithm over n periods, i.e., $\log S_k$ for $0 \leq k \leq n$. The growth rate at time k (denoted by X_k) is defined by

$$X_k := \log S_k - \log S_{k-1}$$

Thus, the growth rate over n periods is the sum of growth rates up to n:

$$\log S_n - \log S_0 = \sum_{k=1}^{n} X_k$$

We assume that $\log S_k$ is described by a random walk with an initial point $\log S_0$, which is equivalent to the following assumption.

Assumption 2.1. Growth rates $X_1, X_2, ..., X_n$ are independent and identically distributed (iid) random variables with a distribution F.

It is worth mentioning two predictions implied by the iid assumption. First, under the iid assumption, a growth rate is independent of its firm size; that is, X_k does not depend on $\log S_{k-1}$. This property is called Gibrat's law, which is widely accepted in the literature as a reasonable approximation for firm growth dynamics. Second, the iid assumption implies no autocorrelation of growth rates; for example, high growth in a period does not imply high growth in the following periods. We will check the empirical validity of no autocorrelation in Section 3.2.

As shown below, the property of a random walk depends on the distribution of X_k, especially on its distribution tail. Rather than specifying the distribution of X_k, we introduce distribution classes characterized by the heaviness of the distribution tail. Since our interest is in high-growth firms, only the right tail of the distribution is considered.

The first distribution class is light-tailed distributions, whose tail is exponentially bounded. More precisely, this distribution class is determined by the existence of the moment generating function:

Definition 2.2. A distribution is light-tailed if its moment generating function exists for some $\lambda > 0$; that is,

$$Ee^{\lambda X_k} < \infty$$

for some $\lambda > 0$.

Examples of light-tailed distributions include the distribution of a bounded random variable (e.g., the uniform distribution), Gaussian, and Laplace distribution. The Laplace distribution is of particular importance in our analysis: since the Laplace distribution has an exponential tail, it can be seen as the boundary of this class. That is, if a distribution has a heavier tail than an exponential, the distribution does not belong to the light-tailed distributions. In such a case (i.e., when the moment generating function does not exist for any $\lambda > 0$), we say that the distribution is heavy-tailed.

Next, we introduce a subclass of heavy-tailed distributions, which is called subexponential distributions. The formal definition is given as follows:

Definition 2.3. A heavy-tailed distribution F on $\mathbb{R}^+$ is subexponential if

$$\lim_{x\to\infty} \frac{\overline{F * F}(x)}{\overline{F}(x)} \tag{1}$$

exists, where $\overline{F}(x) := F[x, \infty)$ and $F * F(x)$ is the convolution of F with itself.

Let F be a distribution on $\mathbb{R}$ and X be a random variable drawn from F. F is subexponential if the distribution of $X^+ := \max\{0, X\}$ is subexponential.

For later purpose, we introduce a subclass of subexponential distributions, which requires a slightly stronger regularity condition on their tails.

Definition 2.4. A heavy-tailed distribution F on $\mathbb{R}$ is strong subexponential if F satisfies the condition that

$$\lim_{x\to\infty} \frac{1}{\overline{F}(x)} \int_0^x \overline{F}(x-y)\overline{F}(y)dy$$

exists.

One can show if the limit in Eq.(1) exists, it equals 2 (see Chapter 3 in Foss et al. (2011)). This means that the tail probability of the sum of two iid random variables is asymptotically equivalent to the maximum of the two iid random variables. That is, when the sum is large, it is generated by either a large value of the first or second random variable (but not both). This property is referred to as the principle of a single jump, as a single large jump of one component determines the sum.

It should be noted that although (strong) subexponential distributions are a proper sub-class of heavy-tailed distributions (i.e., there exists a heavy-tailed distribution that is not subexponential), heavy-tailed distributions that we encounter in practical applications are (strong) subexponential. For example, Pareto,

log-normal, and Weibull distributions with an exponent less than 1 are included in (strong) subexponential distributions. In particular, consider the Weibull distribution with parameter $\alpha > 0$;

$$\overline{F}_\alpha(x) = e^{-x^\alpha}, \quad x \geq 0$$

The parameter α controls the heaviness of the tail: as α is smaller, the tail becomes heavier. Specifically, when $\alpha = 1$, it reduces to the exponential distribution. Thus, the Weibull distribution with $\alpha \geq 1$ (including the exponential case) is light-tailed, and the Weibull distribution with $\alpha < 1$ is (strong) subexponential.

2.2 Summation and maximum

Let us consider the sum of growth rates over n periods, $\sum_{k=1}^n X_k$. As shown below, the tail probability of the sum is qualitatively different depending on whether the distribution of X_k is light-tailed or heavy-tailed.

Before discussing general results, consider a simple case where $X_1, X_2, ..., X_n$ are iid Gaussian random variables with mean 0 and variance σ^2. In this case, the sum is also a Gaussian with mean 0 and variance $n\sigma^2$. Thus, by using Mills' ratio, we obtain

$$\mathbb{P}(\sum_{k=1}^n X_k > u) = 1 - \Phi\left(\frac{u}{\sqrt{n}\sigma}\right) \leq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2n\sigma^2}\right)$$

for a large u. With a fixed n, the tail probability of the sum is controlled by σ^2 and decays rapidly as $u \to \infty$ (a Gaussian decay).

Next, consider the case where the distribution of X_k is light-tailed. In general, the tail probability of a random variable is closely related to how rapidly the moment generating function increases as λ increases. In particular, we impose a condition on the increasing rate of the moment generating function, which is satisfied for the Gaussian and Laplace distributions.

Proposition 2.1. *Suppose that the moment generating function of X_k satisfies*

$$\log E e^{\lambda X_k} \leq \frac{v\lambda^2}{2(1 - c\lambda)} \tag{2}$$

Then, the tail probability of the sum is bounded as follows:

$$\mathbb{P}(\sum_{k=1}^n X_k > u) \leq \exp\left(-\frac{u^2}{2(nv + cu)}\right)$$

Proof. This is a straightforward application of concentration inequalities for sub-Gamma variables. See, e.g., Chapter 2.4 and Corollary 2.11 in Boucheron et al. (2012). □

For a Gaussian distribution, Eq.(2) is satisfied with $v = \sigma^2$ and $c = 0$. The resultant upper bound on the tail probability of the sum is equivalent to the one above (up to a constant). For a Laplace distribution with parameter b (i.e., the variance is $2b^2$), Eq.(2) is satisfied with $v = 2b^2$ and $c = b$. In this case, with fixed n, the probability in the central region exhibits a Gaussian decay, similar to the Gaussian case (i.e., the effect of the central limit theorem). In contrast, when a relatively large value of u is considered, the tail

probability of the sum deviates from Gaussian decay. However, it should be noted that if the distribution of a component X_k is exponentially bounded, the deviation is still bounded exponentially.

When a subexponential distribution is considered, the deviation from a Gaussian is more severe and has another meaning. Suppose that the distribution of X_k is subexponential. One can show that the limit in Eq.(1) can be extended to an arbitrary n, and the tail probability of the sum can be approximated as follows:

$$\mathbb{P}(\sum_{k=1}^{n} X_k > u) \sim n\mathbb{P}(X_k > u) \tag{3}$$

as $u \to \infty$.

Note that the right-hand side of Eq.(3) is the probability of the maximum of the iid random variables $\max\{X_1, ..., X_n\}$ exceeds u. Intuitively, when we have a sum of subexponential random variables, the probability that their combination with moderate values leads to an extreme value of the sum is negligible. Instead, an extreme value of the sum is typically driven by one extreme value among its components. In particular, Eq.(3) means that if the distribution of the sum $\sum_{k=1}^{n} X_k$ has a heavier tail than an exponential, the distribution of each component X_k is also subexponential and shows the same decay rate as $u \to \infty$. We will use this property in Section 3 to test whether the distribution of firms' growth rates is (strong) subexponential.

2.3 Sample path properties

The previous section illustrates how the distribution of the sum depends on the tail of the distribution of its components. Here, we discuss the conditional distribution of $X_1, ..., X_n$ conditional on the event that $\sum_{k=1}^{n} X_k = u$. In particular, we examine the most probable combination of $X_1, ..., X_n$ to generate the given sum u.

For simplicity, let X_k be a non-negative random variable, and let $f(x_k)$ be the probability density function of X_k (i.e., $F(dx) = f(x)dx$).[3] Assume that $f(x)$ can be written as follows:

$$f(x) = e^{-h(x)}, \quad x \geq 0$$

For example, if $f(x)$ is the density function of an exponential distribution, then $h(x) = x$. When n iid random variables with F are considered, the probability that the sum is equal to u is given by

$$\mathbb{P}\left(\sum_{k=1}^{n} X_k = u\right) = \int_{\sum_{k=1}^{n} X_k = u} \exp\left(-\sum_{k=1}^{n} h(X_k)\right) dX_1 \dots dX_n$$

What combination of $X_1, ..., X_n$ is most probable, given that their sum is equal to u? This problem corresponds to the minimization of $\sum_{k=1}^{n} h(X_k)$ subject to $\sum_{k=1}^{n} X_k = u$. Let us consider the case where h is a convex function (e.g., the case of Weibull distribution with $\alpha > 1$). Jensen's inequality implies that the

[3] This example is taken from Sornette (2006).

minimum of the sum $\sum_{k=1}^{n} h(X_k)$ is attained at $X_1 = ... = X_n = u/n$.[4] In other words, the most probable combination of $X_1, ..., X_n$ that produces the sum $\sum_{k=1}^{n} X_k = u$ is the one where all components have the same value of u/n. Thus, it is more likely that each component contributes equally to the sum.

In contrast, when h is a concave function (e.g., a Weibull distribution with $\alpha < 1$), the way that components $X_1, ..., X_n$ generate the sum $\sum_{k=1}^{n} X_k = u$ differs qualitatively. $\sum_{k=1}^{n} h(X_k)$ is minimized when $X_{k^*} = u$ for some $k = k^*$ and $X_1 = \ldots = X_n = 0$ for $k \neq k^*$.[5] This suggests that one component dominates the sum while other components contribute nothing to the sum. Finally, note that the boundary case is the exponential distribution in which h is a linear function. In this case, both types of behavior can occur.

The fact that the sum is generated in two different ways can be seen by considering the ratio of X_1 in the sum. Suppose that $X_1, X_2 \geq 0$ be two independent random variables drawn from a Weibull distribution with parameter α.[6] Let us consider the distribution of the ratio $X_1/(X_1 + X_2)$ conditional on the event that their sum is equal to u (i.e., $X_1 + X_2 = u$). The probability density function of the ratio given u (denoted by $g_{\alpha,u}$) is given by

$$g_{\alpha,u}(r) = c(r(1-r))^{\alpha-1} e^{-u^{\alpha}(r^{\alpha}+(1-r)^{\alpha})} \tag{4}$$

where c is a normalizing constant independent of r.[7]

Figure 2 depicts the density $g_{\alpha,u}$ with three different values of α. As seen in the figure, it is symmetric at $1/2$ for all cases (this is obvious because X_1 and X_2 are two iid random variables). Let us examine the

[4]Indeed, let $\widehat{X}_k$ be the deviation from u/n, i.e., $\widehat{X}_k := X_k - u/n$. Jensen's inequality states that for a real convex function φ,

$$\varphi\left(\frac{\sum_k x_k}{n}\right) \leq \frac{\sum_k \varphi(x_k)}{n}.$$

Thus,

$$\sum_k h(X_k) = h\left(\frac{u}{n} + \widehat{X}_1\right) + \ldots + h\left(\frac{u}{n} + \widehat{X}_n\right) \geq nh\left(\frac{u}{n}\right)$$

where we used $\sum_k \widehat{X}_k = 0$ by definition.

