New composite distributions for modeling industrial income and wealth per employee

Martin Wiegand, Saralees Nadarajah

November 27, 2018

Abstract

Forbes Magazine offers an annual list of the 2000 largest publicly traded companies, shedding light on four different measurements: Sales, profits, market value and assets held. Soriano-Hernández, del Castillo-Mussot, Campirán-Chávez and Montemayor-Aldrete [Physica A, 471, 733-749, 2017] modeled these wealth metrics using composite distributions made up of two parts. In this note, we introduce different composite distributions to more accurately describe the spread of these wealth metrics.

Index terms— Composite distribution, Forbes Global 2000, Income, Wealth

1 Introduction

In this note, we investigate the Global 2000 data set, compiled by Forbes Magazine. It features the 2000 largest companies, with their most important financial indicators, namely annual profits, sales, market value and assets along with employee count and a ranking system based on a combined value.

An analysis was published by Soriano-Hernández et al. [14], where two part models were introduced along with a number of different distribution combinations to predict the percentage of companies below a certain wealth threshold. The tail distribution remained a Pareto distribution of type 1 for all estimations, combined with either a log-normal, gamma or exponential distribution modeling the body part of the sample. These distributions were chosen on the grounds of previous successful modeling in finance [15], [6], [7] or modeling of gas propagation in physics [4].

The basic principle was to divide the data into two parts, and introduce a partial distribution for each part. Both distributions would then be connected by a hard cut-off, leading to a non-continuous, abrupt transition. Formally, the probability density function (PDF) and the cumulative distribution function (CDF) of the composite model can be specified by

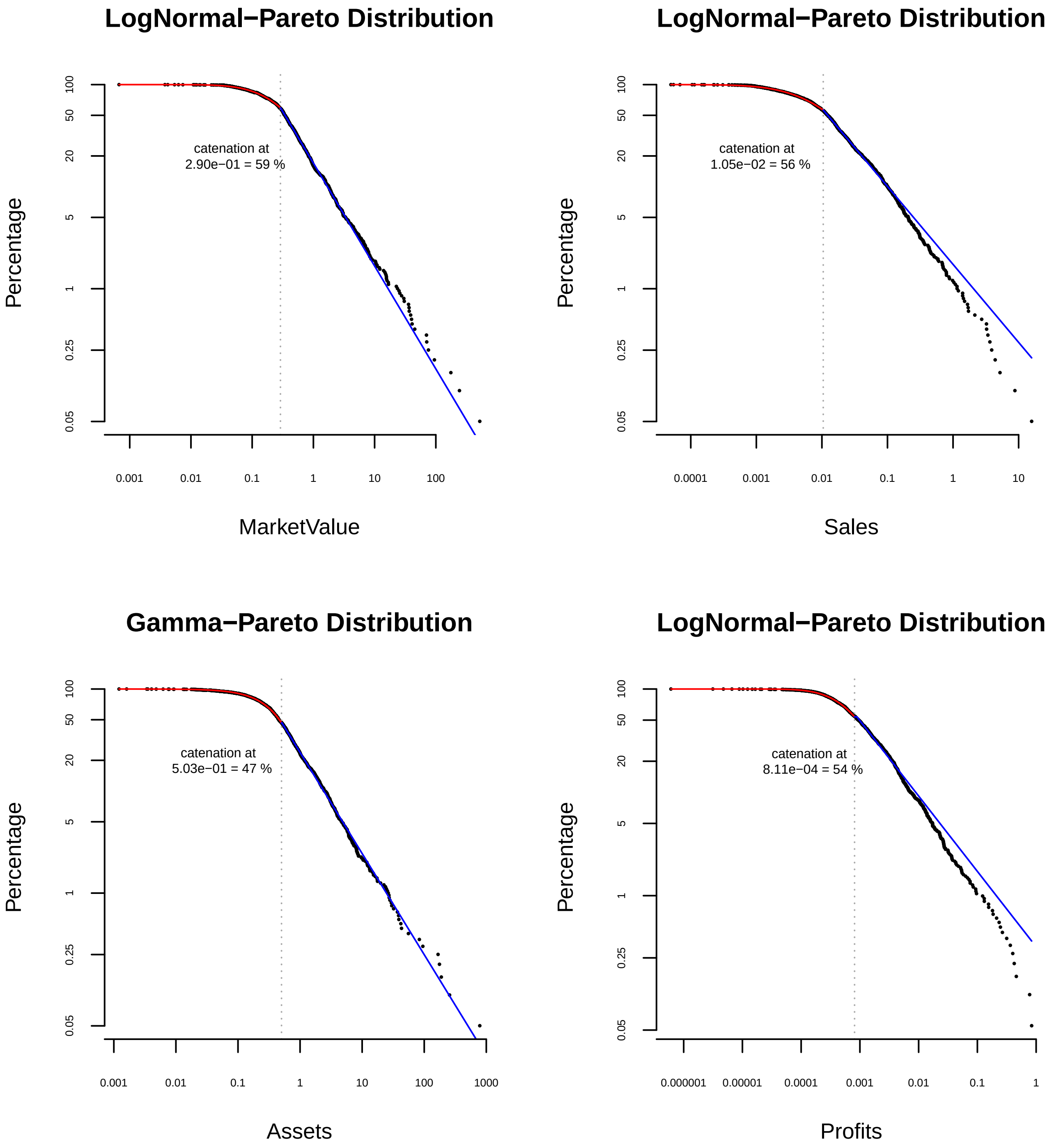
$$f(x) = \begin{cases} f_1(x), & \textit{if } x \leq \theta, \\ \left[1 - F_1(\theta)\right] f_2(x), & \textit{if } \theta < x, \end{cases}$$

