NTC
Joint Research

Discussion Paper Series

The U-shape law for high growth firms

Fellow, Research Institute of Economy, Trade and Industry
Yoshiyuki Arata

Visiting Professor, National Tax College, Japan
Professor, School of Commerce, Waseda University
Daisuke Miyakawa

Assistant Professor, Research Dept., National Tax College, Japan
Katsuki Mori *

230203–01HJ

* Mori was at the National Tax College when this research was conducted and is currently at the National Tax Agency.

税務大学校
National Tax College
https://www.nta.go.jp

The U-shaped law of high-growth firms

Yoshiyuki Arata[*†‡]
Research Institute of Economy, Trade and Industry

Daisuke Miyakawa
Waseda University and National Tax College

Katsuki Mori
National Tax College

Abstract

This paper investigates the growth dynamics of firms using the theory of stochastic processes and corporate tax records covering nearly all firms in Japan. We show that the growth path of high-growth firms (HGFs) is characterized by a single large jump. Specifically, before the occurrence of this jump, the growth path of an HGF resembles that of non-HGFs, but it then increases rapidly in size due to the jump. This growth pattern with jumps is typical (i.e., most likely) for HGFs. To provide further empirical evidence, we consider the ratio that represents how much the growth rate of the first half of a given period contributes to the growth rate over the entire period. The histogram of this ratio shows a U-shaped curve for HGFs, indicating that high growth over the entire period can be explained by high growth in either the first half or the second half of the entire period (but not both). This U-shaped curve serves as further evidence that a single large jump determines the growth path of HGFs.

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

[*]Corresponding author: y.arata0325@gmail.com

[†]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). Mori was at the National Tax College when this research was conducted and is currently at the National Tax Agency. Conflict of interest: none.

[‡]This paper serves as a supplement to "Explaining Zipf's Law by Rapid Growth" by Y. Arata, H. Yoshikawa, and S. Okamoto and is not intended to be submitted to an academic journal on its own. For further theoretical background, refer to the original paper.

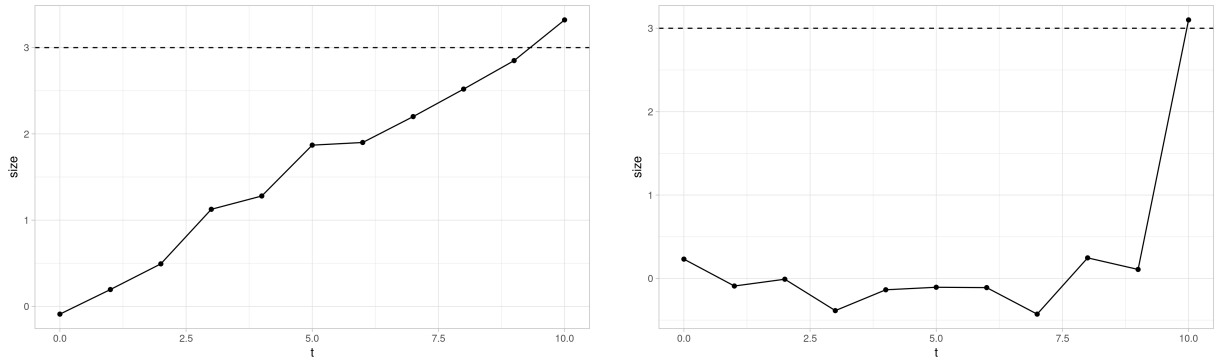
1 Introduction

What drives the growth of a firm? How do firms grow? These questions about the growth dynamics of firms are among the most classic and crucial themes in economics. Especially over the past decade, high-growth firms (HGFs) have been recognized as drivers of job creation, the emergence of new markets, and economic growth (e.g., Haltiwanger et al. (2017)). Understanding the growth dynamics of HGFs has now become important not only for researchers but also for policymakers.

However, despite the importance of such HGFs, many empirical studies on firm growth dynamics have reached the following unpleasant conclusion: we are unable to identify which firms will become HGFs in the future. While certain variables related to firm growth dynamics, such as age and size, have been identified, it is understood that economic models have very weak explanatory power and cannot be used to predict HGFs. For instance, surveying empirical studies, Geroski (2000) concludes as follows: "[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 this mean that firm growth is completely random and nothing can be said about it? Is there no way to improve our understanding of firm growth dynamics?

This paper shows that, even if we are unable to identify potential HGFs, we can still gain meaningful insights into how firms grow. Specifically, our analysis does not rely on any particular optimization model but rather examines the typical growth path of HGFs by analyzing the statistical regularities observed in empirical data. Our analysis reveals the most likely growth paths for HGFs. A key assumption in our analysis, which is extensively tested using empirical data, pertains to the distribution of firm growth rates. It is well known in the literature that the growth rate distribution has a heavier tail than a Gaussian distribution and is closer to a Laplace distribution. In our analysis, by examining the growth rate distribution more closely, we find that the distribution has a tail that is strictly heavier than an exponential. Based on this empirical fact regarding the shape of the growth rate distribution, our analysis reveals that firm growth is characterized not by steady, incremental increases, but rather by abrupt, substantial jumps. Until just before these jumps occur, the growth path of an HGF is indistinguishable from that of non-HGFs, but then, the firm size increases rapidly due to the jump (see **Figure 1**). We find that this type of growth pattern is not an exception but rather the norm.

Our analysis is twofold, consisting of a theoretical analysis using probability theory and an empirical analysis using comprehensive administrative data from Japan. In the theoretical analysis, we consider two classes of distributions: light-tailed and heavy-tailed distributions. Light-tailed distributions are those whose tails are exponentially bounded. This means that the tail probability decreases more rapidly than an exponential function, implying that the probability of the random variable (i.e., the growth rate of firms) taking extremely large or small values is low. This class of distributions includes the Gaussian distribution. The Laplace distribution also belongs to this class but is considered to be on the boundary with the heavy-tailed distributions class described below. Heavy-tailed distributions are those with tails heavier than an



(a) Gradual increase

(b) Increase by a sudden jump

Figure 1: The image of sample paths for HGFs. The horizontal axis represents time, and the vertical axis represents firm size. In Panel (a), a firm gradually increases in size, achieving high growth over the entire period through many small successes. In Panel (b), the firm's growth path is characterized by a sudden, large jump.

exponential function. This means that the probability of the random variable taking extreme values is higher than what would be predicted by an exponential. This class of distributions includes distributions with heavy tails, such as the log-normal distribution and the Pareto distribution. 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.

In our theoretical analysis, we assume that the logarithm of firm size follows a random walk and focus on the sample paths of firms that experience rapid growth over n periods (i.e., firms whose growth rate over n periods exceeds a large threshold u). Since the growth rate over n periods is composed of n individual growth rates, we analyze how these growth rates from each period contribute to the overall growth rate over n periods and achieve u. By utilizing ruin theory (e.g., Asmussen and Albrecher (2010)), we show that when the growth rate distribution is light-tailed, high growth rates over n periods are primarily determined by the cumulative effect of individual growth rates. The growth rate in each period contributes almost equally to the overall high growth rate over n periods, and thus, the sample path is characterized by a gradual increase, as illustrated in **Figure 1**(a). In other words, for a light-tailed distribution of growth rates, the most likely growth path leading to high growth over n periods is a path characterized by a gradual increase.

In contrast, when the growth rate distribution is heavy-tailed, high growth over n periods is driven by the presence of a large individual growth rate, or what we refer to as a *jump*. More precisely, the probability that the growth rate over n periods exceeds a threshold u is asymptotically equal to the probability that the maximum of the n individual growth rates exceeds u. That is, high growth over n periods is dominated by a single burst within one period, making the contributions from growth rates in other periods negligible. The heavy-tailedness of the growth rate distribution implies that a rare, large jump characterizes the growth path of HGFs, as illustrated in **Figure 1**(b).

Given these theoretical backgrounds, our empirical tasks are to (1) verify the random walk assumption

and (2) test the distribution class to which the growth rate distribution belongs. For our empirical analysis, we use data on corporate tax records provided by the National Tax College, which includes annual sales revenues, profits, and the amount of corporate tax paid by firms. This is the population data in Japan and covers 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.

To empirically test the random walk assumption, we focus on the correlation of growth rates over consecutive periods. Specifically, we consider dependencies not only in the entire distribution, as measured by Spearman's ρ_S or Kendall's τ, but also in their tail regions. We find that the dependence between consecutive growth rates, if they exist, is weak across the entire distribution and even weaker in the tail regions. These results support the random walk assumption, especially when focusing on HGFs. For the growth rate distribution, we consider one-year and three-year growth rates and analyze the heaviness of the distribution tails using density estimates, QQ-plots, and mean excess functions. We find that the growth rate distributions have tails that are strictly heavier than an exponential, indicating that they are subexponential. These empirical findings confirm that the two assumptions used in our theoretical analysis are consistent with the data, implying that the growth path of HGFs is driven by a large jump rather than a gradual increase.

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

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

where X_1 and X_2 represent the growth rates for the first half and second half of a certain period, 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 over the two years). The ratio r represents the contribution of the growth rate in the first half to the growth rate over the entire period. We examine how the histogram of r changes depending on the value 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 focusing on HGFs, it is more likely that high growth over the entire period is caused by high growth in either the first half or the second half of the entire period, but not both. This empirical finding aligns with our theoretical analysis, which suggests that the sample path of HGFs is characterized by sudden, large jumps. The U-shaped curve of the histogram of r reflects that these jumps occur either in the first half or the second half but are not evenly distributed across the entire period.