[5]This can be shown as follows: Suppose that the statement does not hold. Thus, there exist at least two k such that $0 < X_k < x$. Take such two k (denoted by k_1, k_2) so that $X_{k_1} \geq X_{k_2}$. The concavity of the function h yields that $\sum_k h(X_k)$ can be lowered by replacing X_{k_1}, X_{k_2} with $X_{k_1} - \varepsilon, X_{k_2} + \varepsilon$. This is a contradiction.

[6]This example is taken from Foss et al. (2011).

[7]This can be proved as follows: Let ξ_1, ξ_2 be random variables such that $\xi_1 := \frac{X_1}{X_1+X_2}, \xi_2 := X_1 + X_2$. Thus, X_1 and X_2 are written as $X_1 = ru, X_2 = u(1-r)$. The probability of interest is

$$\begin{aligned}\Pr(\xi_1 = r|\xi_2 = u) &= \frac{\Pr(\xi_1 = r, \xi_2 = u)}{\Pr(\xi_2 = u)} \\ &= \frac{\Pr(X_1 = ru, X_2 = u(1-r))}{\Pr(\xi_2 = u)}\end{aligned}$$

The numerator is calculated using the fact that X_1 and X_2 are independent of each other. The denominator is determined by u and independent of r. Setting the normalizing constant c, we obtain the result.

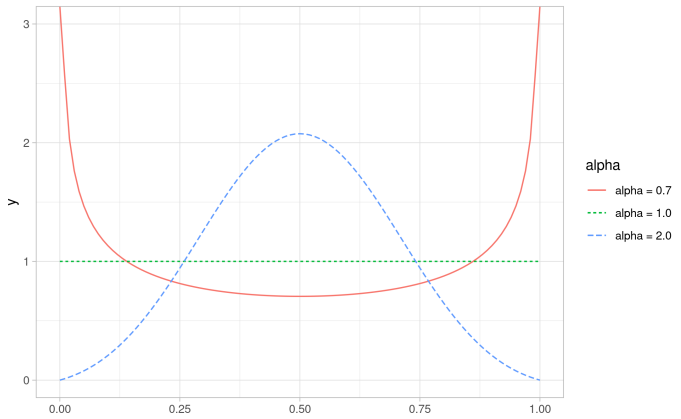


Figure 2: The density $g_{\alpha,u}$. Three values of α are considered: $\alpha = 0.7, 1.0, 2.0$.

density more closely for each value of α. For the case of $\alpha > 1$ (i.e., a light-tailed case), the density is unimodal and peaked at $1/2$. This indicates that the most probable event is that X_1 and X_2 are of similar size (i.e, equal to $u/2$). Especially for a large value of u, the density is concentrated around $1/2$.

In contrast, when $\alpha < 1$ (i.e., a heavy-tailed case), the density peaks at 0 and 1 and exhibits a U-shape curve. Thus, it is more likely that either X_1 or X_2 (but not both) takes a large value and dominates the sum u. Futhermore, as suggested by Eq.(4), the density is concentrated at 0 and 1 as $u \to \infty$. Thus, for a large value of u, it is highly unlikely that both X_1 and X_2 are large and contribute equally to the sum.

Finally, consider the case of $\alpha = 1$ (i.e., an exponential case). The density is uniform, meaning that this case can be seen as the boundary case. Thus, the crucial point is whether the distribution tail is heavier than an exponential.

The rest of this section introduces two general results in probability theory (Asmussen (1982) and Asmussen and Klüppelberg (1996)) that formalize the intuition given above. To do so, we need some technical assumptions. Suppose that we are interested in firms that grow rapidly and outperform others. However, if we consider a large value of $\sum_{k=1}^{n} X_k$ (e.g., $\sum_{k=1}^{n} X_k > u$), this event would happen with probability 1 when EX_k is positive and a longer time period is considered (here, we assume that n is not fixed). Instead of considering this trivial case, we define growth rate as an excess from a positive value c (close to but larger than EX_k) and focus on firms whose sum of excess growth rates is large. Formally, letting $Y_k := X_k - c$, we consider the random walk of $Y_1, ..., Y_n$ with a negative drift and focus on the event

$$\sum_{k=1}^{n} Y_k > u$$

for some n.

The probability of the event is less than 1 (because $EY_k < 0$) and becomes much smaller as u is large (i.e., a rare event). But once this rare event happens, how does the sequence of $Y_1, Y_2...$ reach u? More precisely, we consider the following conditional probability $\mathbb{P}_u := \mathbb{P}(\cdot | \sum_{k=1}^{n} Y_k > u \text{ for some } n)$. Under

P_u, let F_n denote the empirical distribution; that is,

$$F_n(x) := \frac{1}{n}\sum_{k=1}^{n} I(Y_k \leq x)$$

In particular, letting $\nu(u)$ be the time at which $\sum_{k=1}^{n} Y_k$ exceeds u for the first time (i.e., $\nu(u) := \inf\{n : \sum_{k=1}^{n} Y_k > u\}$), $F_{\nu(u)}$ is the empirical distribution of growth rates conditional on the event that the random walk exceeds u.

For light-tailed distributions, Asmussen (1982) identifies the distribution to which the empirical distribution $F_{\nu(u)}$ converges and characterizes the fluctuations of the random walk conditioned on $\nu(u) < \infty$. Let F_γ denote the twisted distribution of F defined by

$$F_\gamma(x) := \int_{-\infty}^{x} e^{\gamma y} dF(dy)$$

for $\gamma > 0$ satisfying $Ee^{\gamma Y_k} = 1$ and $E|Y_k|e^{\gamma Y_k} < \infty$. Note that F_γ has a positive mean. The result relevant to our analysis is the following:

Theorem 2.2 (Theorem 3.1 and Corollary 3.1 in Asmussen (1982))**.** *Suppose that F is light-tailed. Then, as $u \to \infty$*

$$\left\| F_{\nu(u)} - F_\gamma \right\| \xrightarrow{\mathbb{P}_u} 0$$

If properly normalized, $(\sum_{k=1}^{t\nu(u)} Y_k - tu)_{0\leq t\leq 1}$ converges in distribution to a Brownian bridge.

This theorem means that under $\mathbb{P}_u$ (i.e., conditioned on the event that $\sum_{k=1}^{n} Y_k$ exceeds u for some n), the empirical distribution of Y_k is close to F_γ for a large u. Recall that the unconditional mean of Y_k is negative; that is, most firms do not outperform the trend, and $\sum_{k=1}^{n} Y_k$ finally goes to $-\infty$ as $n \to \infty$. The fact that the mean of F_γ is positive suggests that (excess) growth rates for high-growth firms (i.e., firms with $\sum_{k=1}^{n} Y_k > u$ for some n), $Y_1, ..., Y_{\nu(u)}$ are upward drifted, and in turn, their sum reaches u. The latter half of the theorem presents the same picture for the sample path: since the expectation of the Brownian bridge at any t is 0, the sum $\sum_{k=1}^{t\nu(u)} Y_k$ increases its value at the rate tu on average. Thus, when the distribution of growth rate is light-tailed, the typical sample path is a gradual increase, as depicted in **Figure 1**(a).

For subexponential distributions, Asmussen and Klüppelberg (1996) provides the convergence of $F_{\nu(u)}$ and its sample path properties.

Theorem 2.3 (Theorem 1.1 and 1.2 in Asmussen and Klüppelberg (1996))**.** *Suppose that F is strong subexponential and belongs to the maximum domain of attraction of extreme value distributions.*[8] *Then, as $u \to \infty$,*

$$\left\| F_{\nu(u)} - F \right\| \xrightarrow{\mathbb{P}_u} 0$$

[8]More precisely, the condition needed here is F belongs to the maximum domain of attraction of Frechet or Gumbel distributions. This class is broad, including heavy-tailed distributions (such as Pareto, log-normal, and Weibull distributions), and does not impose any restriction in practical applications. For extreme value theory, see Embrechts et al. (1997).

Furthermore, $\{\sum_{k=1}^{\lfloor t\nu(u)\rfloor} Y_k/\nu(u)\}_{0\le t\le 1}$ *converges in distribution to* $\{-\mu t\}_{0\le t\le 1}$*, where* μ *is the mean of* F*.*

This theorem implies that in contrast to the light-tailed case, the conditional empirical distribution of growth rates for HGFs (i.e., firms with $\sum_{k=1}^{n} Y_k > u$ for some n) is essentially the same as the unconditional one. Indeed, the latter half of the theorem means that up to time $\nu(u)$, the conditional random walk decreases as $-\mu t\nu(u)$ on average, which is the same as other non-HGFs. Intuitively, this is equivalent to saying that the sample path for HGFs is the same as that of non-HGFs just before a large jump arrives, then the single large jump results in the upcrossing at u, as described in **Figure 1**(b).

To summarize, the above discussion shows that given the random walk assumption, there are two types of sample paths: gradual increase or sudden increase by a large jump. The type of its sample path is determined by whether the growth rate distribution is light-tailed or subexponential. Thus, our remaining tasks are to empirically examine (1) the random walk assumption (especially autocorrelation of growth rates) and (2) the distribution class of the growth rate distribution. These tasks will be carried in the next section.

3 Empirical Results

This section provides our empirical analysis using data on corporate tax records in Japan. Section 3.1 describes this data. Section 3.2 analyzes the random walk hypothesis by focusing on the autocorrelation of growth rates. Section 3.3 analyzes the heavy-tailedness of the growth rate distribution. Section 3.4 shows that the properties of subexponential distributions discussed in Section 2 are consistent with our data.

3.1 Data description

Our data is based on corporate tax records collected by the National Tax Agency and provided by the National Tax College. All firms in Japan are required to report their profits, which are used to calculate the amount of their corporate tax payments. Since this report is mandatory, this data provides extensive coverage of almost all firms in Japan.[9] Firms also report their basic attributes (e.g., firm's name, location, and industry) and annual sales revenues. Spanning from 2014 to 2020, the data is accompanied by a unique ID for each firm. Using this unique ID, we construct panel data for analysis.

We have two other auxiliary data provided by the National Tax College, which are combined with the panel data. One contains information about a firm's incorporation date, which enables us to identify a firm's age. In the following analysis, we defined a firm's age as the difference between 2014 and the year of its

[9] When firms pay corporate tax, parent firms that own 100% of the stocks of a subsidiary can file taxes as a consolidated parent firm, allowing them to offset the profits and losses between the parent firm and its subsidiary (known as the consolidated tax system). In the data used for this analysis, firms utilizing this system are excluded. As of 2019, the total number of parent firms using this system is $1,721$, and the total number of consolidated subsidiaries included is $12,983$.