and

$$F(x) = \begin{cases} F_1(x), & \textit{if } x \leq \theta, \\ F_2(x) + F_1(\theta) - F_1(\theta) F_2(x), & \textit{if } \theta < x, \end{cases}$$

respectively.

Rather than modeling the values provided directly, it was decided that a quotient of a metric and the employee count would be more expressive, giving insight into how much returns the employees generate. This led to the assumption that all observed businesses can be divided into two categories, depending on the number workforce employed. The first category would be companies in retails like Wal-Mart, which have to rely on large numbers of employees for their services. On the other hand we have companies like Apple, which due to the nature of their products and production processes can perform with comparatively low employee number in relation to their revenue.



LogNormal−Pareto Distribution

LogNormal−Pareto Distribution

Gamma−Pareto Distribution

LogNormal−Pareto Distribution

Figure 1: The four metrics with the respective best fit two part models.

The focus on the application of this model is less in fitting a PDF or CDF to the data, but in a decreasing percentage function $100\left[1-\widehat{F}(x)\right]$ to describe the amount of companies below a certain wealth metric. In this note, we introduce a new approach which suggests a third subpopulation. We then show that it provides a considerable improvement in fits over the old approach. We verify our results with a number of error measures reflecting the goodness of fit and comparative plots, visualising the areas of greatest improvement.

2 Composite models

Contrary to the previously proposed model, we argue that a distinct third population group is present, leading us to the construction of a three part composite model. The best fitting two part models in Figure 1 do not entirely capture the proper curvature of the percentage function. More strikingly, they also differ considerably in the tails. While the parts closer to the catenation point are still somewhat adequately captured, higher values stray further away. This is especially evident in the profit and sales plots, where we hypothesise that a third section at around 0.1 billion and 0.05 billion could diminish this deviation.

Furthermore, we like to introduce a smoother variant of the composite model, which in its two part form has been introduced by Bakar and Nadarajah [2]. The PDF and CDF are provided below for arbitrary distributions with PDFs f_1, f_2 and CDFs F_1, F_2 merged at point $\theta \in \mathbb{R}$ with weight $\Psi \in \mathbb{R}^+$:

$$f(x) = \begin{cases} \frac{1}{1+\Phi} \frac{f_1(x)}{F_1(\theta)}, & \textit{if } x \leq \theta, \\ \frac{\Phi}{1+\Phi} \frac{f_2(x)}{1-F_2(\theta)}, & \textit{if } \theta < x, \end{cases}$$

and

$$F(x) = \begin{cases} \frac{1}{1+\Phi} \frac{F_1(x)}{F_1(\theta)}, & \textit{if } x \leq \theta, \\ \frac{\Phi}{1+\Phi} \left[1 + \Phi \frac{F_2(x)-F_2(\theta)}{1-F_2(\theta)}\right], & \textit{if } \theta < x. \end{cases}$$

We now expand this model by a third partial distribution with PDF f_3 and CDF F_3. Additionally we now mark $\theta_1 < \theta_2$ as the catenation points between the composites and Ψ, Θ as weights. This yields

$$f(x) = \zeta \begin{cases} \frac{f_1(x)}{F_2(\theta_1)}, & \textit{if } x \leq \theta_1, \\ \Phi \frac{f_2(x)}{F_2(\theta_2)-F_2(\theta_1)}, & \textit{if } \theta_1 < x \leq \theta_2, \\ \Psi \frac{f_3(x)}{1-F_3(\theta_2)}, & \textit{if } \theta_2 < x \end{cases} \tag{1}$$

and

$$F(x) = \zeta \begin{cases} \frac{F_1(x)}{F_1(\theta_1)}, & \textit{if } x \leq \theta_1, \\ 1 + \Phi \frac{F_2(x)-F_2(\theta_1)}{F_2(\theta_2)-F_2(\theta_1)}, & \textit{if } \theta_1 < x \leq \theta_2, \\ 1 + \Phi + \Psi \frac{F_3(x)-F_3(\theta_2)}{1-F_3(\theta_2)}, & \textit{if } \theta_2 < x. \end{cases}$$

