Growth and Differences of Log-Normals

Abstract

The growth of natural, social, and economic phenomena including firms, cities, and pandemics is known to be heavy-tailed. Neither a simple explanation nor a well-fitting distributional form for these heavy-tailed growth phenomena is known. Here I show that an extension of the log-linear production function provides both a simple explanation and a *single* well-fitting and theoretically motivated distributional form for *all of them*. I discuss why these results arise as a consequence of the Central Limit Theorem and sketch dynamic models using this production function for the phenomena listed above, yielding remarkable fit between the predicted and observed data distributions. My results include: (i) predicting the distribution of firm cashflows; (ii) providing a well-behaved distribution for equity returns; (iii) sketching a model of increasing-returns-to-scale cities in which more than one city can rationally exist; (iv) proposing an extension to the classical Malthusian "birth-death" model; and (v) rationalizing a variety of observed growth distributions.

JEL classifications: C46, D20, E23, G30, R10
Keywords: Growth, distributions, log-Normal, firm size, city size, COVID-19, income.

Current draft: Thursday 12$^{\text{th}}$ September, 2024
Current status: R&R at ReStud

1 Introduction

Many phenomena arise as the difference between two forces. In the context of firms, consider the fundamental sources and uses equation of the firm: *income = sales - expenses*. In the context of cities, Duranton and Puga (2003) posit that "*The trade-off between agglomeration economies and urban costs [...] is widely accepted as the key explanation behind the existence of cities and provides some important implications for their population growth.*" And in the context of population dynamics (and pandemics), the foundational model remains the birth-death model of Malthus (1798), with population growth the difference between birth and death (or infected and recovered).

A second observation, established since the empirical work of Gibrat (1931) and the theoretical work of Roy (1950, 1951), is that *each* of the two correlated opposing forces is itself likely a *product* of many latent random forces and hence approximately log-Normally distributed. This is "Gibrat's Law," or the multiplicative Central Limit Theorem (CLT). Combining these two observations yields surprising results. I show that it implies a simple production function to use when modelling such difference phenomena, which then: (i) predicts an obscure yet simple and theoretically grounded statistical distribution for these phenomena, confirmed by the data; and (ii) predicts that the heavy-tailed *growth* distributions of such phenomena should follow the *same* distribution, also confirmed by the data.

The implications for growth distributions are notable. That growth rates are heavy-tailed is well-established in several domains, including firms: Ashton (1926), Fama (1965); cities: Eeckhout (2004), Gabaix and Ioannides (2004); and the COVID-19 pandemic: Beare and Toda (2020), Parag, Donnelly, and Zarebski (2022). It is also documented here for various data series in Figures 1–3, discussed later. While recent works such as Guvenen, McKay, and Ryan (2023) and Jaimovich, Terry, and Vincent (2023) aim to devise empirical approximations for such fat-tailed growth phenomena, we lack a first-principles economic explanation for their emergence or a clear statistical representation of said distributions. Here I aim to fill this gap.

The production function proposed here, the *difference-of-log-linears*, henceforth DLL, is a straightforward extension of the log-linear production function, itself a workhorse of neo-classical economics since at least the work of Cobb and Douglas (1928). The difference-of-log-linears models (net) production as the difference between two log-linear functions, one for each opposing force, such that

$$Y_{DLL} = Y_p - Y_n = \exp\left(\boldsymbol{x} \cdot \boldsymbol{\beta}_p\right) - \exp\left(\boldsymbol{x} \cdot \boldsymbol{\beta}_n\right)$$

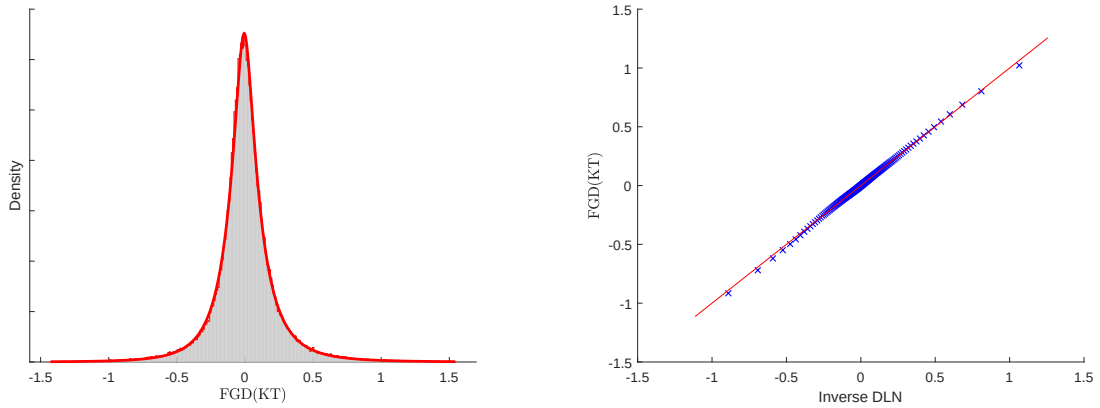
with $Y_{DLL} \in \mathbb{R}$ the net output, $\boldsymbol{x}$ a vector of the (logged) factors of production, and $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_n$ two elasticity vectors. I also show an alternative equivalent representation is

$$\begin{aligned} Y_{DLL} &= 2 \cdot \exp\left(\lambda\right) \cdot \sinh\left(\tau\right) \\ \lambda &= \boldsymbol{x} \cdot \frac{\beta_p + \beta_n}{2} = \log\left(\sqrt{Y_p \cdot Y_n}\right) \\ \tau &= \boldsymbol{x} \cdot \frac{\beta_p - \beta_n}{2} = \log\left(\sqrt{Y_p / Y_n}\right) \end{aligned}$$

which defines the *scale* $\lambda \in \mathbb{R}$ and the *efficiency* $\tau \in \mathbb{R}$ of production, and considerably simplifies the mathematical analysis of models based on this production function.

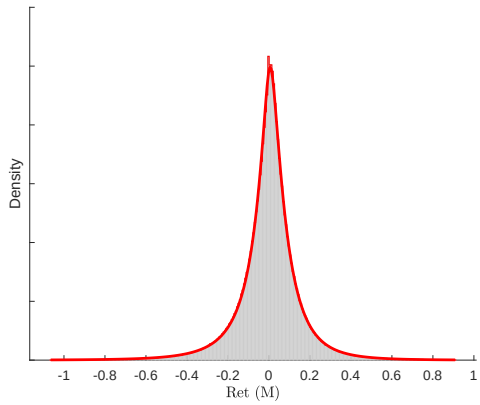
Furthermore, I show that under mild and widely-used assumptions on $\boldsymbol{x}$, the DLL production function implies that (net) production, as well as the growth rate of production, distribute as the difference-of-log-Normals, henceforth DLN. This simple distribution, which the Online Appendix fully characterizes for the first time, is almost completely unexplored.[1] At the time of writing, I was unable to find instances of using it anywhere in the sciences, and only two works tangentially considering it, Lo (2012) and Gulisashvili and Tankov (2016). Both works consider some statistical tail properties of sums of log-Normals, but state their results can apply to differences of log-Normals as well. Nevertheless, here I show the DLN characterizes a plethora of disparate phenomena remarkably well and propose it as an indispensable tool in the economist's toolbox along with its progenitor, the DLL.

[1] The *sum* of log-Normal RVs has been used in several scientific and economic disciplines including telecommunication, actuary, insurance, and derivative valuation.

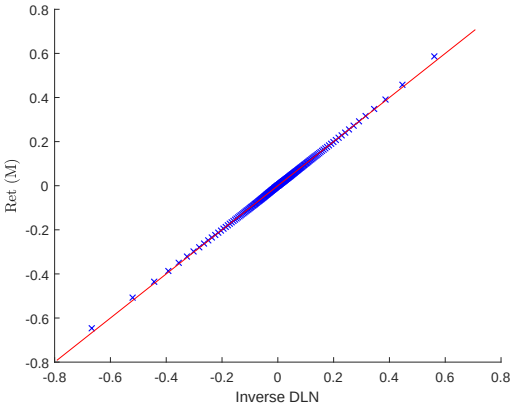


(a) Capital growth

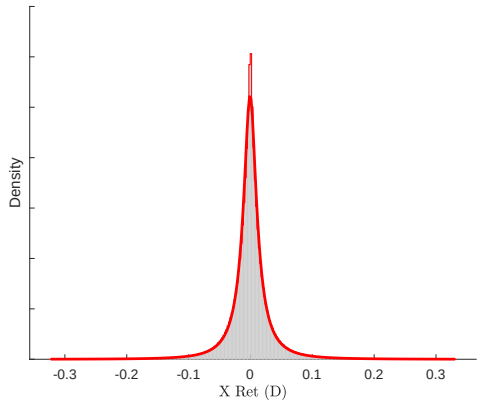
(b) q-q vs. DLN



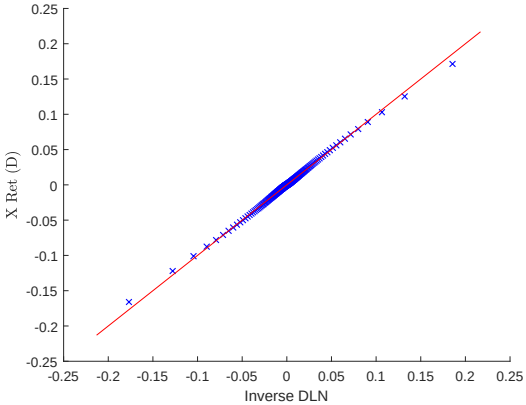
(c) Equity return (raw,M)



(d) q-q vs. DLN

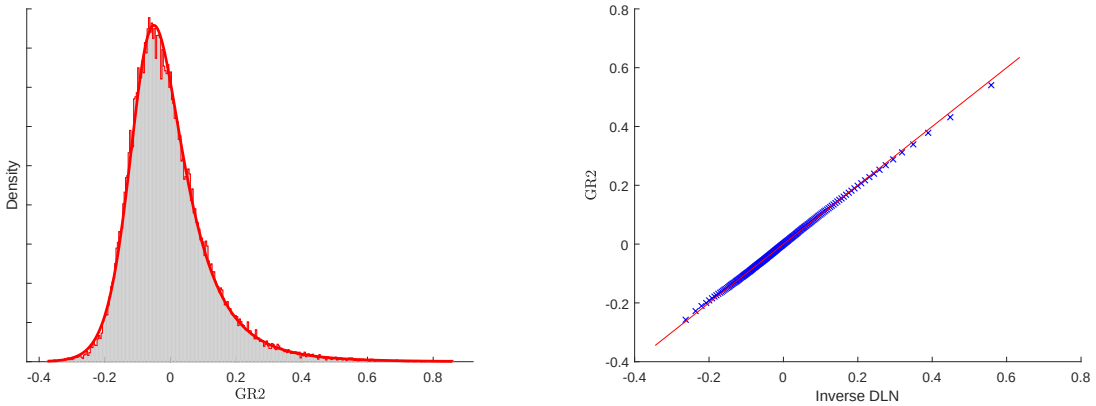


(e) Equity return (FF3,D)



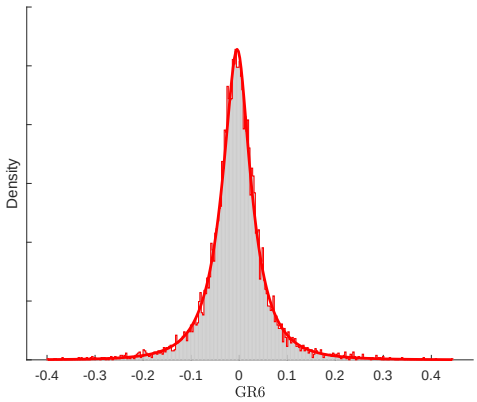
(f) q-q vs. DLN

Fig. 1. Firm growth distributions. Panel (a) presents the growth of capital (total assets) for a set of 143K firm-year observations from 1970-2019. Panels (c) and (e) present the monthly raw and daily excess equity returns for 2M and 5M observations, respectively. The panels include MLE-fitted DLN distributions. Panels (b),(d),(f) present the respective q-q plots vs. the DLN. Full data description is in Data Appendix A.

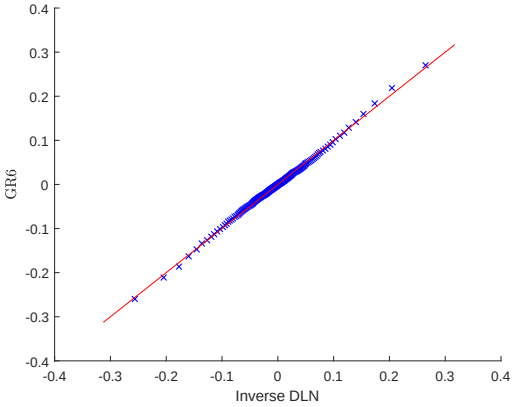


(a) Population growth

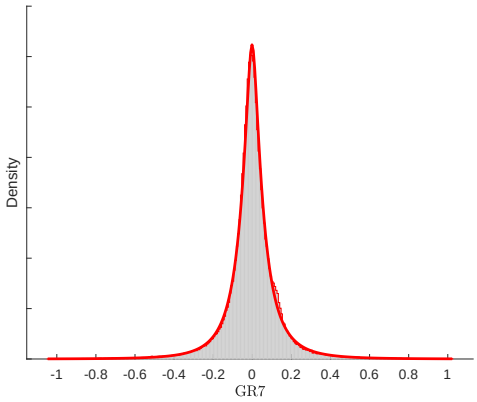
(b) Pop growth (q-q)



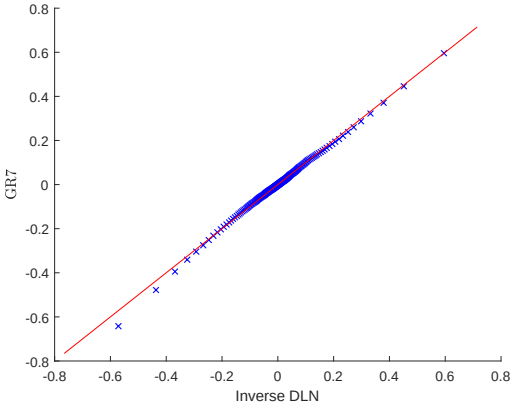
(c) County GDP growth



(d) GDP growth (q-q)

