数理解析研究所講究録
第1954巻 2015年 80-89

Estimation of the tail index for lattice-valued sequences

Muneya Matsui, Nanzan University
joint work with
Thomas Mikosch, University of Copenhagen
Laleh Tafakori, University of Shiraz

Abstract

If one applies the Hill, Pickands or Dekkers-Einmahl-de Haan estimators of the tail index of a distribution to data which are rounded off one often observes that these estimators oscillate strongly as a function of the number k of order statistics involved. We study this phenomenon in the case of a Pareto distribution. We provide formulas for the expected value and variance of the Hill estimator and give bounds on k when the central limit theorem is still applicable. We illustrate the theory by using simulated and real-life data.

1 Empirical example (a data set of word counts)

Numerous real-life data (X_n) have power-like tails in the sense that for some $\alpha > 0$ and large x,

$$P(X_n > x) \approx x^{-\alpha},$$

and therefore the estimation of the tail index α (or equivalently α^{-1}) is important. Although there exists a multitude of estimators of α, if the data are rounded off one often observes that these estimators often oscillate strongly as a function of the number k of order statistics involved.

In this paper, we study the influence of discretization effects (such as round-off, imprecise data) on the estimation of α^{-1}. Our main focus is on the Hill estimator but we also touch on the Pickands and Dekkers-Einmahl-de Haan estimators (see Sec. 3 of Matsui et al. (2013)) and show simulation evidence that the estimator may suffer from the discreteness of the data.

Hill estimator : writing

$$X_{(1)} \leq \cdots \leq X_{(n)}$$

for the order statistics of the observations $X_1, \ldots, X_n$, the Hill estimator of α^{-1} is given by

$$\widehat{\alpha}_k^{-1} = \frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{(n-i+1)}}{X_{(n-k+1)}}, \quad 0 \leq k \leq n.$$

To illustrate the effect of discreteness of data on the Hill estimator we consider a data set of word counts from the English language. The data have Pareto-like tails with an estimated index close to 1; see Figure 1, top left. The data consists of the 10 000 largest counts between 3 000 and 23 million. In the remaining 3 graphs in Figure 1 we show the Hill plots of the same data when the last one, two or three digits in the count data are replaced by zeros. This means that the counting units are tens, hundreds, thousands, respectively. While the effect for units of tens is hardly visible, we see significant changes in the Hill plot for units of hundreds and thousands.

2 Simulated example (discretized Pareto data)

We choose a Pareto distribution as our toy model and therefor consider a Pareto distributed iid sequence which satisfies

$$X_i \stackrel{d}{=} U_i^{-1/\alpha}, \quad i \in \mathbb{N}, \tag{2.1}$$

for an iid $U(0,1)$ sequence (U_i) and some positive α. It is easy to see that

$$\overline{F}(x) = P(U_i^{-1/\alpha} > x) = x^{-\alpha}, \quad x \geq 1, \tag{2.2}$$

and hence the order statistics satisfy the relation $X_{(i)} \stackrel{d}{=} U_{(n-i+1)}^{-1/\alpha}$ for $i \leq n$.

In Figure 2, top left, we show a Hill plot based on a sample of size 10 000 with $\alpha = 1$. The plot nicely shows the trade-off between bias and variance depending on the chosen values k: too small values of k lead to a large variance while too large k lead to a larger bias. In the remaining graphs of Figure 2 we show the Hill plots for the iid sample $10^{-l}[10^l U_i^{-1/\alpha}]$ for $l = 0, 1, 2$, where $[x]$ denotes the integer part of any real number x. (Due to the scale invariance of the Hill estimator, these Hill plots coincide with those based on $([10^l U_i^{-1/\alpha}])$.) This transformation of $U_i^{-1/\alpha}$ turns all digits but the first l ones behind the comma into zeros. In this sense, we obtain a discretization of the random variables $U_i^{-1/\alpha}$ by rounding off.