For convenience, we introduce $\zeta = \frac{1}{1+\Phi+\Psi}$ as scaling parameter. Soriano-Hernández et al. [14] have proposed most prominently the gamma and log-normal distributions for the body (f_1, F_1) and a Pareto type 1 distribution for the tail (f_2, F_2). We mostly agree with the choice of distributions being used for the body component, albeit introducing a beta Weibull (beta Wb) distribution, which we will employ for the body as well as the mid proportion of the data. The beta Wb distribution due to Lee et al. [11] has the PDF and CDF specified by

$$f_{\lambda,k,\alpha,\beta}(x) = k\lambda^k x^{k-1} \exp\left[-\beta(\lambda x)^k\right] \left\{1 - \exp\left[-\beta(\lambda x)^k\right]\right\}^{\alpha-1}$$

and

$$F_{\lambda,k,\alpha,\beta}(x) = I_{1-\exp[-\beta(\lambda x)^k]}(\alpha, \beta),$$

respectively, for $x > 0$, $\lambda > 0$, $k > 0$, $\alpha > 0$ and $\beta > 0$, where $B(a, b)$ and $I_x(a, b)$ are the beta function and incomplete beta function ratio defined by

$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt$$

and

$$I_x(a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1}(1-t)^{b-1} dt,$$

respectively, for $0 < x < 1$, $a > 0$ and $b > 0$.

While the Pareto Type-I distribution has been commonly used to describe the distribution of wealth among the inhabitants of a country or larger region, especially the more prosperous ones, we wish to transfer this approach to the wealth spread among international companies. However we believe that the general shape of the type I distribution cannot accurately capture the most extreme points of the tail. Asymptotic arguments such as the Pickands-Balkema-de Haan theorem mandate the shift to the generalised Pareto distribution of higher order, offering a better fit through its more flexible shape. This argument has already found its way into application (see Degen and Embrechts [5] or Kleiber and Kotz [10]) and it seems sensible to us to update the preexisting approach.

3 Model comparison

We are now left with three body, two midsection and one tail distribution of interest to us. The next step is to fit all six different section combinations onto the four wealth metric data sets and compare the resulting error measures for previous and current approaches. In addition to the squared and absolute aggregate errors, we investigate a number of variations on the Akaike information criterion (AIC) due to Akaike [1], the Kolmogorov-Smirnov (KS) test statistic (here the maximum deviation of the percentage function) and the Anderson-Darling (AD) statistic. Specifically, we use the Bayesian information criterion (BIC) due to Schwarz [13], consistent Akaike information criterion (CAIC) due to Bozdogan [3], corrected Akaike information criterion (AICc) due to Hurvich and Tsai [9] and Hannan-Quinn criterion (HQC) due to Hannan and Quinn [8]. The results can be found in Figure 2.

To ensure the optimal fit for the model and to cut down on computation time, we first computed maximum likelihood estimates (MLEs) for each of the partial distributions and used the resulting parameter estimates as initial values for the MLEs of the composite model. The MLEs obtained in this way were used to compute on the one hand the AIC, BIC, etc measures, and secondly serve as initial values for the minimization of the deviation between the modeled percentage function and empirical percentages (for example, proportion of the data sample below a given value). All computations were performed using the R software [12]. The optimization function in the R software used to find the MLEs is shown in the last column of Figure 2.

If we look into the percentage based measures (for example, squared and absolute aggregate errors), it becomes evident that the tri-composite model provides a better fit, with up to almost 60 percent reduction in squared error and up to 50 percent reduction in absolute error. For most instances the choice of body distribution only plays a minor part, as long as it is sufficiently heavy-tailed (gamma, log-normal, beta Wb distributions tend to perform mostly similarly). We omitted the exponentially distributed body-variant, since it consistently provided the highest error margins in our fitting process. With the introduction of a mid-section and the inherent improved flexibility of the composite distribution, we anticipated lower error measures. Nonetheless by the amount the measures have changed, we see our hypothesis of a third distinct subgroup within the company samples verified.

The density based measures draw a similar picture. What strikes us, is that not only do the composite models possess a lower AIC, BIC etc. value than the hard cut-off model, but the values are within a very tight grouping. We attribute this to the fact that $f(x)$ as described in (1) consists of disjunct partitions, which do not need to be necessarily continuous at the catenation points.

Furthermore, the tail distribution remains identical and the mid distributions only differ slightly from one another. Since we developed the model in order to more accurately determine the percentage function as characterized in [14], we give direct error measures of the percentage deviation more weight.

