An empirical study of the tails of mutual fund size

Yonathan Schwarzkopf[1,2] and J. Doyne Farmer[2,3]

[1] *California Institute of Technology, Pasadena, CA 91125.*
[2] *Santa Fe Institute, Santa Fe, NM 87501.*
[3] *Luiss Guido Carli, ROMA Italy.*
(Dated: May 28, 2010)

The mutual fund industry manages about a quarter of the assets in the U.S. stock market and thus plays an important role in the U.S. economy. The question of how much control is concentrated in the hands of the largest players is best quantitatively discussed in terms of the tail behavior of the mutual fund size distribution. We study the distribution empirically and show that the tail is much better described by a log-normal than a power law, indicating less concentration than, for example, personal income. The results are highly statistically significant and are consistent across fifteen years. This contradicts a recent theory concerning the origin of the power law tails of the trading volume distribution. Based on the analysis in a companion paper, the log-normality is to be expected, and indicates that the distribution of mutual funds remains perpetually out of equilibrium.

PACS numbers: 89.65.Gh,89.75.Da,02.60.Ed

I. INTRODUCTION

As of 2007 the mutual fund industry controlled 23% of household taxable assets in the United States[1]. In absolute terms this corresponded to 4.4 trillion USD and 24% of U.S. corporate equity holdings. Large players such as institutional investors are known to play an important role in the market [1]. This raises the question of who has this influence: Are mutual fund investments concentrated in a few dominant large funds, or spread across many funds of similar size? Are there mutual funds that are so large that they are "too big to fail"?

This question is best addressed in terms of the behavior of the upper tail of the mutual fund size distribution. The two competing hypotheses usually made in studies of firms are Zipf's law vs. a lognormal. Zipf's law means that the distribution of the size s is a power law with tail exponent $\zeta_s \approx 1$, i.e.

$$P(s > X) \sim X^{-\zeta_s},$$

Log-normality means that $\log s$ has a normal distribution, i.e. the density function $p_{LN}(s)$ obeys

$$p(s) = \frac{1}{s\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(s) - \mu_s)^2}{2\sigma_s^2}\right).$$

From the point of view of extreme value theory this distinction is critical, since it implies a completely different class of tail behavior[2]. These are both heavy tailed, but Zipf's law is much more heavy tailed. For a log-normal all the moments exist, whereas for Zipf's law none of the moments exist. For Zipf's law an estimator of the mean fails to converge. In practical terms, for mutual funds this would imply that for any sample size N, with significant probability an individual fund can be so large that it is bigger than all other $N - 1$ firms combined. In contrast, for a log-normal, in the limit as $N \to \infty$ the relative size of a single fund becomes negligible.

This question takes on added meaning because the assumption that mutual funds follow Zipf's law has been argued to be responsible for the observed power law distribution of trading volume [3, 4]. Gabaix et al. have also asserted that the mutual fund distribution follows Zipf's law and have used this in a proposed explanation for the distribution of price returns [5, 6].

We resolve this empirically using the Center for Research in Security Prices (CRSP) dataset and find that the equity fund size distribution is much better described by a log-normal distribution.

Our results are interesting in the broader context of the literature on firm size. Mutual funds provide a particularly good type of firm to study because there are a large number of funds and their size is accurately recorded. It is generally believed that the resulting size distribution from aggregating across industries has a power law tail that roughly follows Zipf's law, but for individual industries the tail behavior is debated[3]. A large number of stochastic process models have been proposed to explain this[4]. Our results add support to the notion that for single industries the distribution is log-normal.

The log-normality of the distribution of mutual funds is also interesting for what it suggests about the underlying processes that determine mutual fund size. In a

---

[1] Data is taken from the Investment Company Institute's 2007 fact book available at www.ici.org.