Summary statistics for firms' growth rates Corporate tax data in Japan						
year	count	mean	sd	q1	median	q3
one-year growth rate						
2014-2015	551559	−0.050	0.430	−0.100	−0.008	0.067
2015-2016	545287	−0.037	0.480	−0.096	−0.010	0.063
three-year growth rate						
2014-2017	529207	−0.092	0.600	−0.180	−0.023	0.113
2017-2020	507624	−0.138	0.690	−0.260	−0.074	0.077
six-year growth rate						
2014-2020	495460	−0.210	0.760	−0.370	−0.105	0.117

Table 1: Summary statistics of growth rates. One-, three, and six-year growth rates are considered. For example, the three-year growth rate in 2017 represents the growth rate from 2014 to 2017, i.e., $X_{2017} = \log S_{2017} - \log S_{2014}$.

incorporation. The other is about records of mergers. It enables us to identify the year when a merger occurs and firm IDs involved in each merger. To focus on firms' internal growth, we exclude firm-year observations where a merger takes place from our main samples.[10]

We add several conditions for sample selection. First, we exclude micro firms from our samples. Since our main variable of interest is the growth rate, which is defined as the log difference in sales (i.e., $X_{i,k} := \log S_{i,k} - \log S_{i,k-1}$ for firm i at period k), the extremely small size of a firm at the initial period (i.e., $S_{i,k-1}$) would generate an extremely large value of its growth rate. As our analysis is based on the iid assumption of growth rates, we consider only firms with sales of larger than 100 million yen in the initial period (i.e., in 2014). Second, we exclude firms in financial and government sectors from our samples. Lastly, we exclude firms with ages less than ten years from our samples. An analysis of these young firms will be given in Appendix 5.1, where we found that the growth dynamics for the young firms are different from those for older firms. Here, we focus on firms that are neither too young nor too small, as the random walk assumption (or Gibrat's law) is likely to hold for them. The sample size and summary statistics of growth rates are given in **Table 1**.

[10] Another detail about our panel data is that certain firms have multiple records within a year. This occurs because these firms have accounting periods of less than a year, resulting in the submission of multiple financial accounts during that year. Each of these accounts is used to calculate the corresponding tax payments. In our analysis, we aggregate a firm's sales revenues within a year to approximate their annual sales revenues. Then, we exclude samples if the duration of the aggregated accounting period (i.e., the difference between the closing date of the latest accounting period and the starting date of the oldest accounting period) is less than 11 months.

3.2 Autocorrelation

The random walk assumption implies that successive growth rates are independent of each other. We examine the empirical validity of the random walk assumption by analyzing the autocorrelation of growth rates. The most commonly used measure to assess the correlation between two random variables is Pearson's correlation coefficient. However, as explained below, it is not suitable for our analysis, and instead, we use alternative measures based on copula theory.[11]

The fundamental idea of copula theory is that any bivariate distribution function F can be decomposed as follows:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

where C is a function called copula and F_1, F_2 are the marginal distributions. The copula function C is independent of the marginal distributions, meaning that the dependence structure is uniquely determined by C.

Since the dependence structure of two variables is determined by C, one might think that Pearson's correlation coefficient is determined solely by C. This is not the case; Pearson's correlation coefficient can vary when the marginal distributions change while C remains unchanged. Moreover, Pearson's correlation coefficient of X_1 and X_2 is not necessary equal to that of $h_1(X_1)$ and $h_2(X_2)$ for strictly increasing functions h_1 and h_2. Copula theory suggests that any measure of dependence should be uniquely determined by copula C. In addition to this limitation, Pearson's correlation coefficient is vulnerable to extremes or outliers. This causes a problem when analyzing high growth rates, which is of particular interest in our analysis. To accurately assess dependence between successive growth rates, we should avoid excluding such extreme values as outliers.

Better alternatives are Kendall's τ, Spearman's ρ, and Pearson's correlation coefficient of normal scores. Kendall's τ is a measure of ordinal association between two variables, meaning it captures the degree to which the variables tend to be ranked in a similar way: if $(X_1, X_2), (X_1', X_2')$ are independent random pairs with a common distribution F, (the population version of) Kendall's τ is defined as

$$\tau := \mathbb{P}\left[(X_1 - X_1')(X_2 - X_2') > 0\right] - \mathbb{P}\left[(X_1 - X_1')(X_2 - X_2') < 0\right]$$

Kendall's τ is within the range $[-1, 1]$ and equal to 0 when the two variables are independent of each other. Spearman's ρ is a measure of rank correlation, which assesses the degree to which two variables are correlated when their values are ranked. This is defined as the correlation coefficient of the transformed variables $F_1(X_1)$ and $F_2(X_2)$:

$$\rho := \text{Cor}[F_1(X_1), F_2(X_2)]$$

Similar to Kendall's τ, ρ is within the range $[-1, 1]$ and equal to 0 when the two variables are independent of each other. The correlation coefficient of normal scores is defined as the correlation coefficient between

[11] Copula theory has been widely used in financial literature. See Joe (2014).

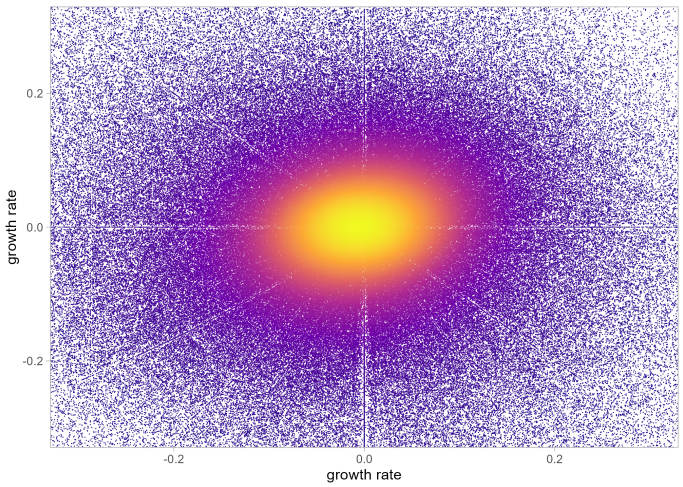


Figure 3: Scatter plot of growth rates in 2015 (y-axis) and 2016 (x-axis). The bright orange color represents the high density of sample points.

two variables transformed into standard Gaussian variables:

$$\rho_N := \text{Cor}[\Phi^{-1}(F_1(X_1)), \Phi^{-1}(F_2(X_2))]$$

where Φ is the distribution function of the standard Gaussian. Since we are familiar with (multivariate) Gaussian distribution, this measure is useful to assess the degree of dependence between two variables.

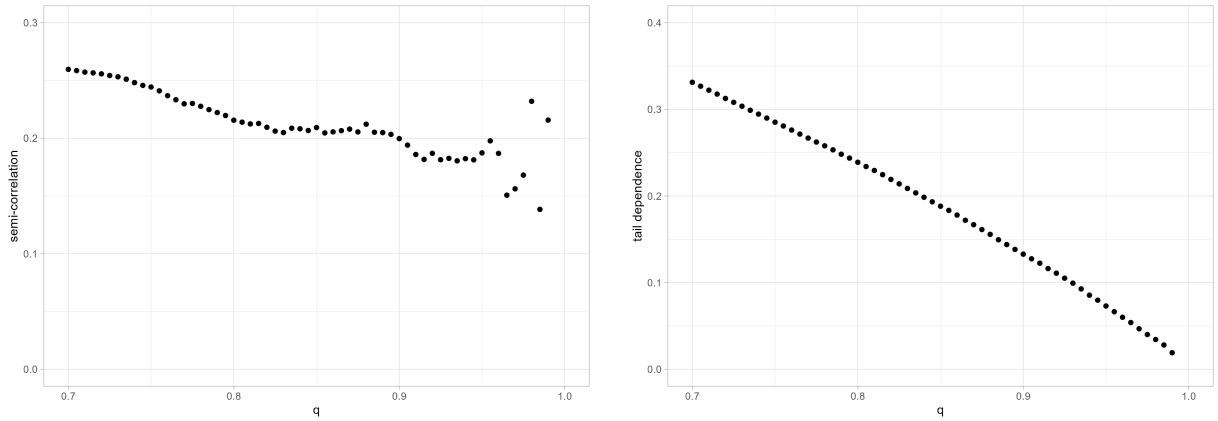
Our empirical data suggests that the dependence between growth rates in 2015 and 2016 is very weak. **Figure 3** depicts the scatter plot of growth rates in 2015 and 2016, showing no clear dependence between them. Kendall's τ and Spearman's ρ correlation coefficients are both close to zero, with values of -0.0061 and -0.017, respectively, suggesting only a slightly negative dependence. Pearson's correlation coefficient of normal scores is -0.070, indicating a weak negative correlation. Overall, these coefficients suggest that dependence between successive growth rates–if any–is weak.

Although the three measures can provide valuable insight into the dependence between growth rates, they tend to be more sensitive to dependence in the central region. This is because samples are more abundant in the central region and therefore have a greater influence on the three measures. In other words, the dependence in the extreme region may not be well-represented by these measures.

To address this concern, we use two additional measures: the semi-correlation of normal scores and the tail dependence coefficient. The semi-correlation of normal scores is defined as the correlation coefficient of the transformed variables conditioned on an event that both variables are large:

$$\rho_N^+(q) := \text{Cor}[\Phi^{-1}(F_1(X_1)), \Phi^{-1}(F_2(X_2)) \mid X_1 > F_1^{-1}(q), X_2 > F_2^{-1}(q)]$$

We are interested in the behavior of $\rho_N^+(q)$ when q is close to 1. The other measure is the tail dependence



(a) Semi-correlation

(b) Tail dependence coefficient

Figure 4: Tail dependence measures.

coefficient, which is defined as

$$\lambda_U := \lim_{q \to 1} \mathbb{P}(X_2 > F_2^{-1}(q) \mid X_1 > F_1^{-1}(q))$$

This is the probability that an extreme of X_2 (i.e., high growth rate at the second period) occurs conditioned on the event the growth rate in the first period X_1 is high. When λ_U converges to a positive value, it is called tail dependence. When these two measures are close to 0, it suggests that extremely high growth does not occur consecutively.

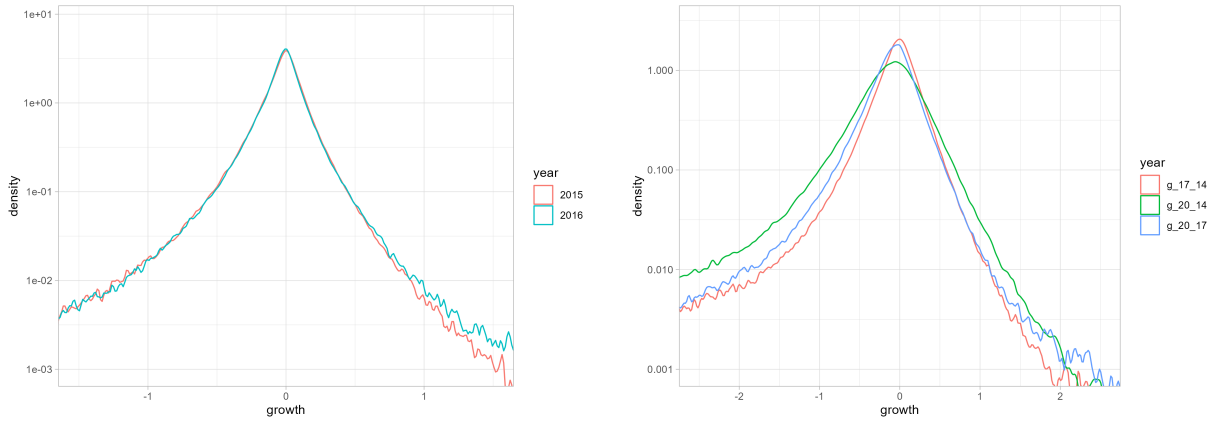
The two tail dependence measures are calculated using growth rates in 2015 and 2016. **Figure 3** shows that both measures decrease as we consider the tail region, i.e., $q \to 1$. Thus, dependence between successive growth rates comes mainly from dependence in the central region, and when growth rates in the tail region are considered, it becomes weakened. Since extremes of growth rates are of main interest, this result suggests that the random walk assumption provides a reasonable approximation for empirical firm growth dynamics.