Model	Metric	Partial Distribution			Optimized via Percentage Function						Optimized via Density					
		Body	Mid	Tail	Absolute	Squared	Rel. Abs.	Rel. Sq.	KS	AD	AIC	BIC	AICc	HQC	CAIC	Function
Two part hard cut-off	Market Value	LogNorm	-	Pareto1	673.20	347.18	0.338	0.174	1.27	1999	7416	7438	7416	7424	7442	optim
		Gamma	-	Pareto1	825.37	690.55	0.414	0.346	2.43	1943	7853	7876	7853	7572	7880	optim
		Expon	-	Pareto1	1593.89	2825.00	0.800	1.417	4.16	1861	10287	10303	10287	10293	10306	optim
	Sales	LogNorm	-	Pareto1	861.61	530.10	0.276	0.266	1.37	1934	4062	4085	4062	4070	4089	optim
		Gamma	-	Pareto1	966.82	1015.43	0.485	0.509	2.86	1875	-2067	-2045	-2067	-2059	-2041	optim
		Expon		Pareto1	1148.76	1058.17	0.576	0.531	2.46	1891	7524	7540	7524	7530	7543	optim
	Assets	LogNorm	-	Pareto1	1189.89	1095.36	0.597	0.550	1.86	2089	9909	9932	9909	9917	9936	optim
		Gamma	-	Pareto1	630.88	309.96	0.317	0.156	0.99	2015	6682	6705	6682	6690	6709	optim
		Expon	-	Pareto1	1889.23	2436.54	0.948	1.223	2.92	1871	11410	11426	11410	11416	11429	optim
	Profits	LogNorm	-	Pareto1	1041.46	801.72	0.574	0.442	1.43	1834	-4150	-4128	-4150	-4142	-4124	optim
		Gamma	-	Pareto1	1345.97	1396.64	0.742	0.769	2.29	1741	501	523	501	509	527	optim
		Expon	-	Pareto1	2375.88	5842.95	1.309	3.219	4.69	1637	2953	2970	2953	2959	2973	optim
Three part composite model	Market Value	LogNorm	Pareto4	Pareto4	595.08	259.95	0.299	0.130	1.05	2006	2456	2534	2456	2485	2548	optim
		Gamma	Pareto4	Pareto4	550.04	252.90	0.276	0.127	1.33	1983	2455	2534	2455	2484	2548	optim
		BetaWb	Pareto4	Pareto4	459.11	173.07	0.230	0.087	0.95	2162	2446	2536	2446	2479	2552	optim
		LogNorm	BetaWb	Pareto4	517.76	224.03	0.260	0.112	1.17	1998	2461	2539	2461	2490	2553	optim
		Gamma	BetaWb	Pareto4	497.78	202.54	0.250	0.102	1.01	1999	2467	2546	2468	2496	2560	optim
		BetaWb	BetaWb	Pareto4	593.58	231.47	0.298	0.116	0.92	2011	2505	2595	2506	2538	2611	optim
	Sales	LogNorm	Pareto4	Pareto4	493.65	190.27	0.248	0.095	0.84	2019	-9878	-9800	-9878	-9849	-9786	optim
		Gamma	Pareto4	Pareto4	514.17	195.64	0.258	0.098	0.96	2008	-10232	-10153	-10231	-10203	-10139	optim
		BetaWb	Pareto4	Pareto4	482.84	186.34	0.242	0.093	0.90	2101	-9740	-9650	-9739	-9707	-9634	optim
		LogNorm	BetaWb	Pareto4	515.18	223.08	0.258	0.112	1.17	1986	-9891	-9813	-9891	-9862	-9799	optim+nlm
		Gamma	BetaWb	Pareto4	426.80	134.63	0.214	0.068	0.72	2022	-9914	-9836	-9914	-9885	-9822	optim+nlm
		BetaWb	BetaWb	Pareto4	502.86	198.86	0.252	0.100	1.03	1982	-9922	-9833	-9922	-9889	-9817	optim+nlm
	Assets	LogNorm	Pareto4	Pareto4	340.82	93.27	0.171	0.047	0.65	2016	3813	3891	3813	3841	3905	optim
		Gamma	Pareto4	Pareto4	349.24	93.86	0.175	0.047	0.58	2017	3784	3862	3784	3813	3876	optim
		BetaWb	Pareto4	Pareto4	353.57	98.92	0.177	0.050	0.63	2002	3824	3914	3824	3857	3930	optim
		LogNorm	BetaWb	Pareto4	348.30	90.91	0.175	0.046	0.57	2013	3811	3890	3812	3840	3904	optim
		Gamma	BetaWb	Pareto4	337.12	88.60	0.169	0.044	0.59	2008	3796	3875	3797	3825	3889	optim
		BetaWb	BetaWb	Pareto4	337.00	89.79	0.169	0.045	0.67	2006	3802	3891	3802	3834	3907	optim
	Profits	LogNorm	Pareto4	Pareto4	547.78	248.90	0.302	0.137	1.03	1926	-18081	-18004	-18080	-18052	-17990	optim
		Gamma	Pareto4	Pareto4	578.11	326.53	0.319	0.180	1.22	1802	-18107	-18030	-18107	-18079	-18016	optim
		BetaWb	Pareto4	Pareto4	602.18	290.41	0.332	0.160	1.07	1857	-18032	-17944	-18032	-18000	-17928	optim
		LogNorm	BetaWb	Pareto4	474.65	193.04	0.262	0.106	0.85	1910	-17998	-17920	-17997	-17969	-17906	optim
		Gamma	BetaWb	Pareto4	432.23	161.55	0.238	0.089	0.85	1903	-17881	-17803	-17880	-17847	-17789	optim
		BetaWb	BetaWb	Pareto4	596.69	286.05	0.329	0.158	1.07	1862	-17831	-17743	-17831	-17799	-17727	optim