Our findings are derived from statistical regularities, that is, the random walk and the growth rate distribution; we do not employ a particular optimization model to describe firm behavior. Following the spirit of Geroski (2000), we assume that firm growth is highly unpredictable—essentially random—and that it is not possible to predict which firms will become HGFs in the future. However, our analysis shows that even if a firm's growth is random, there exist robust empirical features that characterize firm growth dynamics. This is because, due to the randomness (i.e., the sufficient complexity of firm growth), firm growth dynamics are governed by the logic of probability theory. The U-shaped curve of the ratio r, which we call

the U-shaped law, is an example of how randomness gives rise to an empirical law in economic phenomena.

Related literature

This paper belongs to the literature on firm growth dynamics, which aims to understand the observed empirical regularities (see Coad (2009), Coad et al. (2014), Dosi et al. (2017), and Coad et al. (2022b) for a survey). In particular, a series of empirical studies have explored two key assumptions regarding firm growth: the heavy-tailedness of the growth rate distribution and the random walk assumption (i.e., the independence of growth rates over consecutive periods).

Regarding the former, since the seminal work by Stanley et al. (1996), it has been recognized 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 means that compared to a Gaussian distribution, the growth rate distribution has a more peaked center and heavier tails. This distribution shape is one of the most robust empirical regularities, 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 demonstrate that the tail of the growth rate distribution is strictly heavier than that of the Laplace distribution (i.e., an exponential tail). For instance, Bottazzi et al. (2011) reject the null hypothesis that growth rates follow a Laplace distribution and propose the Subbotin family, which includes a Laplace distribution as a special case. In our analysis, consistent with these empirical studies, we confirm that the tail of the growth rate distribution is heavy and strictly heavier than an exponential function. However, we do not specify the functional form of the growth rate distribution. In our analysis, the only requirement is that the tail is heavier than an exponential (i.e., subexponential), and there is no need to specify its functional form. In this respect, our implications regarding the pattern of firm growth dynamics can be considered robust.

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 exhibit negative autocorrelation for small firms, large firms exhibit positive autocorrelation. However, this autocorrelation reported in the literature is generally weak, and in most instances, "lagged growth is a poor signal of future growth" (Coad et al. (2013), p.617).[1] 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), Coad et al. (2018), Hölzl (2014), Esteve-Pérez et al. (2022)). These studies empirically show that the persistence of high growth is quite weak. Notably, Daunfeldt and Halvarsson (2015) show that HGFs are "one-hit wonders,"

[1]Another empirical regularity related to the random walk assumption is Gibrat's law. Considering the sample sizes used in recent empirical studies, the null hypothesis that the growth rates of firms are independent and identically distributed tends to be rejected. However, when considering only matured firms excluding small businesses, it is known that Gibrat's law approximates the empirical growth process well. See Lotti et al. (2009) and Daunfeldt and Elert (2013).

meaning that firms experiencing a high-growth period do not subsequently undergo another high-growth period.[2] Additionally, another empirical finding that makes our stochastic approach more appealing is the lack of firm attributes characterizing growth dynamics. 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 extremely difficult to predict high growth in advance, and therefore, it is reasonable to describe firm growth dynamics as a stochastic process, aligning with the random walk assumption.[3]

The closest paper to our study is Coad et al. (2013), where they model firm growth dynamics as a simple random walk with increments of ± 1. That is, by only considering whether the increment is positive or negative, they compare the frequency of observed growth patterns (such as four consecutive positive growths $++++$ or alternating pattern $+-+-$) with those predicted by the simple random walk. They show that this simple random walk provides a good approximation for growth dynamics.[4] 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 (i.e., the growth rate distribution). Our analysis shows that the sample path properties of the random walk qualitatively differ depending on the heaviness of the distribution tail of increments. By this method, our analysis provides a unified explanation entailing the heaviness of the distribution tail, (non-)persistence, and the sample path 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 examine the effect of firm age on firm growth dynamics.

[2] In analyzing the persistence of firm growth, many previous studies have focused the autocorrelation of the firms' growth rates themselves. In contrast, Capasso et al. (2014) and Bottazzi et al. (2023) utilize quantile values and transition probability matrices of growth rates, focusing on dependence in the tail regions. In line with these studies, our analysis examines dependence not only as represented by correlation coefficients across the entire distribution but also in the tail regions.

[3] One might think that the high growth of a firm is the result of innovation, such as R&D investments, and that there is a close connection between the persistence of a firm's high growth and the persistence of innovation. However, an empirical study by Guarascio and Tamagni (2019) show that the persistence of a firm's high growth is not related to the persistence of innovation. Such results support our approach of describing firm growth as a stochastic process.

[4] As an empirical study on firm growth patterns, Coad et al. (2022a) should be mentioned. They investigate whether the growth patterns of firms (e.g., whether the growth path is smooth or involves significant fluctuations) affect subsequent growth or the probability of exit. In contrast, our analysis focuses solely on the observed growth patterns (i.e., without considering the effect on subsequent growth or exit probabilities), aiming to find empirical laws within those growth patterns.

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 random variables. 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 is defined as $X_k := \log S_k - \log S_{k-1}$. 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 implications derived from the iid assumption. First, under the iid assumption, a firm's growth rate is independent of its initial size; that is, X_k does not depend on $\log S_{k-1}$. This is known as Gibrat's law and is widely accepted in the existing literature as a reasonable approximation for the empirical growth dynamics. Furthermore, the iid assumption implies that there is no autocorrelation of growth rates $X_1, X_2, ..., X_n$. This means that, under the iid assumption, high growth in one period does not affect the probability of high growth in subsequent periods. We will empirically examine this point in Section 3.2.

As will be discussed in Section 2.3, the property of a random walk depends on the distribution of its increments, namely X_k, especially on the tail part. In our analysis, rather than assuming X_k to follow some particular distribution, we introduce distribution classes categorized by the heaviness of the distribution tail. Since our interest lies in HGFs, only the right tail of the distribution is considered. The first distribution class is light-tailed distributions, which are defined by their tails being exponentially bounded. More precisely, this class is defined 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 distribution, and Laplace distribution. The Laplace distribution, in particular, is of great importance in our analysis; since the Laplace distribution has an exponential tail, it can be considered as the boundary of this class. That is, if a distribution has a tail strictly heavier than an

exponential, it is not light-tailed. In such cases (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 known as subexponential distributions, which requires regularity in the asymptotic behavior of the distribution tail.

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 also introduce a subclass of subexponential distributions, which requires another slightly stronger regularity condition on their tails.

Definition 2.4. A heavy-tailed distribution F on $\mathbb{R}$ is strong subexponential if the mean of $X^+ := \max\{0, X\}$ exists and

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

exists.

If the limit in Eq.(1) exists for a heavy-tailed distribution, it is equal to 2 (see Theorem 2.12 in Foss et al. (2011)). Recall that $\overline{F * F}$ represents the tail probability of the sum of two iid random variables, X_1 and X_2, and the tail probability of the maximum of these two random variables is $\mathbb{P}(\max\{X_1, X_2\} > x) = 1 - F^2(x) \sim 2\overline{F}(x)$. Therefore, Eq.(1) means that the tail probability of the sum is asymptotically equivalent to that of the maximum of the two random variables. That is, when the sum $X_1 + X_2$ takes a large value, it is due to either X_1 or X_2 taking a large value, but not both.

While (strong) subexponential distributions are a proper subclass of heavy-tailed distributions (i.e., there exist heavy-tailed distributions that are not subexponential), almost all of the heavy-tailed distributions encountered in practical applications are (strong) subexponential.[5] For instance, Pareto, log-normal, and Weibull distributions with an exponent less than 1 are included in the class of (strong) subexponential distributions (see, e.g., Chapter 3 in Foss et al. (2011)). In particular, consider the Weibull distribution with parameter $\alpha > 0$;

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

[5]The requirement in Definition 2.3 is only that the limit in Eq.(1) exists, and it does not require convergence to a specific value. As long as there is sufficient regularity for the limit in Eq.(1) to exist, these conditions are satisfied, meaning that, except for pathological cases, most heavy-tailed distributions can be considered subexponential. The same applies to Definition 2.4.

The parameter α controls the heaviness of the tail: as α decreases, the tail becomes heavier. When $\alpha = 1$, it reduces to the exponential distribution. Thus, the Weibull distribution with $\alpha \geq 1$ (including the exponential case) is light-tailed, whereas 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 qualitatively differs depending on whether the distribution of X_k is light-tailed or heavy-tailed. Before discussing the 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 also follows a Gaussian distribution with mean 0 and variance $n\sigma^2$. Thus, 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. Here, Φ represents the distribution function of the Gaussian distribution. Thus, with a fixed n, the tail probability of the sum is controlled by σ^2 and decays rapidly as $u \to \infty$ (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. Specifically, 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 a centered random variable X_k satisfies*

$$\log E e^{\lambda X_k} \leq \frac{\nu^2 \lambda^2}{2}, \quad \textit{for all } |\lambda| < \frac{1}{\alpha} \tag{2}$$

for some non-negative constants ν, α. Then, the tail probability of the sum is bounded as follows:

$$\mathbb{P}(\sum_{k=1}^{n} X_k > u) \leq \begin{cases} \exp(-\frac{u^2}{2n\nu^2}) & \textit{for } 0 \leq u \leq \frac{n\nu^2}{\alpha} \\ \exp(-\frac{u}{2\alpha}) & \textit{for } u > \frac{n\nu^2}{\alpha} \end{cases}$$

