한국경제학보 제22권 제2호
The Korean Journal of Economics Vol. 22, No. 2 (Spring 2015)

Fitting the Tail of Firm Size Distribution in Korea[*]

Joocheol Kim[**] · Hyun Yeol Kim[***]

Abstract

This study estimates the tail index of firm size distribution in Korea, by fitting its right tail to a scaled Log phase-type distribution. We use the sample of statutory-audited firms, including both financial and non-financial companies. Regardless of the choice of proxy for firm size, our estimator fits the extreme tail of the distribution better than the previous methods, such as the popular Hill estimator. The estimates of the tail index vary across different firm size measures and the economic statuses, indicating that firm sizes do not necessarily follow the Zipf's law. Furthermore, the upper tail of the firm size distribution in Korea becomes thicker during the financial crises in 1997 and 2007–2008. This adds to the evidence that the allocation of resources across firms is affected by financial crises.

KRF Classification : B030904
Keywords : Firm Size Distribution, Tail Index, Pareto Parameter, Log Phase-type Distribution, Hill Estimator

[*] Joocheol Kim is grateful for the support of the National Research Foundation of Korea (NRF-2014S1A5A2A01011100).

[**] Corresponding Author, Associate Professor, School of Economics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea, e-mail: joocheol@yonsei.ac.kr

[**] Graduate Student, School of Economics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea, e-mail: hyunyeolkim@gmail.com

Ⅰ. Introduction

This study aims to add onto the empirical evidence of firm size distribution in South Korea with respect to its upper tail. Examinations of firm size distribution from data are of interest to many researchers from various fields. In industrial organization literature, firm size distribution comes up as the result of the underlying firm dynamics such as, firms' entry, growth, decline and exit. Beginning with Gibrat (1931) which introduces an elementary firm dynamics model that generates firm sizes following random walk when taken logarithms, many economists from the field of industrial organization proposed models that lead to certain theoretical distributions of firm sizes.[1] An accurate examination of firm size distributions is crucial in this literature as many theories are aimed to explain the developing mechanism of observed firm size distributions.

Firm size distributions are also used in the calibration of firm dynamics models in example, macroeconomics, where such models are considered to be the key to explain not only the stylized facts in firm dynamics but also some important macroeconomic features. For Hopenhayn (1992) proposes a seminal dynamic stochastic model of firm entry and exit to an industry in order to account for the high job turnover rates of firms. A recent work by Clementi and Palazzo (2014) suggests that the endogenous firm entry and exit affect the aggregate dynamics such as the dynamics of output, employment, and wage, as they magnify the effect of productivity shock to the economy. In such literature, researchers use the estimated tail index of firm size distributions in order to see if their models generate results similar to the real world.

Firm size distributions are almost always highly skewed to the

[1] See de Wit (2005) for a review.

right; that is, there are large numbers of small firms and small number of big firms. The upper tail of firm size distribution is particularly of our interest for various reasons. According to di Giovanni and Levchenko (2013), for example, the welfare impact of fixed costs of production, export and the entry decision into the export market is very small when the firm sizes follow the Zipf's law; that is, when the tail index of firm size distribution is equal to 1. This is because large firms at the upper tail of the distribution are rarely affected by such fixed costs. If a firm size distribution is flatter, however, smaller firms have more significant effect to the economy; therefore, fixed costs in such economy have more impact on the welfare from international trade.

On the other hand, Kang et al. (2011) use the Pareto exponent as a measure of inequality among the sample firms and relate it with the economic and political disturbance at the corresponding time. According to their interpretation, the inequality among firms is deteriorating if the tail becomes thicker. Especially in many East Asian countries—including South Korea—where a small number of large firms have huge influences on the economy, it is important to examine the firm size distribution correctly with much focus on its upper tail.

Therefore, many researchers have been investigating the upper tail of firm size distributions from observable data using various techniques. For instance, Axtell (2001) and di Giovanni and Levchenko (2013) use both OLS-CDF and OLS-PDF estimators for the United States and cross-country data, respectively. Gabaix and Landier (2008) apply Hill (1975) estimator for estimating the tail index of the United States' firm distribution. Such empirical results, in general, do not reject the hypothesis of the Zipf's law in the upper tail of firm size distributions.

With respect to the case of Korea, Kang et al. (2011) look at the evolution of the Zipf exponent of the firm sizes in 1980-2007. Using the OLS-Rank estimator that is similar to the one used by Fujiwara et al. (2003) and Podobnik et al. (2010), they estimate the Pareto exponent of the distribution of non-financial companies listed on Korean stock markets. Among their findings is that the Pareto exponent increases before the 1997 financial crisis and decreases afterwards.

Bottazzi, Pirino and Tamagni (2013) provide a summary of several recent empirical studies on firm size distribution. They further suggest that the Hill (1975) estimator is most recommended for estimating the tail index among many popular estimators, for the better theoretical properties and for its preferable performance in their Monte Carlo analysis. However, they only consider the property of the Hill point estimates under the assumption that the researchers already know where the power tail starts in the firm size distribution —i.e. the tail threshold. A critical shortfall in the usage of the Hill estimator is that it does not give any suggestion regarding the selection of the tail threshold. In fact, the estimate is extremely sensitive to the choice of the threshold; an inappropriate choice of the threshold may lead to a substantial bias in the estimated value of the tail index.

Recently, Kim and Kim (2015) propose an alternative parametric tail index estimator using a scaled Log phase-type (PH) distribution. Numerical exercises show that the eigenvalue plot suggested by them is more stable and easier to use compared to the Hill counterpart.

In this paper, we examine the firm size distribution in Korea using both the classical Hill plot and the eigenvalue plot proposed by Kim and Kim (2015). The results are consistent with previous works including that of Kang et al. (2011); the firm sizes do not always follow the Zipf's law, and the tail index tends to increase in the period of financial crisis in 1997 and 2008. In addition, the comparison between the eigenvalue plot and the Hill plot show that the eigenvalue plot performs better in general, suggesting that the Log PH distribution is an appropriate choice of parametric model in estimating the tail index of firm size distributions.