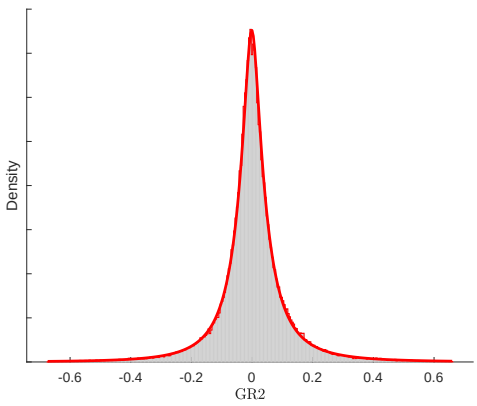


(e) MA-Ind GDP growth

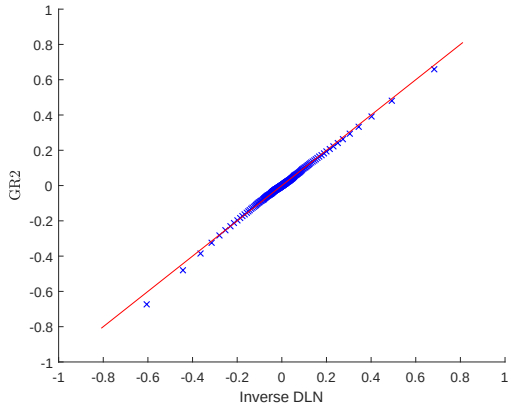


(f) GDP growth (q-q)

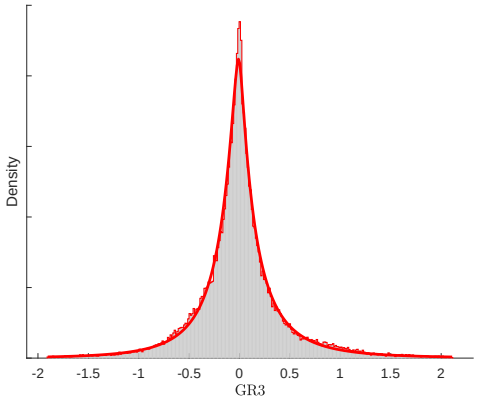
Fig. 2. City distributions stylized facts. Panels (a),(c),(e) present the distributions of population growth, GDP growth by county, and GDP growth by metropolitan area and industry, respectively. Panels (b),(d),(f) present the respective q-q plots. Data for Panel (a) are from Rozenfeld et al. (2011), and Data for panels (c),(e) are from the BEA. Full data description is in Data Appendix A.



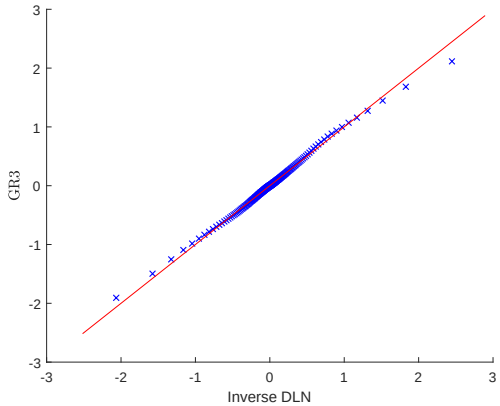
(a) Case growth



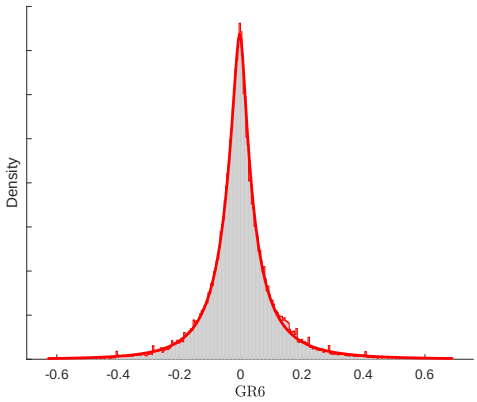
(b) Case growth (q-q)



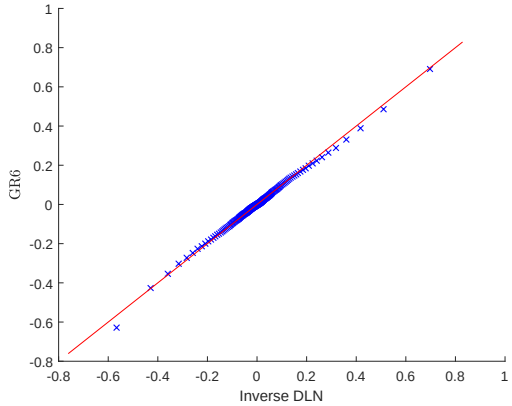
(c) Tests growth



(d) Tests growth (q-q)



(e) Vaccination growth



(f) Vacc. growth (q-q)

Fig. 3. Covid distributions stylized facts. Panels (a),(c),(e) present the distributions of daily new cases growth, new tests conducted growth, and new vaccinations growth, respectively, per country. Panels (d)-(f) present the respective q-q plots. Data are from Our World In Data Covid-19 depository. Full data description is in Data Appendix A.

2 The rise of DLL and DLN in economic data

2.1 The DLL production function

The canonical neo-classical log-linear (LL) production function can be written in general form as

$$Y_{LL} = \prod_{i=1}^{M} X_i^{\beta_i} = \exp\left(\boldsymbol{x} \cdot \boldsymbol{\beta}\right) \tag{1}$$

with $Y_{LL} > 0$ the output, $X_i > 0$ factors of production, $\beta_i \geq 0$ the corresponding elasticities, $\boldsymbol{x}$ a vector of the logged factors, $\boldsymbol{\beta}$ a vector of the elasticities, and $\boldsymbol{x} \cdot \boldsymbol{\beta}$ their inner product. For example, setting $\boldsymbol{x} = \{\log(A), \log(L), \log(K)\}$ and $\boldsymbol{\beta} = \{1, \alpha, 1-\alpha\}$ will yield the well known constant-returns-to-scale (CRS) Cobb-Douglas,

$$Y = \exp(a + \alpha \cdot l + (1-\alpha) \cdot k) = A \cdot L^{\alpha} \cdot K^{1-\alpha} \tag{2}$$

in which A is TFP, L is labor, K is capital, α is the labor share, and lower case letters denote logged values as usual.

The difference-of-log-linears (DLL) is defined, as the name implies, to be

$$Y_{DLL} = Y_p - Y_n = \exp\left(\boldsymbol{x} \cdot \boldsymbol{\beta}_p\right) - \exp\left(\boldsymbol{x} \cdot \boldsymbol{\beta}_n\right) \tag{3}$$

with $Y_{DLL} \in \mathbb{R}$ the net output and $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_n$ two elasticity (or weight) vectors. Note that because elasticities may be 0, this form is general enough to allow different factors to affect the positive and negative components. Further note that Y_{DLL} is *not* limited to be positive, and a useful feature of the DLL production function is its ability to yield negative net output values, e.g. firm losses.

A second useful feature of the DLL production function, making it amenable to mathematical manipulation, is that it can be factored into the multiplication of an exponent function and a Hyperbolic Sine (sinh) function — the hyperbolic equivalent of moving from Cartesian to Polar coordinates. We can hence alternatively write the DLL as

$$Y_{DLL} = 2 \cdot \exp\left(\lambda\right) \cdot \sinh\left(\tau\right) \tag{4}$$

with

$$\begin{aligned}\lambda &= \boldsymbol{x} \cdot \frac{\beta_p + \beta_n}{2} = \log\left(\sqrt{Y_p \cdot Y_n}\right) \\ \tau &= \boldsymbol{x} \cdot \frac{\beta_p - \beta_n}{2} = \log\left(\sqrt{Y_p / Y_n}\right)\end{aligned} \tag{5}$$

which defines the *scale* $\lambda \in \mathbb{R}$ and the *efficiency* $\tau \in \mathbb{R}$ of production. Both can be easily calculated if one separately observes Y_p and Y_n, as in e.g. the case of firm sales and expenses. Note that λ is the mid-point between $\log(Y_p)$ and $\log(Y_n)$, and τ is the (equal) distance from λ to the logs of Y_p and Y_n, with the appropriate sign. The inverse mapping is hence

$$\begin{aligned}Y_p &= \exp(\lambda + \tau) \\ Y_n &= \exp(\lambda - \tau)\end{aligned} \tag{6}$$

Clearly, the sign of (net) production depends on the sign of τ and its magnitude primarily depends on λ, with a smaller role for τ.

Section 4 below discusses the growth of Y_{DLL}, complicated by the fact it may take negative values. It shows that a generalized measure of growth for Y_{DLL} can be written as:

$$\frac{dY_t/dt}{|Y_t|} = \text{sgn}(\tau_t) \cdot \left[\frac{d\lambda_t}{dt} + \frac{d\tau_t}{dt} \cdot \frac{1}{\tanh(\tau_t)}\right] \approx \text{sgn}(\tau_t) \cdot \left[(\lambda_{t+1} - \lambda_t) + \frac{\tau_{t+1} - \tau_t}{\tau_t}\right] \tag{7}$$

with sgn() the sign function. The approximation is due to two reasons: using the forward discrete difference for the continuous time derivatives, and replacing $\tanh(\tau_t)$ with τ_t, which is a valid approximation when $|\tau_t| \ll 1$. The first term in the brackets is difference in *scales* (i.e. a usual difference in logs measure) and the second term is *percent* difference in efficiency. The second term then leads to the heavy tails of growth due to a "low base effect" in efficiency, when $|\tau_t|$ is close to zero.

Finally, while the elasticity of factor of production i in the LL production function is simply β_i, and is fixed regardless of the values of $\boldsymbol{x}$ or of the other elasticities in $\boldsymbol{\beta}$, the same

is not true for the DLL. It is easy to see that

$$\frac{\partial Y_{DLL}}{\partial X_i} / \frac{Y_{DLL}}{X_i} = \frac{\beta_{p,i} \cdot Y_p - \beta_{n,i} \cdot Y_n}{Y_{DLL}} = \frac{\beta_{p,i} \cdot Y_p - \beta_{n,i} \cdot Y_n}{Y_p - Y_n} \tag{8}$$

with $\beta_{p,i}, \beta_{n,i}$ the i^{th} elements of β_p, β_n. Put differently, the DLL production function yields elasticities that depend on the current levels of production and on their breakdown, with important implication to e.g. returns-to-scale in capital at the firm level.

2.2 The DLN as a consequence of the CLT

The DLN arises due to a straightforward set of statistical facts: (i) both the sum and difference of two Normal RVs are Normal; (ii) the sum of two log-Normal RVs is approximately log-Normal; but (iii) the difference of two log-Normal RVs is decidedly not log-Normal. For one, the log-Normal is strictly positive while the DLN is supported on the entire real line $\mathbb{R}$. The DLN also exhibits exponential (i.e., heavy) tails in both the positive and negative directions, inherited from the log-Normals composing it, quite different from the Normal Gaussian.

The DLN is simple to describe. To define the DLN, consider an RV W such that

$$W = Y_p - Y_n = \exp(X_p) - \exp(X_n) \quad \text{with} \quad \boldsymbol{X} = (X_p, X_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{9}$$

in which $\boldsymbol{X}$ is a bi-variate Normal with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_p \\ \mu_n \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_p^2 & \sigma_p \cdot \sigma_n \cdot \rho_{pn} \\ \sigma_p \cdot \sigma_n \cdot \rho_{pn} & \sigma_n^2 \end{bmatrix} \tag{10}$$

and hence $\boldsymbol{Y}$ is a bi-variate log-Normal RV. We say that W follows the five-parameter DLN distribution and denote $W \sim \text{DLN}(\mu_p, \sigma_p, \mu_n, \sigma_n, \rho_{pn})$.

Recall that the additive central limit theorems (CLTs) state that

$$Y^{+} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} X_i^{+} \sim \mathcal{N} \tag{11}$$

for $X_i^{+} \sim \Omega_i^{+}$ under mild regularity conditions on the Ω_i^{+} depending on the version of the CLT used. Put differently, the additive CLTs state that a phenomenon in nature which is an additive combination of many latent random forces will tend to distribute Normally. This is specifically the case for AR(1) processes in levels, whose ergodic distribution is Normal, under mild conditions, and is true regardless of the statistical distribution of the AR(1) disturbances which are *not* required to be Normal — see e.g. Hamilton (1994) Ch. 7.

Consider next the multiplicative CLT, sometimes known as "Gibrat's law" following Gibrat (1931), which states that

$$Y^{*} = \lim_{M \to \infty} \left(\prod_{i=1}^{M} X_i^{*} \right)^{\frac{1}{M}} \sim \text{log-}\mathcal{N} \tag{12}$$

for $X_i^{*} > 0 \sim \Omega_i^{*}$ under similarly mild regularity conditions. Put differently, the multiplicative CLT states that a phenomenon in nature which is a *product* of many latent random forces will tend to distribute log-Normally. An early discussion of the implications and reasoning for assuming multiplicative impact is the seminal work of Roy (1950, 1951). This result specifically holds for AR(1) processes in logs, whose ergodic distribution is log-Normal. Many physical and economic non-negative quantities, such as mass, population count, epidemic spread, interest rates, firm sales, firm value, and individual income have been modeled as the products of latent random factors (e.g., random growth models) and are approximately log-Normally distributed.