Figure 2: Error measure compilation for composite models, loop tolerance=10^{-8} for percentage function fitting.

Next we turn our attention to the comparative plots in Figure 3. All variants of the newer model indeed capture the distribution of the wealth metrics more closely. While the logarithmic nature of the plots overemphasizes the tail deviations, where especially sales and profits data can seen with increased accuracy, the greatest improvement has been achieved in the body and mid-sections of the data (for example, when the sample size is less than 1500). We therefore have provided a plot of the absolute error propagation throughout the data sample, with the previous model in gray and our approach in red once more. See Figure 4.

As we can see for the sales and assets metric, the sharp error peaks through the entire sample have been trimmed considerably (Fig. 4). It seems that the improvement of the 3-part model streches through all sections for those two particular attributes. The samples of the market values and especially profits draw a somewhat different picture. Most of the improvement of the new model over the previous two part approach is accumulated within the last section, or more accurately the third tercile.

The improvement in the lower two parts are due to higher flexibility of the composite distribution as a whole and a more accurate choice of partial distributions, based on empirical findings with a large number of parametric functions. Contrarily the reason the tail distribution achieves a superior fit lies within the expected asymptotic convergence of the population tail towards a generalized pareto distribution [5]. While the historically applied Pareto type-I distribution is an appropriate starting point, the shape remains too restrictive to provide an accurate fit. It is expected as stated earlier that the sample tail converges against a Pareto distribution of high order, as the tail threshold is moved higher towards the last sample point. The introduction of the additional section thus concentrates this partial distribution better on the respective data, e.g. a higher quantile of samples, to capture this effect.