Our results impose different implication toward firm size distribution in South Korea in comparison to the previous work of Kang et al. (2011). First, all the yearly data are pooled cross-sectionally in our analysis, whereas Kang et al. (2011) used the panel data. Using panel data naturally includes only the firms that survived throughout the whole sample years; it may have excluded firms that newly came in or exited from the market during that period. Therefore, our sample choice of firms better reflects the firms' distribution for each year as we allow for such entrances and exits of firms that consist the samples. Second, while Kang et al. (2011) examine samples of non-financial firms only, this paper draws similar conclusion from the pooled sample of both financial and non-financial firms. Third, we include the observations from years 2007-2008 and show that the tail of the distribution in the global financial crisis behaves similarly to that in the Asian financial crisis. As Kang et al. (2011) do not cover these years, this paper is providing additional evidence regarding the dynamics in firm size distribution during financial crises. Lastly, we fit the tail part of the distribution to a scaled Log phase-type distribution, while Kang et al. (2011) used OLS-Rank estimator in their analysis.

The paper is organized as follows. In Section 2, we briefly introduce the logic behind the tail index estimator derived from scaled Log phase-type distribution. Section 3 overviews the data I used for the analysis. Section 4 provides the estimation results and their implications. Section 5 concludes the present paper.

Ⅱ. Tail Index Estimator from a Scaled Log PH Distribution

Introduced by Neuts (1975), a phase-type (PH) distribution is the distribution of the time till absorption, denoting by X, in Markov process with m transient states and one absorbing state. m is a finite number, and is called as the phase size. The PH distribution has an initial probability vector $(\alpha,0)$ with α being an m-dimensional row vector such that the sum of all its elements equals to 1, i.e. $\alpha 1 = 1$. Its infinitesimal generator is an $(m+1)\times(m+1)$ matrix, i.e.

$$Q = \begin{pmatrix} T & t \\ 0 & 0 \end{pmatrix}, \qquad t = -T1, \tag{1}$$

where 0 is an m-dimensional row vector of zeros and 1 is an m -dimensional column vector of ones. The subgenerator matrix T is an $m\times m$ matrix with the off-diagonal elements non-negative and the diagonal elements strictly negative. The two parameters for a PH distribution is α and T; if X follows a PH distribution, we denote if by $X \sim PH(\alpha, T)$.

Introducing the PH distribution, Neuts (1975) additionally assumes that this non-singular matrix T is stable, i.e. the real parts of its all eigenvalues are negative. Let $-\eta_T$ be the largest eigenvalue of T. That is, η_T takes the absolute value of the eigenvalue that is the closest to zero. One of the properties of PH distribution is that if $X \sim PH(\alpha, T)$, then its excess loss $X - d|X > d$ also follows a PH distribution, i.e.

$$X - d|X > d \sim PH(\alpha_d, T), \text{ where } \alpha_d = \frac{\alpha e^{dT}}{\alpha e^{dT} 1} \tag{2}$$

for any constant $d > 0$. d is called a threshold.

Derived from a PH distribution, the Log phase-type (LogPH) variable is defined as a random variable that follows a PH distribution when the natural logarithm is taken. Kim and Kim (2015) generalize this ordinary LogPH distribution to a scaled LogPH distribution by multiplying some constant to an ordinary LogPH distributed variable. Let us define $Y=\exp(X-c)=e^{-c}e^{X}$ for some finite constant c, where $X\sim PH(\alpha,T)$. Then it is denoted by $Y\sim LogPH(\alpha,T;-c)$. Shifting of a LogPH distribution by choosing an appropriate value of c allows for a wider support containing values below 1.[2)]

The scaled LogPH distribution has properties resembling those of ordinary LogPH distribution. Suppose $Y=\exp(X-c)=e^{-c}e^{X}\sim LogPH(\alpha,T;-c)$. Among the properties of the scaled LogPH random variable Y is:

$$\lim_{y\to\infty}\frac{\overline{F_Y}(y\lambda)}{\overline{F_Y}(y)}=\lambda^{-\eta_T},\quad \forall\,\lambda>0, \tag{3}$$

where $\overline{F_Y}(y)=1-F_Y(y)$ is the tail function of the distribution of Y.[3)] If we express (3) in other words, the tail function $\overline{F_Y}$ is regularly varying with index $-\eta_T<0$. Hence, the distribution of Y belongs to the maximum domain of attraction of the Frechet distribution. By the Balkema-de Haan-Pickands theorem, the excess loss $Y-d|Y>d$ from such a distribution with a sufficiently large positive threshold d converges to the Generalized Pareto distribution (GPD) with a positive Pareto parameter denoted as ξ.[4)] The shape

[2)] Note that an ordinary phase-type distribution has a limited support of $[1,\infty]$.

[3)] $F_Y(y)$ is the cumulative distribution function of Y.

[4)] The GPD distribution function is given by $G_{\xi,\sigma}(y)=1-(1+\frac{\xi}{\sigma}y)^{-1/\xi},\ y>0$. for some constant $\xi>0$.

parameter ξ taking a positive value corresponds to the heavy tailed distribution with a tail index of $1/\xi$. As we can see in (3), the tail index $1/\xi$ is equal to η_T in this case. Hence, it is enough to estimate the parameter η_T to obtain the Pareto parameter for a scaled LogPH distributed variable.

In addition, the mean excess function of the scaled LogPH distribution converges to a linear form as the threshold gets large. That is,

$$E[Y-d|Y>d] = -d[\alpha_{\log d+c}(\boldsymbol{I}+T)^{-1}t+1], \quad d \geq e^{-c} \qquad (4)$$

$$= -d[\frac{\alpha e^{Tc}e^{T\log d}}{\alpha e^{Tc}e^{T\log d}1}\boldsymbol{I}+T)^{-1}t+1] \sim a \bullet d, \quad d \to \infty, \qquad (5)$$

for some positive number a, where $\alpha_{\log d+c} = \frac{\alpha e^{(\log d+c)T}}{\alpha e^{(\log d+c)T}1}$.[5] This implies that the mean excess function of a scaled LogPH distribution becomes linear as the threshold takes a sufficiently large value, just as that of a GPD.[6]

These theoretical properties imply any heavy-tailed loss distribution can be approximated by a suitably scaled LogPH distribution with some parameter set $(\alpha, T; -c)$, which has a regularly varying distribution with a positive tail index of η_T. Suppose the variable Y follows a LogPH distribution such that $Y \sim LogPH(\alpha, T; -c)$. Then, its log excess loss $\log Y - d|\log Y > d$ follows an ordinary PH distribution with its location parameter 0, i.e.

$$\log Y - d|\log Y > d \sim PH(\alpha_{c+d}, T; 0),$$

$$\text{where } \alpha_{c+d} = \frac{\alpha e^{(c+d)T}}{\alpha e^{(c+d)T}1}, \qquad (6)$$

[5] This is from (2). Replacing d with $\log d + c$ renders the result.