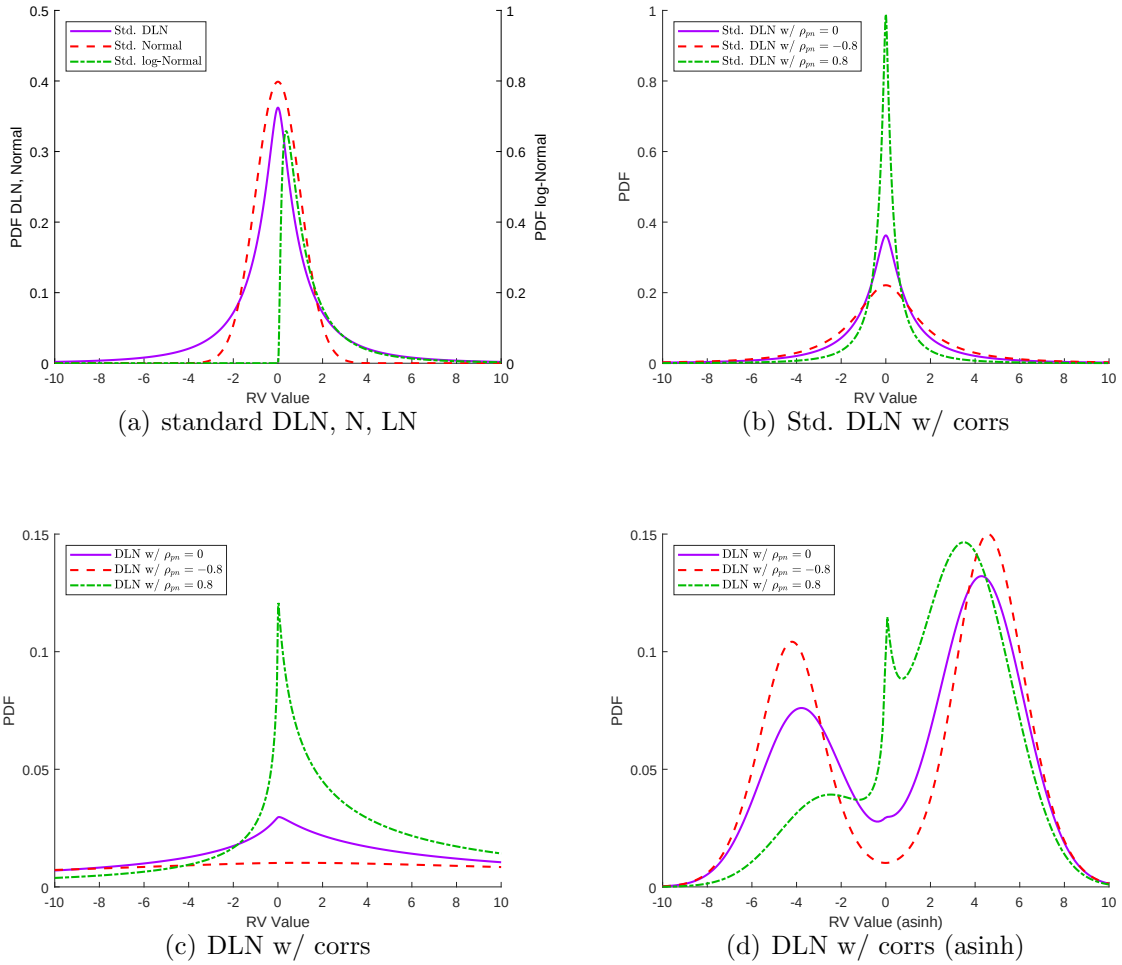
Finally, consider a natural phenomenon impacted by two main forces operating in opposite directions, i.e., $W = Y_p - Y_n$. If the two main forces are additive combinations of many latent random forces, then the natural phenomenon will tend to distribute Normally as well.

In this case, the importance of modelling the forces separately is diminished because aggregating them yields a model with similar distributional predictions. The same is not true, however, if the two forces are multiplicative combinations. In this case, failing to explicitly model both forces will yield markedly different predictions, because the difference between two log-Normal RVs does not collapse to a log-Normal RV.

It is further easy to see that the DLN is closed under multiplication or division by a log-Normal RV. This property is inherited from the closure of the Normal distribution to addition and subtraction, which in turn guarantees the closure of the log-Normal distribution to multiplication and division, and when combined with the distributive property yields the closure of DLN. This is true even if the log-Normal RV in question is correlated with Y_p, Y_n.

Figure 4 presents several instances of the DLN distribution. Panel (a) presents and contrasts the standard Normal, standard DLN, and standard log-Normal. The standard DLN is defined as DLN(0,1,0,1,0), i.e. the difference between two exponentiated uncorrelated standard Normal RVs. Panel (b) shows the role of the correlation coefficient ρ_{pn}, controlling tail-weight vs. peakedness. Panel (c) repeats the analysis of Panel (b) for a different parametrization common in practical applications, exhibiting the problem of dealing with the DLN's characteristic heavy tails in both the positive and negative directions. Panel (d) then presents the data of panel (c) but the X-axis is transformed using the Inverse Hyperbolic Sine (asinh). While previous work (Bellemare and Wichman, 2020; Aihounton and Henningsen, 2021; Mullahy and Norton, 2022) discusses the asinh transform as an ad-hoc solution, its use here is theoretically motivated by Equation 4. It acts as a log transform in both the positive and negative directions, and allows us to observe the characteristic "double Normal" shape of the (asinh-transformed) DLN.

To clarify the use of the DLL production function and the DLN distribution arising from it, the next section considers the DLL in the context of three examples, beginning with the most straightforward one: firms.



(a) standard DLN, N, LN

(b) Std. DLN w/ corrs

(c) DLN w/ corrs

(d) DLN w/ corrs (asinh)

Fig. 4. DLN Examples. Panel (a) graphs the PDFs of the standard Normal, log-Normal, and DLN. Panel (b) graphs the PDFs of standard DLN with different correlation coefficients ρ_{pn}. Panel (c) presents the PDFs of a DLN with parameters $(3, 2, 2, 2)$, common in practice, and varying correlation coefficients ρ_{pn}. Panel (c) presents the PDF for the range ± 10, which is a significant truncation due to the long tails of this DLN. Panel (d) presents the same PDFs as Panel (c), but the x-axis is asinh-transformed, such that it spans the range sinh(-10) $\approx$ -11,000 to sinh(10) $\approx$ 11,000.

3 Examples of using the DLL and DLN

3.1 Example 1: Firms

What is the statistical distribution of firm income? Firm income, often called cashflows, is of utmost importance in both major branches of financial economics research: corporate finance and asset pricing. Cashflows are both the means to growth — providing money for investments, and the ends of growth — providing money for dispensations (e.g., dividends). It is hence quite surprising that the statistical distribution of income has seen such scant interest in the finance literature, and is hitherto unspecified.

In typical neo-classical "q-theory" models income is modeled using an LL production function:

$$Y_{LL} = Z \cdot K^{\theta_z} = \exp\left(z + \theta_z \cdot k\right) \tag{13}$$

with $Z = \exp(z) > 0$ firm productivity, $K = \exp(k) > 0$ firm capital, and $0 < \theta_z < 1$ returns-to-scale coefficient. The q-theory models generally abstract from labor, assuming it is elastically adjustable within period. Wages and other expenses are already accounted for, as Y_{LL} directly models income, or sales minus expenses. It is also ubiquitously assumed that firm (log) productivity z_t follows an AR(1) process, i.e. that Z_t follows an AR(1) in logs. The main issue with this modeling choice is that it counter-factually yields firms with strictly positive income, distributing log-Normally due to the AR(1) properties discussed above. Furthermore, growth is counter-factually Normally distributed. The lack of negative income in such models ignores a critical feature of the profit-and-loss mechanism of firm dynamics — namely, losses.

If, instead of modeling income directly, we were to model sales and expenses *separately*, and model income as their *difference*, the DLL production function naturally arises,

$$Y_{DLL} = \underbrace{\exp\left(s + \theta_s \cdot k\right)}_{Sales \equiv \mathbb{S}} - \underbrace{\exp\left(x + \theta_x \cdot k\right)}_{Expenses \equiv \mathbb{X}} = 2 \cdot \exp\left(\lambda\right) \cdot \sinh\left(\tau\right) \tag{14}$$

with $0 < \theta_x, \theta_s < 1$ returns to scale parameters in sales and expenses, respectively. In a slight abuse of notation, firm sales are denoted $\mathbb{S}$ and firm expenses $\mathbb{X}$. The production function Y_{DLL} is now a function of three variables — the endogenous capital stock k and two stochastic exogenous variables, s and x, controlling the dynamics of sales and expenses and following a VAR(1) process. We can again represent the function in terms of firm income scale and efficiency, λ and τ, both of which are strictly observable due to the observability of sales and expenses, and are always defined, even if the firm suffers losses.

Before proceeding, it is worth contemplating the economic meaning of Z, S, X. What determines the factor-productivity of the firm, Z (or its sales and expenses productivity, S, X)? they are functions of the "skill, dexterity, and judgment with which labor is applied," as in Smith (1776), or of the firm's production technology, cost structure, managerial talent, market power, and a host of other components, including luck. In that sense, the stochastic variables are partly endogenous. Of course, all firms would prefer to produce as much income as possible from a given amount of capital K. Put differently, all firms would like to have as high a Z as possible, or equivalently as high an S and as low an X as possible. Firms hence optimize the components of Z, S, X under their control, and as a result, achieve (log) productivity μ_Z (or μ_S, μ_X) on average. But firms differ in their abilities, and the differences are persistent. Z_t, S_t, X_t hence represent the current productivities of the representative firm, given its optimizing behavior on their components. In this way, Z and S, X are the usual measures of our ignorance regarding the firm, as in Abramovitz (1956).

Importantly, the simple choice of modelling sales and expenses explicitly, described in Equation 14, *predicts* that firm income should distribute DLN. Figure 5 presents the relevant data distributions. Panels (a) and (d) of Figure 5 present truncated views of firm cashflows (the income of the firm) and firm dispensations (the income of firm owners, both equity and bond holders), respectively. Negative cashflows are losses, and negative dispensations are capital infusions from owners into the firm. Income clearly exhibits exponential tails in both the positive and negative directions, explaining the need for truncation. To deal with

the double-exponential nature of the tails, Panels (b) and (e) then present the same data, untruncated but with the x-axis transformed to asinh scale. Panels (a), (b), (d), (e) are overlaid with MLE-fitted DLN distributions, exhibiting excellent fit. The accompanying q-q plots in Panels (c) and (f) confirm this observation.

These two income measures are tightly related via the other side of the sources and uses equation of the firm, $income = investment + dispensations$, with firm investment itself an object of intense interest in the finance literature. Panel (g) presents the distribution of firm investment, defined as change in total capital net of depreciation between periods. The fit to the DLN is again remarkable. While it is common in the relevant literature to consider investment rates rather than magnitudes, dividing investment by total beginning-of-period capital, the conclusions remain unchanged. This is because capital is approximately log-Normally distributed, as Gibrat (1931) first showed, and the closure of the DLN to division by a log-Normal discussed above. Similar reasoning implies that the average product of capital (APK), defined as income divided by beginning-of-period capital, distributes DLN as well. Both facts are visually confirmed in Panels (h),(i).

Table 1 present formal goodness-of-fit tests versus the DLN, Stable, and Laplace distributions. The Stable and Laplace distributions both exhibit double-exponential tails and are the main distributions previously considered in the relevant literature on heavy-tailed firm growth.[2] The three goodness-of-fit tests I use are the Kolmogorov-Smirnov (K-S), Chi-square (C-2), and Anderson-Darling (A-D) tests, each of which being sensitive to different distributional deviations. Firm income, dispensations, investment, investment rate, and APK are all rejected as Stable or Laplace, but not as DLN. The DLN is a 5-parameter distribution, while the Stable and (asymmetric) Laplace have 4 and 3 parameters, respectively. I hence further present log-likelihood-based "horse races" between the distributions using the Akaike and Bayesian Information Criteria (AIC and BIC), which penalize for extra parameters. Throughout, the DLN handily beats the other distributions.

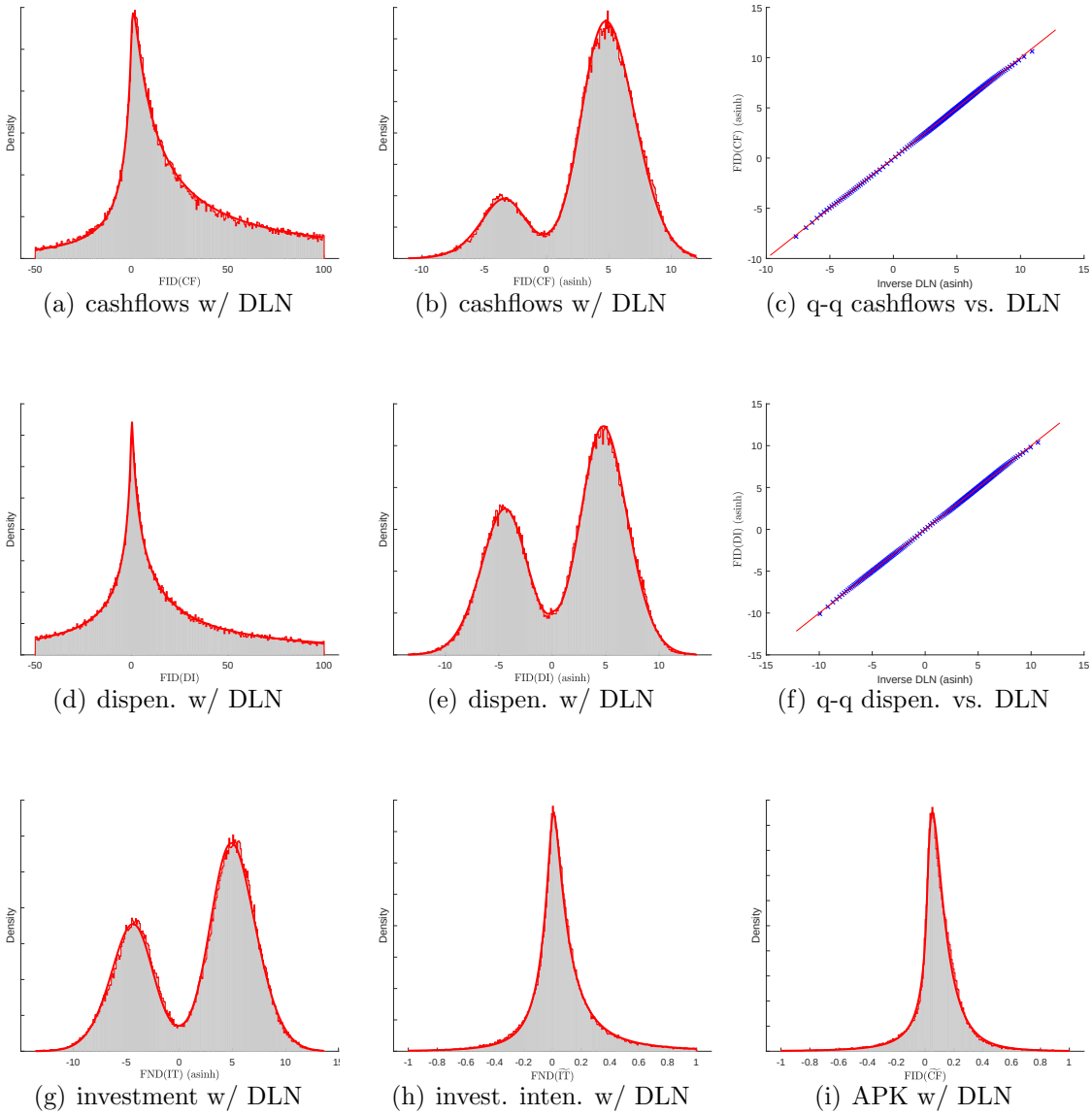
[2]See e.g. Mandelbrot (1961), Fama (1963), Fama and Roll (1971), Stanley, Amaral, Buldyrev, Havlin, Leschhorn, Maass, Salinger, and Stanley (1996), Bottazzi and Secchi (2003), Luttmer (2011).