3.3 Growth rate distribution

Here, we examine the growth rate distribution, with a specific focus on its distribution tail. The first method is density estimation in the log scale (y-axis). Recall that if growth rates follow a Laplace distribution, the density function in the log scale would exhibit a triangular shape; that is, the density exhibits a straight line in the tail (i.e., an exponential tail). Thus, the deviation from the straight line in the tail region can be seen as evidence that the growth rate distribution deviates from an exponential tail.

Figure 5 shows the density estimate in the log scale for one- and three-year growth rates. Both figures show that in the right tail, the density does not follow a straight line but rather exhibits a curved shape. This provides evidence that the growth rate distribution has a heavier tail than an exponential and belongs to subexponential distributions.

The second method is the mean excess function over threshold u (denoted by $e(u)$). This function is



(a) One-year growth rates

(b) Three-year growth rates

Figure 5: Density estimates of growth rates. The y-axis is in the log-scale.

defined as follows:

$$e(u) := E[X_k - u \mid X_k > u] \quad \text{for } u > 0.$$

$e(u)$ is the conditional expectation of overshoot $X_k - u$ given that X_k exceeds u. The advantage of $e(u)$ is that the increasing or decreasing rate of $e(u)$ over u reflects the tail-heaviness of the underlying distribution of X_k. For example, if X_k is drawn from an exponential distribution with parameter λ, then $e(u) = \lambda^{-1}$; that is, $e(u)$ is a constant. Intuitively, if $e(u)$ is an increasing function of u, the distribution of X_k has a heavier tail than an exponential.

Consider the empirical counterpart of the mean excess function:

$$\widehat{e}(u) := \frac{1}{\overline{F}_n(u)} \int_u^\infty \overline{F}_n(y)dy \quad \text{for } u > 0,$$

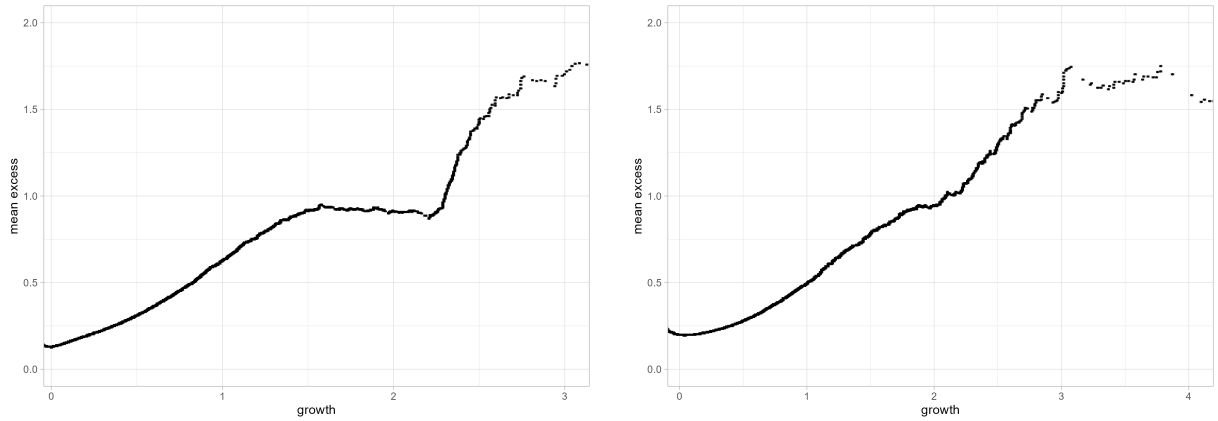
where F_n is the empirical distribution of growth rates for n samples. **Figure 6** depicts $\widehat{e}(u)$ over u for one-year growth rates in 2015 and three-year growth rates in 2017. Both figures show that $\widehat{e}(u)$ is an increasing function of u, though the increasing rate is not constant especially for one-year growth rates. This is another evidence that the growth rate distribution has a heavier tail than an exponential. Consistent with the density estimates in **Figure 5**, the growth rate distribution belongs to subexponential distributions.

The third method is the parametric estimation of the Weibull tail with a shape parameter α. The Weibull tail with α means that for some x^* and constant C, the tail of the growth rate distribution is described as follows:

$$\overline{F}_\alpha(x) = Ce^{-(x/b)^\alpha}, \quad x \geq x^*$$

where b is a scale parameter. The advantage of the Weibull tail is that it nests an exponential tail (i.e., $\alpha = 1$) and that it belongs to subexponential distributions when α is strictly less than 1.

Assuming that growth rates follow a distribution with a Weibull tail, we estimate the shape parameter



(a) One-year growth rates

(b) Three- and six-years growth rates

Figure 6: Mean excess function over threshold u.

α using the Hill's type method developed by Gardes et al. (2011).[12] Since the estimation relies only on an upper part of samples, we use the top 1% samples, i.e., $X_k > \overline{F}_n^{-1}(0.99)$. The estimate of the shape parameter, $\widehat{\alpha}$, is 0.49 for one-year growth rate in 2015 and 0.56 for three-year growth rates in 2017. Both estimates suggest that the tail of the growth rate distribution is heavier than an exponential tail (i.e., $\alpha = 1$). Thus, this provides further evidence that the growth rate distribution is subexponential.

We also assess the goodness of fit of the Weibull tail with the estimate $\widehat{\alpha}$. **Figure 7** illustrates the comparison between the Weibull tail and the empirical counter cumulative distribution function (CCDF) of growth rates. Both figures in **Figure 7** show that the Weibull tail provides reasonable fit up to $X_k \approx 1.5$, but the empirical CCDF appears to be heavier than the Weibull tail for $X_k \gg 1.5$. Although further exploitation is needed for precise characterization of such extremely large values of X_k, these figures align with our assumption that the tail of the growth rate distribution is heavier than an exponential.[13]

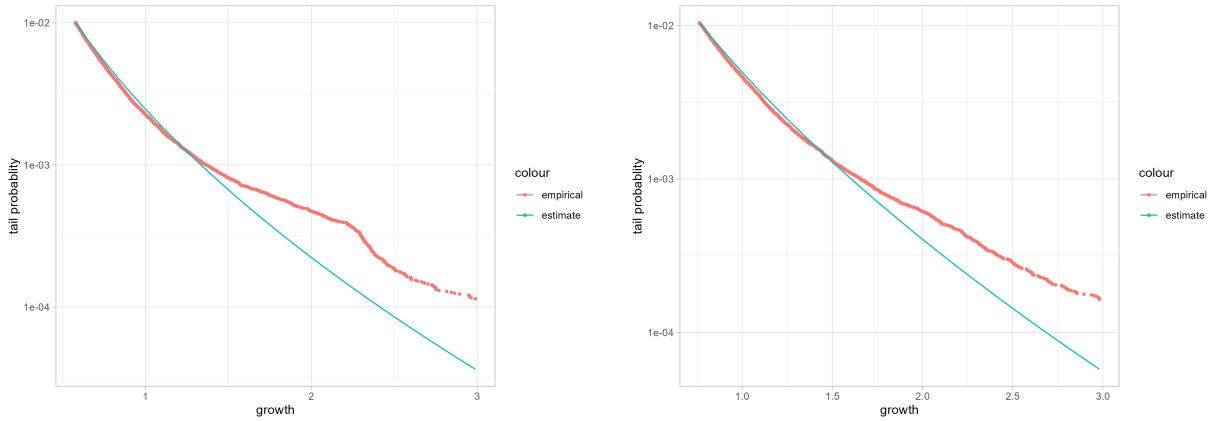
Finally, we consider the distribution of six-year growth rates, i.e., $\sum_{k=15}^{20} X_k$. Recall that when the distribution of X_k is light-tailed, the distribution of the sum $\sum_{k=15}^{20} X_k$ is bounded by an exponential tail, as discussed in Section 2.2. In other words, when the distribution of $\sum_{k=15}^{20} X_k$ has a heavier tail than an exponential, it indicates that the distribution of X_k is subexponential.

[12] More precisely, in Gardes et al. (2011), the Weibull tail is defined as

$$\lim_{t\to\infty} \frac{\log(\overline{F}(\lambda t))}{\log \overline{F}(t)} = \lambda^{\alpha},$$

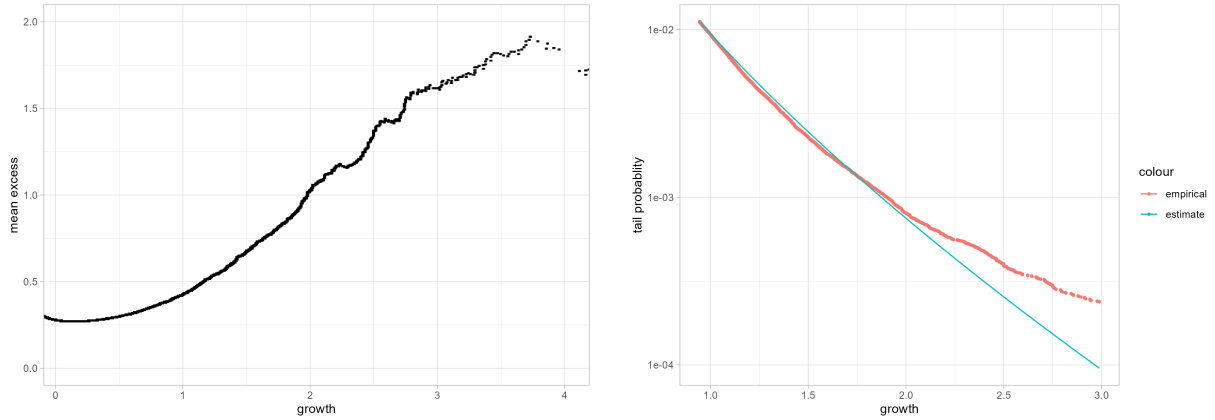
and the main idea for estimation is that when taking the logarithm of both sides of the equation, it is a linear function of λ with a slope α. The asymptotic properties of the estimate of α are also given in Gardes et al. (2011).

[13] While the properties of the distribution tail are defined as ones in the limit $x \to \infty$, we need to set a criterion above which samples are considered as part of the tail when dealing with empirical data. As we will see in Section 3.4, the property of a single big jump for subexponential distributions becomes evidence when $X_k \approx 0.8$ is considered. We can expect that the tail of the distribution starts around $X_k \approx 0.8$, and therefore, the deviation from the Weibull tail at $X_k \gg 1.5$ is not relevant to our analysis.



(a) One-year growth rate **(b)** Three-year growth rate

Figure 7: Estimated Weibull tail with $\widehat{\alpha}$. The counter cumulative distribution function for empirical growth rates is also given for comparison.



(a) Mean excess function **(b)** Weibull tail

Figure 8: Mean excess function and Weibull tail for six-year growth rates. The estimate $\widehat{\alpha}$ for the Weibull tail is 0.63.