[2] According to extreme value theory a probability distribution can have only four possible types of tail behavior. The first three correspond to distributions with finite support, thin tails, and tails that are sufficiently heavy that some of the moments do not exist, i.e. power laws. The fourth category corresponds to distributions that in a certain sense do not converge; it is remarkable that most known distributions fall into one of the first three categories [2].

[3] Some studies have found that the upper tail is a log-normal [7–13] while others have found a power law [12–14]

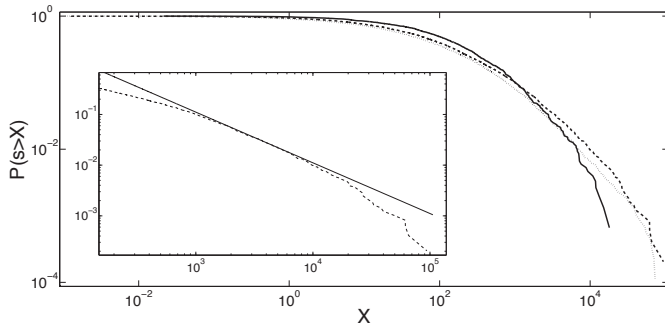[4] For past stochastic models see [5, 7, 9, 15–19]

FIG. 1. The CDF for the mutual fund size s (in millions of 2007 dollars) is plotted with a double logarithmic scale. The cumulative distribution for funds existing at the end of the years 1993, 1998 and 2005 are given by the full, dashed and dotted lines respectively.
Inset: The upper tail of the CDF for the mutual funds existing at the end of 1998 (dotted line) is compared to an algebraic relation with exponent -1 (solid line).

companion paper [21] we develop a model for the random process of mutual fund entry, exit and growth under the assumption of market efficiency, and show that this gives a good fit to the data studied here. We show that while the steady-state solution is a power law, the timescale for reaching this solution is very slow. Thus given any substantial non-stationarity in the entry and exit processes the distribution will remain in its non-equilibrium log-normal state. See the discussion in Section V.

II. DATA SET

We analyze the Center for Research in Security Prices (CRSP) Survivor-Bias-Free US Mutual Fund Database[5]. The database is survivor bias free as it contains historical performance data for both active and inactive mutual funds. We study monthly data from 1991 to 2005[6] on all reported equity funds. We define an equity fund as one whose portfolio consists of at least 80% stocks. The results are not qualitatively sensitive to this, e.g. we get essentially the same results even if we use all funds. The data set has monthly values for the Total Assets Managed (TASM) by the fund and the Net Asset Value (NAV). We define the size s of a fund to be the value of the TASM, measured in millions of US dollars and corrected for inflation relative to July 2007. Inflation adjustments are based on the Consumer Price Index, published by the BLS.

[5] The US Mutual Fund Database can be purchased from the Center for Research in Security Prices (www.crsp.com).
[6] There is data on mutual funds starting in 1961, but prior to 1991 there are very few entries. There is a sharp increase in 1991, suggesting incomplete data collection prior to 1991.

III. IS THE TAIL A POWER LAW?

Despite the fact that the mutual fund industry offers a large quantity of well-recorded data, the size distribution of mutual funds has not been rigorously studied. This is in contrast with other types of firms where the size distribution has long been an active research subject. The fact that the distribution is highly skewed and heavy tailed can be seen in Figure 1, where we plot the cumulative distribution of sizes $P(s > X)$ of mutual fund sizes in three different years.

A visual inspection of the mutual fund size distribution suggests that it does not follow Zipf's law[7]. In the inset of Figure 1 we compare the tail for funds with sizes $s > 10^2$ million to a power law $s^{-\zeta_s}$, with $\zeta_s = -1$. Whereas a power law corresponds to a straight line when plotted on double logarithmic scale, the data show substantial and consistent downward curvature. The main point of this paper is to make more rigorous tests of the power law vs. the log-normal hypothesis. These back up the intuitive impression given by this plot, indicating that the data are not well described by a power law.