Proof. This is a straightforward application of concentration inequalities (see, e.g., Wainwright (2019) and Boucheron et al. (2012) for reviews). Proposition 2.9 in Wainwright (2019) states that using Chernoff's inequality, we obtain the upper bound on the tail probability of X_k:

$$\mathbb{P}(X_k > u) \leq \begin{cases} \exp(-\frac{u^2}{2\nu^2}) & \text{for } 0 \leq u \leq \frac{\nu^2}{\alpha} \\ \exp(-\frac{u}{2\alpha}) & \text{for } u > \frac{\nu^2}{\alpha} \end{cases}$$

Due of the independence of $X_1, ..., X_n$, the logarithm of the moment generating function exhibits additivity; that is,

$$\log E e^{\lambda \sum_{k=1}^{n} X_k} \leq \frac{n\nu^2 \lambda^2}{2}, \quad \text{for all } |\lambda| < \frac{1}{\alpha}$$

From this and the above upper bound, the desired result is derived. □

For the Gaussian distribution, Eq.(2) is satisfied with $\nu = \sigma, \alpha = 0$. This result is equivalent to the

upper bound of the tail probability derived from Mills' ratio, except for a constant factor. For the Laplace distribution with parameter b (i.e., the variance is $2b^2$), Eq.(2) is satisfied with $\nu = 2b, \alpha = 2b$. In this case, for a fixed n, the probability in the central region (i.e., $0 \leq u \leq 2nb$) exhibits a Gaussian decay, similar to the case of the Gaussian distribution. This aligns with the central limit theorem, suggesting that the rapid decay of probability occurs because $X_1, ..., X_n$ cancel each other out. On the other hand, when considering a large value of u, the tail probability of the sum $\sum_{k=1}^{n} X_k$ deviates from Gaussian decay. This is because, for a fixed n, the central limit theorem holds only near the center of the distribution and does not extend to the region where $u > 2nb$. Moreover, note that if the distribution of a component X_k is exponentially bounded, the tail probability of the sum $\sum_{k=1}^{n} X_k$ is also exponentially bounded. That is, if the tail probability of the sum $\sum_{k=1}^{n} X_k$ is heavier than an exponential, the tail probability of the element X_k is not exponentially bounded.

When subexponential distributions are considered, the deviation from Gaussian distributions is more significant and provides a different implication. If the distribution of X_k is subexponential, the limit in Eq.(1) can be extended to any n, the tail probability of the sum can be approximated as follows (see Corollary 3.20 in Foss et al. (2011)): as $u \to \infty$,

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

Note that the right-hand side of Eq.(3) represents the probability that the maximum of the n iid random variables $\max\{X_1, ..., X_n\}$ exceeds u; that is, $\mathbb{P}(\max\{X_1, ..., X_n\} > u) = 1 - F^n(u) \sim n\overline{F}(u)$. This means that an extreme value of the sum $\sum_{k=1}^{n} X_k$ is typically caused by one extreme value among its components. In other words, when dealing with subexponential random variables, the probability that a combination of moderate values of components results in an extreme value of the sum is negligible. Moreover, Eq.(3) implies 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 exhibits the same decay rate as $u \to \infty$. In Section 3, we use these properties to test whether the growth rate distribution is (strong) subexponential.

2.3 Sample path properties

Here, we discuss the distribution of the growth rates $X_1, \ldots, X_n$ given that a high growth rate is achieved over n periods. In other words, we examine which combinations of growth rates $X_1, \ldots, X_n$ are most likely to occur given this rare event. On this point, we introduce two general results from probability theory (Asmussen (1982) and Asmussen and Klüppelberg (1996)).

To accomplish this, several technical assumptions are necessary. We are interested in firms that grow rapidly and outperform others. However, if $\mathbb{E}X_k$ is positive and a longer time period is considered (i.e., $n \to \infty$), the condition $\sum_{k=1}^{n} X_k > u$ is satisfied with probability 1. This means almost all firms would meet this condition. Rather than considering such a trivial case, we define the growth rate as an excess from

a constant c (a value close to $\mathbb{E}X_k$) and focus on firms whose sum of excess growth rates is large. This is equivalent to focusing on firms that significantly outperform (i.e., exceed) the average growth rate of other firms. More precisely, letting $Y_k := X_k - c$ and considering the random walk of $Y_1, ..., Y_n$ with a negative drift, we focus on the event $\sum_{k=1}^{n} Y_k > u$ for some n.

The probability of this event is less than 1 (because $\mathbb{E}Y_k < 0$), and it becomes smaller as u increases (i.e., a rare event). Once this rare event occurs, how does the sequence $Y_1, Y_2, ...$ reach u? We consider the following conditional probability $\mathbb{P}_u := \mathbb{P}(\cdot | \sum_{k=1}^{n} Y_k > u \text{ for some } n)$. Let F_n be the empirical distribution of $Y_1, ..., Y_n$ under $\mathbb{P}_u$:

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

Letting $\nu(u)$ be the time when $\sum_{k=1}^{n} Y_k$ 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 $Y_1, ..., Y_{\nu(u)}$ conditional on the event that the random walk exceeds u at $\nu(u)$.

For the case of light-tailed distributions, Asmussen (1982) identifies the distribution towards 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

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

where $\gamma > 0$ is chosen such that $Ee^{\gamma Y_k} = 1$ and $E|Y_k|e^{\gamma Y_k} < \infty$. Note that F_γ has a positive mean. Letting $\| \cdot \|$ denote the supreme norm, the result relevant to our analysis is as follows:

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 indicates 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, meaning that for most firms, the sum $\sum_{k=1}^{n} Y_k$ will eventually trend towards $-\infty$ as $n \to \infty$. The fact that F_γ has a positive mean suggests that the growth rates for HGFs (i.e., firms for which $\sum_{k=1}^{n} Y_k > u$ for some n), $Y_1, ..., Y_{\nu(u)}$ are upward drifted, and consequently, their sum reaches u. The latter half of the theorem provides a similar image for the sample paths of HGFs. 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 of tu on average. Therefore, when the growth rate distribution 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.[6] *Then, as* $u \to \infty$*,*

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

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

This theorem implies that, in contrast to the light-tailed case, the conditional distribution of growth rates for HGFs (i.e., firms that satisfy $\sum_{k=1}^{n} Y_k > u$ for some n) is essentially the same as the unconditional one, i.e., non-HGFs. Indeed, the latter half of the theorem means that the random walk of HGFs up to time $t\nu(u)$ (i.e., $Y_1, ..., Y_{t\nu(u)}$) decreases on average by $-\mu t\nu(u)$, which is the same as that for other non-HGFs. Intuitively, this is equivalent to saying that the sample path for HGFs is the same as that for non-HGFs just before a large jump arrives, then a single large jump leads to 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: a gradual increase and a sudden increase due to 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 class of the growth rate distribution. These tasks will be carried out 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 our data. Section 3.2 analyzes the random walk assumption by focusing on the correlation of growth rates over consecutive periods. Section 3.3 analyzes the heavy-tailedness of the growth rate distribution. Section 3.4 shows that the sample path properties of the random walk 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. Since the amount of corporate tax for each firm is calculated based on the firm's profits, firms are required to report their profits annually. As this report is mandatory for all firms in Japan, this data covers almost all companies in Japan.[7] Firms also report their basic attributes (e.g., firm's name,

[6] More precisely, the condition required here is that 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 more on extreme value theory, see Embrechts et al. (1997).

[7] While our analysis defines a firm as an unconsolidated entity, and indeed most firms pay corporate taxes on this basis, parent firms that own 100% of the stocks of a subsidiary can file taxes as a consolidated firm, allowing them to offset the profits and

location, and industry) as well as annual sales revenues. Spanning from 2014 to 2020, the data includes a unique ID for each firm, which is used to construct panel data.

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 incorporation. The other is about records of mergers. It enables us to identify the year when a merger occurs and the firm IDs involved in each merger. To focus on firms' internal growth, we exclude firm-year observations where a merger takes place from our samples.[8]

We impose several conditions on the sample used in our analysis. First, we exclude firms in the financial and government sectors from our sample. Next, we exclude micro firms with extremely small sales. This is because our main variable of interest, the growth rate, is defined as the logarithmic difference in sales (i.e., for firm i in period k, $X_{i,k} := \log S_{i,k} - \log S_{i,k-1}$), and if the size of a firm in the initial period is extremely small, it would result in an extremely large value for its growth rate. Since our analysis assumes that growth rates are iid random variables, we only consider firms with sales of more than 100 million yen in the initial period (i.e., in 2014). Finally, we exclude firms established less than ten years before 2014 from our sample. This is because the growth dynamics of a firm are related to its age, and specifically, the random walk assumption (or Gibrat's law) that we impose in our analysis is less likely to hold for newly established firms (see, e.g., Lotti et al. (2009)). In other words, our analysis focuses on firms that are neither too young nor too small.[9]

As a result of the aforementioned procedures, the sample size is reduced to $548,657$ for growth rates from 2014 to 2015 (denoted by X_{15}).[10] The sample sizes and summary statistics of growth rates for other years are provided in **Table 1**.

---

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 $12,983$.

[8] 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 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.

[9] The analysis of young firms with age less than ten years is provided in the Appendix. We find that the growth dynamics of young firms differ from those of older firms.

[10] We denote the one-year growth rate by X_k and the three- and six-year growth rates by X_{k-l}, where k is the end period and l is the initial period. For example, X_{17-14} represents the three-year growth rate from 2014 to 2017, i.e., $X_{17-14} := \log S_{2017} - \log S_{2014}$.