It immediately follows that if firm investment rates distribute DLN then firm capital growth rates distribute DLN as well, as Panels (a),(b) of Figure 1 have already shown, and the appropriate column of Table 1 confirms. The intuitive explanation is that a value maximizing firm will increase its scale if its expected income efficiency is higher than its user-cost of capital, and vice-versa. While this is by no means a rigorous argument, in a companion paper Parham (2023) writes a standard q-theory model of firm dynamics, but using a DLL production function to separately model sales and expenses. It shows that in this case, firm income, income growth, capital growth, and value growth are all DLN-distributed. The core effect at play in creating heavy-tailed capital and value growth is the operational leverage of the firm, or the "low base effect" in τ discussed above, which can be captured by the following example.

Consider a firm with $\$1B$ in sales and $\$950M$ in expenses during period t (these values approximately correspond to the median firm in the data). If the firm increases both sales and expenses by 10%, income will grow by 10% as well, from $\$50M$ to $\$55M$. If, however, the firm increases sales by 10% but *decreases* expenses by 10%, income will grow by 390% to $\$245M$. With income exhibiting heavy-tailed growth, we would expect firm value to exhibit heavy-tailed growth as well, because firm value is simply the NPV of future income, and rapid growth in income should propagate to rapid growth in value.

A corollary of the results above is that equity returns should distribute DLN as well. Equity returns are merely another measure of firm growth, when one takes firm size to be defined as market value of equity, and adjusts for dispensations to equity holders. Panels (c)-(f) of Figure 1 present the fit to the DLN for two examples, monthly raw returns and daily excess returns relative to the Fama-French 3-factor model, and the appropriate columns of Table 1 confirm. In unreported results, the DLN is not rejected for yearly, monthly, or daily returns, both raw and excess (relative to the 3,4,5-factor Fama-French models). The fit of equity returns to the DLN is especially noteworthy given the voluminous literature on the determinants, fat-tails, and statistical properties of returns.

The importance of this result should be taken in context. The heavy tails of returns were studied by Mandelbrot (1960, 1961) and Fama (1963, 1965) who propose the family of Stable distributions (also known as Stable-Paretian or Pareto-Lévy) as a statistical model of returns. The Stable distribution remains a workhorse of empirical asset pricing work, despite a critical flaw — it lacks finite second (or higher order) moments — our ubiquitous measures of risk — thus providing a precarious basis for Modern Portfolio Theory (MPT). The Stable was later rejected by Officer (1972), who shows that the second moment of returns in the data is well-behaved and concludes that "It may be that a class of fat-tailed distributions with finite second moments will be found [...] but as yet this remains to be clearly demonstrated." The DLN has finite moments of all orders.



(a) cashflows w/ DLN (b) cashflows w/ DLN (c) q-q cashflows vs. DLN

(d) dispen. w/ DLN (e) dispen. w/ DLN (f) q-q dispen. vs. DLN

(g) investment w/ DLN (h) invest. inten. w/ DLN (i) APK w/ DLN

Fig. 5. Income - Stylized facts. Panels (a),(d) present the truncated distributions of income (cashflows, w/ negatives indicating losses) and dispensations (to equity and debt holders, w/ negatives indicating cash infusions). Panels (b),(e) present the untruncated distributions but with asinh-scaled X-axis, and Panels (c),(f) present the respective q-q plots vs. the DLN. Panels (g),(h),(i) present the distributions of: investment (increase in assets net of depreciation, w/ negatives indicating asset decreases, in asinh-scale), investment rate (investment divided by beginning-of-period capital), and APK (cashflows divided by beginning-of-period capital), respectively. All panels except (c),(f) are overlaid with ML-fitted DLN distributions. Full data description is in the appendix.

Table 1
Distributional tests — Firms

This table presents the results of tests of distributional form for firm income (CF), dispensations (DI), investment (IT), investment intensity ($\widetilde{\text{IT}}$=IT/L.KT), APK (=CF/L.KT), income growth (dCF), capital growth (dKT), yearly total value growth adjusted for dispensations (dVL), monthly raw equity returns (dEQ_M), and daily excess equity returns (dEQ_D^{ex}). K-S is a Kolmogorov–Smirnov test; C-2 is a binned χ^2 test with 50 bins; A-D is an Anderson-Darling test. Panels (a)-(c) report the test statistics and their p-values *rejecting the distribution* for the Stable, Laplace, and DLN, respectively. Panel (d) reports the relative likelihoods using the AIC and BIC. Full data description is in the appendix.

	CF	DI	IT	$\widetilde{\text{IT}}$	APK	dCF	dKT	dVL	dEQ_M	dEQ_D^{ex}
Panel (a): Stable										
K-S	0.034	0.028	0.026	0.020	0.038	0.014	0.011	0.015	0.015	0.015
p-val	0.022	0.027	0.028	0.036	0.019	0.046	0.055	0.044	0.045	0.043
C-2	647	379	380	193	295	120	91	179	195	204
p-val	0.012	0.018	0.018	0.027	0.021	0.035	0.040	0.028	0.027	0.027
A-D	19.48	19.99	13.75	12.25	19.32	5.86	3.79	6.50	6.67	6.70
p-val	0.025	0.025	0.029	0.031	0.025	0.041	0.047	0.039	0.039	0.039
Panel (b): Laplace										
K-S	0.364	0.288	0.279	0.079	0.098	0.073	0.027	0.017	0.026	0.027
p-val	0.000	0.000	0.000	0.004	0.000	0.006	0.027	0.041	0.028	0.027
C-2	>999	>999	>999	924	>999	>999	290	118	244	182
p-val	0.000	0.000	0.000	0.008	0.004	0.002	0.022	0.035	0.024	0.028
A-D	>999	>999	>999	93.14	123	105.75	18.72	7.75	19.38	17.40
p-val	0.000	0.000	0.000	0.008	0.006	0.007	0.025	0.037	0.025	0.026
Panel (c): DLN										
K-S	0.003	0.005	0.006	0.008	0.003	0.007	0.006	0.004	0.003	0.007
p-val	0.138	0.104	0.086	0.071	0.149	0.074	0.087	0.108	0.148	0.080
C-2	8	15	27	42	12	59	17	14	5	21
p-val	0.142	0.096	0.070	0.057	0.111	0.049	0.089	0.099	0.353	0.078
A-D	0.21	0.39	0.840	1.84	0.18	1.11	0.50	0.45	0.10	0.47
p-val	0.117	0.095	0.076	0.059	0.123	0.070	0.089	0.091	0.148	0.090
Panel (d): Relative likelihood tests										
AIC R.L.:										
Stable	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Laplace	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DLN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
BIC R.L.:										
Stable	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Laplace	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DLN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

3.2 Example 2: Cities

The size and growth distributions of cities is another field of intense economic study due in part to the out-sized economic value created by cities. Modelling the city's "production function" (or data-generating process) goes at least as far back as Henderson (1974). Later, Gabaix (1999b,a) discuss why the observed Zipf's law for city sizes is a natural results of Gibrat's Law and propose a random growth model with a production function in which amenity shocks increase the utility derived from consumption in a multiplicative way.

In an overview on the growth of cities, however, Duranton and Puga (2014) highlight the role of the interaction between agglomeration benefits and congestion costs in determining city growth. While Gabaix (1999b) accounts for congestion, he does so within the context of a LL production function. This modelling choice then yields the conclusion that agglomeration benefits must be asymptotically CRS (or else technological differences must be unbounded), contrary to intuition and evidence that they are in fact IRS — see e.g. West (2017). This section briefly sketches the structure of a DLL-based city model, in which amenities and congestion interact to yield DLN-shaped city growth while also allowing for IRS behavior.

That city growth distributes DLN is evident in Figure 2. Panel (a) presents data on population growth using the data on city populations from Rozenfeld et al. (2011), who use a clustering algorithm to define the boundaries of cities. Panel (c) and (e) present data on the growth of economic activity (i.e., local GDP) by county and by metropolitan area and industry from the US Bureau of Economic Analysis. The DLN arises again, as can be seen visually in the respective q-q plots in Panels (b),(d),(f). Formal tests vs. the Stable, Laplace, and DLN are provided in the first three columns of Table 2. As in the firms case, the Stable and Laplace are rejected. While the DLN is also rejected for the data in panels (e)+(f), it nevertheless still handily beats the other distributions in AIC- and BIC-based "horse-races" for all three data distributions.

This finding is a natural outcome of a simple model of city dynamics in which cities grow subject to the interplay between agglomeration benefits and congestion costs, both of which

exert exponential (i.e., multiplicative) influence on the flow of aggregate economic value created by cities. The economic surplus is captured by city inhabitants, firms operating in the city, the government, or in general the social planner. Both benefits and costs increase with the number of inhabitants, just as both sales and expenses increase with firm capital. But the interplay between them may give rise to positive surplus (i.e., $Y_{DLL} > 0$), leading to immigration into the city, or to negative surplus (i.e., $Y_{DLL} < 0$) leading to emigration out of the city, subject to adjustment costs. Importantly, this surplus is normalized to the alternative available in other cities, as usual, which here is set to 0 for simplicity.

A city in this model is hence indexed by (log) population n, amenity index a, and congestion index c. Both a and c are stochastic, i.e., exogenous to the model. Cities of course influence a and c with their policies, and would prefer a to be as high as possible and c to be as low as possible. They are however limited in their ability to do so, adjust both slowly (thus making them persistent), and their moves to jointly optimize a, c given their budget constraints are considered stochastic in the model, as is the case with the firm model. The surplus created by the city can then be written using the DLL production function as

$$Y_{DLL} = \underbrace{\exp\left(a + \theta_a \cdot n\right)}_{Amenities \equiv \mathbb{A}} - \underbrace{\exp\left(c + \theta_c \cdot n\right)}_{Congestion \equiv \mathbb{C}} \tag{15}$$

in which θ_a,θ_c are the returns to scale in amenities and congestion.

The economic surplus can then be used by the social planner to attract inhabitants from other, less prosperous cities, which is captured by an adjustment cost to be paid when increasing n, similar to the convex adjustment costs in the firm investment literature. Unlike in the firm case, however, it is likely that for cities $\theta_a, \theta_c > 1$, i.e. cities face IRS in both amenities and congestion: ceteris paribus, a doubling of the population will more than double the total benefits, but will also more than double the total costs. The model can accommodate this by requiring $\theta_c > \theta_a$ to avoid the degenerate result of a single, infinitely large city, hitherto preventing the writing of LL-based models with IRS.

In this model, a shock to a, c can reshuffle the economic surplus in cities and lead to mass migration. If all restaurants in NYC close down (decreasing its a) and the subway becomes unsafe (increasing its c) due to a pandemic, the surplus of NYC decreases rapidly. This will then lead to mass exodus. Conversely, the utility of the outdoors opportunities around Charlottesville, VA is higher (increasing its a) without a corresponding increase in congestion because public transit is not a common form of transportation (leaving c unchanged or only increasing it mildly). This then leads to increased demand for housing in Charlottesville (whose supply is quasi-fixed in the short term) and large property price increases.

3.3 Example 3: Populations and pandemics

Another way of illustrating the emergence of the DLN distribution is in the context of the simple population dynamics ("birth-death") model of Malthus (1798). The birth-death model remains the foundational model of population dynamics and is also the basis for more complex epidemiological models like the SIR (Susceptible-Infected-Removed) or SEIR (Susceptible-Exposed-Infected-Removed) models of pandemics.[3]

Denote $N(t)$ the size of the population in some closed natural habitat (with no immigration or emigration) at time t. The population dynamics of the system are described by the ordinary differential equation:

$$\frac{dN\left(t\right)}{dt} = b\left(t\right) \cdot N\left(t\right) - d\left(t\right) \cdot N\left(t\right) = \left[b\left(t\right) - d\left(t\right)\right] \cdot N\left(t\right) = r\left(t\right) \cdot N\left(t\right) \tag{16}$$

in which $b\left(t\right) \geq 0$ and $d\left(t\right) \geq 0$ are the instantaneous birth and death rates at time t. Both are stochastic and depend on underlying latent forces such as food availability, climate, predation, etc.

The difference between the birth and death rates $r\left(t\right) \in \mathbb{R}$ is the instantaneous growth rate of the population. Stochastic Malthusian models concentrate on this difference $r\left(t\right)$,

[3] See e.g. Voigtländer and Voth (2013), Lindenstrand and Svensson (2013), Chladná, Kopfová, Rachinskii, and Rouf (2020), Gatto and Schellhorn (2021), Fernández-Villaverde and Jones (2022).

and are often described using the stochastic differential equation:

$$dr\left(t\right)=\phi_r\cdot\left(\mu_r-r\left(t\right)\right)\cdot dt+\sigma_r\cdot dW_r\left(t\right) \tag{17}$$

in which $dW_r\left(t\right)$ is a Wiener process, $\phi_r>0$ the mean-reversion parameter, $\mu_r\in\mathbb{R}$ the long-term mean, and $\sigma_r\geq 0$ the volatility of the Wiener process. Put differently, the stochastic Malthusian model implicitly assumes the population growth rate is Normally distributed.

One may however ask: what is the distribution of the birth and death rates? If both distribute Normally, then their difference will distribute Normally as well and the stochastic model assumption is vindicated. From an information-theoretic perspective, assuming an RV is Normal is the least restrictive assumption because the Normal is the continuous distribution with the maximum entropy for a given mean and variance — see e.g. Cover and Thomas (2006). Nevertheless, because the birth and death rates are required to be non-negative and the Normal is supported on all of $\mathbb{R}$, this assumption is inadmissible. The next least restrictive assumption (in the information theoretic sense) is then to assume that these rates distribute log-Normally, because the log-Normal is the maximum entropy continuous *positive* distribution for a given mean and variance.