To test the validity of the power law hypothesis we use the method developed by Clauset et al. [20]. They use the somewhat strict definition[8] that the probability density function $p(s)$ is a power law if there exists an s_{min} such that for sizes larger than s_{min}, the functional form of the density $p(s)$ can be written

$$p(s) = \frac{\zeta_s}{s_{min}} \left(\frac{s}{s_{min}} \right)^{-(\zeta_s+1)}, \qquad (1)$$

where the distribution is normalized in the interval $[s_{min}, \infty)$. There are two free parameters s_{min} and ζ_s. This crossover size s_{min} is chosen such that it minimizes the Kolmogorov-Smirnov (KS) statistic D, which is the distance between the CDF of the empirical data $P_e(s)$ and that of the fitted model $P_f(s)$, i.e.

$$D = \max_{s \geq s_{min}} |P_e(s) - P_f(s)|.$$

Using this procedure we estimate ζ_s and s_{min} for the years 1991- 2005 as shown in Table I. The values of ζ_s computed in each year range from 0.78 to 1.36 and average $\bar{\zeta}_s = 1.09 \pm 0.04$. If indeed these are power laws this is consistent with Zipf's law. But of course, merely

[7] Previous work on the size distribution of mutual funds by Gabaix et al. [5, 6, 19] argued for a power law while we argue here for a log-normal.
[8] In extreme value theory a power law is defined as any function that in the limit $s \to \infty$ can be written $p(s) = g(s)s^{-(\zeta_s+1)}$ where $g(s)$ is a slowly varying function. This means it satisfies $\lim_{s \to \infty} g(ts)/g(s) = C$ for any $t > 0$, where C is a positive constant. The test for power laws in reference [20] is too strong in the sense that it assumes that there exists an s_0 such that for $s > s_0$, $g(s)$ is constant.

computing an exponent and getting a low value does not mean that the distribution is actually a power law.

To test the power law hypothesis more rigorously we follow the Monte Carlo method utilized by Clauset et al. Assuming independence, for each year we generate $10,000$ synthetic data sets, each drawn from a power law with the empirically measured values of s_{min} and ζ_s. For each data-set we calculate the KS statistic to its best fit. The p-value is the fraction of the data sets for which the KS statistic to its own best fit is larger than the KS statistic for the empirical data and its best fit.

The results are summarized in Table I. The power law hypothesis is rejected with two standard deviations or more in six of the years and rejected at one standard deviation or more in twelve of the years (there are fifteen in total). Furthermore there is a general pattern that as time progresses the rejection of the hypothesis becomes stronger. We suspect that this is because of the increase in the number of equity funds. As can be seen in Table I, the total number of equity funds increases roughly linearly in time, and the number in the upper tail N_{tail} also increases.

We conclude that the power law tail hypothesis is questionable but cannot be unequivocally rejected in every year. Stronger evidence against it comes from comparison to a log-normal, as done in the next section.

IV. IS THE TAIL LOG-NORMAL?

A visual comparison between the two hypotheses can be made by looking at the Quantile Quantile (QQ) plots for the empirical data compared to each of the two hypotheses. In a QQ-plot we plot the quantiles of one distribution as the x-axis and the other's as the y-axis. If the two distributions are the same then we expect the points to fall on a straight line. Figure 2 compares the two hypotheses, making it clear that the log-normal is a much better fit than the power law. For the log-normal QQ plot most of the large values in the distribution fall on the dashed line corresponding to a log-normal distribution, though the very largest values are somewhat above the dashed line. This says that the empirical distribution decays slightly faster than a log-normal. There are two possible interpretations of this result: Either this is a statistical fluctuation or the true distribution really has slightly thinner tails than a log-normal. In any case, since a log-normal decays faster than a power law, it strongly suggests that the power law hypothesis is incorrect and the log-normal distribution is a better approximation.