Summary statistics for firms' growth rates

Corporate tax data in Japan

year	count	mean	sd	q1	median	q3
one-year growth rate						
2014-2015	548657	−0.049	0.430	−0.100	−0.008	0.067
2015-2016	537760	−0.038	0.470	−0.096	−0.011	0.062
three-year growth rate						
2014-2017	526652	−0.092	0.600	−0.180	−0.023	0.113
2017-2020	501034	−0.140	0.680	−0.260	−0.074	0.075
six-year growth rate						
2014-2020	493294	−0.210	0.760	−0.360	−0.105	0.116

Table 1: Summary statistics of growth rates. One-, three-, and six-year growth rates are considered. Here, only firms that are not in the financial and government sectors, with sales of more than 100 million yen in 2014, and that have been established for more than ten years are included.

3.2 Autocorrelation

The random walk assumption implies that the growth rates in consecutive periods are independent of each other. In this section, we examine the empirical validity of the random walk assumption by analyzing the dependence structure between growth rates in consecutive periods. The most commonly used measure to evaluate the dependence between two random variables is Pearson's correlation coefficient. However, it is well known that Pearson's correlation coefficient is not independent of their marginal distributions. For example, even if the dependence between the two random variables remains unchanged, Pearson's correlation coefficient can vary when the marginal distributions change. In other words, it is not clear whether Pearson's correlation coefficient truly reflects the dependence between the two random variables or is influenced by their marginal distributions, making it a less reliable measure of dependence (for more details, see Embrechts et al. (2002)). To avoid these issues, we use Kendall's τ and Spearman's ρ_S as alternative measures.

Kendall's τ measures the ordinal association between two random variables, assessing 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, Kendall's τ is defined as follows:

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

Kendall's τ measures the probability that the order of two random variables aligns, and thus, it does not depend on the marginal distributions of X_1, X_2. Kendall's τ ranges between $[-1, 1]$, and equals 0 when the two random variables are independent of each other.

Spearman's ρ_S is a measure of rank correlation, assessing the degree to which the order of two variables

is correlated. It is defined as the correlation coefficient of the transformed variables $F_1(X_1)$ and $F_2(X_2)$:

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

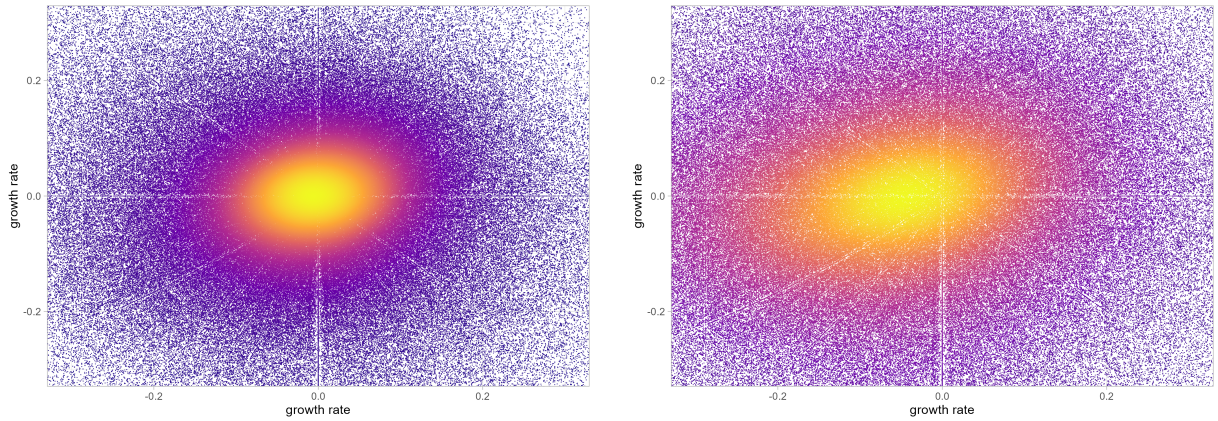
Here, F_1 and F_2 are the distribution functions of X_1 and X_2, respectively. When transformed by these functions, both $F_1(X_1)$ and $F_2(X_2)$ follow a uniform distribution in the $[0, 1]$ interval, thus ρ_S does not depend on the marginal distributions of X_1, X_2. Similar to Kendall's τ, ρ_S ranges between $[-1, 1]$, and equals 0 when the two variables are independent of each other. For a later purpose, we provide the values of Kendall's τ and Spearman's ρ_S when X_1 and X_2 follow a bivariate Gaussian distribution with parameter ρ (see Section 4.3 in Joe (2014)):

$$\tau = 2\pi^{-1} \arcsin(\rho), \quad \rho_S = 6\pi^{-1} \arcsin(\rho/2)$$

Using these dependence measures, we examine the dependence of one-year growth rates (i.e., X_{15}, X_{16}) as well as the dependence of three-year growth rates (X_{20-17}, X_{17-14}). The results show that for one-year growth rates, the estimates of Kendall's τ and Spearman's ρ_S are -0.0059 and -0.017 respectively. For the three-year growth rates, the estimates of Kendall's τ and Spearman's ρ_S are 0.033 and 0.045 respectively. In both cases, the estimates of τ and ρ_S are both close to 0, suggesting that the dependence between consecutive growth rates is very weak. This can also be confirmed through the scatter plots of the two growth rates. **Figure 2**(a) and (b) show the scatter plots for the one-year growth rates X_{15}, X_{16} and for the three-year growth rates X_{20-17}, X_{17-14}, respectively. As is evident from the figures, there is no clear dependence between them. Consistent with the estimates of the correlation coefficients, this suggests that any dependence between consecutive growth rates (if it exists) is very weak.[11]

The results of weak dependencies above suggest the validity of treating growth rates as independent random variables. However, Spearman's ρ and Kendall's τ capture dependencies across the entire range of the distribution and thus may not necessarily reflect dependencies in the tail regions (i.e., high growth). To analyze dependencies in the tail regions, we count how many times high growth is experienced by each firm within the entire sample period (i.e., across six growth rates: $X_1, X_2, ..., X_6$). Consider the following two extreme scenarios. The first scenario is that one group of firms experiences consecutive high growth throughout the entire period, while the other group does not experience any high growth at all. The second scenario is that the random walk assumption holds, and therefore, the frequency of high growth occurrences follows a binomial distribution. To be more precise, when considering growth rates higher than the pth

[11] Using the estimates of Kendall's τ and Spearman's ρ_S, it is possible to test whether X_{15} and X_{16} are independent of each other. Given that $\tau = \rho_S = 0$ under the independence assumption (i.e., the null hypothesis), we test whether the estimates significantly deviate from 0. We confirmed that the null hypothesis can be rejected with p-values below 0.01. The same applies to X_{20-17} and X_{17-14}. Since our analysis assumes that growth rates are iid random variables, one might think that this result contradicts our assumption. However, as emphasized in the Introduction, what is necessary for our analysis is that firm growth dynamics are sufficiently *random*. Considering the magnitude of the estimates of Kendall's τ and Spearman's ρ_S, this assumption seems to be empirically valid. Indeed, as we will see later, the statistical regularities predicted by the iid assumption are actually observed in the data.



(a) X_{15} (y-axis) and X_{16} (x-axis)

(b) X_{17-14} (y-axis) and X_{20-17} (x-axis)

Figure 2: Scatter plot of growth rates. The bright orange color represents areas of high sample point density.

percentile as high growth, the probability distribution of the number of occurrences would be following: for $0 \leq m \leq 6$

$$\mathbb{P}(\text{Number of high growth} = m) = \binom{6}{m} p^m (1-p)^{6-m}$$

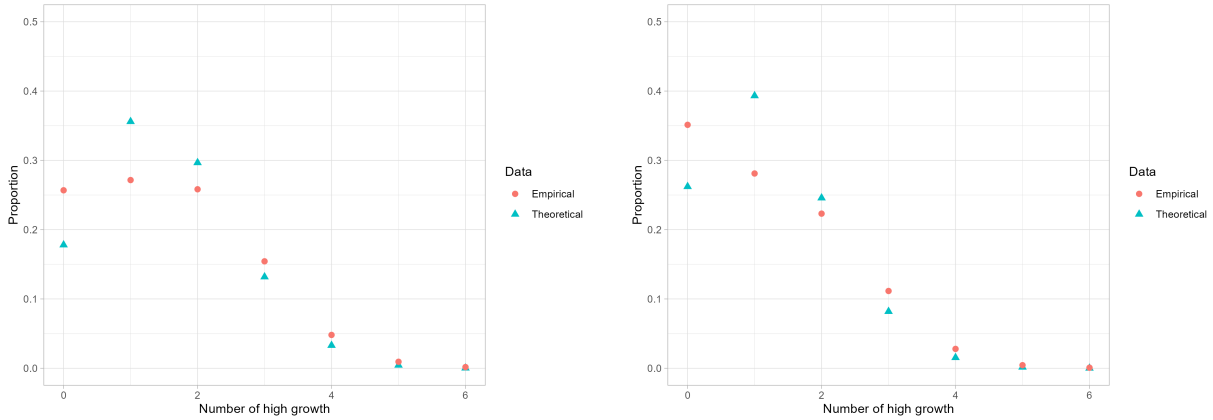
Below, we consider cases with $p = 0.75, 0.80, 0.97, 0.99$ and compare the empirical frequency of high-growth occurrences with a binomial distribution.