[6] See Kim and Kim (2015) and Ahn et al. (2012) for more detailed proofs.

because of (2). Hence, a standard EMPHT (expectation-maximization for phase-type) algorithm proposed by Asumssen et al. (1996) can be applied to fit a PH distribution to the log excess loss values of the data for any value that the location parameter c takes. In addition, note that the matrix T is invariant to both the parameters d and c. This implies that fitting the log excess loss to a PH distribution will generate the same T and thus η_T for any threshold d one takes.

Kim and Kim (2015) uses this second fact to propose a new way of estimating tail index for a heavy tailed distribution, which we make use of in this paper. The specific procedure to apply this method is as follows:

First, transform the original samples $(y_1,\cdots,y_n)$ by taking logarithms as $x_i = \log y_i, i = 1,\cdots,n$. Second, take the threshold d as the smallest observation from the sample, and obtain excess loss for each, $x_i - d|x_i > d, i = 1,\cdots,n$. Third, we obtain the corresponding estimate of T and η_T with the excess loss data by using the EMPHT-program. The EMPHT-program fits the PH distribution using a so-called EM (expectation-maximization) algorithm, which basically performs maximum likelihood estimation iteratively. In each step of iteration, the program updates the estimate so that the information divergence (the Kullback-Leibler information) becomes smaller in each step; if the number of iteration set by the user is large enough, we expect to have an estimate of T that has converged to a value that maximizes the likelihood function (See Asumssen et al. (1996) for detailed explanation on the EMPHT procedure. Also note that it is enough to assume that the phase size m is equal to 2; see Kim and Kim (2015) for the justification of such a setting). Keeping the estimated η_T from the third step, we then go back to the original sample, eliminate the smallest observation, and repeat the first to third steps. We take all possible values for the excess loss threshold d, by eliminating the samples from smallest to largest one by one, and obtain the resulting $\hat{\eta}_T$ after each process. After eliminating the second largest observation, we can stop and plot the number of deletions $n-n_d$[7] and its corresponding $\hat{\eta}_T^{-1}$. This figure $(n-n_d, \hat{\eta}_T^{-1}(d))$ is used to find the appropriate threshold where the heavy tail begins. It is called an eigenvalue plot, as $\hat{\eta}_T^{-1}$ is the inverse of the eigenvalue of T that is closest to 0. The last step is to look for the range of threshold values where the resulting $\hat{\eta}_T$'s take similar values in the eigenvalue plot, because that is where the (scaled) LogPH distribution is fitting well.[8]

This estimation of the tail index by fitting the distribution to a scaled LogPH distribution has many practical advantages. First, any heavy tailed loss distribution can be approximated by an appropriately scaled LogPH distribution and moreover, there is no burden of estimating the location parameter c when the only estimation target is the tail index. Second, the eigenvalue plot suggests where the tail threshold is. The previous estimation methods including the Hill plot have been chosen rather an arbitrary way of

[7] n is the sample size before any deletions and n_d is the sample size after deletion. As we delete the sample from the smallest to the largest one by one, $n-n_d$ naturally ranges from 0 to $n-1$.

[8] It is important to understand how to read the eigenvalue plot and where to look for its stable area for its practical use. The plot can usually be divided into three parts based on its behavior. First, the plot is unstable at the first segment where the threshold is too small, as the log excess loss of a scaled LogPH distribution becomes convex decreasing for any loss data only when the threshold is sufficiently large enough. The estimated $\widehat{\eta_T}$ is unreliable for this region. Second, the middle part of the plot is rather stable and this is where the tail index is identified. The last part of the plot becomes highly unstable since the observations used in the fitting are too few, and the extreme quantiles suffer from substantial volatility. Hence, one should look at the middle part of an eigenvalue plot where it is stable. If the plot takes different values in that segment, the right end region of the interval is considered to be closer to the true value, as this is the region where the GPD phenomenon takes place.

defining where the power tail begins in the sample distribution[9]. However, the eigenvalue plot suggests a better way of finding the true tail threshold, which is to look at the beginning of the stable area in the mid-range. Third, the eigenvalue plot performs better－that is, a scaled LogPH distribution is often fitted to the data better－than the Hill plot. According to Kim and Kim (2015), the eigenvalue plot is a generalized version of the Hill plot, and it particularly shows better performance than the Hill counterpart when the targeted sample distribution is far from an ordinary Pareto distribution. Indeed, a firm size distribution seems to be the very example of this. In the following sections, we show that the upper tail of firm size distributions is well fitted by a scaled LogPH distribution.

Ⅲ. Data

As mentioned previously, there is no unanimity about how to measure firm sizes; different proxies and measures are used depending on the research purposes. For this study, we examine the distribution of each firms' employment, sales, and assets separately.

We use the annual data obtained from KISVALUE DB provided by Korea Information Service, Inc., which collects the information on statutory audited corporations in Korea. Statutory audited corporations are the firms in certain criteria such that their accounts are required to be audited by the law; thus the database clearly does not represent the entire population of Korean firms.[10] As we are

[9] For example, a typical way of using the Hill plot is to look at the 5－10% upper tail to determine the tail index. This number is not backed by a theory but has been commonly used.

[10] In the standard of 2013, statutory audited firms are among: (i) listed corporations in the stock market, (ii) firms with total assets greater than 10

interested in the right tail of the distribution, however, such a sample selection will hardly affect the conclusion we draw.

In choosing the sample years, we examine the years 1997-1998 and 2007-2008 in order to observe the change in firm size distribution around the times of financial crisis, along with the year 2013 in which is the most recent sample available. Unlike the work of Kang et al. (2011), all the available yearly data from financial and non-financial firms are pooled cross-sectionally.[11]

The first column of <Figure 1> plots log of rank (the rank of the largest firm is 1) to log of size in different measures. The upper extreme of the distribution is linear, indicating that the tail is Pareto distributed. If it is always possible to find the tail threshold, one can estimate the tail index by simply obtaining the OLS estimate to the coefficient β in the regression below:

$$\log size = \log A - \beta \log Rank \tag{7}$$

However, the threshold point where the firm sizes start to be Pareto distributed is often ambiguous. The OLS estimation result is sensitive to the choice of threshold; an inappropriate choice of the threshold will lead to a serious bias in the estimated value of the tail index.