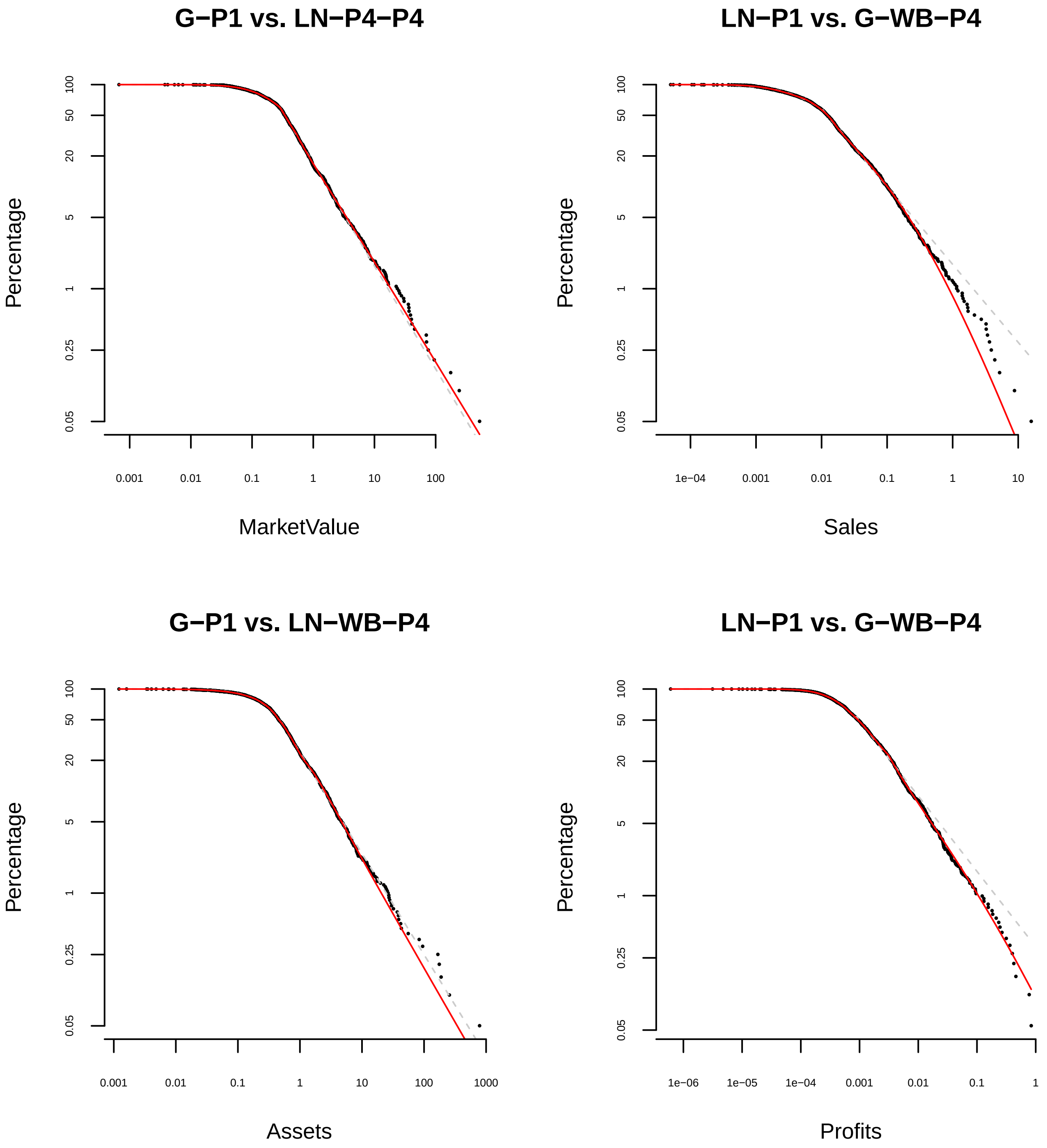


Figure 3: A comparative log-scale plot between the best fit variants of old (grey) and new (red) models.

In figure 4 we have singled out this fact. We can clearly see how the standard Pareto distribution fails to accurately describe the behaviour of the very last data points. Especially in the latter three metrics this circumstance presents itself quite dramatically. Due to the double logarithmic scale which is used, the differences in the higher quantiles are somewhat exaggerated.

We therefore provide a table of the deviation measures to compare both tail modeling approaches numerically.