The results for $p = 0.75, 0.80$ are provided in **Figure 3**, that is, we consider the number of occurrences of *moderate* positive growth within the sample period. The histogram of the number of occurrences deviates both qualitatively and quantitatively from the binomial distribution. For example, in the case of $p = 0.8$, the mode of the binomial distribution is given at $m = 1$, whereas the histogram of occurrences has its mode at $m = 0$. Therefore, for $p = 0.75, 0.80$, the independence of growth rates (i.e., the random walk hypothesis) does not hold, suggesting the existence of dependence between growth rates in consecutive periods.

However, when considering cases of high growth, such as $p = 0.97$ and 0.99, the implication is totally different. The results for these cases are provided in **Figure 4**. As clearly shown in **Figure 4**, the histogram of occurrences of high growth is very close to the binomial distribution. This means that there are no special groups of firms that experience high growth consecutively, and even among firms that have experienced high growth during the sample period, most experience only once. This serves as additional evidence that, when considering the tail region such as $p = 0.97, 0.99$, the random walk assumption provides a good approximation for the empirical firm growth dynamics.

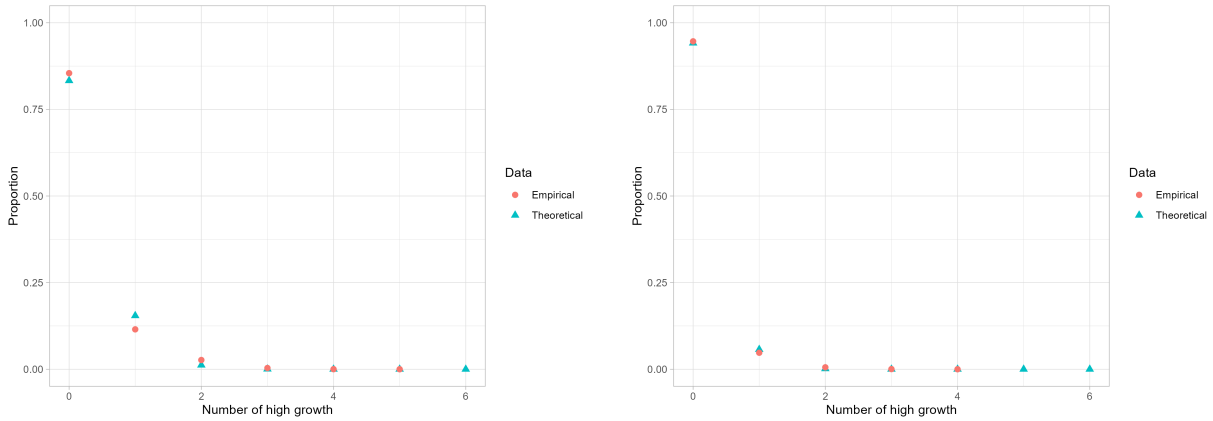
3.3 Growth rate distribution

In this section, we test the other assumption in our analysis, namely, whether the growth rate distribution is subexponential. We investigate whether the growth rate distribution has a heavier tail than an exponential function. It should be noted that our analysis does not require that the growth rate distribution or its tails



(a) $p = 0.75$ **(b)** $p = 0.80$

Figure 3: Histogram of the number of occurrences of $X_k > F_k^{-1}(p)$ within the sample period compared to the binomial distribution. For panels (a) and (b), the binomial distribution with parameters $p = 0.75, 0.80$ is calculated, respectively. "Empirical" represents the histogram of the number of occurrences of $X_k > F_k^{-1}(p)$, and "Theoretical" represents the theoretical values calculated from the binomial distribution.



(a) $p = 0.97$ **(b)** $p = 0.99$

Figure 4: Histogram of the number of occurrences of $X_k > F_k^{-1}(p)$ within the sample period compared to the binomial distribution. For panels (a) and (b), the binomial distribution with parameters $p = 0.97, 0.99$ is calculated, respectively. "Empirical" represents the histogram of the number of occurrences of $X_k > F_k^{-1}(p)$, and "Theoretical" represents the theoretical values calculated from the binomial distribution.

to follow any particular functional form but only that it has a tail heavier than an exponential. Hence, in the following analysis, we focus on comparing the estimated tail of the growth rate distribution with an exponential function.

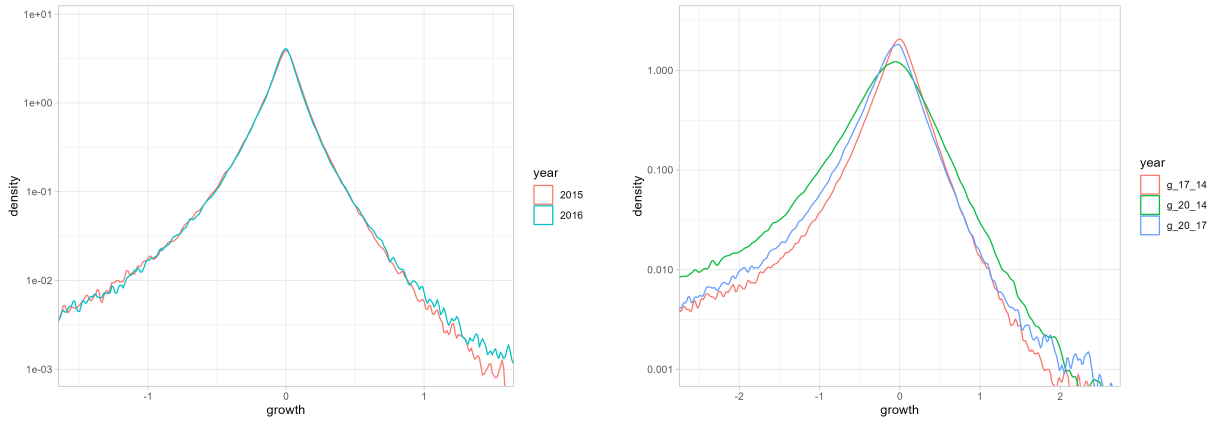
First, we perform a density estimation of the growth rate distribution. Here, we plot the density function with the y-axis on a log scale. The reason for using a log scale on the y-axis is that an exponential function appears as a straight line, and the density function of the Laplace distribution forms a triangle: that is, in the tail region (i.e., the exponential tail region), the density function exhibits a straight line if growth rates follow the Laplace distribution. Therefore, any deviation from a straight line in the tail region can be considered evidence that the growth rate distribution deviates from an exponential tail. We also consider the complementary cumulative distribution function (CCDF). This function, which is defined as $1 - F_n(x)$ where F_n is the empirical distribution derived from the sample, represents the probability of observing values greater than x, i.e., an estimate of the tail probability. Similar to the density estimate, we examine whether the CCDF derived from the data deviates from a straight line.

Figure 5 plots the density estimates for one-year and three-year growth rates with the y-axis on a logarithmic scale. As is evident from both figures, consistent with previous studies, the growth rate distribution is more peaked at the center and has heavier tails compared to the Gaussian distribution. Moreover, in both figures, the density function does not follow a straight line but instead curves upwards in the tail regions. Similarly, the CCDF shown in **Figure 6** does not follow a straight line but curves upwards as larger growth rates are considered. This serves as evidence that the growth rate distribution has a tail heavier than an exponential function, implying that it is subexponential.[12]

Additionally, in **Figure 5**(b) and **Figure 6**(b), we also consider the six-year growth rate distribution. Compared to the distributions for one-year and three-year growth rates, the central part appears smoother, while in the tails regions, it remains heavier, similar to those of one-year and three-year growth rates distributions. As seen in Section 2.2, within a short sample period of $n = 6$, the central limit theorem applies only to a limited central part, and the tails remain heavier than both the Gaussian and exponential. Thus, in the tail regions, the distribution of the sum $\sum_{k=1}^{n} X_k$ reflects the characteristics of the distribution of individual growth rates X_k.

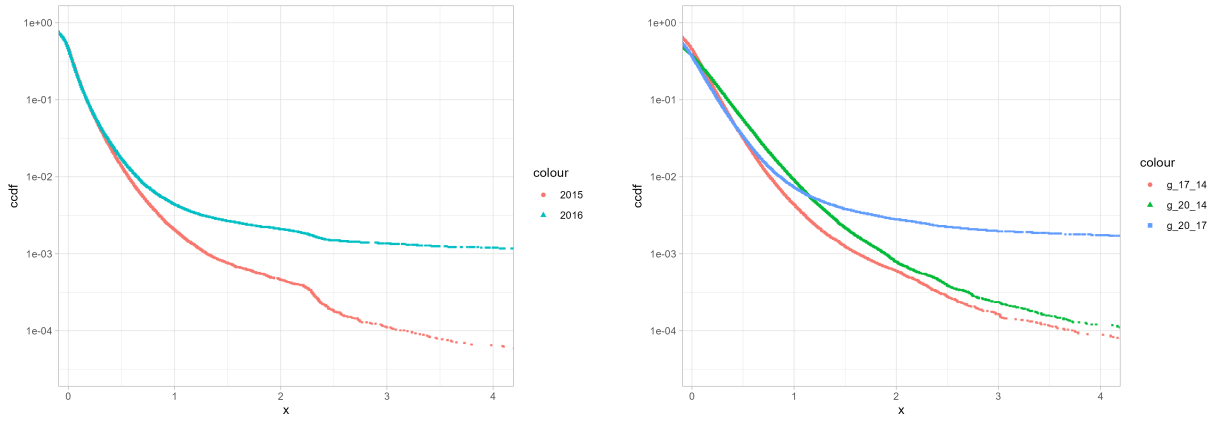
Next, to further verify that the growth rate distribution has tails heavier than an exponential, we use QQ-plots. The main idea behind QQ-plots is that if a random variable X_k follows a distribution F, then the transformed random variable $F(X_k)$ should follow a uniform distribution (for more details, see Section

[12] **Figure 5** and **Figure 6** show differences in the shapes of the growth rate distribution by year (e.g., between X_{15} and X_{16}), which can be attributed to differences in their sample. Specifically, our analysis considers only firms with sales of over 100 million yen in *2014*, which applies equally to both X_{15} or X_{16}. In other words, when considering X_{16}, we are not imposing a condition that firms have sales over 100 million yen in *2015*, thus including smaller firms that fall below this threshold in 2015 in our sample. In our analysis, especially as in Section 2.3, we focus on firms selected based on their 2014 sales and then examine their subsequent growth paths, thus employing this sample selection.