Using the same methods above, we show the mean excess function and the estimated Weibull tail for six-year growth rates in **Figure 8**. Here, the estimate $\widehat{\alpha}$ for the Weibull tail is 0.63, which is lower than $\alpha = 1$, i.e., an exponential tail. Similar to the cases of one- and three-year growth rates, both figures show that the distribution of six-year growth rates has a heavier tail than an exponential. Thus, the deviation from an exponential tail for six-year growth rates implies that the distribution of one- and three-year growth rates are not exponentially bounded, providing further evidence that the growth rate distribution is subexponential.

3.4 Sample path properties

The empirical results given in Sections 3.2 and 3.3 imply that the two assumptions in our analysis (i.e., the random walk assumption and subexponential growth rate distribution) hold, and therefore, the growth dynamics for HGFs is characterized by jumps. Here, we provide additional direct evidence to further support

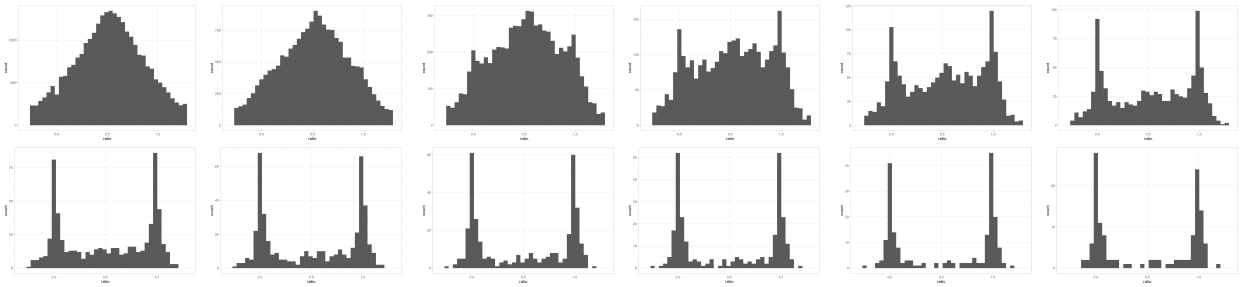


Figure 9: A series of the histograms of r_1 conditional on $X_{15} + X_{16} > u$. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

the importance of this jump-type process.

Let us consider the contribution of the growth rate in the first period to the overall growth rate, which is defined as follows:

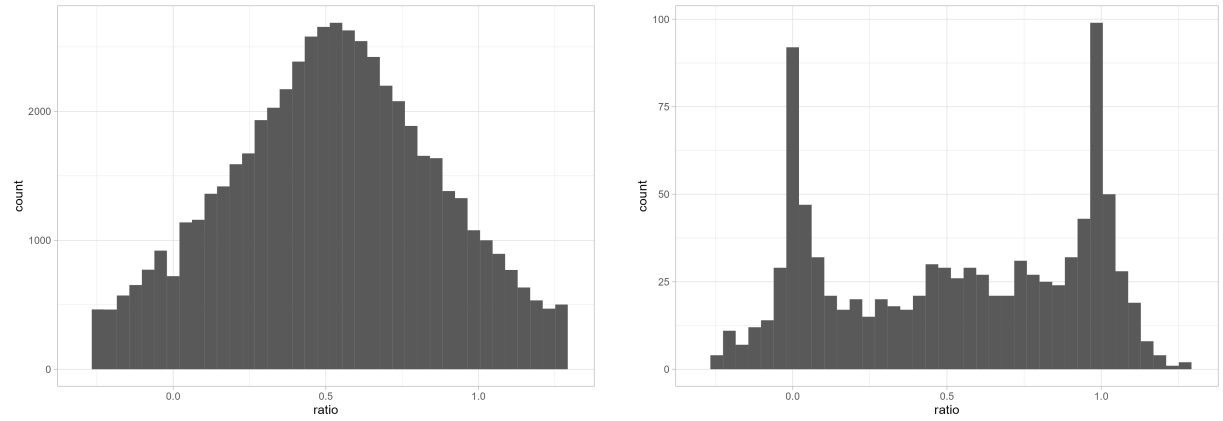
$$r_1 := \frac{X_{15}}{X_{15} + X_{16}}, \quad r_3 := \frac{\sum_{k=15}^{17} X_k}{\sum_{k=15}^{17} X_k + \sum_{k=18}^{20} X_k}$$

Ratio r_1 represents the contribution of the growth rate in 2015 to the growth rate over the two years (i.e., $n = 2$). Similarly, ratio r_3 represents the contribution of the growth rate in the first three years to the growth rate over the six years (i.e., $n = 6$). For example, when a firm's growth rate is 3% in 2015 and 3% in 2016, r_1 is equal to $1/2$, meaning that both growth rates in 2015 and 2016 contribute equally to the overall two-year growth rate. Since X_k's are assumed to be independent and identically distributed random variables in our analysis, the distributions of r_1 and r_3 are symmetric at $1/2$. The question to address here is whether the case of $r_1 = 1/2$ (or $r_3 = 1/2$) is a likely event or not.

Using the growth rates in 2015 and 2016 in our samples, we provide in **Figure 9** the histogram of r_1 conditional on the event $X_{15} + X_{16} > u$, where u varies from 0.2 to 2.4.[14] Specifically, **Figure 10** provides the histograms of r_1 for $u = 0.2$ and $u = 1.2$. These figures shows that when $X_{15} + X_{16}$ is relatively small (e.g., $u = 0.2$), the histogram of r_1 exhibits a mountain shape with a peak at $1/2$. That is, when the growth rate over the entire period is not high, it is more likely that both growth rates contribute equally to the growth rate over the entire period. In contrast, as the value of u increase (e.g., $u = 1.2$), the mountain shape collapses. Instead, the histogram exhibits a U-shaped curve with peaks at 0 and 1, meaning that high growth over the entire period is caused by a single large value of either X_{15} or X_{16} (but not both). Thus, when HGFs (over the two years) are considered, it is more likely that a HGF has extremely high growth in a year, which determines high growth over the entire period.

A similar U-shaped curve is observed for the histogram of r_3. **Figure 11** gives the histograms of r_3

[14] When calculating the histograms of r_1, we exclude samples where r_1 is exactly equal to 0 or 1. This is because for some firms, the values of current sales are exactly the same as the previous ones. To make our analysis conservative, the histograms in **Figure 10** exclude these samples. Even without these samples exactly equal to 0 or 1, spikes at 0 and 1 are still clearly observed. We apply the same procedure to r_3 as well.



(a) $X_{15} + X_{16} > 0.2$ **(b)** $X_{15} + X_{16} > 1.2$

Figure 10: Histogram of r_1.

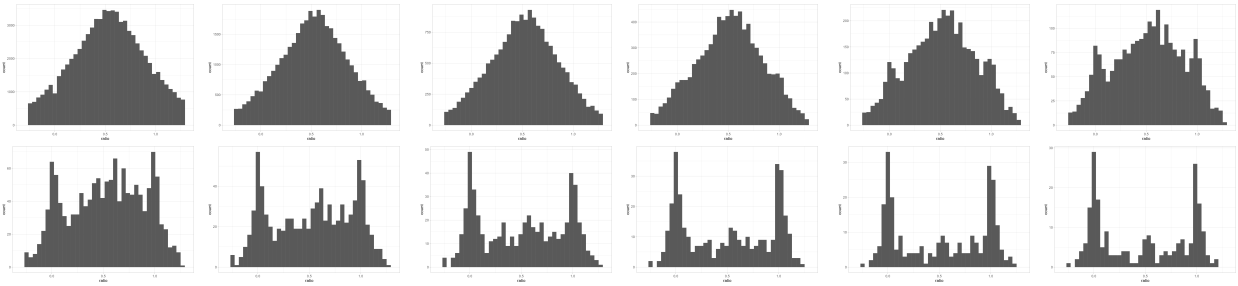
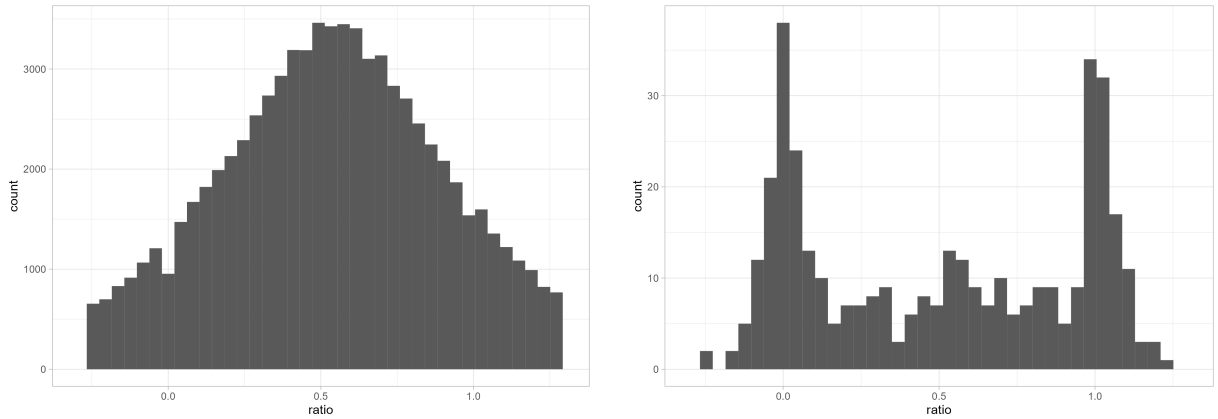


Figure 11: A series of the histogram of r_3 conditional on $\sum_{k=15}^{20} X_k > u$. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

conditional on the event $\sum_{k=15}^{20} X_k > u$, where u varies from 0.2 to 2.4. Specifically, **Figure 12** provides the histograms of r_3 for $u = 0.2$ and $u = 2.0$. As in the case of r_1, when u is relatively small, the histogram peaks at $1/2$. The contribution of growth in the first-half period is approximately equal to the contribution of growth in the second-half period. In contrast, as u becomes larger, the U-shaped curve with spikes at 0 and 1 emerges. This means that high growth in the entire period (i.e., a large value of $\sum_{k=15}^{20} X_k$) is explained by high growth either in the first-half or second-half period (but not both). In other words, it is more likely that HGFs have a short period during which they grow rapidly.

The observed U-shaped curve for the histograms of r_1 and r_3 is direct empirical evidence supporting the implication given in Section 2: for HGFs, the most typical (or likely) path is not a gradual increase over the entire period but a path characterized by a large jump. Since this U-shape curve captures the essence of firm growth dynamics for HGFs, we refer to it as the U-shaped law of HGFs.



(a) $\sum_{k=15}^{20} X_k > 0.2$ **(b)** $\sum_{k=15}^{20} X_k > 2.0$

Figure 12: Histogram of r_3.

4 Conclusion

The understanding of firm growth dynamics is a fundamental topic in economics, and indeed, numerous studies have been conducted on this subject so far. However, predicting firm growth, especially for HGFs, is a formidable task, and the empirical growth process looks completely random. This paper attempts to characterize the seemingly random dynamics using probability theory and shows that there exists a robust empirical law governing dynamics.