A more quantitative method to address the question of which hypothesis better describes the data is to compare the likelihood of the observation in both hypotheses [20]. We define the likelihood for the tail of the distribution to be
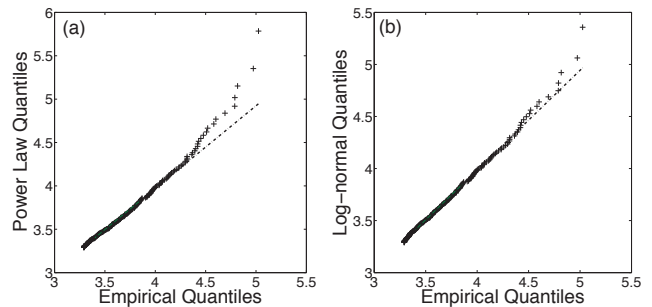
$$L = \prod_{s_j \geq s_{min}} p(s_j).$$

FIG. 2. A Quantile-Quantile (QQ) plot for the upper tail of the size distribution of equity funds. The quantiles are the base ten logarithm of the fund size, in millions of dollars. The empirical quantiles are calculated from the size distribution of funds existing at the end of the year 1998. The empirical data were truncated from below such that only funds with size $s \geq s_{min}$ were included in the calculation of the quantiles. (a) A QQ-plot with the empirical quantiles as the x-axis and the quantiles for the best fit power law as the y-axis. The power law fit for the data was done using the maximum likelihood described in Section III, yielding $s_{min} = 1945$ and $\alpha = 1.107$. (b) A QQ-plot with the empirical quantiles as the x-axis and the quantiles for the best fit log-normal as the y-axis, with the same s_{min} as in (a). The log-normal fit for the data was done used the maximum likelihood estimation given s_{min} (2) yielding $\mu = 2.34$ and $\sigma = 2.5$.

We define the power law likelihood as $L_{PL} = \prod_{s_j \geq s_{min}} p_{PL}(s_j)$ with the probability density of the power law tail given by (1). The lognormal likelihood is defined as $L_{LN} = \prod_{s_j \geq s_{min}} p_{LN}(s_j)$ with the probability density of the lognormal tail given by

$$p_{LN}(s) = \frac{p(s)}{1 - P(s_{min})} \tag{2}$$
$$= \frac{\sqrt{2}}{s\sqrt{\pi}\sigma} \left[\text{erfc} \left(\frac{\ln s_{min} - \mu}{\sqrt{2}\sigma} \right) \right]^{-1} \exp \left[-\frac{(\ln s - \mu)^2}{2\sigma^2} \right].$$

The more probable that the empirical sample is drawn from a given distribution, the larger the likelihood for that set of observations. The ratio indicates which distribution the data are more likely drawn from. We define the log likelihood ratio as

$$\mathcal{R} = \ln \left(\frac{L_{PL}}{L_{LN}} \right). \tag{3}$$

For each of the years 1991 to 2005 we computed the maximum likelihood estimators for both the power law fit and the log-normal fit to the tail, as explained above and in Section III. Using the fit parameters, the log likelihood ratio was computed and the results are summarized graphically in Figure 3 and in Table I. The ratio is always negative, indicating that the likelihood for the log-normal hypothesis is greater than that of the power law hypothesis in every year. It seems clear that tails of the mutual fund data are much better described by a log-normal than by a power law.