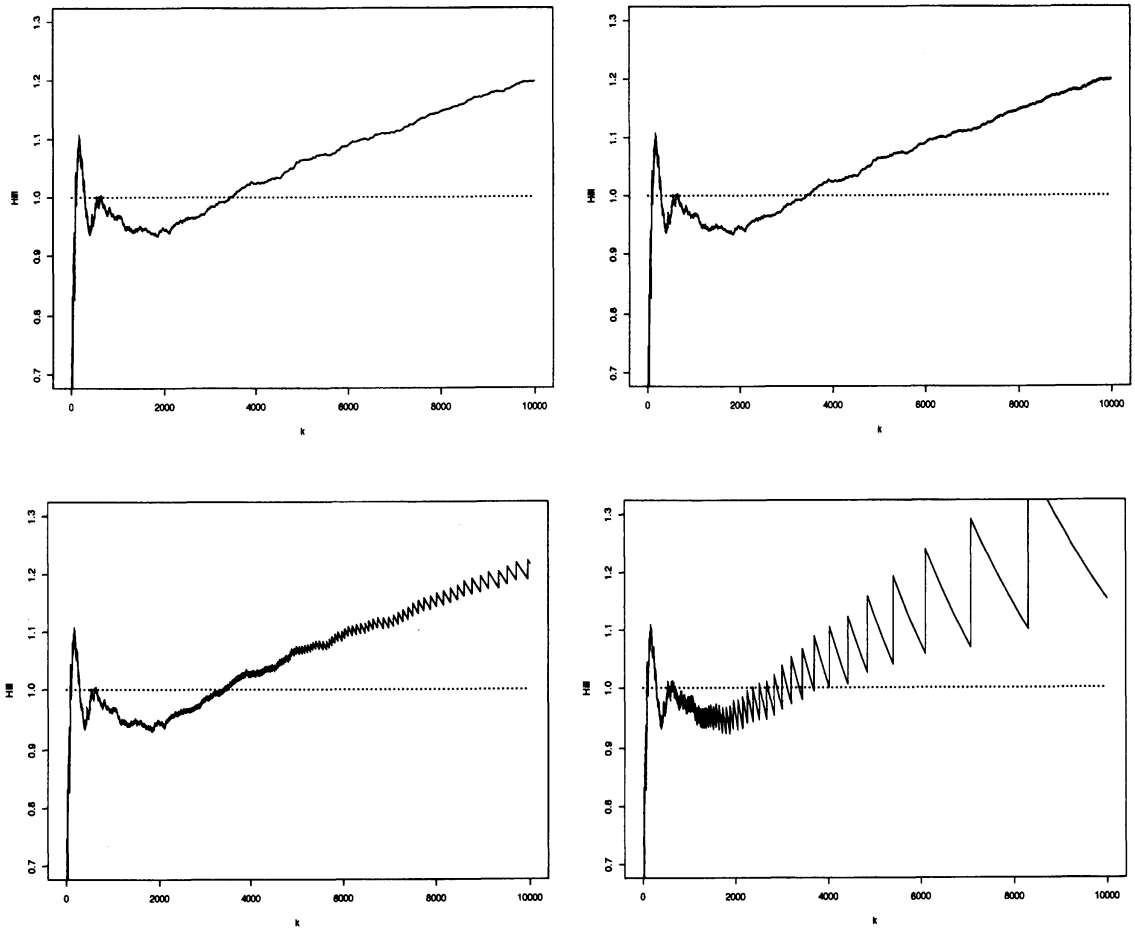

Figure 1: Hill plot $(k, \widehat{\alpha}_k^{-1})$, $k = 2, \ldots, n-1$, based on the counts of 10 000 words which are used most frequently in the English language. Notice that the Hill plot yields a reliable estimator only for small k, up to 1000 say. *Top left.* The estimated tail index of the count data is close to one. *Top right.* The last digit in the counts is replaced by zero. *Bottom left.* The last two digits are replaced by zeros. *Bottom right.* The last three digits are replaced by zeros.