Both the second and the third columns of <Figure 1> imply that the right tail of the distributions may be well fitted by the LogPH distribution or the GPD. The figures in the middle plots survival functions on log scale; for all of the three measures, the plot is almost linear for sufficiently large firms. The last column shows

billion won, (iii) firms with total assets greater than 7 billion won and total liabilities greater than 7 billion won, and (iv) firms with total assets greater than 7 billion won and employment greater than 300.

[11] Kang et al. (2011) used the panel data for non-financial firms listed in Korean stock markets.

sample values of mean excess function for various thresholds. As the mean excess function of LogPH distribution is approximately linear for thresholds large enough, the fact that these figures are almost linear also suggests that it is worth trying fitting the distributions to a scaled LogPH distribution.

Ⅳ. Results

1. Firm Size Distribution in 2013

<Figure 2> shows the eigenvalue plot for distribution of firm employment, sales, and asset values in 2013, along with corresponding Hill plots for comparison. The stable area of the eigenvalue plots in the middle is where the tail index is identified.[12] The ranges of stable areas along with the ranges of estimated Pareto parameters, or the inverse of tail indices, are provided in <Table 1>.

Two aspects are notable from these results. First, the estimated Pareto parameters vary across firm size measures, approximately ranging from 0.9 to 1.4. Although the estimated Pareto parameter when firm sizes are measured in sales is taking value close to 1, the results from other firm size proxies are rather far from being 1. This is consistent with the results from Kang et al. (2011) that firm sizes in Korea do not necessarily follow the Zipf's law. As we have mentioned in the introduction, some studies in the field of industrial organization stems from an assumption that the tail index of firm size distributions is 1 in general. However, the results here suggest

[12] The upper and lower bounds for the stable area are superimposed in each figure. When the number of deletions from the sample is too small, the plot is unstable as the distribution has not yet converged to the GPD. The last part of the plot becomes highly unstable since the observations used in the fitting are too few.

that one should not hastily assume that the tail index of firm size distribution is always equal to one if, for example, he or she tries to develop a firm dynamics model involving firms' labor and capital resources.

Second, the eigenvalue plots show better performances in general than the Hill plots. The stable area for eigenvalue plots are relatively wide, which means that the Log phase-type distributions are fitting well. The counterpart Hill plots, especially when sales and assets are used as size measures, are not stable in their upper 10%. Although researchers using the Hill plot in practice usually look for its stable area only in the upper 5 to 10 percent tail, our result imply that Hill plots are not recommended in tail index estimation for firm size distributions. Even if one ignores such convention and look for the stable region in the entire range of thresholds, the eigenvalue plot is performing well whenever its corresponding Hill plot is stable. This demonstrates that eigenvalue plots can be used as a superior alternative to the Hill plots, as Kim and Kim (2015) argue.[13)]

2. Changes in Firm Size Distribution along Business Cycles

In this section, we examine the tail index of firm size distributions in years 1997 and 1998 along with those in years 2007 and 2008 in order to see the change in its behavior around the 1997 Asian financial crisis and the global financial crisis of 2007-2008. In short, our results add to the evidence that the upper tail of firm size distribution becomes thicker in the advent of financial crises.

<Figure 3> illustrates the eigenvalue plot of firms' employment distribution, and <Table 2> the estimation result. Although the plots

[13)] In fact, the Hill plot is a special case of the eigenvalue plot corresponding to phase size 1.

for years 1997-1998 do not show great performance as in other samples[14], it is sufficient enough to see that the tail became thicker after experiencing 1997 financial crisis. To be specific, the estimated Pareto parameter has increased by 5.1%. One could interpret this result as the dispersion in the allocation of labor available to firms has increased; more labor became available for larger firms and less for smaller firms. This is possibly because big firms have more resources to deal with aggregate economic shock and thus were able to maintain the current employment level in the face of financial crisis. In contrast, smaller firms with less means of adjustment may have had to fire many workers or even close down. For the years 2007-2008, the Pareto parameter has slightly increased by 1.2% after the global financial crisis occurred; this indicates that firms' employment was less affected by the financial crisis during this period compared to the Asian financial crisis.

Similar phenomena can be found from distributions of sales, as <Figure 4> and <Table 3> demonstrate. After the Asian financial crisis in 1997, the estimated Pareto parameter increases by 2.8%, and it increases very much by 7.4% after the global financial crisis. The economy of South Korea is highly dependent on export, and hence, the sales of firms must have been much affected by the global financial crisis. According to the result here, the inequality among firms in terms of sales was also hit by the global financial crisis. Our result here also implies that financial crises not only on the allocation of resources across firms, but also on the dispersion of firm's performances.

For the distribution of firms' asset values, the inverse of tail indices again increases by 6% in 1998, although the eigenvalue plot in 1997 does not show a great performance as in other results. The Pareto parameter increases by 2.5% in 2008. These results again imply that

[14] This is possibly due to small sample sizes.

the dispersion of firms′ asset distribution has some correlation with financial crises. To be specific, the firms′ asset distribution was particularly hit by the Asian financial crisis when there was a huge outflow in the foreign investments. The results are presented in <Figure 5> and <Table 4>.

Ⅴ. Conclusion

This paper estimates the tail index of firm size distribution in Korea, by fitting the right tail of the distribution to a scaled LogPH distribution. The tail index of the distribution varies depending on the choice of firm size measure and on the status of the economy. In general, the Pareto parameter tends to increase after experiencing financial crises. Our results also confirm that the usage of eigenvalue plot is usually effective for estimating tail indices of firm size distribution.