Data on the COVID-19 pandemic supports this assumption. Panels (a) and (d) of Figure 6 present the daily infection rate and death rate per country, respectively, along with ML-fitted log-Normal distributions. Panels (b),(e) present the logs of the rates, along with ML-fitted skew-Normal distributions, and panels (c),(f) present the respective q-q plots vs. the skew-Normal. In unreported results, the log rates are generally not rejected as skew-Normal using the K-S, C-2, and A-D distributional tests, and their kurtosis are close to 3. The skewness likely stems from under-reporting in some countries or from the "slow start" of the pandemic, which is better captured by a more complex SIR/SEIR model.

If both birth and death rates are log-Normally distributed then their difference — the growth rate of the population (or the spread of the pandemic) — should distribute as the

difference-of-log-Normals, or DLN. To confirm, Figure 3 presents the daily growth rates, per country, in the number of: new cases, tests given, and vaccinations given, along with ML-fitted DLN distributions and the accompanying q-q plots. Formal tests vs. the Stable, Laplace, and DLN are provided in the last three columns of Table 2. As in the firms and cities cases, the Stable and Laplace are rejected, the DLN is generally not rejected, and it handily beats the other distributions in AIC- and BIC-based "horse-races."

The evidence presented here then proposes a simple adjustment to the stochastic Malthusian model, in which the birth and death (log) rates are explicitly exponentiated, and follow a joint Ornstein-Uhlenbeck process:

$$\begin{aligned} dN(t) &= \left[\exp\left(\widehat{b}(t)\right) - \exp\left(\widehat{d}(t)\right)\right] \cdot N(t) \cdot dt \\ d\widehat{b}(t) &= \phi_b \cdot \left(\mu_b - \widehat{b}(t)\right) \cdot dt + \sigma_b \cdot dW_b \\ d\widehat{d}(t) &= \phi_d \cdot \left(\mu_d - \widehat{d}(t)\right) \cdot dt + \sigma_d \cdot dW_d \\ \mathbb{E}\left[dW_b \cdot dW_d\right] &= \rho_{bd} \cdot dt \end{aligned} \tag{18}$$

with $\phi_b, \phi_d > 0$ the mean-reversion parameters, $\mu_b, \mu_d \in \mathbb{R}$ the long-term means, $\sigma_b, \sigma_d \geq 0$ the volatility of the Wiener processes, and $\rho_{bd} \in [-1, 1]$ the correlation between them. This model yields Normal distributions of the log-rates $\widehat{b}(t), \widehat{d}(t)$, log-Normal distributions of the birth and death rates $b(t), d(t)$, and DLN distribution of population growth $r(t)$.

Returning to the context of epidemiology, the value R_0, which has gained notoriety during the COVID-19 pandemic, represents the average number of new infections generated by one infected individual in a completely susceptible population at the SIR/SEIR models. Lindenstrand and Svensson (2013) describe how the R_0 value is related to the Malthusian model and arises as a simple transform of $r(t)$. The process described in Equation 18 above can hence be useful in more accurately estimating R_0 during early stages of pandemics by following the method described by Lindenstrand and Svensson (2013) with slight adjustments.

3.4 Other examples

Once one sees the DLN distribution in growth data, one may have difficulty un-seeing it. This section presents, without exploring the economic or natural mechanism, three other examples of DLN-distributed growth data encountered "in the wild."

Panel (a) of Figure 7 presents a replication of Figure 1 from Guvenen, Karahan, Ozkan, and Song (2021). The panel presents the 5-year wage income growth distribution, from 2010 to 2015, based on a 10% sample of the U.S. population from the Master Earnings File of the US Social Security Administration. Income growth is DLN-distributed, as the q-q plot in Panel (d) shows. The appropriate column of Table 2 provides the usual statistical tests rejecting the Stable and Laplace but generally failing to reject the DLN, along with the horse-race results strongly favoring the DLN.

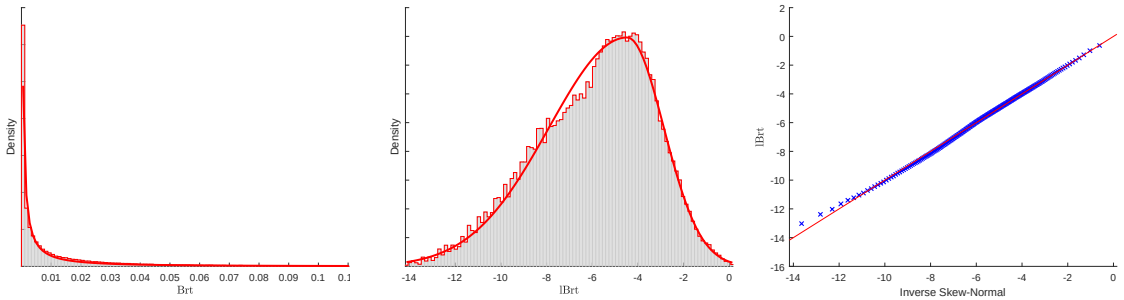
Panel (b) of the same figure presents the growth in daily average temperatures for 321 cities worldwide, from Abidjan (Ivory Coast) to Zurich (Switzerland), for the period 1995-2020. Panel (e) presents the q-q plot, and Table 2 provides the relevant statistical tests, with the usual results. It seems remarkable that a natural physical phenomenon of unique importance in the current discourse fits the DLN distribution as well as it does.

Finally, Panels (c),(f) of Figure 7 repeat the presentation for the growth in the yearly number of scientific and technical journal articles published per year by country. The data are again from Our World In Data, and span the period 2000-2018. Despite the sparsity of the data (spanning only 3.5K observations), the statistical results in the last column of Table 2 remain strong: this growth measure is rejected as Stable and Laplace but not as DLN, and the DLN handily beats the others in the AIC- and BIC-based horse races.

Table 2
Distributional tests — Cities, populations, pandemics, and other phenomena

This first three columns of this table present the results of tests of distributional form for population growth (dPOP), GDP growth by county (dGDPc), and GDP growth by metropolitan area and industry (dGDPm) for US locales. The next three columns present results for daily new cases growth (dCASE), new tests conducted growth (dTEST), and new vaccinations growth (dVACC) per country. The last three tests are for 5-year wage growth (dWAGE), daily temperature growth per city (dTEMP), and growth in scientific and technical journal articles per country (dJOUR). Tests are described in Table 1 and full data description is in the appendix.

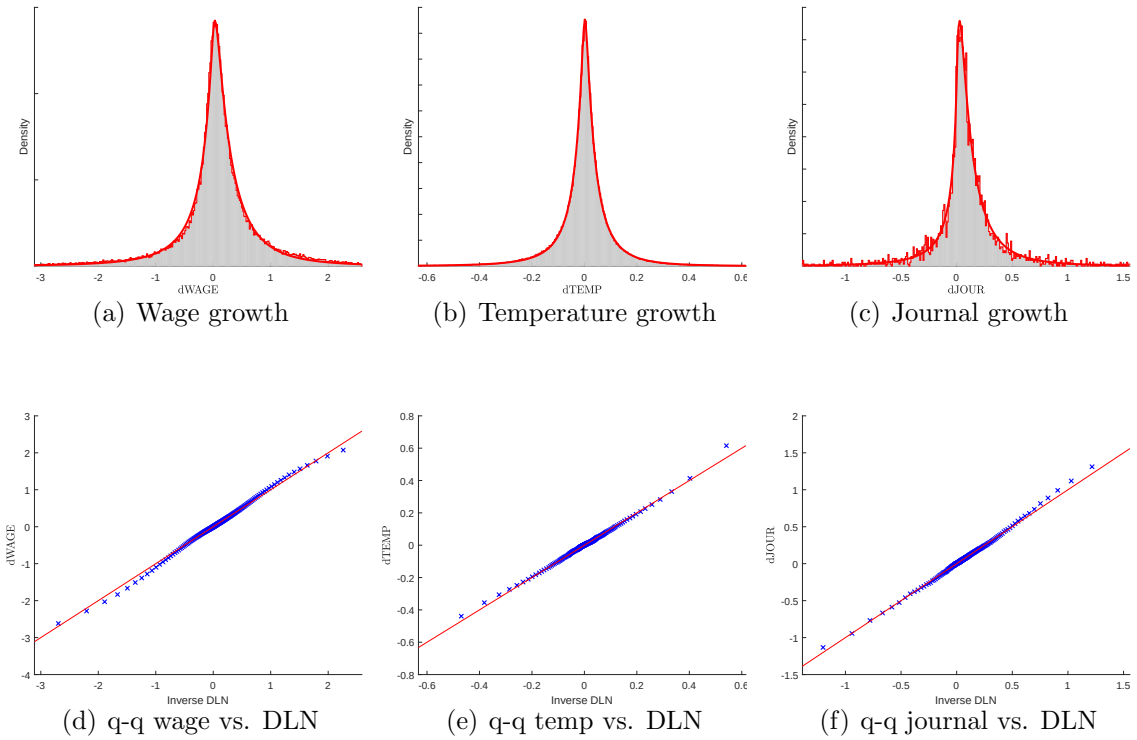
	dPOP	dGDPc	dGDPm	dCASE	dTEST	dVACC	dWAGE	dTEMP	dJOUR
Panel (a): Stable									
K-S	0.045	0.012	0.015	0.020	0.029	0.025	0.020	0.013	0.027
p-val	0.015	0.054	0.045	0.035	0.026	0.030	0.036	0.050	0.028
C-2	131	98	126	315	444	519	296	237	380
p-val	0.033	0.038	0.034	0.021	0.016	0.014	0.021	0.024	0.018
A-D	34.8	2.7	5.5	10.4	19.9	17.0	11.5	5.9	9.8
p-val	0.018	0.052	0.041	0.033	0.025	0.027	0.031	0.041	0.033
Panel (b): Laplace									
K-S	0.027	0.026	0.040	0.033	0.054	0.081	0.065	0.063	0.081
p-val	0.027	0.029	0.018	0.022	0.011	0.004	0.008	0.008	0.004
C-2	252	258	406	347	686	2224	1173	1016	1892
p-val	0.023	0.023	0.017	0.019	0.011	0.000	0.006	0.007	0.001
A-D	13.1	12.2	31.3	23.0	46.2	114.4	65.1	87.9	102.3
p-val	0.030	0.031	0.020	0.023	0.015	0.007	0.012	0.009	0.008
Panel (c): DLN									
K-S	0.006	0.007	0.007	0.007	0.009	0.016	0.009	0.009	0.010
p-val	0.087	0.077	0.075	0.077	0.063	0.042	0.063	0.064	0.060
C-2	25	61	42	33	70	174	75	44	167
p-val	0.072	0.048	0.057	0.064	0.045	0.029	0.044	0.056	0.030
A-D	0.835	0.773	1.214	0.951	1.714	5.345	2.794	0.433	1.349
p-val	0.076	0.077	0.068	0.073	0.061	0.042	0.052	0.092	0.065
Panel (d): Relative likelihood tests									
AIC R.L.:									
Stable	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Laplace	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DLN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
BIC R.L.:									
Stable	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Laplace	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DLN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000



(a) infection rate (b) log infection rate (c) q-q log-rate vs. SN

(d) death rate (e) log death rate (f) q-q log-rate vs. SN

Fig. 6. Birth and death rates — stylized facts. Panels (a),(d) present the distributions of new daily COVID19 infections and deaths, respectively, per million people, along with ML-fitted log-Normals. Panels (b),(e) repeat for the log-rates, along with ML-fitted skew-Normals. Panels (c),(f) present the respective q-q plots vs. the skew-Normal. Data are from Our World In Data Covid-19 depository. Full data description is in the appendix.



(a) Wage growth (b) Temperature growth (c) Journal growth

(d) q-q wage vs. DLN (e) q-q temp vs. DLN (f) q-q journal vs. DLN

Fig. 7. Other growth phenomena. Panel (a) presents the distribution of 5-year wage income growth for US workers, adjusted for age and time fixed-effects. Panel (b) presents the growth in daily average temperatures for 321 cities worldwide. Panel (c) presents the growth in the yearly number of scientific and technical journal articles published per year by country. All three panels are overlaid with ML-fitted DLN curves. Panels (d)-(f) present the respective q-q plots. Full data description is in the appendix.

4 Growth of DLN RVs

How can one measure growth in DLN-distributed RVs? A firm that had \$100 of income in year 1 and \$120 of income in year 2 has certainly grown its income. One can argue whether it is preferable to say the firm grew by $120/100 - 1 = 0.2 = 20\%$ or by $\log(120) - \log(100) = 0.182$ log-points, yet the question itself is well-formed. But what if the firm had $-\$100$ of income (i.e. \$100 in losses) in year 1, and then \$120 of income (i.e. profit) in year 2? What was its growth?

To begin, we require a definition of growth. Barro and Sala-I-Martin (2003) define instantaneous growth of a time-continuous and *strictly positive* RV $Z(t) > 0$, here denoted $\Gamma\{Z(t)\}$, as

$$\Gamma\{Z(t)\} \equiv \frac{dZ(t)/dt}{Z(t)} = \frac{Z'(t)}{Z(t)} \approx \frac{Z_{t+1} - Z_t}{Z_t} = \frac{Z_{t+1}}{Z_t} - 1 \equiv \Gamma_{\%}\{Z_{t+1}\} \tag{19}$$

with the approximation stemming from using the first-difference of discrete variables to approximate the derivative $Z'(t)$. This yields the well-known formulation of percentage growth in discrete variables, denoted $\Gamma_{\%}\{Z_{t+1}\}$.

Before proceeding, it is instructive to place some structure on the problem and consider the growth of a log-Normally distributed RV, as most measures of size are approximately log-Normally distributed. To that end, consider the following setting:

$$X_{t+1} = X_t + \epsilon_t^X \quad ; \quad \epsilon_t^X \sim \mathcal{N}(0, \sigma_X^2) \quad ; \quad Y_t = \exp(X_t) \tag{20}$$

In which X_t is a simple random walk and hence distributes Normally in the limit, and $Y_t > 0$ is hence log-Normally distributed. What is the growth in Y_t?