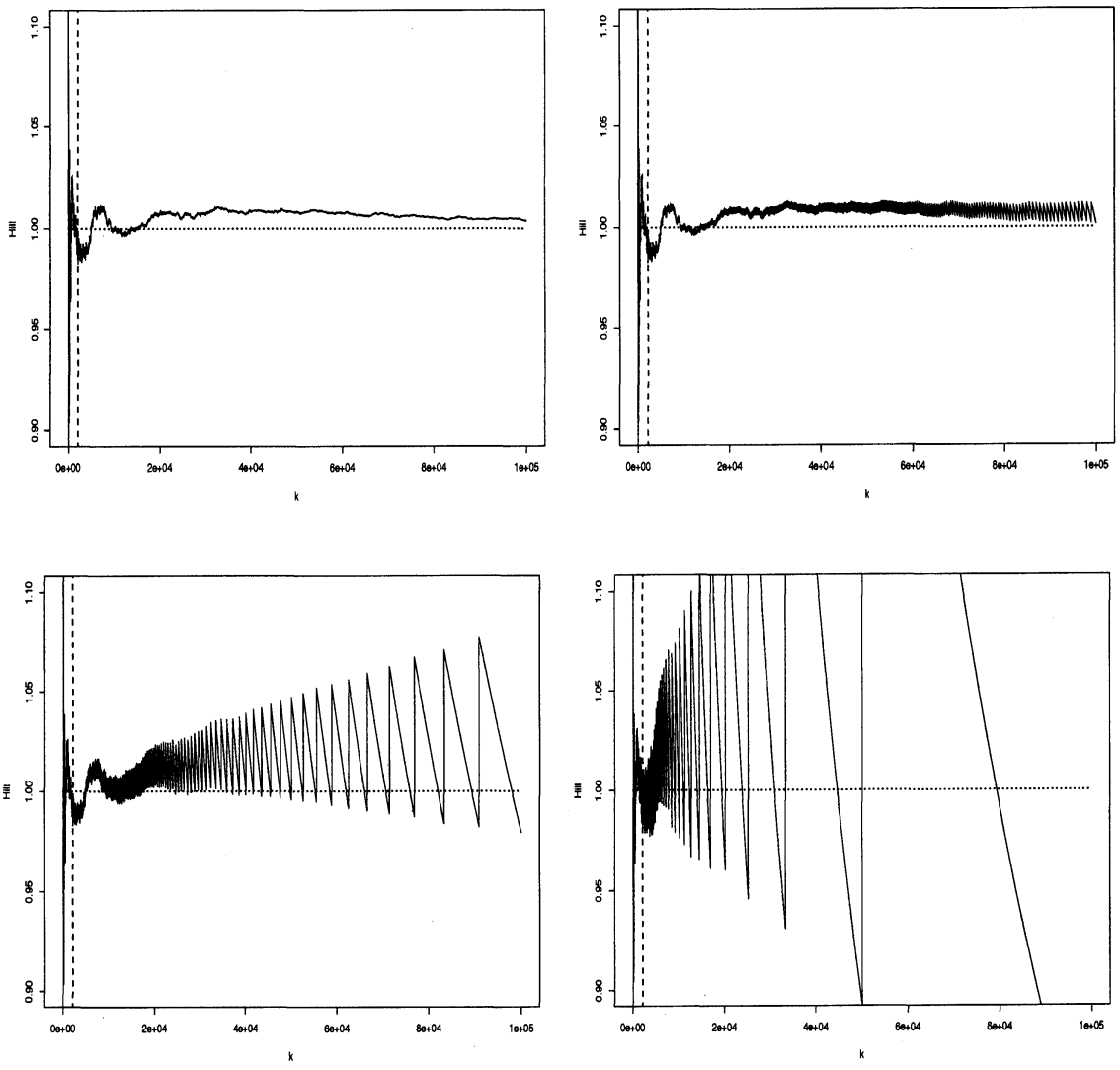

Figure 2: Hill plot $(k, \widehat{\alpha}_k^{-1})$, $k = 2, \ldots, n-1$, for a sample of size $n = 10\,000$. Top left. The data have a Pareto distribution (2.2) with $\alpha = 1$. In the other figures all digits but the first l behind the comma are set equal to zero. Top right. $l = 2$. Bottom left. $l = 1$. Bottom right. $l = 0$. The vertical line at $k = n^{2/3}$ is an upper limit for those k for which the central limit theorem is still valid; see Corollary 4.3.

3 Theoretical explanation (what is the singularity ?)

The tail of the transformed random variable is given by

$$\begin{aligned} \overline{F}(x) &= 1 - P([10^l U_i^{-1/\alpha}] \leq [10^l x]) \\ &= 1 - P(10^l U_i^{-1/\alpha} < [10^l x] + 1) \\ &= \frac{(10^l)^\alpha}{([10^l x] + 1)^\alpha}. \end{aligned} \tag{3.3}$$

Since $[y] \in (y-1, y]$ for $y \in \mathbb{R}$ one immediately concludes that

$$P(10^{-l}[10^l U_i^{-1/\alpha}] > x) \sim P(U_i^{-1/\alpha} > x) = x^{-\alpha}, \quad x \to \infty.$$

Therefore the standard theory (see Mason (1982) or Theorem 3.2.2 in de Haan and Ferreira (2006)) yields that the Hill estimator is consistent:

$$\widehat{\alpha}_k \xrightarrow{P} \alpha, \quad \text{if } k \to \infty \text{ and } k/n \to 0$$

and even strongly consistent (i.e. $\widehat{\alpha}_k \xrightarrow{\text{a.s.}} \alpha$) if $k/n \to 0$ and $k/\log\log n \to \infty$; see Deheuvels et al. (1988).