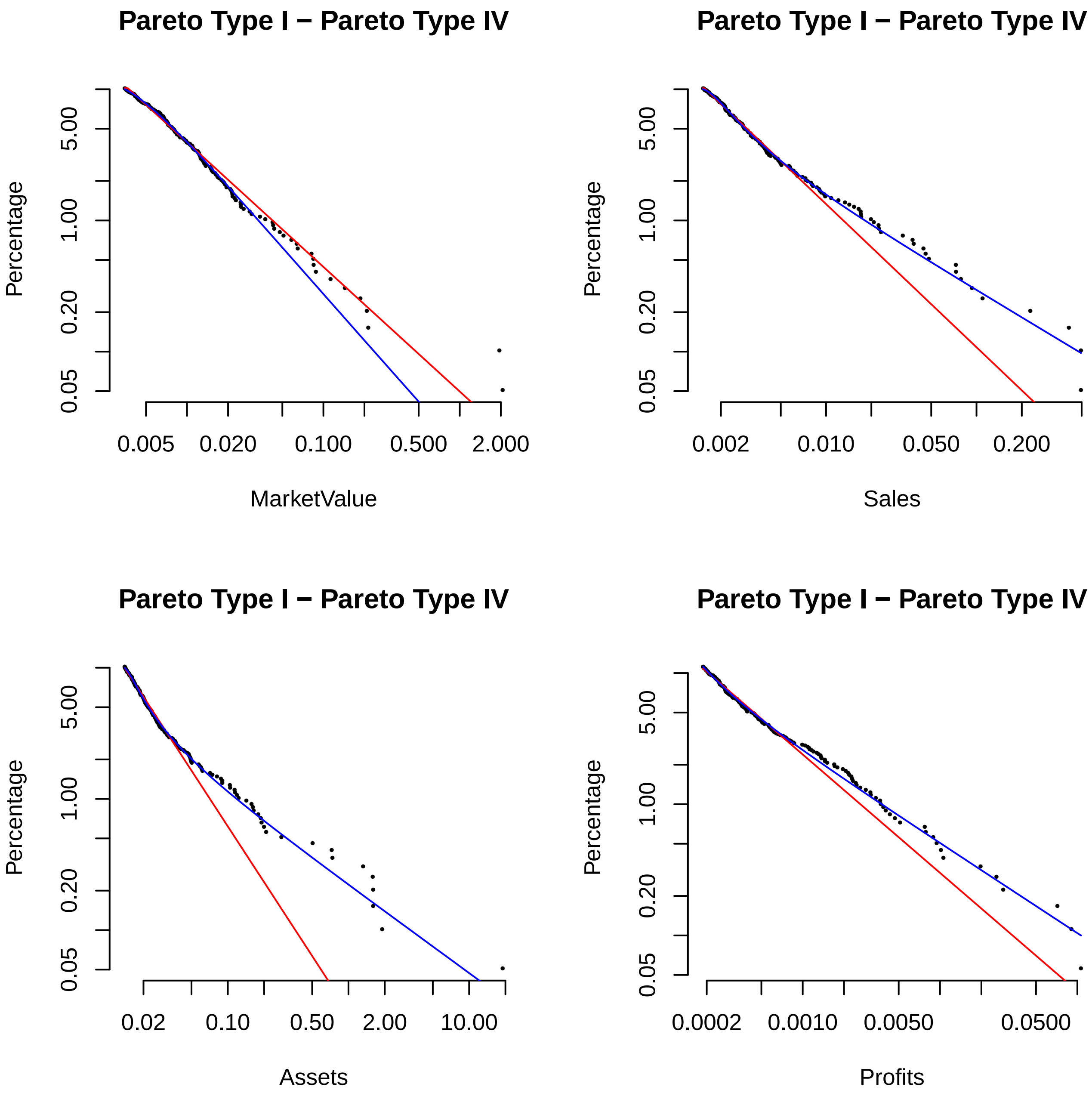


Figure 4: A log-scale plot of the top 10% of the respective samples, with only Pareto-I and Pareto-IV distributions fitted.

In figure 5 we can see the improvement of the type IV distribution over the standard Pareto is considerable. As anticipated we can see how the more general distribution is better suited to capture the extreme tails of all 4 metrics.

	Absolute Error		Squared Error		Kolmogorov - Smirnov	
Pareto Distribution	Type I	Type IV	Type I	Type IV	Type I	Type IV
Assets	14.2098	1.650316	42.37595	15.14181	0.690889	0.241456
Market Value	9.353292	2.781869	34.75241	18.93321	0.530084	0.329805
Profits	15.17699	4.587656	47.47998	23.92047	0.542113	0.343262
Sales	6.682243	1.578278	30.39588	14.12695	0.450729	0.213245

Figure 5: Deviations of the Pareto Type distributions

4 Conclusions

As our initial observations in the original model suggested, a third distinct subgroup in the sample seems likely, and its addition into the composite model greatly enhanced the distribution fit, more precisely the proposed percentage function $100\left[1-\widehat{F}(x)\right]$. Not only did the errors in the newly introduced section drop, but it also allowed a more precise fit for the body and tail sections, which now stretch over a significantly smaller sample.