One can simply measure the growth in percentage terms and say that the growth is

$$\Gamma_{\%}\{Y_{t+1}\} \equiv \frac{Y_{t+1}}{Y_t} - 1 = \exp(X_{t+1} - X_t) - 1 = \exp(\epsilon_t^X) - 1 \tag{21}$$

or one can apply the definition in Equation 19 above and say

$$\frac{Y'(t)}{Y(t)} = \frac{Y(t) \cdot X'(t)}{Y(t)} = X'(t) \approx X_{t+1} - X_t = \log(Y_{t+1}) - \log(Y_t) \equiv \Gamma_{log}\{Y_{t+1}\} = \epsilon_t^X \quad (22)$$

which yields the well-known formulation of growth as a difference in logs between consecutive values, denoted $\Gamma_{log}\{Y_{t+1}\}$. The difference between the two statements stems from whether one differentiates the function Y before applying the first-difference approximation. Put differently, measuring growth in log-Normal RVs using percentage growth introduces a "convexity bias". Hence, the concept of growth used is closely related to the distribution whose growth is being considered.

Consider next the following definition for *generalized* growth of $Z(t) \in \mathbb{R}\backslash\{0\}$, denoted $\widetilde{\Gamma}\{Z(t)\}$,

$$\widetilde{\Gamma}\{Z(t)\} \equiv \frac{dZ(t)/dt}{|Z(t)|} = \frac{Z'(t)}{|Z(t)|} \approx \frac{Z_{t+1} - Z_t}{|Z_t|} \equiv \widetilde{\Gamma}_{\%}\{Z_{t+1}\} \quad (23)$$

in which we can again apply the discrete difference to get generalized (also known as "signed") percent growth, denoted $\widetilde{\Gamma}_{\%}\{Z_{t+1}\}$. This definition has several desirable properties. Clearly, when $Z(t) > 0$ this definition coincides with Equation 19 and is hence a proper extension. The absolute value further guarantees that $dZ(t)/dt > 0$ will yield positive growth, regardless of the sign of $Z(t)$. To test the intuition of the definition, consider several examples and their respective generalized percent growths, denoted $\widetilde{\%}$.

$Z_t \backslash Z_{t+1}$	-150	-100	-50	0	50	100	150
-100	$-50\widetilde{\%}$	$0\widetilde{\%}$	$50\widetilde{\%}$	$100\widetilde{\%}$	$150\widetilde{\%}$	$200\widetilde{\%}$	$250\widetilde{\%}$
$+100$	$-250\widetilde{\%}$	$-200\widetilde{\%}$	$-150\widetilde{\%}$	$-100\widetilde{\%}$	$-50\widetilde{\%}$	$0\widetilde{\%}$	$+50\widetilde{\%}$

With a generalized definition of growth in hand, consider the following setting:

$$\begin{aligned} X_{t+1}^p &= X_t^p + \epsilon_t^p \;\; ; \;\; Y_t^p = \exp\left(X_t^p\right) \\ X_{t+1}^n &= X_t^n + \epsilon_t^n \;\; ; \;\; Y_t^n = \exp\left(X_t^n\right) \end{aligned} \qquad (\epsilon_t^p, \epsilon_t^n)^T \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right) \quad (24)$$

in which X^p, X^n are Normally distributed and Y^p, Y^n are log-Normally distributed. Can we characterize the growth of the DLN-distributed RV $W_t = Y_t^p - Y_t^n$? Applying the generalized growth definition of Equation 23, we have

$$\begin{aligned} \widetilde{\Gamma}\{W(t)\} \equiv \frac{W'(t)}{|W(t)|} &= \frac{Y^p(t) \cdot dX^p(t)/dt - Y^n(t) \cdot dX^n(t)/dt}{|W(t)|} \\ &\approx \frac{Y_t^p \cdot \Gamma_{log}\{Y_{t+1}^p\} - Y_t^n \cdot \Gamma_{log}\{Y_{t+1}^n\}}{|Y_t^p - Y_t^n|} \equiv \widetilde{\Gamma}_{dln}\{W_{t+1}\} \end{aligned} \tag{25}$$

which implies the growth of a DLN RV can be defined as a function of the levels and log-point growth rates of its two component log-Normal RVs, and denoted $\widetilde{\Gamma}_{dln}\{W_{t+1}\}$.

If we instead write $W(t) = 2 \cdot \exp(\lambda(t)) \cdot \sinh(\tau(t))$ using the hyperbolic notation of the DLL, we have

$$\begin{aligned} \widetilde{\Gamma}\{W(t)\} \equiv \frac{W'(t)}{|W(t)|} &= \frac{2 \cdot \exp(\lambda(t)) \cdot \left(\frac{d\lambda(t)}{dt} \cdot \sinh(\tau(t)) + \frac{d\tau(t)}{dt} \cdot \cosh(\tau(t))\right)}{2 \cdot \exp(\lambda(t)) \cdot |\sinh(\tau(t))|} \\ &= \text{sgn}(\tau(t)) \cdot \left[\frac{d\lambda(t)}{dt} + \frac{d\tau(t)}{dt} \cdot \frac{1}{\tanh(\tau(t))}\right] \\ &\approx \text{sgn}(\tau_t) \cdot \left[(\lambda_{t+1} - \lambda_t) + \frac{\tau_{t+1} - \tau_t}{\tau_t}\right] \end{aligned} \tag{26}$$

which yields Equation 7 from Section 2.1.

To conclude, the example firm from the beginning of this section can be said to have had a growth rate of $(120 - (-100))/|-100| = 2.2 = 220\widetilde{\%}$. To calculate its DLN-based growth rate, however, we will need to know not just its income in each of the periods, but their respective components, namely the sales and expenses it incurred in each period. Finally note that if $\tau_t, \tau_{t+1} > 0$, the dln growth measure of Equations 25 and 26 continues to hold under the traditional growth definition of Barro and Sala-I-Martin (2003).

5 Concluding remarks

A variety of phenomena in nature arise at the balance of two opposing forces. Examples explored here include sales and expenses, agglomeration and congestion, and birth and death. I further show these forces are exponential, and hence approximately log-Normally distributed, and how these two facts in turn lead to the rise of the difference-of-log-Normals distribution from first principles by relying on the CLT.

I characterize the distribution, which to my knowledge has not been characterized or used previously, and provide evidence that several important data series indeed distribute as the difference-of-log-Normals. These include: firm cashflows, dispensations, investment, investment intensity, average product of capital, capital growth, cashflow growth, and equity returns at various intervals; population growth and economic output of US locales; the growth in COVID-19 cases, tests conducted, and vaccinations given; and the growth of labor wages, city temperatures, and journal publications.

I formalize the process giving rise to these phenomena into the difference-of-log-linears production function. This production function endogenously proposes novel definitions for production growth and efficiency, as well as yields novel implications about production growth and returns-to-scale. It does so in a mathematically tractable manner, owing to Hyperbolic Geometry. I further discuss how this production function can be used to model firms, cities, populations, and pandemics.

Finally, and building on the difference-of-log-linears production function's ability to model negative income in a tractable manner, I propose an extension of growth measures to sometimes-negative RVs. I show how, when an RV is distributed as a difference-of-log-Normals and generated by a difference-of-log-linears process, a generalization of log-point growth and of percent growth allows us to coherently discuss growth from negative values.

While simple, this line of inquiry appears quite useful. I show how it: predicts the distribution of firm cashflows, a core object of interest in finance; provides a well-behaved distribution for equity returns, thus mending a long-standing puzzle at the heart of Modern

Portfolio Theory; allows writing models of increasing-returns-to-scale cities in which more than one city can rationally exist; proposes an extension to the classical Malthusian "birth-death" model; and rationalizing a variety of observed growth distributions.

There is a sense by which this work is simply an extension of "Gibrat's Law" (Gibrat, 1931). Gibrat's core insight was that growth is a proportional process, leading to the multiplicative CLT and a log-Normal limiting distribution of size. To support his theory, Gibrat analyzed a broad range of data on incomes, firms, and industry sizes, and found a remarkable fit to the log-Normal distribution. Gibrat's goal was to "*convince his readers that this was a statistical regularity sufficiently sharp to provide a basis for serious mathematical modeling*" (Sutton, 1997).

Similarly, the core insight explored here is that growth phenomena are balanced between opposing multiplicative forces, or that "every marginal benefit must have a marginal cost," a lá Newton's third law. Due to the first-principles reliance on the CLT, it is likely that difference-of-log-Normals growth would be observed in a variety of natural phenomena not analyzed here. I hence posit that the difference-of-log-Normals is the general distribution of growth phenomena — a hypothesis which certainly requires further investigation. This in turn would mean the difference-of-log-Normals is a *fundamental distribution in nature*, in the sense that it arises in disparate settings, similar to the repeated occurrence of its better-known peers, the Normal and log-Normal distributions.

References

Abramovitz, M., 1956. Resource and Output Trends in the United States since 1870. In: *Resource and Output Trends in the United States since 1870*, NBER, pp. 1–23.

Aihounton, G. B. D., Henningsen, A., 2021. Units of measurement and the inverse hyperbolic sine transformation. The Econometrics Journal 24, 334–351.

Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. Second International Symposium on Information Theory pp. 267–281.

Ashton, T. S., 1926. The Growth of Textile Businesses in the Oldham District, 1884-1924. Journal of the Royal Statistical Society 89, 567–583.

Barro, R. J., Sala-I-Martin, X. I., 2003. Economic Growth. The MIT Press, Cambridge, Mass, second ed.

Beare, B. K., Toda, A. A., 2020. On the emergence of a power law in the distribution of COVID-19 cases. Physica D: Nonlinear Phenomena 412, 132649.

Bellemare, M. F., Wichman, C. J., 2020. Elasticities and the Inverse Hyperbolic Sine Transformation. Oxford Bulletin of Economics and Statistics 82, 50–61.

Bottazzi, G., Secchi, A., 2003. Why are distributions of firm growth rates tent-shaped? Economics Letters 80, 415–420.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York, second ed.

Cambanis, S., Huang, S., Simons, G., 1981. On the theory of elliptically contoured distributions. Journal of Multivariate Analysis 11, 368–385.

Chladná, Z., Kopfová, J., Rachinskii, D., Rouf, S. C., 2020. Global dynamics of SIR model with switched transmission rate. Journal of Mathematical Biology 80, 1209–1233.

Cobb, C. W., Douglas, P. H., 1928. A Theory of Production. The American Economic Review 18, 139–165.

Cover, T. M., Thomas, J. A., 2006. Elements of Information Theory 2nd Edition. Wiley-Interscience, Hoboken, N.J, second ed.

Duranton, G., Puga, D., 2003. Micro-Foundations of Urban Agglomeration Economies. In: *Handbook of Regional and Urban Economics*, National Bureau of Economic Research, Cambridge, MA, vol. 4, pp. 2063–2117.

Duranton, G., Puga, D., 2014. The Growth of Cities. In: *Handbook of Economic Growth*, Elsevier, vol. 2, pp. 781–853.

Eeckhout, J., 2004. Gibrat's Law for (All) Cities. The American Economic Review 94, 1429–1451.

Fama, E. F., 1963. Mandelbrot and the Stable Paretian Hypothesis. The Journal of Business 36, 420–429.

Fama, E. F., 1965. The Behavior of Stock-Market Prices. The Journal of Business 38, 34–105.

Fama, E. F., Roll, R., 1971. Parameter Estimates for Symmetric Stable Distributions. Journal of the American Statistical Association 66, 331–338.

Fang, K.-T., Kotz, S., Ng, K. W., 1990. Symmetric Multivariate and Related Distributions. CRC Press.

Fernández-Villaverde, J., Jones, C. I., 2022. Estimating and simulating a SIRD Model of COVID-19 for many countries, states, and cities. Journal of Economic Dynamics and Control 140, 104318.

Gabaix, X., 1999a. Zipf's Law and the Growth of Cities. The American Economic Review 89, 129–132.

Gabaix, X., 1999b. Zipf's Law for Cities: An Explanation. The Quarterly Journal of Economics 114, 739–767.

Gabaix, X., Ioannides, Y. M., 2004. Chapter 53 - The Evolution of City Size Distributions. In: Henderson, J. V., Thisse, J.-F. (eds.), *Handbook of Regional and Urban Economics*, Elsevier, vol. 4 of *Cities and Geography*, pp. 2341–2378.

Gatto, N. M., Schellhorn, H., 2021. Optimal control of the SIR model in the presence of transmission and treatment uncertainty. Mathematical Biosciences 333, 108539.

Gibrat, R., 1931. Les Inégalités Économiques. Librairie du Recueil Sirey, Paris.

Gubner, J., 2006. A New Formula for Lognormal Characteristic Functions. IEEE Transactions on Vehicular Technology 55, 1668–1671.

Gulisashvili, A., Tankov, P., 2016. Tail behavior of sums and differences of log-normal random variables. Bernoulli 22, 444–493.

Guvenen, F., Karahan, F., Ozkan, S., Song, J., 2021. What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics? Econometrica 89, 2303–2339.

Guvenen, F., McKay, A., Ryan, C. B., 2023. A Tractable Income Process for Business Cycle Analysis.

Hamilton, J. D., 1994. Time Series Analysis. Princeton Univ. Press, Princeton, NJ.

Henderson, J. V., 1974. The Sizes and Types of Cities. The American Economic Review 64, 640–656.

Jaimovich, N., Terry, S., Vincent, N., 2023. The Empirical Distribution of Firm Dynamics and Its Macro Implications. NBER Working Paper .

Lindenstrand, D., Svensson, Å., 2013. Estimation of the Malthusian parameter in an stochastic epidemic model using martingale methods. Mathematical Biosciences 246, 272–279.

Lo, C. F., 2012. The Sum and Difference of Two Lognormal Random Variables. Journal of Applied Mathematics 2012, 1–13.

Luttmer, E. G. J., 2011. On the Mechanics of Firm Growth. The Review of Economic Studies 78, 1042–1068.

Malthus, T., 1798. An Essay on the Principle of Population. J. Johnson, London.

Mandelbrot, B., 1960. The Pareto-Lévy Law and the Distribution of Income. International Economic Review 1, 79–106.

Mandelbrot, B., 1961. Stable Paretian Random Functions and the Multiplicative Variation of Income. Econometrica 29, 517–543.