Standard results about asymptotic normality of the Hill estimator are not available in this case since such a theory requires that a second order condition on $\overline{F}$ must be satisfied. According to de Haan and Ferreira (2006), Theorem 3.2.5, asymptotic normality of $\widehat{\alpha}_k$ can be achieved if the following second order condition holds as $x \to \infty$ for $t > 0$

$$\frac{\overline{F}(tx)}{\overline{F}(x)} - t^{-\alpha} \sim b(t)a(x),$$

where $|a(x)|$ is regularly varying[1] with a non-positive index and $b(t)$ is a positive function of t. We observe that ($\{x\}$ denotes the fractional part of x)

$$\frac{\overline{F}(tx)}{\overline{F}(x)} - t^{-\alpha} = \frac{([10^l x] + 1)^\alpha}{([10^l tx] + 1)^\alpha} - t^{-\alpha} \tag{3.5}$$

[1]Recall that the distribution function F of a positive random variable X has a *regularly varying tail* if it can be written in the form

$$\overline{F}(x) = P(X > x) = \frac{L(x)}{x^\alpha}, \quad x > 0, \tag{3.4}$$

where $\alpha \geq 0$ is the *tail index* and L is a *slowly varying function*, i.e., for every $c > 0$, $\lim_{x\to\infty} L(cx)/L(x) = 1$. It is known that the definition (3.4) is equivalent to $\lim_{x\to\infty} \overline{F}(tx)/\overline{F}(x) = t^{-\alpha}$ for all $t > 0$.

$$\sim \; t^{-\alpha}(10^l x)^{-1}\alpha\big(-\{10^l x\} + (1-t^{-1}) + t^{-1}\{10^l tx\}\big)\,.$$

The right-hand side exhibits very erratic behavior. For irrational $10^l t$, the sequence $(\{10^l tx\})_{x=1,2,\ldots}$ is uniformly distributed in the number theoretical sense; see Weyl (1916). In particular, it visits any interval $(a,b) \subset (0,1)$ infinitely often. Then the sequence

$$-\{10^l x\} + (1-t^{-1}) + t^{-1}\{10^l tx\} \;=\; 1 + t^{-1}(-1 + \{10^l tx\})\,, \quad x = 1,2,\ldots\,,$$

is uniformly distributed on $(1-t^{-1},1)$. If x assumes the integers $1,2,\ldots$ and $10^l t$ is an integer, the last term of right-hand side vanishes. Hence $|\overline{F}(tx)/\overline{F}(x) - t^{-\alpha}|$ is not a regularly varying function as required.

4 Main result (asymptotic normality)

However, asymptotic normality of $\widehat{\alpha}_k^{-1}$ can still be derived from the corresponding results for $(U_i^{-1/\alpha})$ if $k_n = o(n^{2/(2+\alpha)})$; see Corollary 4.3 below. We show a similar result for the Pickands and Dekkers-Einmahl-de Haan estimators. Note that in this report we present only final results and all proofs are found in Matsui et al. (2013).

Let the Hill estimator based on $X_1^{(l)},\ldots,X_n^{(l)}$, $n \geq 3$, $2 \leq k \leq n-1$, denoted by $\widehat{\alpha}_{k,l}^{-1}$. In the next result we measure the deviation of $E\widehat{\alpha}_{k,l}^{-1}$ from $E\widehat{\alpha}_k^{-1} = \alpha^{-1}$.

Proposition 4.1 *Consider the sequence* $X_i^{(l)} = 10^{-l}[10^l U_i^{-1/\alpha}]$, $i = 1,2,\ldots$, *for a fixed integer* $l \geq 0$. *Then, for the Hill estimator* $\widehat{\alpha}_{k,l}^{-1}$ *based on* $X_1^{(l)},\ldots,X_n^{(l)}$, $n \geq 3$, $2 \leq k \leq n-1$,

$$-10^{-l}E\big(U_{(k)}^{1/\alpha}\big) \leq E\widehat{\alpha}_{k,l}^{-1} - \alpha^{-1} \leq 10^{-l}(1+10^{-l})E\big(U_{(k)}^{1/\alpha}\big)\,. \tag{4.6}$$

Moreover, writing $D_k^{(1)} = \widehat{\alpha}_k^{-1} - \widehat{\alpha}_{k,l}^{-1}$, *we have for any* $p > 0$,

$$E|D_k^{(1)}|^p \leq 10^{-lp}(1+10^{-l})^p E\big(U_{(k)}^{p/\alpha}\big)\,. \tag{4.7}$$

The left and right hand sides in (4.6) *and* (4.7) *converge to zero as* $n \to \infty$ *if* $k = k_n \to \infty$ *and* $k/n \to 0$ *or* k *is fixed and* $l \to \infty$.