There are some limitations to this study. First, confidence intervals for the estimated Pareto parameters are not available. Second, the computing time of EMPHT algorithm is burdensome for some data sets. This second restriction has limited us to choose only a few sample years for the analysis instead of estimating the tail indices for all the years available. Yet these qualifications can be viewed as providing a natural set of issues to be addressed by future work. Our estimation results present some useful information regarding the relationship between financial crises and firm size distribution. If possible, estimating the tail index for all the years available will clarify the relationship between firm size distribution and financial crises even further. To be specific, some modifications on the computer codes for EMPHT algorithm such that the iteration stops when convergence has achieved for estimating $\hat{\eta}_T$ will help reducing the computing time. In addition, our methodology can be applied to parameterizing firm distributions of Korea in various economic models. Lastly, further study on the comparison between financial and non-financial firms as well as among different industries in terms of the Pareto parameter will also draw interesting implications about the economic system in South Korea and its business cycle movements. In particular, measuring the tail index of firm size distribution by industries could more specifically suggest where in the Korean economy is mainly amplifying the inequality among firms, and may give practical advice regarding its economic policy implication.

Received: February 4, 2015. Revised: March 10, 2015. Accepted: March 18, 2015.

◈ *References* ◈

Ahn, S., J. H. T. Kim, and V. Ramaswami (2012), "A New Class of Models for Heavy Tailed Distributions in Finance and Insurance Risk," *Insurance: Mathematics and Economics*, 51, pp. 43-52.

Asmussen, S., O. Nerman, and M. Olsson (1996), "Fitting Phase-type Distributions Via the EM Algorithm," *Scandinavian Journal of Statistics*, 23, pp. 419-441.

Axtell, R. L. (2001), "Zipf Distribution of US Firm Sizes," *Science*, 293, pp. 1818-1820.

Bottazzi, G., D. Pirino, and F. Tamagni (2013), "Zipf Law and the Firm Size Distribution: A Critical Discussion of Popular Estimators," *LEM Working Paper Series*, No. 2013/17.

Clementi, G. L., and B. Palazzo (2013), "Entry, Exit, Firm Dynamics, and Aggregate Fluctuations," NBER; No. w19217.

De Wit, G. (2005), "Firm Size Distributions: An Overview of Steady-state Distributions Resulting from Firm Dynamics Models," *International Journal of Industrial Organization*, 23, pp. 423-450.

Di Giovanni, J., and A. A. Levchenko (2013), "Firm Entry, Trade, and Welfare in Zipf's World," *Journal of International Economics*, 89: pp. 283-296.

Fujiwara, Y., C. Di Guilmi, H. Aoyama, M. Gallegati, and W. Souma (2004), "Do Pareto - Zipf and Gibrat Laws Hold True? An Analysis with European Firms," *Physica A: Statistical Mechanics and its Applications*, 335, pp. 197-216.

Gabaix, X., and A. Landier (2008), "Why has CEO Pay Increased so Much?" *Quarterly Journal of Economics*, 123, pp. 49-100.

Gibrat, R. (1931), "Les Inégalités Economiques," *Sirey, Paris*.

Hill, B. (1975), "A Simple General Approach to Inference about the Tail of a Distribution," *The Annals of Statistics*, 3, pp. 1163-1174.

Hopenhayn, H. A. (1992), "Entry, Exit, and Firm Dynamics in Long Run Equilibrium," *Econometrica*, 60, pp. 1127-1150.

Kang, S. H., Z. Jiang, C. Cheong, and S. M. Yoon (2011), "Changes of Firm Size Distribution: The Case of Korea," *Physica A: Statistical Mechanics and its Applications*, 390, pp. 319-327.

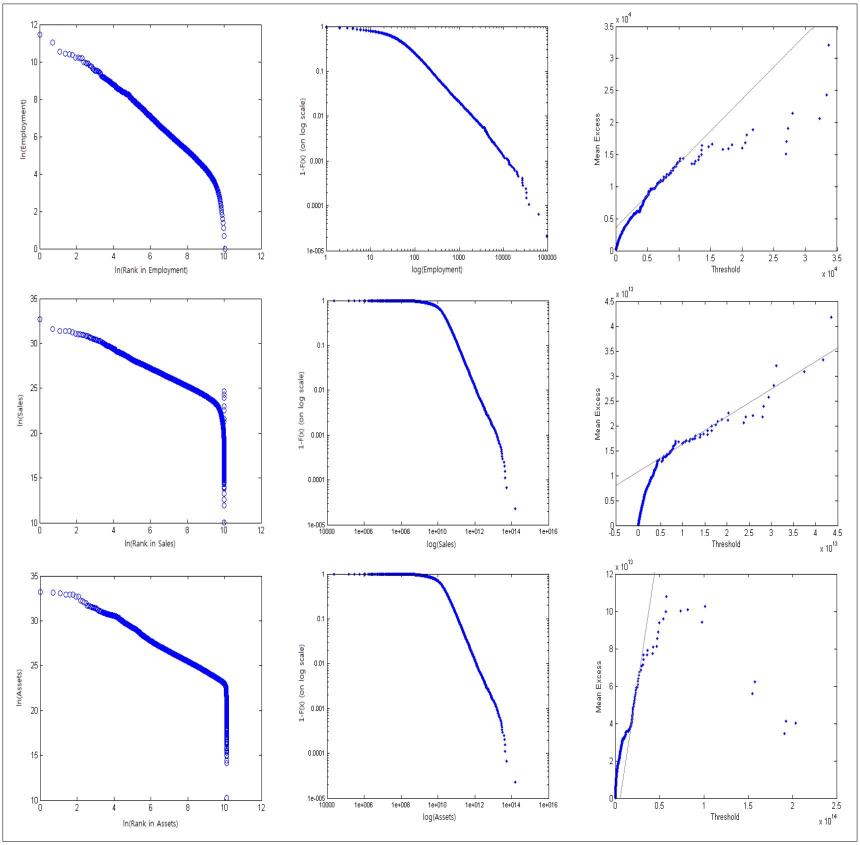
Kim, J. H. T. and J. Kim (2015), "A Parametric Alternative to the Hill Estimator for Heavy-tailed Distributions," *Journal of Banking and Finance*, 54, pp. 60-71.

Neuts, M. F. (1975), "Computational Uses of the Method of Phases in the Theory of Queues," *Computers and Mathematics with Applications*, 1, pp. 151-166.

Podobnik, B., D. Horvatic, A. M. Petersen, B. Urošević, and H. E. Stanley (2010), "Bankruptcy Risk Model and Empirical Tests," *Proceedings of the National Academy of Sciences*, 107, pp. 18325-18330.