variable	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	mean	std
$\mathcal{R}$	-0.50	-1.35	-1.49	-1.71	-3.29	-18.42	-2.25	-1.29	-6.57	-4.96	-2.63	-2.95	-2.00	-1.05	-0.99	-3.43	4.45
N	372	1069	1509	2194	2699	3300	4253	4885	5363	5914	6607	7102	7794	8457	8845	-	-
E[s] (mn)	810	385	480	398	448	527	559	619	748	635	481	335	425	458	474	519	134
Std[s] (bn)	1.98	0.99	1.7	1.66	1.68	2.41	2.82	3.38	4.05	3.37	2.69	1.87	2.45	2.64	2.65	2.42	0.8
E[ω]	5.58	4.40	4.40	3.86	3.86	3.91	3.84	3.85	4.06	3.97	3.60	3.37	3.55	3.51	3.59	3.96	0.54
Std[ω]	1.51	1.98	2.09	2.43	2.50	2.46	2.50	2.51	2.46	2.45	2.63	2.42	2.49	2.59	2.50	2.34	0.29
ζ_s	1.33	1.36	1.19	1.15	1.11	0.78	1.08	1.10	0.95	0.97	1.01	1.07	1.07	1.10	1.14	1.09	0.14
s_{min}	955	800	695	708	877	182	1494	1945	1147	903	728	836	868	1085	1383	974	408
N_{tail}	81	129	232	256	280	1067	290	283	557	662	717	494	652	630	550	-	-
p-value	0.58	0.48	0.02	0.45	0.07	0	0.01	0.11	$5\,10^{-4}$	0.04	0.03	0.07	0.08	0.08	0.15	0.15	0.19

TABLE I. Table of monthly parameter values for equity funds defined such that the portfolio contains a fraction of at least 80% stocks. The values for each of the monthly parameters (rows) were calculated for each year (columns). The mean and standard deviation are evaluated for the monthly values in each year.
$\mathcal{R}$ - the base 10 log likelihood ratio of a power law fit relative to a log-normal fit as given by equation (3). A negative value of $\mathcal{R}$ indicates that the log-normal hypothesis is a likelier description than a power law. For all years the value is negative meaning that the log-normal distribution is more likely.
N - the number of equity funds existing at the end of each year.
$E[\omega]$ - the mean log size of funds existing at the end of each year.
$Std[\omega]$ - the standard deviation of log sizes for funds existing at the end of each year.
$E[s]$ - the mean size (in millions) of funds existing at the end of each year.
$Std[s]$ - the standard deviation of sizes (in billions) for funds existing at the end of each year.
ζ_s - the power law tail exponent (1).
s_{min} - the lower tail cutoff (in millions of dollars) above which we fit a power law (1).
N_{tail} - the number of equity funds belonging to the upper tail s.t. $s \geq s_{min}$.
p-value - the probability of obtaining a goodness of fit at least as bad as the one calculated for the empirical data, under the null hypothesis of a power law upper tail.
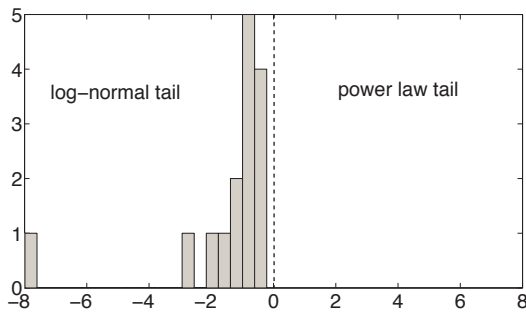
FIG. 3. A histogram of the base 10 log likelihood ratios $\mathcal{R}$ computed using (3) for each of the years 1991 to 2005. A negative log likelihood ratio implies that it is more likely that the empirical distribution is log-normal then a power law. The log likelihood ratio is negative in every year, in several cases strongly so.

V. IMPLICATIONS OF LOG-NORMALITY

The log-normal nature of the size distribution has important implications on the role investor behavior plays in the mutual fund industry. Is the size distribution of mutual funds, i.e. the concentration of assets, determined through investor choice or is it just a consequence of the random nature of the market? In a companion paper [21] we propose that the size distribution can be explained by a simple random process model. This model, characterizing the entry, exit and growth of mutual funds as a random process, is based on market efficiency, which dictates that fund performance is size independent and fund growth is essentially random. This model provides a good explanation of the concentration of assets, suggesting that other effects, such as transaction costs or the behavioral aspects of investor choice, play a smaller role.