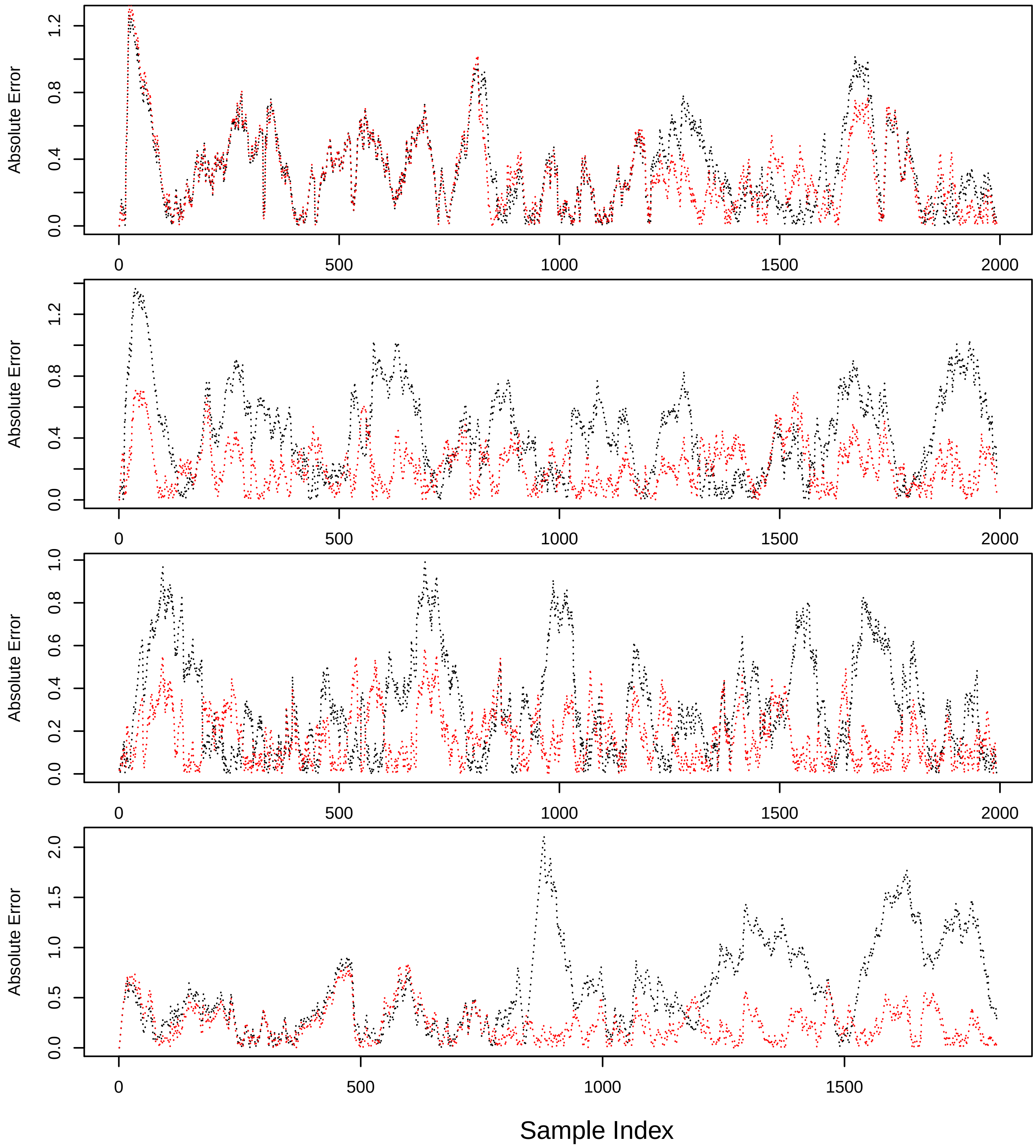


Figure 6: From top to bottom: market value, sales, assets and profits with the absolute error of the two part (black) and three part (red) models.

While the parameter count has considerably swollen up to 16, which leads to a significant increase in computation time, the smooth tri-composite distribution does give a more accurate impression of the wealth distribution between major international companies as all tested error measures verify. In our opinion this handicap is negligible, since the data set is released annually, hence the actual fitting procedure only has to be done sporadically. Furthermore, the multi step optimization approach we used (parameter estimates of section distributions used as initial values) narrows down the search process significantly, such that a practical use of the composite model seems likely to us.

Since we noticed that the actual choice of the body distributions does in most cases not affect the accuracy of the model too greatly, one might go as far as suggesting that for most practical purposes a single model (for example, the gamma-beta Wb-Pareto type 4 model, which has consistently performed well) can be utilized to replace all incarnations of the previous two part model.

Acknowledgements

The authors would also like to thank the Editor and the referee for careful reading and comments which greatly improved the paper.

References

[1] H. Akaike; *A new look at the statistical model identification* IEEE Transactions on Automatic Control, 19, 716-723, 1974.

[2] S.A.A. Bakar, S. Nadarajah; *CompLognormal: An R package for composite lognormal distributions* The R Journal, 5, 2013.

[3] H. Bozdogan; *Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions* Psychometrika, 52, 345-370, 1987.

[4] B.K. Chakrabarti, A. Chakrabarti, S.R. Chakravarty, A. Chatterjee; *Econophysics of Income and Wealth Distributions* Cambridge University Press, 2013.

[5] M. Degen, P. Embrechts; *EVT-based estimation of risk capital and convergence of high quantiles* Advances in Applied Probability, 40, 696-715, 2008.

[6] R. Gibrat; *Les égalités économiques* Paris, 1931.

[7] C. Gini; *Measurement of inequality of incomes* Econometric Journal, 31, 124-126, 1921.

[8] E.J. Hannan, B.G. Quinn; *The determination of the order of an autoregression* Journal of the Royal Statistical Society, B, 41, 190-195, 1979.

[9] C.M. Hurvich, C.-L. Tsai; *Regression and time series model selection in small samples* Biometrika, 76, 297-307, 1989.

[10] C. Kleiber, S. Kotz; *Statistical Size Distribution in Economics and Actuarial Sciences* Wiley, 2003.

[11] C. Lee, F. Famoye, O. Olumolade; *Beta-Weibull distribution: Some properties and applications to censored data* Journal of Modern Applied Statistical Methods, 6, Article 17, 2007.

[12] R Development Core Team; *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2017.

[13] G.E. Schwarz; *Estimating the dimension of a model* Annals of Statistics, 6, 461-464, 1978.

[14] P. Soriano-Hernández, M. del Castillo-Mussot, I. Campirán-Chávez, J.A. Montemayor-Aldrete; *Wealth of the world's richest publicly traded companies per industry and per employee: Gamma, log-normal and Pareto power-law as universal distributions?* Physica A, 471, 733-749, 2017.

[15] P. Soriano-Hernández, M. del Castillo-Mussot, O. Córdoba-Rodrígez, R.M. Mansilla-Corona; *Non-stationary individual and household income of poor, rich and middle classes in Mexico* Physica A, 465, 403-413, 2017.