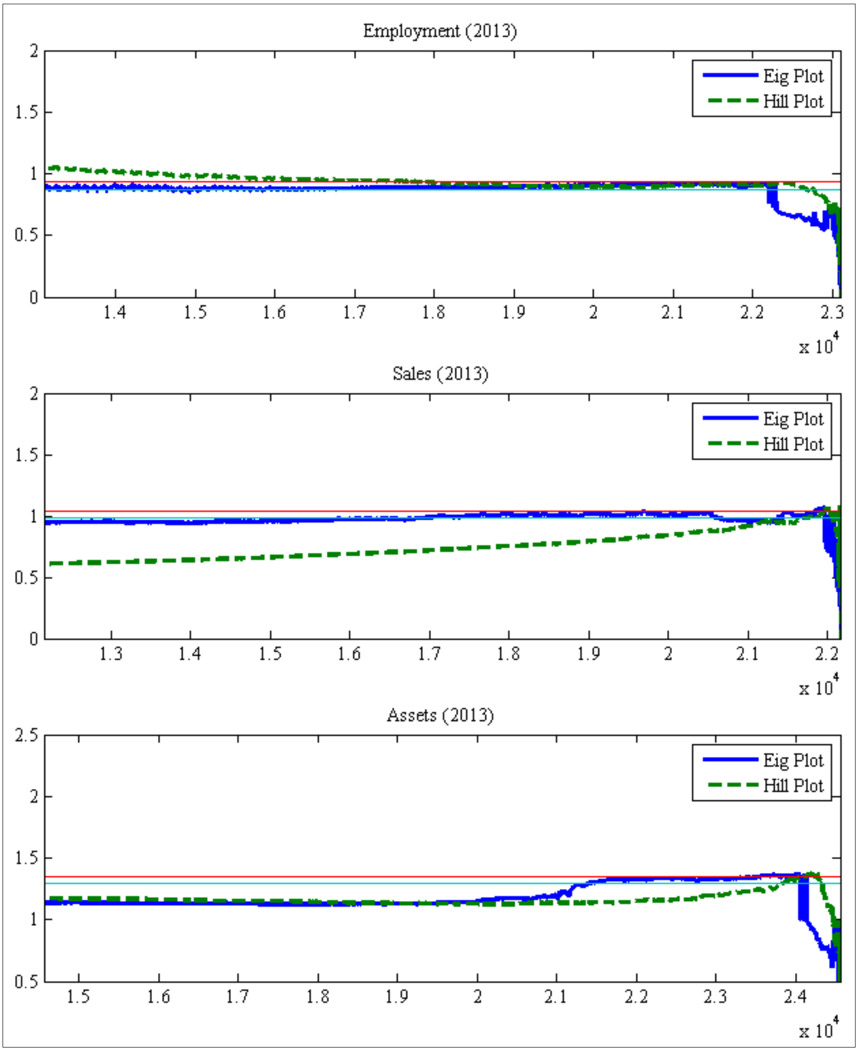
Appendix

【Figure 1】 Log of Rank to Log of Firm Sizes, Log of Survival Function, and Mean Excess Function



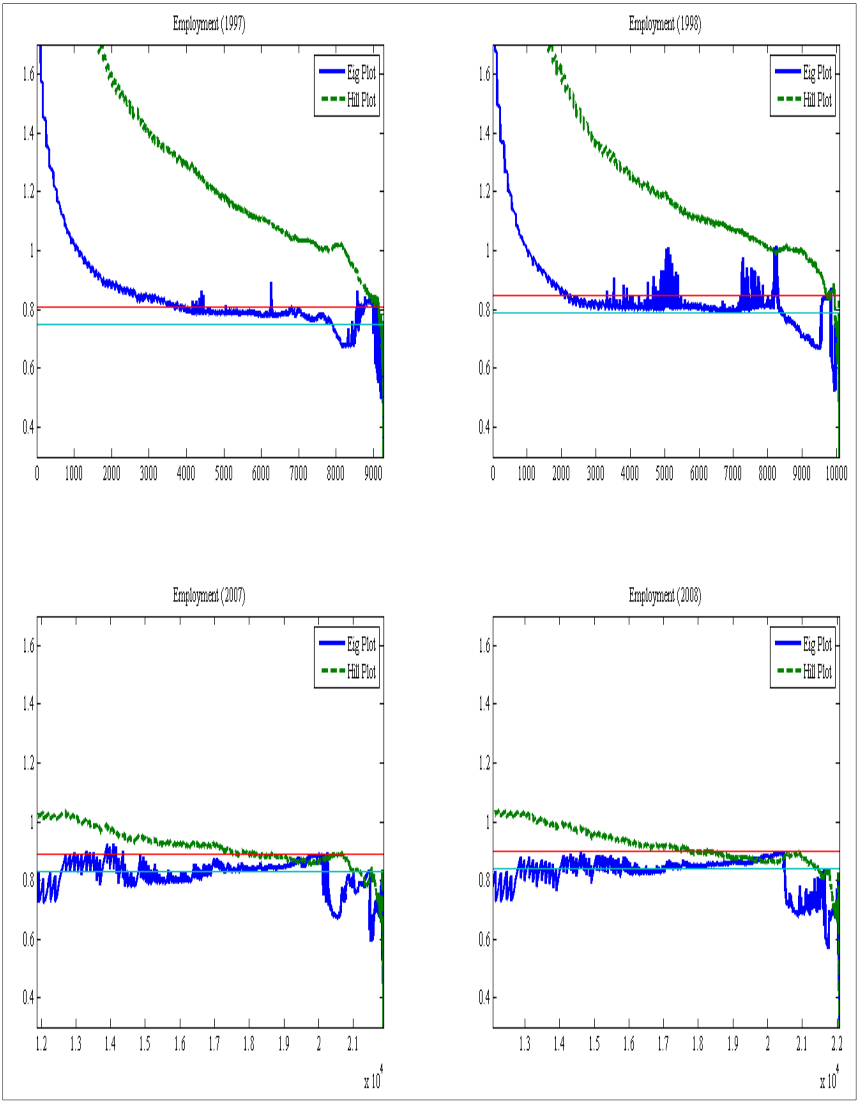
Each row uses different measure for firm sizes; the first row corresponds to employment, the second row to sales, and the third row to assets. The first column plots log of rank (the rank of the largest firm is 1) to log of size. The second column plots survival functions on log scale. The last column shows sample values of mean excess function for various thresholds; the linearity of each mean excess function is guided by the thin grey line.