My analysis is based only on two empirically testable assumptions: the random walk assumption and the subexponential distribution of growth rates. Using comprehensive data based on corporate tax records in Japan, we confirmed that these two assumptions hold; in particular, the empirical fact that the distribution of growth rates has a heavier tail than an exponential has far-reaching implications for firm growth dynamics. These empirical facts and the probability theory imply that the sample path of HGFs looks like a jump-type process. The essence of the process is not a gradual increase but a single large jump. The U-shaped curve of the histogram of r is direct evidence of this property.

It is worth mentioning that in our analysis, we do not specify any economic models for firm growth but derive our implications solely from statistical regularities, such as the subexponential distribution of growth rates. This approach is appealing, especially when the growth process is too complicated to be described by an explicit model, and only its probabilistic features are known. Furthermore, due to this inherent complexity, firm growth dynamics are governed by the logic of probability theory. Our finding suggests that even when firm growth is random and unpredictable (or because of this randomness), there exists an empirical law governing its dynamics, especially for HGFs.

Summary statistics for firms' growth rates Young firms in Japan						
year	count	mean	sd	q1	median	q3
one-year growth rate						
2014-2015	137225	−0.026	0.580	−0.093	0.019	0.140
2015-2016	133834	−0.034	0.590	−0.100	0.006	0.120
three-year growth rate						
2014-2017	127190	−0.051	0.830	−0.180	0.033	0.260
2017-2020	114455	−0.160	1.030	−0.390	−0.025	0.320
six-year growth rate						
2014-2020	118602	−0.140	0.850	−0.300	−0.051	0.150

Table 2: Summary statistics of growth rates for young firms.

5 Appendix

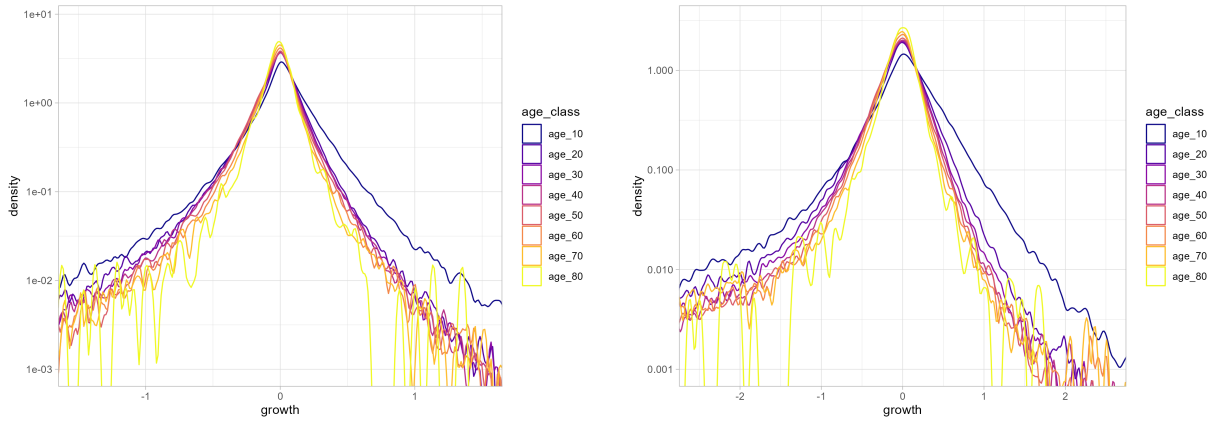
This appendix section provides additional empirical results regarding the U-shaped law. Section 5.1 considers young firms in Japan, which are excluded from our samples in the main text, and shows that the U-shaped law does not hold for this group. Using Orbis data, Section 5.2 shows that when such young firms are excluded, the U-shaped law holds for other countries as well.

5.1 Young firms

Here, we consider firm growth dynamics for young firms with ages less than ten years in 2014, which are excluded from our main analysis in Section 3. The summary statistics of their growth rates are given in **Table 2**.

As expected, the dispersion (e.g., the difference between Q1 and Q3 in **Table 2**) is high for younger firms, compared with **Table 1**. The high dispersion of young firms is also observed using density estimates of growth rates given in **Figure 13**, where the samples are decomposed into age groups. Only the density for the group of firms with ages less than ten years deviates from the densities for other age groups. One might think that this high dispersion of growth rates for young firms is due to the fact that young firms are likely to be small, and the dispersion of growth rates is higher for smaller firms. To mitigate this concern, in **Figure 14**, we restrict our samples to firms with sales of less than 10^9 yen (and larger than 10^8 yen). **Figure 14** shows that even when a firm's size is controlled, the deviation of the density for young firms from other densities is still observed. These figures suggest that the dependence of growth rates on firms' age is relevant for young firms with ages less than ten but weak for other age groups.

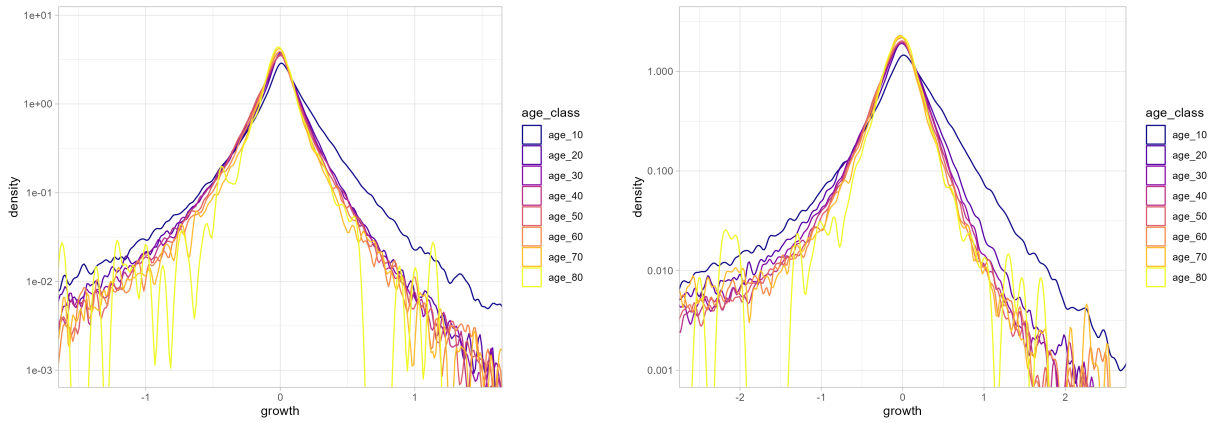
We also find the weak but positive autocorrelation of growth rates for young firms. The Spearman's ρ and Kendall's τ are 0.101 and 0.078, respectively. The correlation coefficient of normal scores is 0.062.



(a) One-year growth rate

(b) Three-year growth rate

Figure 13: Density estimates of growth rates for age groups. Samples are divided into groups of 10-year intervals based on their ages. For example, "age 20" represents the group of firms with age older than 10 and less than 20. In Panel (a), one-year growth rates in 2015 (i.e., X_{15}) are considered. In Panel (b), three-year growth rates in 2017 (i.e., $X_{15} + X_{16} + X_{17}$) are considered.



(a) One-year growth rate

(b) Three-year growth rate

Figure 14: Density estimates of growth rates for age groups. Only firms with sales less than 10^9 yen are considered. See the explanation given in **Figure 13**.

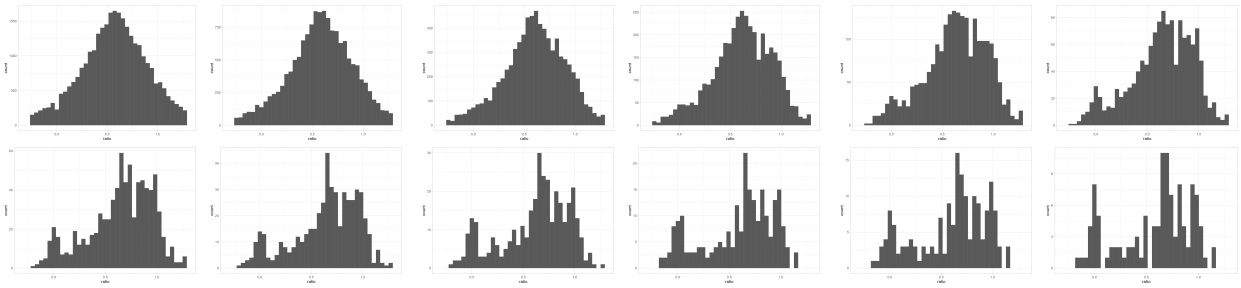


Figure 15: A series of the histograms of r_1 conditional on $X_{15} + X_{16} > u$ for young firms. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

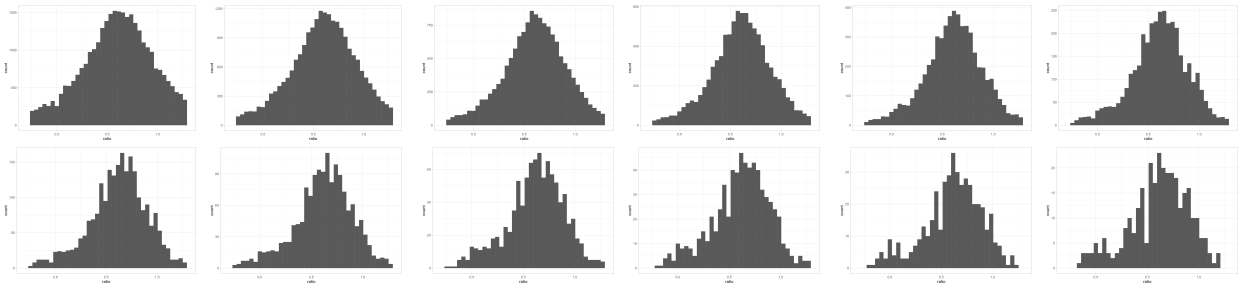


Figure 16: A series of the histogram of r_3 conditional on $\sum_{k=15}^{20} X_k > u$ for young firms. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

Although these estimates are still small, the autocorrelation nature of growth rates for young firms seems to be different from that for other older firms. Indeed, in the context of the empirical validity of Gibrat's law, the previous literature suggests that Gibrat's law is more likely to hold for old and mature firms (see, e.g., Lotti et al. (2009)). Thus, we expect that the U-shape curve for the histograms of ratios r_1 and r_3 would not be observed for younger firms.[15]

We confirm that this is the case. **Figure 15** depicts the histograms of r_1 for young firms using growth rates in 2015 and 2016. As in **Figure 9**, we consider the values of u in $X_{15} + X_{16} > u$ to vary from 0.2 to 2.4. However, we cannot observe the U-shaped curve of the histograms of r_1 for these young firms. Even when a large value of u is considered, the spikes at 0 and 1 are unclear. We also consider r_3 but cannot observe the U-shaped curve for young firms, as shows in **Figure 16**. Thus, we conclude that the U-shaped law does not hold for young firms.

These results imply that the growth paths for young firms are different from those for older firms and seem to be more complicated. The random walk assumption and the subexponentiality of the growth rate distribution are not sufficient to characterize the firm growth dynamics for young firms. Although this issue is worth exploiting, it is beyond the scope of this paper.

[15] Since it is assumed that growth rates are independent and identically distributed in our analysis, the dependence of growth rates on firms' age is not consistent with our assumption. This is why we excluded young firms with age less than ten years from our main samples in Section 3.