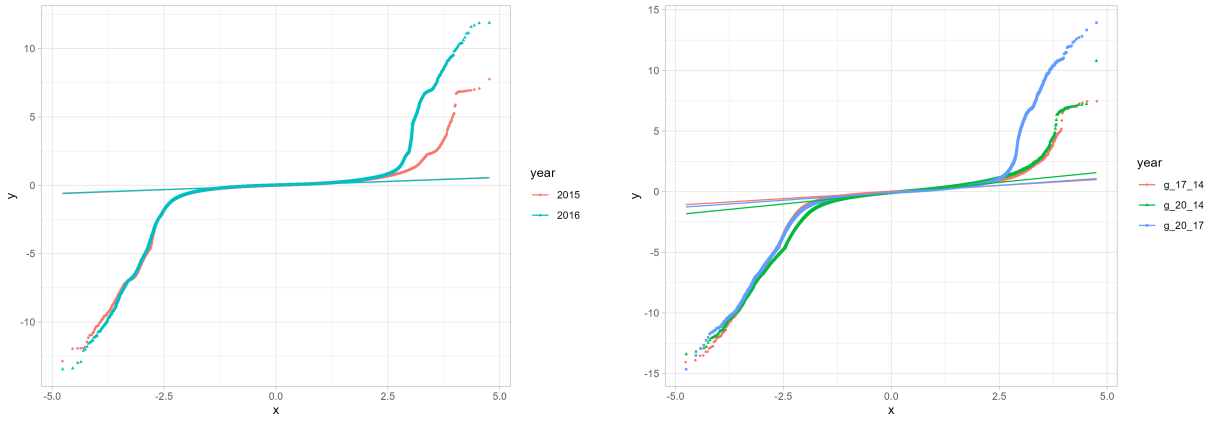
(a) One-year growth rates **(b)** Three-, six-year growth rates

Figure 5: Density estimates of growth rates. In both panels, the y-axis is on a logarithmic scale. Panel (a) provides the density estimates for the one-year growth rates, X_{15} and X_{16}. Panel (b) provides the density estimates for the three-year growth rates, X_{17-14} and X_{20-17}, as well as for the six-years growth rate X_{20-14}.



(a) One-year growth rates **(b)** Three-, six-year growth rates

Figure 6: CCDF of growth rates. Here, we focus solely on positive growth rates. Panel (a) provides the CCDF for the one-year growth rates, X_{15} and X_{16}. Panel (b) provides the CCDF for the three-year growth rates, X_{17-14} and X_{20-17}, as well as for the six-year growth rate X_{20-14}.



(a) One-year growth rates

(b) Three-, six-year growth rates

Figure 7: QQ-plot against the Gaussian distribution. Both x-axis and y-axis are normalized. Panel (a) provides the QQ-plot for the one-year growth rates, X_{15} and X_{16}. Panel (b) provides the QQ-plot for the three-year growth rates, X_{17-14} and X_{20-17}, as well as for the six-year growth rate X_{20-14}.

6.2.1 in Embrechts et al. (1997)). Thus, letting the ordered samples be $X_{n,n} \leq ... \leq X_{1,n}$, if X_k actually follows F, the points

$$\left(F(X_{k,n}), \frac{n-k+1}{n+1}\right), \ k = 1, ..., n, \quad \text{or} \quad \left(X_{k,n}, F^{-1}\left(\frac{n-k+1}{n+1}\right)\right), \ k = 1, ..., n$$

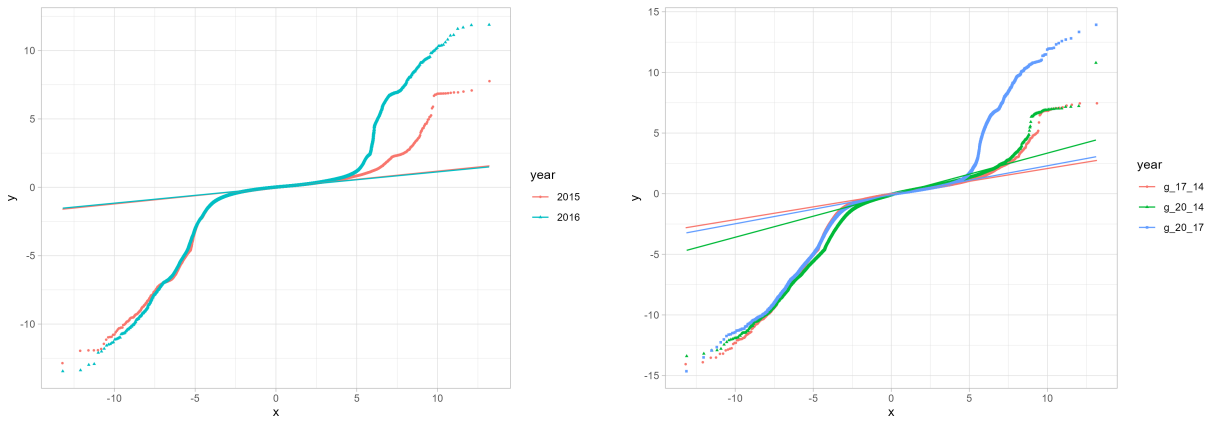
should plot on the 45-degree line (in particular, the latter is called a QQ-plot). Any distribution can be used as the reference distribution F depending on the hypothesis being tested. In the following analysis, we consider Gaussian and Laplace distributions as the reference distribution and verify whether they plot on the 45-degree line.

The results of the QQ-plots are presented in **Figure 7** for the Gaussian distribution and **Figure 8** for the Laplace distribution as the reference distribution. It is clear from **Figure 7** that the QQ-plot does not align with the straight line. Notably, the plot curves upwards in the right tail and downwards in the left tail, indicating that both tails are heavier than those of a Gaussian distribution. Similarly, in **Figure 8**, with the reference distribution being the Laplace distribution, the QQ-plot does not follow the straight line, curving upwards in the right tail and downwards in the left tail. This indicates that both tails are heavier than those of the Laplace distribution. These results, aligning with the results from the density estimates and CCDF, serve as evidence that the growth rate distribution is subexponential.

Finally, we use the mean excess function as a method to characterize the heaviness of the tail of the growth rate distribution. This function represents the expected value of the excess $X_k - u$ given that the growth rate exceeds u. It is defined as a function of u as follows:

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

We investigate how $e(u)$ changes as u increases. The advantage of $e(u)$ lies in the fact that the rate at which $e(u)$ increases or decreases with u reflects the tail heaviness of the distribution of growth rates X_k. For



(a) One-year growth rates

(b) Three-, six-year growth rates

Figure 8: QQ-plot against the Laplace distribution. Both x-axis and y-axis are normalized. Panel (a) provides the QQ-plot for the one-year growth rates, X_{15} and X_{16}. Panel (b) provides the QQ-plot for the three-year growth rates, X_{17-14} and X_{20-17}, as well as for the six-year growth rate X_{20-14}.

instance, if X_k follows an exponential distribution with parameter λ, then $e(u) = \lambda^{-1}$; that is, $e(u)$ remains constant. If $e(u)$ is an increasing (or decreasing) function of u, it implies that the distribution of X_k has a tail that is heavier (or lighter) than that of an exponential function. Hence, when $e(u)$ is an increasing function of u, it serves as further evidence that the growth rate distribution is subexponential.

Figure 9 provides estimates of the mean excess function $e(u)$ for both one-year and three-year growth rates. From this figure, it is clear that $e(u)$ is not constant but rather an increasing function with respect to u, though the rate of increase is not constant, particularly for one-year growth rates. This indicates that the growth rate distribution possesses a tail that is heavier than an exponential. These findings are consistent with results from density estimation, QQ-plot, and Cox-Oakes test, all suggesting that the growth rate distribution is subexponential.

Note that the shapes of the mean excess function $e(u)$ for one-, three-, and six-year growth rates are remarkably similar. This similarity is not a coincidence but rather a consequence of the distributions being subexponential. As seen in Eq.(3), even as n increases, the tail probability of the sum $\sum_{k=1}^{n} X_k$ is determined by the tail probability of the one-year growth rate $\mathbb{P}(X_k > x)$, except for a multiplier n. Given the definition of the mean excess function, the effect of the multiplier n is removed when considering conditional expectations, hence the mean excess function remains the same regardless of the value of n for sufficiently large u. Therefore, the resemblance of $e(u)$ for one-, three-, and six-year growth rates in **Figure 9** serves as evidence supporting that the growth rate distribution is subexponential.