The fact that the fund distribution is a log-normal is interesting because, as we argue in the companion paper, this indicates a very slow convergence toward equilibrium. There we find a time-dependent solution for the underlying random process of mutual fund entry, exit, and growth, and show that the size distribution evolves from a log-normal towards a Zipf power law distribution. However, the relaxation to the steady-state solution is extremely slow, with time scales on the order of a century or more. Given that the mutual fund industry is still young, the distribution remains in its non-equilibrium state as a log-normal. Furthermore, given that the properties of the entry and exit processes are not stable over long periods of time, the non-equilibrium log-normal state will very likely persist indefinitely.

VI. CONCLUSIONS

We have shown in unequivocal terms that the mutual fund size distribution is much closer to a log-normal than to a power law. Thus, while the distribution is concentrated, it is not nearly as concentrated as it might be. Among other things this suggests that that the power law distribution observed for trading volume by Gopikrishnan et al. [22] cannot be explained based on a power law distribution for funds. The companion paper discussed in the previous section [21] constructs a theory that explains the log-normality based on the random nature of the mutual fund entry, exit and growth, and the very long-time scales required for convergence to the steady-state power law solution.

[1] G. Corsetti, P. A. Pesenti, and N. Roubini, SSRN eLibrary(2001).

[2] J. P. Embrechts, C. Kluppelberg, and T. Mikosch, *Modeling Extremal Events for Insurance and Finance* (Springer-Verlag, Berlin, 1997).

[3] M. Levy, S. Solomon, and G. Ram, International Journal of Modern Physics C **7**, 65 (1996).

[4] S. Solomon and P. Richmond, Physica A Statistical Mechanics and its Applications **299**, 188 (Oct. 2001), arXiv:cond-mat/0102423.

[5] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, Nature **423**, 267 (May 2003).

[6] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, The Quarterly Journal of Economics **121**, 461 (May 2006), available at http://ideas.repec.org/a/tpr/qjecon/v121y2006i2p461-504.html.

[7] Simon, Herbert A. and Bonini, Charles P., The American Economic Review **48**, 607 (Sep 1958), ISSN 0002-8282.

[8] M. H. R. Stanley, S. V. Buldyrev, S. Havlin, R. N. Mantegna, M. A. Salinger, and H. E. Stanley, Economics Letters **49**, 453 (Oct 1995).

[9] I. Y. and S. H. A., *Skew Distribution and the Size of Business Firms* (North Holland, 1977).

[10] M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger, and H. E. Stanley, Nature **379**, 804 (Feb. 1996).

[11] L. A. N. Amaral, S. V. Buldyrev, S. Havlin, P. Maass, M. A. Salinger, H. E. Stanley, and M. H. R. Stanley, Physica A **244**, 1 (Oct 1997).

[12] G. Bottazzi and A. Secchi, Review of Industrial Organization **23** (Dec 2003).

[13] G. Dosi, *Statistical Regularities in the Evolution of Industries. A Guide through some Evidence and Challenges for the Theory*, LEM Papers Series 2005/17 (Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy, 2005) available at http://ideas.repec.org/p/ssa/lemwps/2005-17.html.

[14] R. L. Axtell, Science **293**, 1818 (Sep 2001).

[15] R. Gibrat, "Les inegalites economiques," (1931).

[16] H. A. Simon, Biometrika **42**, 425 (Dec 1955).

[17] B. Mandelbrot, The Journal of Business **36**, 394 (Oct 1963).

[18] J. Sutton, Journal of Economic Literature **35**, 40 (Mar 1997), ISSN 0022-0515.

[19] X. Gabaix, R. Ramalho, and J. Reuter, *Power laws and mutual fund dynamics*, Tech. Rep. (Massachusetts Institute of Technology, 2003).

[20] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Reviews(2007).

[21] Y. Schwarzkopf and J. D. Farmer, *What drives mutual fund asset concentration?*, Tech. Rep. (Santa Fe Institute, 2010) http://ssrn.com/abstract=1173046.

[22] P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley, Phys. Rev. E **62**, R4493 (Oct 2000).