Summary statistics for firms' growth rates
The cases of France and Italy

	year	count	mean	sd	q1	median	q3
France							
	2017-2016	185109	−0.019	0.482	−0.064	0.016	0.099
	2018-2017	163385	−0.025	0.530	−0.062	0.019	0.099
	2018-2016	147635	−0.022	0.653	−0.083	0.038	0.160
Italy							
	2017-2016	315432	−0.097	0.792	−0.107	0.012	0.117
	2018-2017	304160	−0.089	0.763	−0.101	0.011	0.113
	2018-2016	301621	−0.116	0.927	−0.145	0.027	0.180

Table 3: Summary statistics of one-year growth rates for France and Italy.

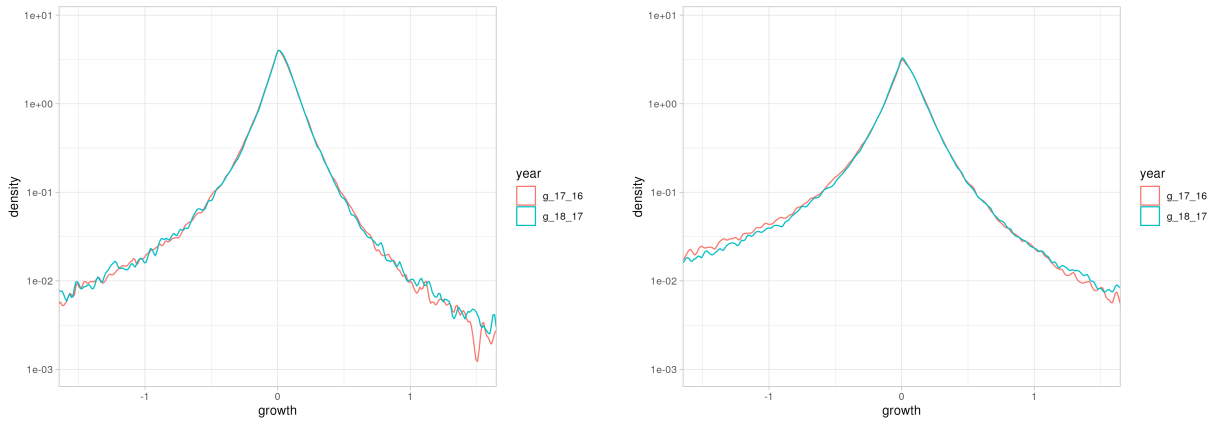
5.2 Orbis data

Here, we check the universality of the U-shaped law using firm-level data for France and Italy. Specifically, we examine the shape of the growth rate distributions and the histograms of the ratio r_1. We use the Orbis data compiled by Bureau van Dijk. The sample period ranges from 2016 to 2018.

As in Section 3, we impose several conditions on our samples. First, we exclude micro firms from our samples and consider only firms with sales of larger than 100 thousands euro. Second, we consider firms with age older than 10 years old. Here, a firm's age is defined by the firm's incorporation date. Finally, using Zephyr data, which identify the date and firms' IDs involved in a M&A, we exclude firm-year observations where a M&A occurs. For these samples, firm i's growth rate is defined as the log difference of firm i's sales. The summary statistics of growth rates are given in **Table 3**.

First, we consider the distribution of growth rates for the two countries. The density estimates of growth rates are shown in **Figure 17**, where the y-axis is the log scale. Both figures show that the density in the tail region deviates from an straight line, that is, an exponential tail. As in Japan's case given in Section 3.3, this deviation suggests that the growth rate distribution is subexponential for both countries.

Given the subexponentiality of the growth rate distribution, one can expect that the sample paths of HGFs are characterized by a large jump. We consider the histogram of the ratio r_1 and examine whether a U-shaped curve appears when HGFs are considered. **Figure 18** and **Figure 19** provide the histograms of r_1 for France and Italy, respectively. In both cases, the histogram exhibits a mountain shape with peak at $1/2$; that is, the equal contribution to the overall (i.e., two-year) growth rate is the most likely to occur. However, as the criteria u increases, the mountain shape collapses and the U-shaped curve with peaks at 0 and 1 emerges. Thus, consistent with Japan's case, the U-shaped law of HGFs holds for these two countries, leading to the universality of this law.



(a) France **(b)** Italy

Figure 17: Density estimates of growth rates. The y-axis is in the log-scale.

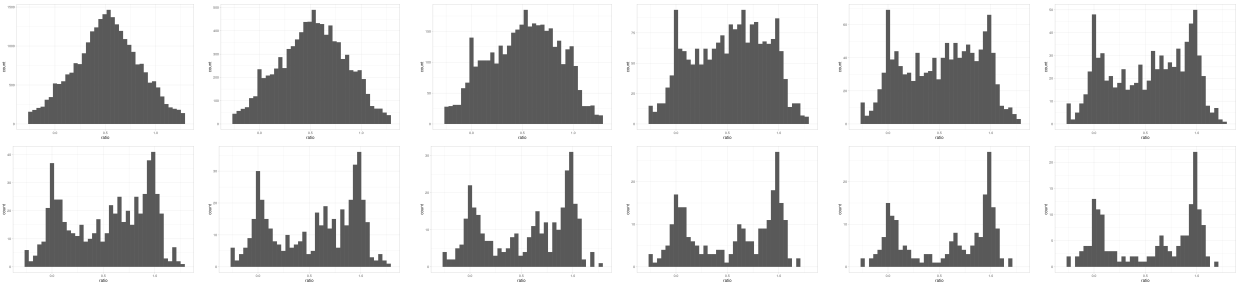


Figure 18: A series of the histograms of r_1 conditional on $X_{17} + X_{18} > u$ for France. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

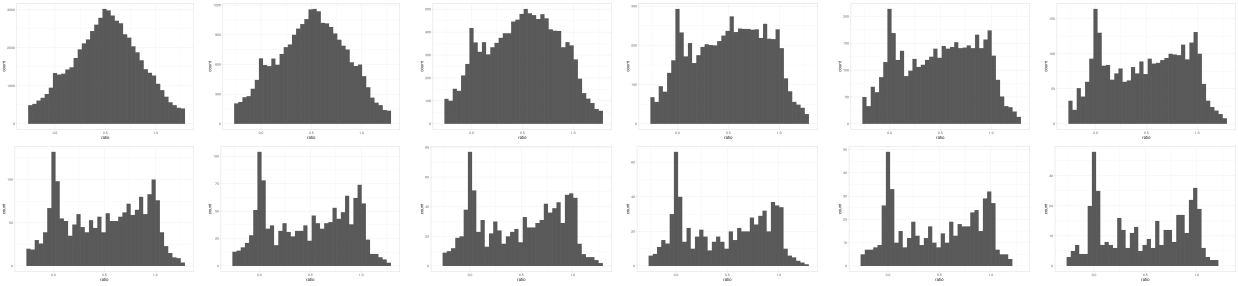


Figure 19: A series of the histograms of r_1 conditional on $X_{17} + X_{18} > u$ for Italy. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

6 Online Appendix

Here, we provide additional figures that demonstrate how the histograms of r_1 and r_3 change as u in $X_{15} + X_{16} > u$ and $\sum_{k=15}^{20} X_k > u$ increases. We examine the following cases:

- firms with sales of larger than 300 million yen
- firms with sales of larger than 500 million yen
- firms in the manufacturing sector
- firms in the service sector

For the third and forth cases, we consider firms with sales of larger than 100 million yen. For each case, we provide the histograms of r_1 and r_3 with different values of u. For r_1 and r_3, u increases from 0.2 (top-left panel) to 2.4 (bottom-right) by 0.2.

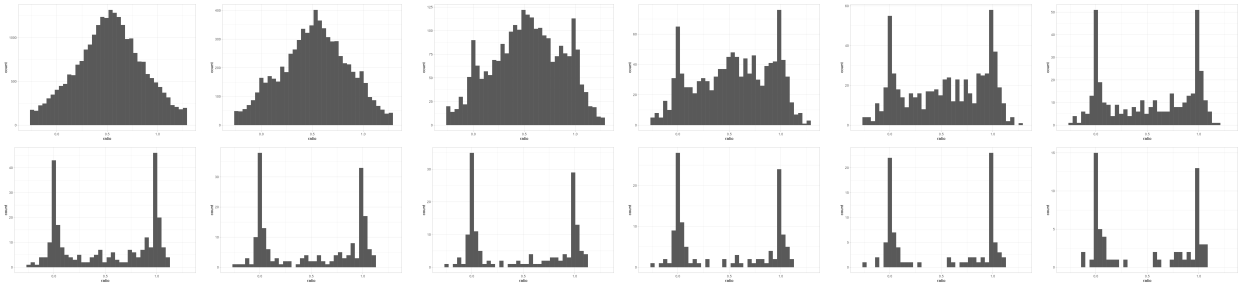


Figure 20: Histograms of r_1 for firms with sales of larger than 300 million yen.

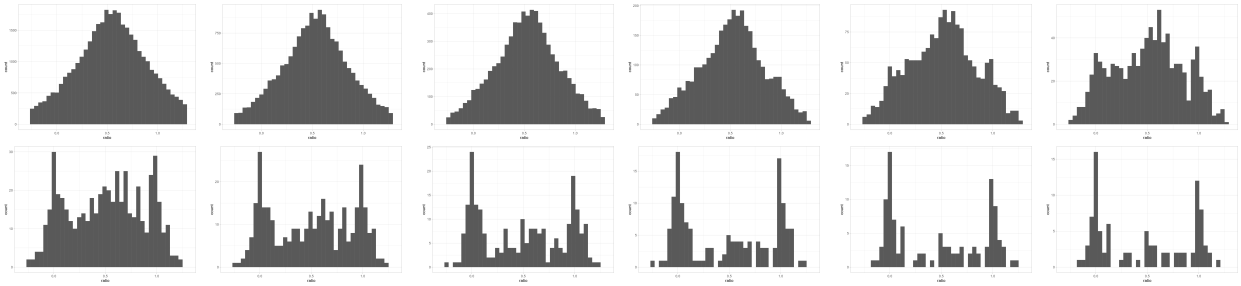


Figure 21: Histograms of r_3 for firms with sales of larger than 300 million yen.

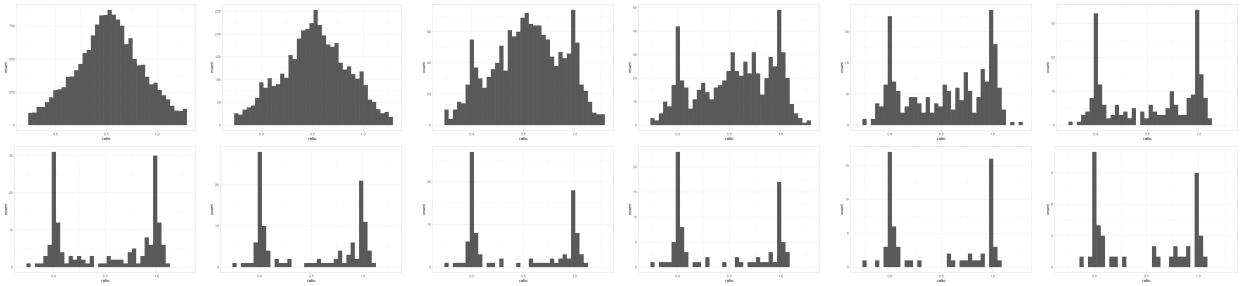


Figure 22: Histograms of r_1 for firms with sales of larger than 500 million yen.