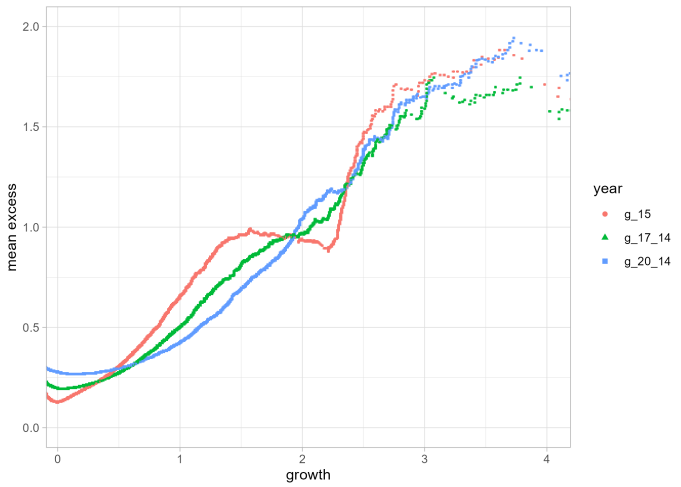


Figure 9: Mean excess function over threshold u. The x-axis represents the threshold u. It shows the mean excess function for the one-year growth rate X_{15}, the three-year growth rate X_{17-14}, and the six-year growth rate X_{20-14}.

3.4 Sample path properties

The empirical results presented in Sections 3.2 and 3.3 indicate that the two assumptions in our analysis, namely the random walk assumption and the subexponentiality of the growth rate distribution, are empirically valid. Therefore, in light of the discussion in Section 2, it implies that the growth dynamics of HGFs are characterized by jumps. Here, we provide direct evidence to support the significance of this jump-type process.

To this end, considering the two-year growth rates (i.e., $X_{15} + X_{16}$) and six-year growth rates (i.e., $X_{20-14} = X_{17-14} + X_{20-17}$), we examine how the growth rate of the first half of the period contributes to the growth rate over the entire period. Specifically, we explore the ratios defined as follows:

$$r_1 := \frac{X_{15}}{X_{15} + X_{16}}, \quad r_3 := \frac{X_{17-14}}{X_{17-14} + X_{20-17}}$$

The ratio r_1 represents how much the growth rate of 2015, X_{15}, contributes to the two-year growth rate $X_{15} + X_{16}$. Similarly, the ratio r_3 represents how much the growth rate of the first three years, X_{17-14}, contributes to the six-year growth rate X_{20-14}. For instance, if a firm's growth rates are 3% in 2015 and 3% in 2016, then r_1 equals $1/2$, indicating that the growth rates of both 2015 and 2016 contribute equally to the growth rate over the two years. Note that, since X_k's are assumed to be iid random variables in our analysis, the distributions of r_1 and r_3 should be symmetric around $1/2$. The question to be addressed here is whether the event of $r_1 = 1/2$ (or $r_3 = 1/2$), where the contributions of the growth rates in the first half and the second half of the period are equal, is the most likely to occur.

Figure 10 presents the histogram of r_1 conditional on the event $X_{15} + X_{16} > u$. This means that we focus on firms whose two-year growth rate exceeds u and examine how the histogram of r_1 changes as u increases. Here, u is varied from 0.2 to 2.4. These figures show that when $X_{15} + X_{16}$ is relatively small (e.g.,

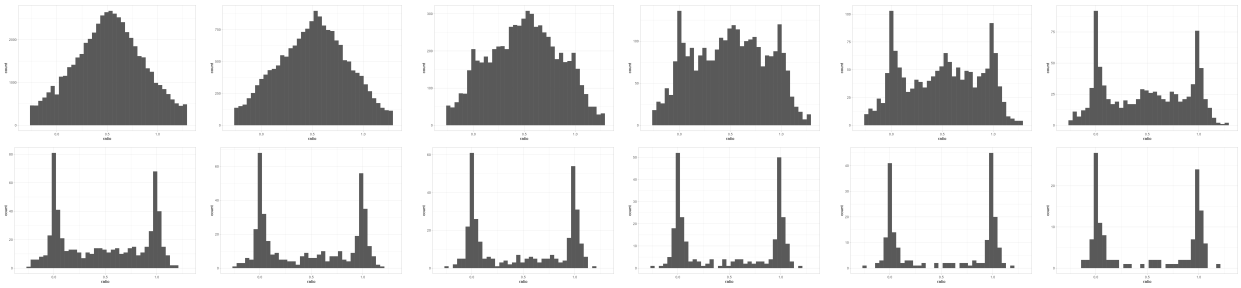


Figure 10: A series of the histograms of r_1 conditional on $X_{15}+X_{16} > u$. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2. We exclude samples where r_1 is exactly equal to 0 or 1. This is because some firms report that the values of their current sales are exactly the same as the previous ones. Even without these samples exactly equal to 0 or 1, spikes at 0 and 1 are still clearly observed.

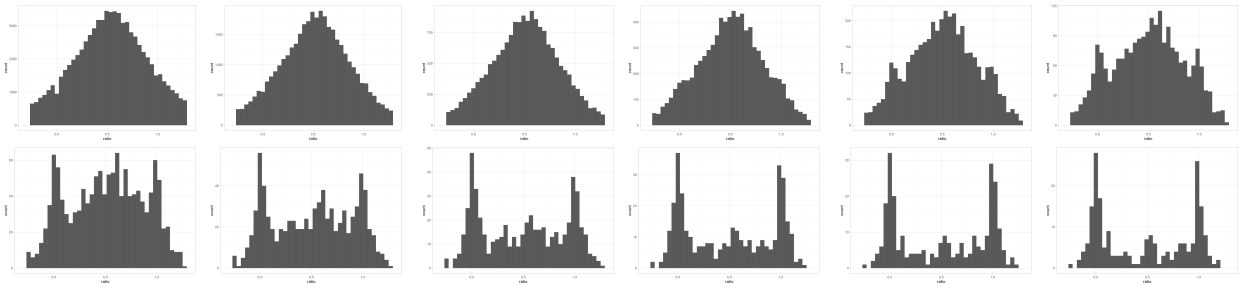


Figure 11: A series of the histogram of r_3 conditional on $X_{20-14} > u$. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

$u = 0.2$), the histogram of r_1 exhibits a mountain shape with a peak at $1/2$. That is, when the growth rate over the two years is not high, it is more likely that the growth rates from both periods contribute equally to the growth rate over the two years. In contrast, as u increases (e.g., $u = 1.2$), the mountain shape collapses. Instead, the histogram exhibits a U-shaped curve with peaks at 0 and 1, which implies that high growth over the two years is caused by a large value of either X_{15} or X_{16}, but not both. Thus, when considering HGFs over the two years, it is more likely that a HGF experiences extremely high growth during one of the periods, which dominates the high growth over the entire two years.

A similar U-shaped curve is observed for the histogram of r_3. **Figure 11** gives the histograms of r_3 conditional on the event $X_{20-14} > u$, where u varies from 0.2 to 2.4. As in the case of r_1, when u is relatively small, the histogram peaks at $1/2$, indicating the contributions from the first and second halves of the entire period are approximately equal. However, as u increases, the U-shaped curve with peaks at 0 and 1 becomes clear. This means that high growth over the entire period (i.e., a large value of X_{20-14}) is explained by high growth in either the first half or the second half of the entire 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 in the histograms of r_1 and r_3 provide direct empirical evidence supporting our 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. Given that this U-shape curve captures the essence of firm growth dynamics for HGFs, we refer to it as the U-shaped law of HGFs.

4 Conclusion

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

Our analysis is based on two empirically testable assumptions: the random walk assumption and the subexponentiality of the growth rate distribution. Using comprehensive data from corporate tax records in Japan, we confirmed that these two assumptions are empirically valid. In particular, the empirical fact that the growth rate distribution has a heavier tail than an exponential has significant implications for firm growth dynamics. We show that the sample path of HGFs is characterized not by a gradual increase in firm size but by jumps. The U-shaped curve of the histogram of r, which represents the contribution of the growth rate in the first half to the entire period, serves as direct evidence of this characteristic of firm growth dynamics.

In our analysis, we do not specify any economic models for firm growth but derive our implications solely from statistical regularities, such as the subexponentiality of the growth rate distribution. This approach is appealing, especially when the growth process is too complex to be explained by an explicit model, and only its probabilistic features are available. Due to this complexity, the firm growth dynamics are governed by the logic of probability theory. Our findings suggest 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	133742	−0.032	0.560	−0.094	0.017	0.130
2015-2016	128911	−0.035	0.570	−0.100	0.006	0.110
three-year growth rate						
2014-2017	124179	−0.058	0.810	−0.180	0.029	0.250
2017-2020	114532	−0.140	0.840	−0.300	−0.052	0.150
six-year growth rate						
2014-2020	111948	−0.160	1.020	−0.390	−0.029	0.300

Table 2: Summary statistics of growth rates for the young firms. Only firms with age less than ten years as of 2014 are considered. One-, three-, and six-year growth rates are considered. Here, only firms that are not in the financial and government sectors and have sales of more than 100 million yen in 2014 are included.