【Figure 2】Firm Size Distribution in Year 2013



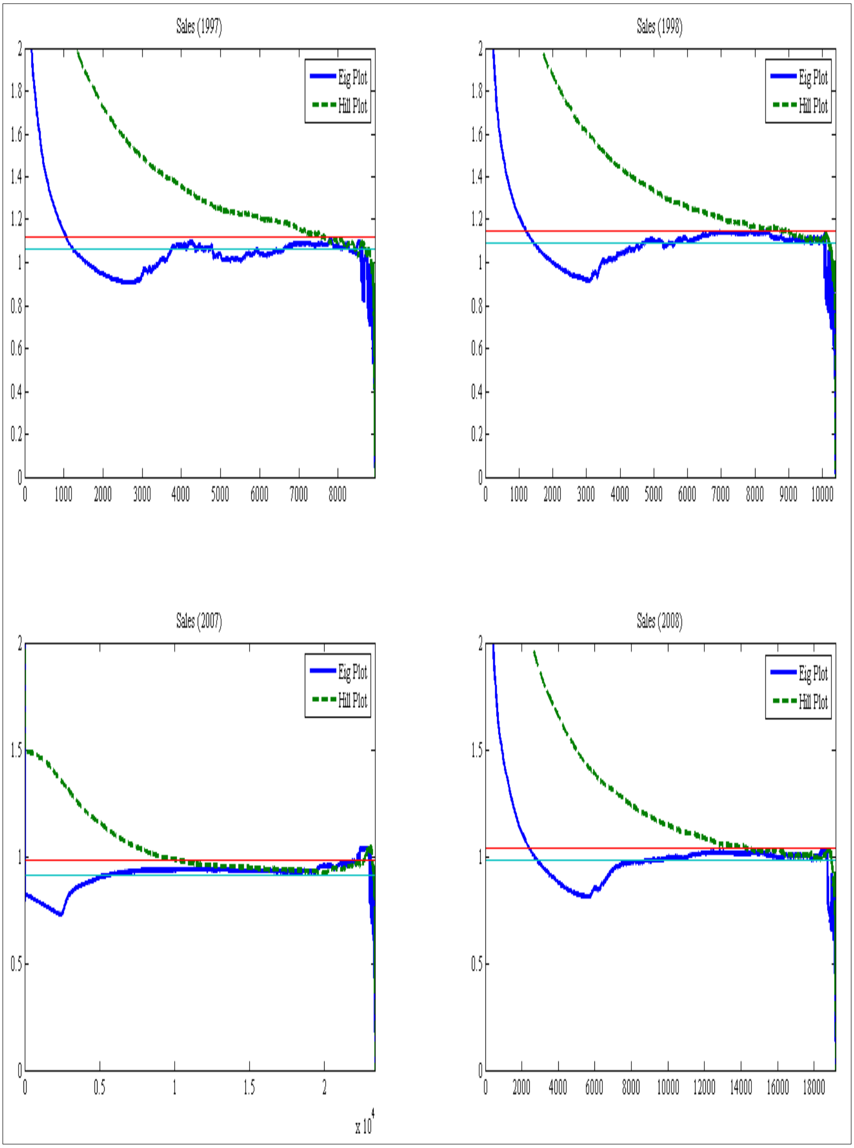
<Figure 2> shows the eigenvalue plot for distribution of firm employment, sales, and asset values in 2013, along with corresponding Hill plots for comparison. The stable area of the eigenvalue plots in the middle—indicated by thin red and green lines—is where the tail index is identified. The estimated value of Pareto parameters vary across the measures of firm size. See <Table 1> for detailed numbers.

【Figure 3】 Changes in Pareto Parameters when Firm Size is Measured in Employment



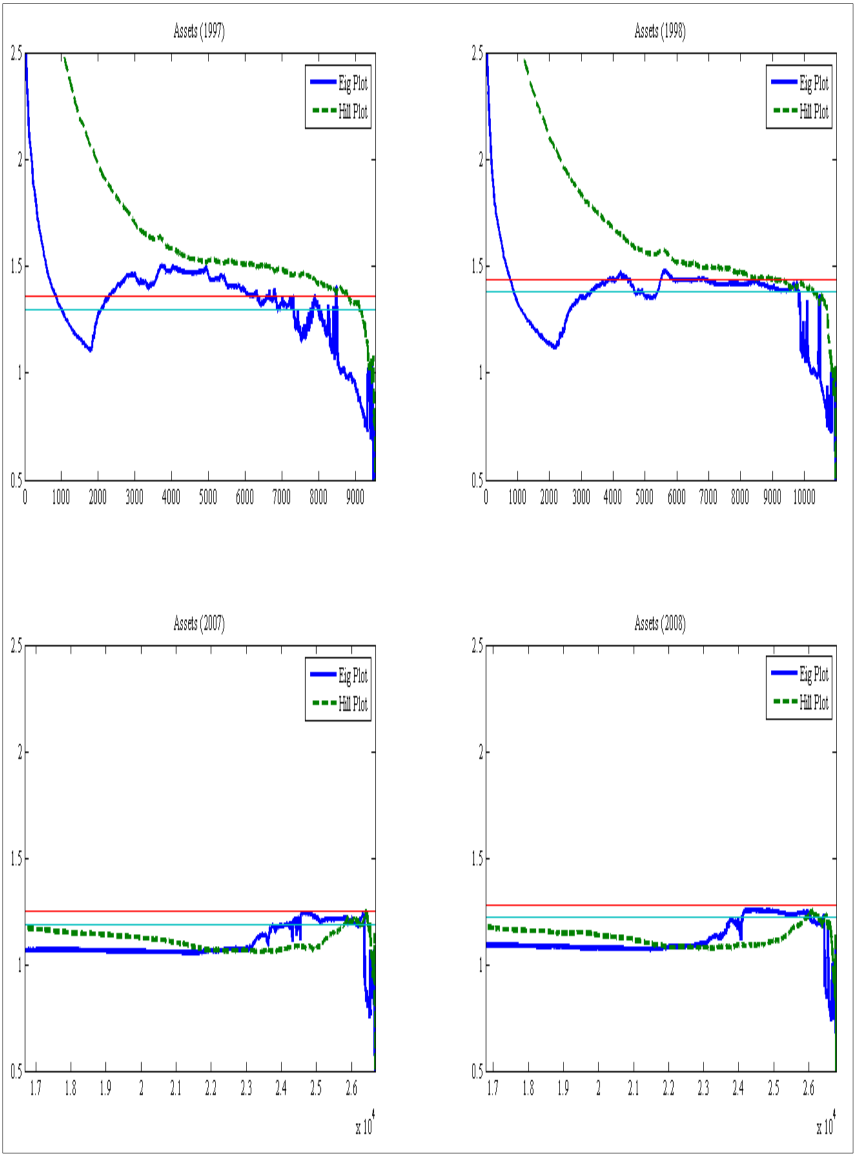
<Figure 3> illustrates the eigenvalue plot of firms' employment distribution. The tail became thicker after experiencing the Asian financial crisis by 5.1%, on average. For the years 2007-2008, the Pareto parameter has slightly increased by 1.2%. See <Table 2> for detailed numbers.