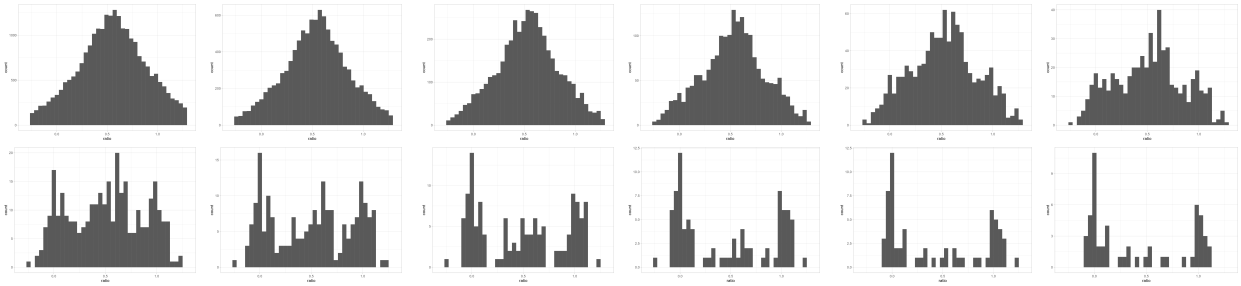


Figure 23: Histograms of r_3 for firms with sales of larger than 500 million yen.

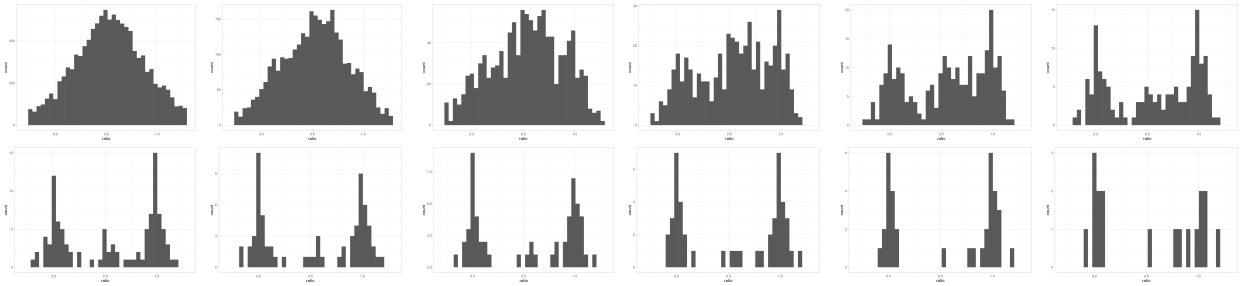


Figure 24: Histograms of r_1 for manufacturing sectors.

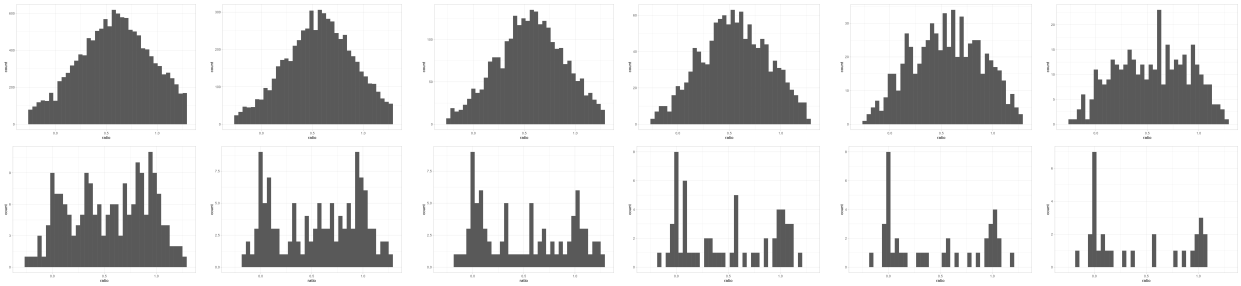


Figure 25: Histograms of r_3 for manufacturing sectors.

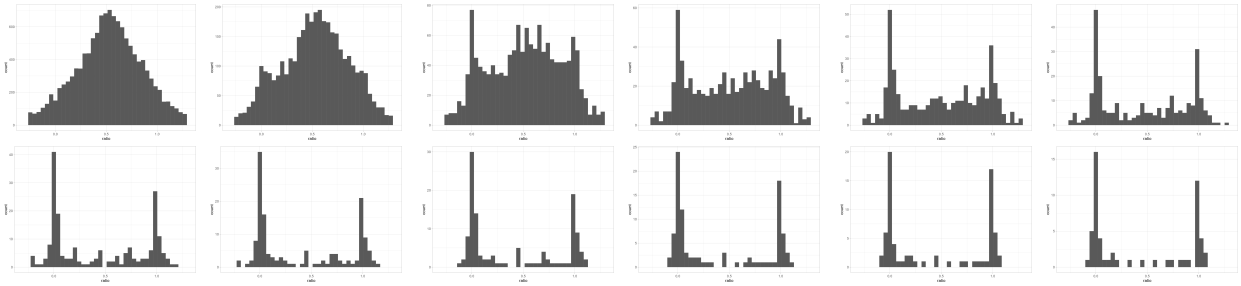


Figure 26: Histograms of r_1 for service sectors.

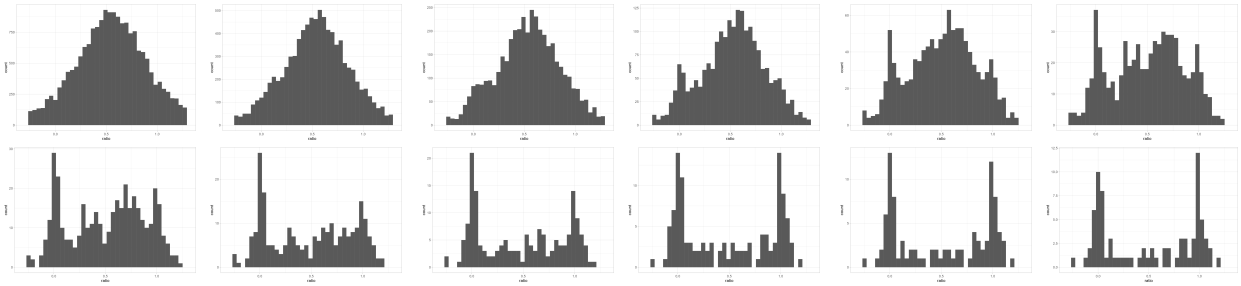


Figure 27: Histograms of r_3 for service sectors.

References

Arata, Y. (2019). Firm growth and laplace distribution: The importance of large jumps. *Journal of Economic Dynamics and Control*, 103:63–82.

Asmussen, S. (1982). Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the gi/g/1 queue. *Advances in Applied Probability*, 14(1):143–170.

Asmussen, S. and Albrecher, H. (2010). *Ruin probabilities*, volume 14. World scientific.

Asmussen, S. and Klüppelberg, C. (1996). Large deviations results for subexponential tails, with applications to insurance risk. *Stochastic processes and their applications*, 64(1):103–125.

Bianchini, S., Bottazzi, G., and Tamagni, F. (2017). What does (not) characterize persistent corporate high-growth? *Small Business Economics*, 48(3):633–656.

Bottazzi, G., Coad, A., Jacoby, N., and Secchi, A. (2011). Corporate growth and industrial dynamics: Evidence from french manufacturing. *Applied Economics*, 43(1):103–116.

Bottazzi, G., Dosi, G., Lippi, M., Pammolli, F., and Riccaboni, M. (2001). Innovation and corporate growth in the evolution of the drug industry. *International journal of industrial organization*, 19(7):1161–1187.

Bottazzi, G. and Secchi, A. (2006). Explaining the distribution of firm growth rates. *The RAND Journal of Economics*, 37(2):235–256.

Boucheron, S., Lugosi, G., and Massart, P. (2012). *Concentration inequalities A nonasymptotic theory of independence*.

Buldyrev, S. V., Growiec, J., Pammolli, F., Riccaboni, M., and Stanley, H. E. (2007). The growth of business firms: Facts and theory. *Journal of the European Economic Association*, 5(2-3):574–584.

Coad, A. (2007). A closer look at serial growth rate correlation. *Review of Industrial Organization*, 31(1):69–82.

Coad, A. (2009). *The growth of firms: A survey of theories and empirical evidence*. Edward Elgar Publishing.

Coad, A., Daunfeldt, S.-O., Hölzl, W., Johansson, D., and Nightingale, P. (2014). High-growth firms: introduction to the special section. *Industrial and Corporate Change*, 23(1):91–112.

Coad, A. et al. (2022). Lumps, bumps and jumps in the firm growth process. *Foundations and Trends® in Entrepreneurship*, 18(4):212–267.

Coad, A., Frankish, J., Roberts, R. G., and Storey, D. J. (2013). Growth paths and survival chances: An application of gambler's ruin theory. *Journal of business venturing*, 28(5):615–632.

Coad, A. and Hölzl, W. (2009). On the autocorrelation of growth rates. *Journal of Industry, Competition and Trade*, 9(2):139–166.

Daunfeldt, S.-O. and Halvarsson, D. (2015). Are high-growth firms one-hit wonders? evidence from sweden. *Small Business Economics*, 44(2):361–383.

Delmar, F., Davidsson, P., and Gartner, W. B. (2003). Arriving at the high-growth firm. *Journal of business venturing*, 18(2):189–216.

Dosi, G., Grazzi, M., Moschella, D., Pisano, G., and Tamagni, F. (2020). Long-term firm growth: an empirical analysis of us manufacturers 1959–2015. *Industrial and Corporate Change*, 29(2):309–332.

Dosi, G., Pereira, M. C., and Virgillito, M. E. (2017). The footprint of evolutionary processes of learning and selection upon the statistical properties of industrial dynamics. *Industrial and Corporate Change*, 26(2):187–210.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

Esteve-Pérez, S., Pieri, F., and Rodriguez, D. (2022). One swallow does not make a summer: episodes and persistence in high growth. *Small Business Economics*, 58(3):1517–1544.

Foss, S., Korshunov, D., and Zachary, S. (2011). *An Introduction to Heavy-Tailed and Subexponential Distributions*.

Frankish, J. S., Roberts, R. G., Coad, A., Spears, T. C., and Storey, D. J. (2013). Do entrepreneurs really learn? or do they just tell us that they do? *Industrial and Corporate Change*, 22(1):73–106.

Gardes, L., Girard, S., and Guillou, A. (2011). Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of*

Statistical Planning and Inference, 141(1):429–444.

Geroski, P. A. (2000). *Competence, Governance, and Entrepreneurship-Advances in Economic Strategy Research*, chapter The growth of firms in theory and in practice. Oxford University Press Oxford and New York.

Guarascio, D. and Tamagni, F. (2019). Persistence of innovation and patterns of firm growth. *Research Policy*, 48(6):1493–1512.

Joe, H. (2014). *Dependence modeling with copulas*. CRC press.

Lotti, F., Santarelli, E., and Vivarelli, M. (2009). Defending gibrat's law as a long-run regularity. *Small business economics*, 32(1):31–44.

Moschella, D., Tamagni, F., and Yu, X. (2019). Persistent high-growth firms in china's manufacturing. *Small Business Economics*, 52(3):573–594.

Sornette, D. (2006). *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media.

Stanley, M. H., Amaral, L. A., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. A., and Stanley, H. E. (1996). Scaling behaviour in the growth of companies. *Nature*, 379(6568):804–806.