Miller, M. B., 2013. Mathematics and Statistics for Financial Risk Management. Wiley.

Mullahy, J., Norton, E. C., 2022. Why Transform Y? A Critical Assessment of Dependent-Variable Transformations in Regression Models for Skewed and Sometimes-Zero Outcomes.

Officer, R. R., 1972. The Distribution of Stock Returns. Journal of the American Statistical Association 67, 807–812.

Parag, K. V., Donnelly, C. A., Zarebski, A. E., 2022. Quantifying the information in noisy epidemic curves. Nature Computational Science 2, 584–594.

Parham, R., 2022. Facts of US Firm Scale and Growth 1970-2019: An Illustrated Guide. arXiv:2302.02485 [q-fin.GN] .

Parham, R., 2023. Why Are Firm Growth Distributions Heavy-Tailed? Mimeo .

Roy, A. D., 1950. The Distribution of Earnings and of Individual Output. The Economic Journal 60, 489–505.

Roy, A. D., 1951. Some Thoughts on the Distribution of Earnings. Oxford Economic Papers 3, 135–146.

Rozenfeld, H. D., Rybski, D., Gabaix, X., Makse, H. A., 2011. The Area and Population of Cities: New Insights from a Different Perspective on Cities. American Economic Review 101, 2205–2225.

Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. W. Strahan and T. Cadell, London.

Stanley, M. H. R., Amaral, L. A. N., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. A., Stanley, H. E., 1996. Scaling behaviour in the growth of companies. Nature 379, 804.

Sutton, J., 1997. Gibrat's Legacy. Journal of Economic Literature 35, 40–59.

Voigtländer, N., Voth, H.-J., 2013. The Three Horsemen of Riches: Plague, War, and Urbanization in Early Modern Europe. The Review of Economic Studies 80, 774–811.

West, G., 2017. Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies. Penguin Press, New York, first edition ed.

A Data Appendix

A.1 Data

Firm data analyzed in Figures 1 and 5 and in Table 1 are from the Compustat/CRSP-combined dataset, accessed via Wharton's WRDS. The accounting-based data cover 164K firm-year observations on 15,797 firms between 1970-2019. Year fixed-effects were taken using the ratio of each year's nominal GDP and the GDP in 2019, from the St. Louis Fed (fred.stlouisfed.org). This means the secular growth trend of GDP was removed, and all data are as a percentage of a fixed economy and in terms of 2019 dollars. Adjusting by the standard GDP deflator does not change any of the conclusions and merely acts to move the mean of the growth distributions slightly to the right. The equity return panels (c) and (e) of Figure 1 use data on the entire universe of 2M monthly return observations and on a random sample of 5M daily return observations from CRSP. Excess equity returns are relative to the Fama-French 3-factor model, using factor loadings from Ken French's data library. Exact definitions of each data variables in Table 1 in terms of Compustat/CRSP mnemonics are available in Table 1 of Parham (2023). Extensive stylized facts and descriptive statistics for these data are available in Parham (2022).

City data presented in 2 and the first three columns of Table 2 are from three sources. The data in Panel (a) are from Rozenfeld et al. (2011), and pertain to population growth from 1991 to 2000 in 46K locales identified by the clustering algorithm of Rozenfeld et al. (2011). The data in Panel (c) are from the US Bureau for Economic Analysis (Gross Domestic Product by County, 2017-2020), and pertain to 9,333 county-year observations on the growth of per-county GDP for 3,111 counties from 2017 to 2020. The data in Panel (e) are again from the US BEA (Gross Domestic Product by Metropolitan Area and Industry, 2001-2017), and pertain to 270K growth observations on 87 industries within 384 MAs over 17 years.

COVID-19 data analyzed in Figures 3 and 6, as well as the middle three columns of Table 2, are from the Our World In Data COVID depository (ourworldindata.org/coronavirus),

downloaded on 2/5/2022. I use daily worldwide data covering 143K observations for per-country case growth, 71K observations for per-country tests growth, and 112K observations for per-country vaccination growth. To avoid low base rate effects and stale data, I only use growth observations when the base value is higher than 10 (e.g., more than 10 infections per day or more than 10 vaccines given), and the growth rate is different from 0 (as growth being exactly 0 usually indicates stale data).

Wage growth data in Panel (a) of Figure 7 is based on the distribution of wage income growth, corrected for age and year fixed-effects. The distribution is provided by the Global Repository of Income Dynamics (GRID, available at https://www.grid-database.org/) which us led by Fatih Guvenen, Luigi Pistaferri and Gianluca Violante, based in part on the work at Guvenen et al. (2021). Data on the US, which I use, are from a 10% panel sample of the U.S. population from the Master Earnings File of the Social Security Administration, concentrating on total annual labor earnings. The GRID project does not provide the data, but rather the distribution itself (i.e., bin midpoints and bin weights). I use the most recent 5-year growth data (from 2015 to 2020), though I verify that pooling the data on 5-year growth for all available US periods does not change the outcomes.

Temperature data in Panel (b) of Figure 7 is based on the average daily temperatures for 321 cities worldwide, from the Weather Project at the University of Dayton, provided by Kelly Kissock. The data were downloaded from Kaggle (https://www.kaggle.com/datasets/sudalairajkumar temperature-of-major-cities) on 9/15/2023, and span 2.7M daily temperature growth observations between 1995-2020.

Journal publication data in Panel (c) of Figure 7 are from Our World in Data (https://ourworldindata.org and-technical-journal-articles) and are originally source from the National Science Foundation via The World Bank. The data cover yearly journal publications from 197 countries during the period 2000-2018, for a total of 3.5K growth observations.

A.2 Analysis

For each growth measure, the data are first fit to the DLN distribution using the MLE estimator described in the Online Appendix. The empirical distribution of the data, along with the fitted DLN (in red) are presented as figures, along with a q-q plot of the empirical CDF vs. the theoretical DLN CDF, also developed in the Online Appendix.

Three statistical goodness-of-fit tests are used to verify whether the empirical data indeed stem from the DLN distribution. The three goodness-of-fit distributional tests I use are the Kolmogorov-Smirnov (K-S), the Chi-square (C-2), and the Anderson-Darling (A-D) tests. The three tests are sensitive to different distributional deviations — K-S has uniform power throughout, C-2 is more powerful around the center-mass, and A-D is more powerful around the tails — hence I report results for all three tests.

Because the alternative distributions I consider — the Laplace distribution (itself a difference of exponentially distributed variates) and the Levy-Stable (aka Pareto-Stable or Power-Law) distribution — have different number of parameters, a simple likelihood ratio test may tilt in favor of the DLN, which has more degrees of freedom (5 for the DLN, 4 for the Stable, and 3 for the asymmetric Laplace). To account for the degrees of freedom, I use the relative likelihood test, derived from the AIC statistic of Akaike (1973). The relative likelihood is a non-nested version of the likelihood ratio test, accounting for the number of parameters. I also report relative likelihood tests using the BIC statistic, which penalizes extra degrees of freedom more heavily. For a review of the information-theoretic approach to model selection see, e.g., Burnham and Anderson (2002).

OA Online Appendix

Equations 9, 10 of the main text have defined the basic structure of a DLN RV,

$$W = Y_p - Y_n = \exp(X_p) - \exp(X_n) \quad \text{with} \quad \boldsymbol{X} = (X_p, X_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and denoted $W \sim \text{DLN}(\mu_p, \sigma_p, \mu_n, \sigma_n, \rho_{pn})$. The next section begins by characterizing its PDF and CDF, as well as discussing the simplified case when $\rho_{pn} = 0$ and the PDF can be derived as a simple convolution using a Fourier transform.

OA.1 PDF and CDF

The PDF for the bi-variate Normal (BVN) RV $\boldsymbol{X}$ is well-known to be

$$f_{BVN}(\boldsymbol{x}) = \frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{2\pi} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) = \frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{2\pi} \cdot \exp\left(-\frac{1}{2}||\boldsymbol{x}-\boldsymbol{\mu}||_{\boldsymbol{\Sigma}}\right) \quad \text{(OA.1)}$$

with $|\boldsymbol{\Sigma}|$ the determinant of $\boldsymbol{\Sigma}$ and $||\boldsymbol{x}||_{\boldsymbol{\Sigma}}$ the Euclidean norm of $\boldsymbol{x}$ under the Mahalanobis distance induced by $\boldsymbol{\Sigma}$.

The PDF for the bi-variate log-Normal (BVLN) RV $\boldsymbol{Y} = (Y_p, Y_n)^T$ can be obtained by using the multivariate change of variables theorem. If $\boldsymbol{Y} = g(\boldsymbol{X})$ then

$$f_Y(\boldsymbol{y}) = f_X(g^{-1}(\boldsymbol{y})) \cdot ||J_{g^{-1}}(\boldsymbol{y})|| \quad \text{(OA.2)}$$

with $J_{g^{-1}}$ the Jacobian matrix of $g^{-1}(\cdot)$ and $||J_{g^{-1}}||$ the absolute value of its determinant. Applying the theorem for $\boldsymbol{Y} = g(\boldsymbol{X}) = (\exp(X_p), \exp(X_n))^T$ we have $g^{-1}(\boldsymbol{y}) = (log(y_p), log(y_n))^T$ and $||J_{g^{-1}}(\boldsymbol{y})|| = (y_p \cdot y_n)^{-1}$. The PDF of a BVLN RV is then

$$f_{BVLN}(\boldsymbol{y}) = \frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{2\pi y_p y_n} \exp\left(-\frac{1}{2}||\log(\boldsymbol{y}) - \boldsymbol{\mu}||_{\boldsymbol{\Sigma}}\right) \quad \text{(OA.3)}$$

We can now define the cumulative distribution function (CDF) of the DLN distribution

using the definition of the CDF of the difference of two RV

$$\begin{aligned} F_{DLN}(w) &= P[W \leq w] = P[y_p - y_n \leq w] = P[y_p \leq y_n + w] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y_n+w} f_{BVLN}(y_p, y_n) dy_p dy_n \end{aligned} \tag{OA.4}$$

which can be differentiated w.r.t w to yield the PDF

$$f_{DLN}(w) = \int_{-\infty}^{\infty} f_{BVLN}(y + w, y) dy = \int_{-\infty}^{\infty} f_{BVLN}(y, y - w) dy \tag{OA.5}$$

but because $f_{BVLN}(\boldsymbol{y})$ is non-zero only for $\boldsymbol{y} > 0$, we limit the integration range

$$f_{DLN}(w) = \int_{\max(0,w)}^{\infty} f_{BVLN}(y, y - w) dy \tag{OA.6}$$

which yields the PDF of the DLN distribution.

It is well-known, however, that the integral in equation OA.6 does not have a closed-form solution. The accompanying code suite evaluates it numerically, and also numerically evaluates the CDF using its definition

$$F_{DLN}(w) = \int_{-\infty}^{w} f_{DLN}(y) dy \tag{OA.7}$$

For the simpler case with difference of uncorrelated log-Normals, i.e. $\rho_{pn} = 0$, we can derive the PDF of the DLN via a characteristic function (CF) approach as well. In this case, we can write the CF of the DLN as $\varphi_{DLN}(t) = \varphi_{LN}(t) \cdot \varphi_{LN}(-t)$ with $\varphi_{LN}(t)$ the CF of the log-Normal. Next, we can apply a Fourier transform to obtain the PDF,

$$f_{DLN}(w) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i \cdot t \cdot w} \cdot \varphi_{DLN}(t) dt \tag{OA.8}$$

Unfortunately, the log-Normal does not admit an analytical CF, and using Equation OA.8 requires a numerical approximation for $\varphi_{LN}(t)$ as well. Gubner (2006) provides a fast and

accurate approximation method for the CF of the log-Normal which I use in the calculation of $f_{DLN}(w)$ when using this method.

OA.2 Moments

OA.2.1 MGF

The moment generating function (MGF) of the DLN can be written as

$$M_W(t) = \mathbb{E}\left[e^{tW}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{tw} f_{BVLN}(y+w, y)dydw \tag{OA.9}$$

but this formulation has limited usability due to the lack of closed-form solution for the integrals. Instead, it is useful to characterize the moments directly, as we can obtain them in closed-form.

OA.2.2 Mean and variance

Using the definitions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in 10, define the mean and covariance of the BVLN RV, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ (element-wise) as

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{(i)} &= \exp\left(\boldsymbol{\mu}_{(i)} + \frac{1}{2}\boldsymbol{\Sigma}_{(i,i)}\right) \\ \hat{\boldsymbol{\Sigma}}_{(i,j)} &= \exp\left(\boldsymbol{\mu}_{(i)} + \boldsymbol{\mu}_{(j)} + \frac{1}{2}\left(\boldsymbol{\Sigma}_{(i,i)} + \boldsymbol{\Sigma}_{(j,j)}\right)\right) \cdot \left(\exp\left(\boldsymbol{\Sigma}_{(i,j)}\right) - 1\right) \end{aligned} \tag{OA.10}$$

Note that if $\boldsymbol{\Sigma}$ is diagonal (i.e., X_p and X_n are uncorrelated) then $\hat{\boldsymbol{\Sigma}}$ will be diagonal as well. We are however interested in the general form of the DLN distribution. The identities regarding the expectation and variance of a sum of RV yield

$$\mathbb{E}[W] = \mathbb{E}[Y_p] - \mathbb{E}[Y_n] = \hat{\boldsymbol{\mu}}_{(1)} - \hat{\boldsymbol{\mu}}_{(2)} = \exp(\mu_p + \frac{\sigma_p^2}{2}) - \exp(\mu_n + \frac{\sigma_n^2}{2}) \tag{OA.11}$$

and

$$\begin{aligned}\text{Var}\left[W\right] &= \mathbb{C}\left[Y_p, Y_p\right] + \mathbb{C}\left[Y_n, Y_n\right] - 2 \cdot \mathbb{C}\left[Y_p, Y_n\right] = \hat{\mathbf{\Sigma}}_{(1,1)} + \hat{\mathbf{\Sigma}}_{(2,2)} - 2 \cdot \hat{\mathbf{\Sigma}}_{(1,2)} \\ &= \exp\left(2\mu_p + \sigma_p^2\right) \cdot \left(exp\left(\sigma_p^2\right) - 1\right) + \exp\left(2\mu_n + \sigma_n^2\right) \cdot \left(exp\left(\sigma_n^2\right) - 1\right) \\ &- 2\exp\left(\mu_p + \mu_n + \frac{1}{2}(\sigma_p^2 + \sigma_n^2)\right) \cdot \left(\exp\left(\sigma_p \sigma_n \rho_{pn}\right) - 1\right)\end{aligned} \tag{OA.12}$$

with $\mathbb{C}$ the covariance operator of two general RV U_1, U_2

$$\mathbb{C}\left[U_1, U_2\right] = \mathbb{E}\left[(U_1 - \mu_1)(U_2 - \mu_2)\right] \tag{OA.13}$$