5 Appendix

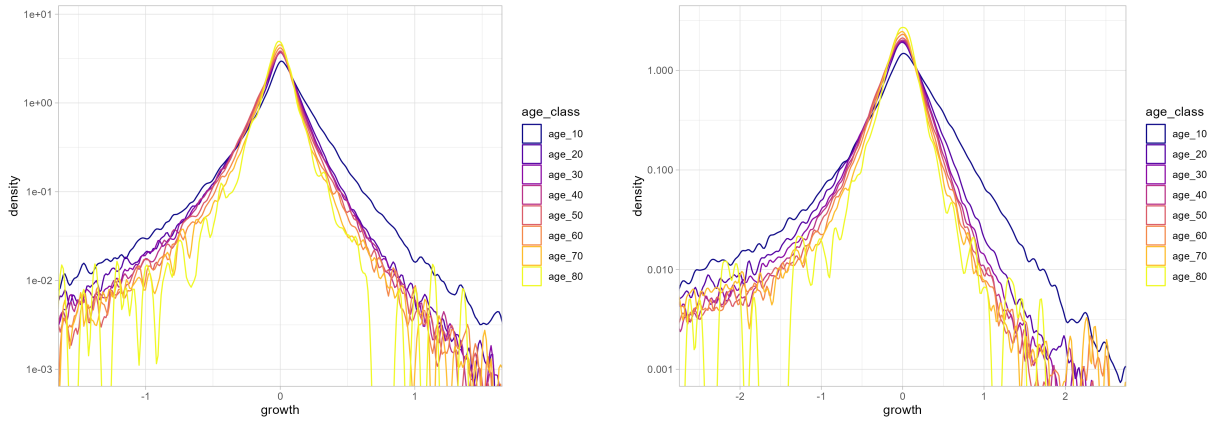
Here, we consider the growth dynamics of young firms established less than ten years before 2014, which are excluded from the analysis in Section 3.[13] We demonstrate that the characteristics of firm growth dynamics observed in Section 3, especially the U-shaped law, do not apply to this group. The summary statistics of the growth rates for these young firms are given in **Table 2**.

Compared to **Table 1**, **Table 2** exhibits a higher dispersion of growth rates among the young firms (e.g., the difference between Q1 and Q3). This high dispersion is also evident in **Figure 12**, where the samples are divided into age groups at ten-year intervals and the densities of growth rates are estimated for each age group. As shown in **Figure 12**, the density for the young firms deviates from those of other age groups.[14] These figures suggest that the impact of firm age on growth rates is significant for the young firms but weaker for other age groups.

Moreover, the shape of the growth rate distribution for the young firms appears to differ from those analyzed in Section 3. **Figure 13** presents the CCDF, QQ-plot (with Laplace distribution as the reference), and the mean excess function of six-year growth rates for the young firms. As discussed in Section 2.2, if the annual growth rates follow a subexponential distribution and the iid assumption holds, then the distribution

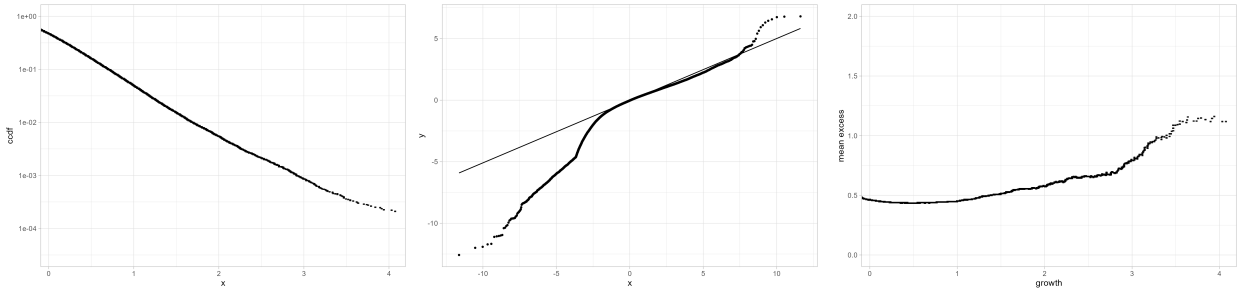
[13] In the following, firms that were established less than ten years before 2014 are referred to as *young firms*.

[14] One might think that the high dispersion of growth rates in the young firms is because these firms are smaller and hence exhibit higher dispersion in growth rates. To address this concern, we consider only firms with sales between 10^8 yen and 10^9 yen and analyze their density function. The results are quite similar to those shown in **Figure 12**, indicating that even when size is controlled for, the density function for the young firms deviates from that of other age groups.



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

Figure 12: 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_{17-14}) are considered.



(a) CCDF **(b)** QQ-plot **(c)** Mean Excess Function

Figure 13: Growth rates for the young firms. The CCDF, QQ-plot, and the mean excess function of six-year growth rates are considered. Panel (b) uses the Laplace distribution as the reference distribution.

of six-year growth rates should also have a heavier tail than an exponential; however, **Figure 13** shows that the right tail of the distribution of six-year growth rates are rather close to an exponential. Thus, the assumption of subexponential growth rates is less likely to hold for the young firms. It is expected that the U-shape curve for the histograms of ratios r_1 and r_3 would not be observed for these young firms.[15]

Figure 14 shows histograms of the r_1 ratio for the young firms, which corresponds to **Figure 10**. It displays how the histogram of r_1 changes as the value of u in the condition $X_{15} + X_{16} > u$ is varied from 0.2 to 2.4. Unlike **Figure 10**, for these young firms, a U-shaped curve in the histograms of r_1 cannot be

[15] We also find the weak but positive autocorrelation of growth rates for the young firms. Kendall's τ and Spearman's ρ are 0.081 and 0.105, respectively. 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 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)), which aligns with our findings.

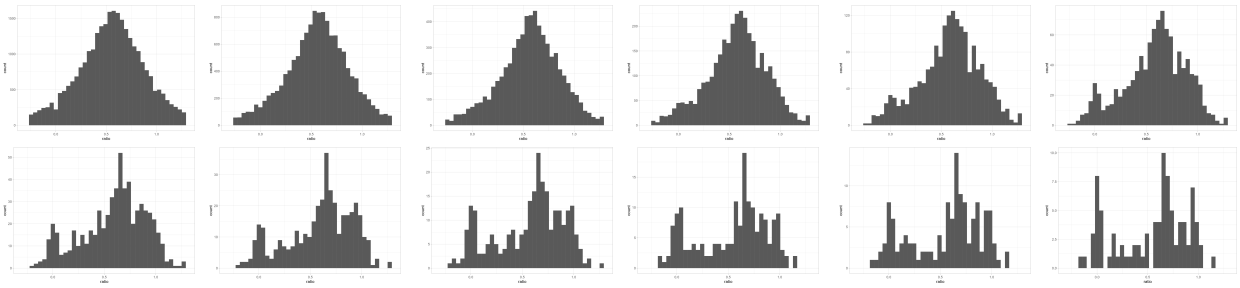


Figure 14: A series of the histograms of r_1 conditional on $X_{15} + X_{16} > u$ for the young firms. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

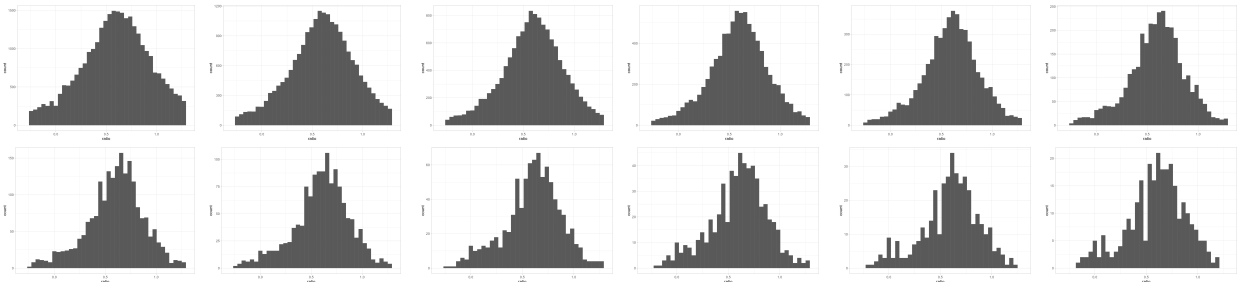


Figure 15: A series of the histogram of r_3 conditional on $X_{20-14} > u$ for the young firms. The value of u increases from 0.2 (top-left) to 2.4 (bottom-right) by 0.2.

observed. Even with large values of u, peaks at 0 and 1 do not appear. In **Figure 15**, r_3 is considered for these young firms, but similar to the case with r_1, a U-shaped curve is not observed. From these results, we conclude that the U-shaped law does not apply to these young firms. These results suggest that the growth paths of the young firms differ from those of older firms and are more complex. Our two assumptions—the random walk assumption and the subexponential distribution of growth rates—seem insufficient to fully characterize the growth dynamics of young firms. This issue is worth further exploration, but it is beyond the scope of this paper.

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*. 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., Kang, T., and Tamagni, F. (2023). Persistence in firm growth: inference from conditional quantile transition matrices. *Small Business Economics*, 61(2):745–770.

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.

Capasso, M., Cefis, E., and Frenken, K. (2014). On the existence of persistently outperforming firms. *Industrial and Corporate Change*, 23(4):997–1036.

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., and Halvarsson, D. (2018). Bursting into life: firm growth and growth persistence by age. *Small Business Economics*, 50(1):55–75.

Coad, A., Daunfeldt, S.-O., and Halvarsson, D. (2022a). Amundsen versus Scott: Are growth paths related to firm performance? *Small Business Economics*, 59(2):593–610.

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. (2022b). 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 Elert, N. (2013). When is Gibrat's law a law? *Small Business Economics*, 41:133–147.

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.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223.

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.

Geroski, P. A. (2000). *Competence, Governance, and Entrepreneurship-Advances in Economic Strategy Research*. 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.

Haltiwanger, J., Jarmin, R. S., Kulick, R., and Miranda, J. (2017). High growth young firms: contribution to job, output, and productivity growth. In *Measuring entrepreneurial businesses: current knowledge and challenges*, pages 11–62. University of Chicago Press.

Hölzl, W. (2014). Persistence, survival, and growth: a closer look at 20 years of fast-growing firms in Austria. *Industrial and Corporate Change*, 23(1):199–231.

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.

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.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press.