【Figure 4】 Changes in Pareto Parameters when Firm Size is Measured in Sales



<Figure 4> illustrates the eigenvalue plot of firms' sales distribution. After the Asian financial crisis in 1997, the estimated Pareto parameter increases by 2.8%, and it increases by 7.4% after the global financial crisis. See <Table 3> for detailed numbers.

【Figure 5】Changes in Pareto Parameters when Firm Size is Mesured in Assets



<Figure 5> shows the eigenvalue plot of the distribution of firms' asset values. The inverse of tail indices again increases by 6% in 1998, and it increases by 2.5% in 2008. See <Table 4> for detailed numbers.

【Table 1】 Estimation Results from Eigenvalue Plots in 2013

Year	Measure	Obs.	Estimated Pareto parameter		Number of deletions from the sample	
			Lower bound	Upper bound	Lower bound	Upper bound
2013	Employment	23102	0.87	0.93	16341	22199
	Sales	22155	0.98	1.04	16868	20580
	Assets	24557	1.29	1.35	21418	23559

<Table 1> shows the ranges of estimated Pareto parameters as well as the ranges of stable areas for the year 2013. Throughout this paper, the eigenvalue plots are defined to be "stable" if the estimated Pareto parameter comes within the margin of some constant of plus or minus 0.3 for sufficiently many consecutive number of deletions. The upper and lower bounds of the number of deletions indicate the starting and ending points of such a "stable" area. See <Figure 2> for graphical illustrations.

【Table 2】 Estimation of the Pareto Parameters when Size is Measured in Employment

Year	Measure	Obs.	Estimated Pareto parameter		Number of deletions from the sample	
			Lower bound	Upper bound	Lower bound	Upper bound
1997	Employment	9265	0.75	0.81	6265	7880
1998		10086	0.79	0.85	9580	9740
2007		21871	0.83	0.89	17340	20112
2008		22071	0.84	0.90	17497	20485

<Table 2> shows the ranges of estimated Pareto parameters as well as the ranges of stable areas for firm size distribution measured in employment. The tail became thicker in the years 1998 and 2008 in comparison with the years 1997 and 2007, respectively. See <Figure 3> for graphical illustrations.

【Table 3】 Estimation of the Pareto Parameters when Size is Measured in Sales

Year	Measure	Obs.	Estimated Pareto parameter		Number of deletions from the sample	
			Lower bound	Upper bound	Lower bound	Upper bound
1997	Sales	8939	1.06	1.12	6624	8277
1998		10391	1.09	1.15	5644	9682
2007		23398	0.91	0.97	5539	22169
2008		19178	0.98	1.04	9093	18734

<Table 3> shows the ranges of estimated Pareto parameters as well as the ranges of stable areas for firm size distribution measured in sales. Again, the Pareto parameter increased in the years 1998 and 2008 in comparison with the years 1997 and 2007, respectively. The lower and upper bounds of the number of deletionsimplies that the sales' distribution is particularly well-fitted by the scaled LogPH distribution. See <Figure 4> for graphical illustrations.

【Table 4】 Estimation of the Pareto Parameters when Size is Measured in Assets

Year	Measure	Obs.	Estimated Pareto parameter		Number of deletions from the sample	
			Lower bound	Upper bound	Lower bound	Upper bound
1997	Assets	9529	1.30	1.36	6934	7308
1998		10999	1.38	1.44	6841	9801
2007		26680	1.19	1.25	24552	25898
2008		26781	1.22	1.28	24122	26027

<Table 4> shows the ranges of estimated Pareto parameters as well as the ranges of stable areas for firm size distribution measured in assets. The estimated value of Pareto parameter increased in the years 1998 and 2008 in comparison with the years 1997 and 2007, respectively. See <Figure 5> for graphical illustrations.

한국 기업크기분포의 우측 꼬리에 대한 추정*

김 주 철** · **김 현 열*****

논문초록

본 연구에서는 한국의 기업크기분포의 우측 꼬리 부분을 scaled Log phase-type 분포로 추정함으로써 그 tail index 값을 추정하고자 한다. 금융기업과 비 금융기업 모두를 포함한 외부 감사의 대상이 되는 기업들을 대상으로 연구를 수행한 결과, 기업 크기의 대리변수를 어떤 것으로 선택하든지 간에 본 연구에서 사용된 방법이 Hill estimator를 포함한 이전의 방법보다 tail index를 효과적으로 추정하는 것으로 나타났다. Tail index의 추정치는 기업 크기의 대리변수에 따라, 그리고 경기에 따라 다르게 계산되었으며, 이는 기업의 크기 분포가 반드시 Zipf's law를 따르지는 않는다는 것을 보여준다. 나아가 1997년과 2007년 전후로 기업크기분포의 우측 꼬리가 더욱 두꺼워진 것을 확인할 수 있는데, 이는 금융위기가 기업들 간의 자원 분포에 분명한 영향을 준다는 것으로 해석할 수 있다.

주제분류 : B030904
핵심 주제 : 기업크기분포, Tail index, 파레토 파라미터, Log phase-type 분포, Hill estimator

* 본 연구는 2015년도 한국연구재단의 지원으로 수행된 연구입니다 (2014S1A5A2A01011100).

** 교신저자, 연세대학교 상경대학 경제학부 부교수, e-mail: joocheol@yonsei.ac.kr
*** 연세대학교 일반대학원 경제학과 석사, e-mail: hyunyeolkim@gmail.com