OA.2.3 Skewness and kurtosis

Skewness and kurtosis of the DLN can similarly be established using coskewness and cokurtosis — see e.g. Miller (2013) for an overview. Coskewness of three general RV U_1, U_2, U_3 is defined as

$$\mathbb{S}\left[U_1, U_2, U_3\right] = \frac{\mathbb{E}\left[(U_1 - \mu_1)(U_2 - \mu_2)(U_3 - \mu_3)\right]}{\sigma_1 \sigma_2 \sigma_3} \tag{OA.14}$$

and cokurtosis of four general RV U_1, U_2, U_3, U_4 is defined as

$$\mathbb{K}\left[U_1, U_2, U_3, U_4\right] = \frac{\mathbb{E}\left[(U_1 - \mu_1)(U_2 - \mu_2)(U_3 - \mu_3)(U_4 - \mu_4)\right]}{\sigma_1 \sigma_2 \sigma_3 \sigma_4} \tag{OA.15}$$

with the property that $\mathbb{S}\left[U, U, U\right] = \text{Skew}\left[U\right]$ and $\mathbb{K}\left[U, U, U, U\right] = \text{Kurt}\left[U\right]$. More importantly, it is simple to show that

$$\text{Skew}\left[U - V\right] = \frac{\sigma_U^3 \mathbb{S}\left[U, U, U\right] - 3\sigma_U^2 \sigma_V \mathbb{S}\left[U, U, V\right] + 3\sigma_U \sigma_V^2 \mathbb{S}\left[U, V, V\right] - \sigma_V^3 \mathbb{S}\left[V, V, V\right]}{\sigma_{U-V}^3} \tag{OA.16}$$

and similarly

$$\begin{aligned}\text{Kurt}\left[U-V\right] = \frac{1}{\sigma_{U-V}^4}[&\sigma_U^4\mathbb{K}\left[U,U,U,U\right] - 4\sigma_U^3\sigma_V\mathbb{K}\left[U,U,U,V\right] \\ &+ 6\sigma_U^2\sigma_V^2\mathbb{K}\left[U,U,V,V\right] - 4\sigma_U\sigma_V^3\mathbb{K}\left[U,V,V,V\right] + \sigma_V^4\mathbb{K}\left[V,V,V,V\right]]\end{aligned} \tag{OA.17}$$

with $\sigma_{U-V} = \text{Var}\left[U-V\right]^{\frac{1}{2}}$ calculated using Equation OA.12. Evaluating the operators $\mathbb{S}$ and $\mathbb{K}$ for the case of DLN requires evaluating expressions of the general form $\mathbb{E}\left[Y_p^i Y_n^j\right]$, which can be done via the MGF of the BVN distribution

$$\mathbb{E}\left[Y_p^i Y_n^j\right] = \mathbb{E}\left[e^{iX_p}e^{jX_n}\right] = \text{MGF}_{BVN}\left(\begin{bmatrix} i \\ j \end{bmatrix}\right) = \mathbb{E}\left[Y_p^i\right]\mathbb{E}\left[Y_n^j\right]e^{ij\boldsymbol{\Sigma}_{(1,2)}} \tag{OA.18}$$

with $\mathbb{E}\left[Y_p^i\right] = \exp\left(i\mu_p + \frac{1}{2}i^2\sigma_p^2\right)$. This concludes the technical details of the derivation.

The method presented can be extended to higher central moments as well. The accompanying code suite includes functions that implement the equations above and use them to calculate the first five moments of the DLN given the parameters $(\mu_p, \sigma_p, \mu_n, \sigma_n, \rho_{pn})$. The appendix describes the results of Monte-Carlo experiments testing the empirical variance and bias of the moments as functions of sample size.

OA.3 Estimation

Given data $\boldsymbol{D} \sim \text{DLN}(\boldsymbol{\Theta})$ with $\boldsymbol{\Theta} = (\mu_p, \sigma_p, \mu_n, \sigma_n, \rho_{pn})$, we would like to find an estimate $\hat{\boldsymbol{\Theta}}$ to the parameter vector $\boldsymbol{\Theta}$. Experiments show that given an appropriate initial guess, the MLE estimates of $\boldsymbol{\Theta}$ perform well in practice. The main parameter of difficulty is ρ_{pn}. This parameter is akin to the shape parameter in the Stable distribution, which plays a similar role and is similarly difficult to estimate, see e.g. Fama and Roll (1971). It hence requires special care in the estimation.

The estimation code provided minimizes the negative log-likelihood of the data w.r.t the DLN PDF using a multi-start algorithm. The starting values for the first four parameters

are fixed for all start points as:

$$\begin{bmatrix} \mu_p \\ \sigma_p \\ \mu_n \\ \sigma_n \end{bmatrix} = \begin{bmatrix} \text{Median}\left[\log\left(\boldsymbol{D}\right)\right] & \text{for} & \boldsymbol{D} > 0 \\ \text{IQR}\left[\log\left(\boldsymbol{D}\right)\right]/1.35 & \text{for} & \boldsymbol{D} > 0 \\ \text{Median}\left[\log\left(-\boldsymbol{D}\right)\right] & \text{for} & \boldsymbol{D} < 0 \\ \text{IQR}\left[\log\left(-\boldsymbol{D}\right)\right]/1.35 & \text{for} & \boldsymbol{D} < 0 \end{bmatrix} \tag{OA.19}$$

while the initial guesses for ρ_{pn} are $(-0.8, -0.3, 0, 0.3, 0.8)$. The estimator $\hat{\boldsymbol{\Theta}}$ is then the value which minimizes the negative log-likelihood in the multi-start algorithm. The estimator inherits asymptotic normality, consistency, and efficiency properties from the general M-estimator theory, as the dimension of $\hat{\boldsymbol{\Theta}}$ is fixed, the likelihood is smooth, and is supported on $\mathbb{R}\ \forall\hat{\boldsymbol{\Theta}}$. A better estimation procedure for the parameters of the DLN might be merited, but is left for future work.

OA.4 The elliptical multi-variate DLN

Practical applications of the DLN require the ability to work with multi-variate DLN RVs. I hence present an extension of the DLN to the multi-variate case using elliptical distribution theory, with the standard reference being Fang, Kotz, and Ng (1990).

The method of elliptical distributions requires a symmetric baseline distribution. We will therefore focus our attention on the symmetric DLN case in which $\mu_p = \mu_n \equiv \mu$ and $\sigma_p = \sigma_n \equiv \sigma$, yielding the three parameter uni-variate symmetric distribution $\text{SymDLN}(\mu, \sigma, \rho) = \text{DLN}(\mu, \sigma, \mu, \sigma, \rho)$. I begin by defining a standardized N-dimensional elliptical DLN RV using SymDLN and the spherical decomposition of Cambanis, Huang, and Simons (1981), and later extend it to a location-scale family of distributions.

Let $\mathbf{U}$ be an N-dimensional RV distributed uniformly on the unit hyper-sphere in $\mathbb{R}^N$ and arranged as a column vector. Let $R \geq 0$ be a univariate RV independent of $\mathbf{U}$ with PDF $f_R(r)$ to be derived momentarily, and let $\mathbf{Z} = R \cdot \mathbf{U}$ be a standardized N-dimensional elliptical DLN RV. A common choice for $\mathbf{U}$ is $\widehat{\mathbf{U}}/||\widehat{\mathbf{U}}||_2$ with $\widehat{\mathbf{U}} \sim MVN(\mathbf{0}_N, \mathbf{1}_N)$. $\mathbf{U}$ captures

a direction in $\mathbb{R}^N$, and we have $\sqrt{\mathbf{U}^T \cdot \mathbf{U}} = ||\mathbf{U}||_2 \equiv 1$, which implies $\sqrt{\mathbf{Z}^T \cdot \mathbf{Z}} = ||\mathbf{Z}||_2 = R$. We further know that the surface area of an N-sphere with radius R is given by

$$S_N(R) = \frac{2 \cdot \pi^{\frac{N}{2}}}{\Gamma\left(\frac{N}{2}\right)} \cdot R^{N-1} \tag{OA.20}$$

and can hence write the PDF of $\mathbf{Z}$ as

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{f_R(||\mathbf{z}||_2)}{S_N(||\mathbf{z}||_2)} = \frac{\Gamma\left(\frac{N}{2}\right) \cdot f_R(||\mathbf{z}||_2)}{2 \cdot \pi^{\frac{N}{2}} \cdot ||\mathbf{z}||_2^{N-1}} \tag{OA.21}$$

We require $f_R(r)$ and $f_{\mathbf{Z}}(\mathbf{z})$ to be valid PDFs, which yields the conditions

$$\begin{aligned} &f_R(r) \geq 0 \ \forall \ r \in \mathbb{R} \\ &f_{\mathbf{Z}}(\mathbf{z}) \geq 0 \ \forall \ \mathbf{z} \in \mathbb{R}^N \\ &\int_{-\infty}^{\infty} f_R(r) \ dr = 1 \\ &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{Z}}(\mathbf{z}) \ d\mathbf{z}_{(N)} \cdots d\mathbf{z}_{(1)} = 1 \end{aligned} \tag{OA.22}$$

to those, we can add the condition that the properly normalized distribution of $f_R(r)$ will be SymDLN,

$$f_R(r) = \widetilde{M}_N(r) \cdot f_{DLN}(r) \tag{OA.23}$$

with $\widetilde{M}_N(r)$ chosen such that the conditions in Equation OA.22 hold. Solving for this set of conditions yields

$$f_R(r) = \frac{r^{N-1}}{\int_0^{\infty} \widetilde{r}^{N-1} \cdot f_{DLN}(\widetilde{r}) \ d\widetilde{r}} \cdot f_{DLN}(r) \tag{OA.24}$$

and

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{\Gamma\left(\frac{N}{2}\right)}{2 \cdot \pi^{\frac{N}{2}} \cdot \int_0^{\infty} \widetilde{r}^{N-1} \cdot f_{DLN}(\widetilde{r}) \ d\widetilde{r}} \cdot f_{DLN}(||\mathbf{z}||_2) = M_N \cdot f_{DLN}(||\mathbf{z}||_2) \tag{OA.25}$$

with M_N a normalization constant depending only on the dimension N and the parameters

of the baseline SymDLN(μ, σ, ρ) being used. We can use $\mathbf{Z}$'s CDF definition to write

$$\begin{aligned} F_{\mathbf{Z}}(\mathbf{z}) &= \int_{-\infty}^{\mathbf{z}_{(1)}} \cdots \int_{-\infty}^{\mathbf{z}_{(N)}} f_{\mathbf{Z}}(\widehat{\mathbf{z}}) \; d\widehat{\mathbf{z}}_{(N)} \cdots d\widehat{\mathbf{z}}_{(1)} \\ &= \int_{-\infty}^{\mathbf{z}_{(1)}} \cdots \int_{-\infty}^{\mathbf{z}_{(N)}} M_N \cdot f_{\mathbf{DLN}}(||\mathbf{z}||_2) \; d\widehat{\mathbf{z}}_{(N)} \cdots d\widehat{\mathbf{z}}_{(1)} \end{aligned} \tag{OA.26}$$

which concludes the characterization of the standardized N-dimensional elliptical DLN RV.

Extending the standardized N-dimensional DLN to a location-scale family of distributions is now straightforward. Let $\widetilde{\boldsymbol{\mu}} = (\mu_1, \mu_2, ..., \mu_N)^T$ be a column vector of locations and let $\widetilde{\boldsymbol{\Sigma}}$ be a positive-semidefinite scaling matrix of rank N. Define

$$\mathbf{W} = \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \cdot \mathbf{Z} \tag{OA.27}$$

with $\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}$ denoting the eigendecomposition of $\widetilde{\boldsymbol{\Sigma}}$. The PDF of $\mathbf{W}$ is then given by

$$\begin{aligned} f_{\mathbf{W}}(\mathbf{w}) &= |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \cdot f_{\mathbf{Z}}\left(\widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \cdot (\mathbf{w} - \widetilde{\boldsymbol{\mu}})\right) \\ &= |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \cdot M_N \cdot f_{DLN}\left(\sqrt{(\mathbf{w} - \widetilde{\boldsymbol{\mu}})^T \cdot \widetilde{\boldsymbol{\Sigma}}^{-1} \cdot (\mathbf{w} - \widetilde{\boldsymbol{\mu}})}\right) \\ &= |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \cdot M_N \cdot f_{DLN}\left(||\mathbf{w} - \widetilde{\boldsymbol{\mu}}||_{\widetilde{\boldsymbol{\Sigma}}}\right) \end{aligned} \tag{OA.28}$$

The CDF of $\mathbf{W}$ can similarly be written as

$$F_{\mathbf{W}}(\mathbf{w}) = |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \cdot M_N \cdot \int_{-\infty}^{\mathbf{w}_{(1)}} \cdots \int_{-\infty}^{\mathbf{w}_{(N)}} f_{\mathbf{DLN}}\left(||\mathbf{w} - \widetilde{\boldsymbol{\mu}}||_{\widetilde{\boldsymbol{\Sigma}}}\right) \; d\widehat{\mathbf{w}}_{(N)} \cdots d\widehat{\mathbf{w}}_{(1)} \tag{OA.29}$$

which characterizes a general elliptical multi-variate DLN RV.

Finally, note that the scaling matrix $\widetilde{\boldsymbol{\Sigma}}$ is not the covariance matrix of $\mathbf{W}$ due to the heavy-tails of $\mathbf{W}$, similar to other heavy-tailed elliptical distributions such as the multi-variate Stable, t, or Laplace distributions. Further note that the normalization integral in Equation OA.24 is numerically unstable for high values of N (e.g., $N \geq 5$), and care should be taken when deriving the PDF of high-dimensional DLN RVs in practice.