Before we proceed further we recall the following benchmark result. Its proof is an immediate consequence of non-discretized case.

Lemma 4.2 *Let (X_i) be an iid sequence with common Pareto distribution defined in (2.2). Assume that $k = k_n \to \infty$ and $k_n \le n$. Then*

$$\sqrt{k}\,(\widehat{\alpha}_k^{-1} - \alpha^{-1}) \xrightarrow{d} Y \quad \text{and} \quad \sqrt{k}\,(\widehat{\alpha}_k - \alpha) \xrightarrow{d} Z,$$

where Y and Z are normally distributed with mean zero and variances α^{-2} and α^2, respectively.

A combination of Proposition 4.1 and Lemma 4.2 yields the following result.

Corollary 4.3 *Consider the sequence $X_i^{(l)} = 10^{-l}[10^l U_i^{-1/\alpha}]$, $i = 1, 2, \ldots$, for a fixed integer $l \ge 0$. Let the Hill estimator $\widehat{\alpha}_{k,l}^{-1}$ based on $X_1^{(l)}, \ldots, X_n^{(l)}$, $n \ge 3$, $2 \le k \le n-1$. Assume $k = k_n \to \infty$ and $k = o(n^{2/(\alpha+2)})$. Then for a normal $N(0, \alpha^{-2})$ distributed random variable Y,*

$$\begin{aligned} \sqrt{k}\,(\widehat{\alpha}_{k,l}^{-1} - \alpha^{-1}) &\xrightarrow{d} Y\,, \\ E\Big(\big|\sqrt{k}\,(\widehat{\alpha}_{k,l}^{-1} - \alpha^{-1})\big|^p\Big) &\to E\big(|Y|^p\big)\,. \end{aligned}$$

We could not prove whether Corollary 4.3 is optimal in the sense that it does not hold for $k \ge cn^{2/(\alpha+2)}$. However, simulations indicate that the Hill estimator $\widehat{\alpha}_{k,l}^{-1}$ is very unreliable for such k-values.

We make an excursion to two other classical estimators of the extreme value index (see Sec. 3 of Matsui et al. (2013)). Also in these cases, $k = n^{2/(\alpha+2)}$ appears as a borderline case for central limit behavior of the corresponding estimators.

5 Moment of the Hill estimator for an integer-valued sequence

Recall that for an iid uniform $U(0,1)$ distributed sequence (U_i), the ith order statistic $U_{(i)}$ has a $\beta(i, n-i+1)$ density (see e.g. Embrechts et al. (1997), Proposition 4.1.2) given by

$$\beta(i, n-i+1)(x) = \frac{n!}{(n-i)!(i-1)!} x^{i-1}(1-x)^{(n-i+1)-1}\,, \quad x \in (0,1)\,. \tag{5.1}$$

Although the Hill estimator $\widehat{\alpha}_k^{-1}$ is known to be an unbiased estimator of α^{-1}, the situation changes in the case of round-off effects:

Lemma 5.1 *Consider the sequence* $X_i^{(l)} = 10^{-l}[10^l U_i^{-1/\alpha}]$, $i = 1, 2, \ldots$, *for a fixed integer* $l \geq 0$. *Then the following relation holds for the Hill estimator* $\widehat{\alpha}_{k,l}^{-1}$ *based on* $X_1^{(l)}, \ldots, X_n^{(l)}$, $n \geq 3$, $2 \leq k \leq n-1$:

$$
\begin{aligned}
& E\widehat{\alpha}_{k,l}^{-1} \\
&= \sum_{s=10^l+1}^{\infty} \log \frac{s}{s-1} \frac{1}{k-1} \sum_{i=1}^{k-1} \int_0^{(10^l/s)^\alpha} \big(\beta(i, n-i+1)(x) \\
&\qquad\qquad -\beta(k, n-k+1)(x)\big)\, dx\,, \\
&= \frac{n}{k-1} \sum_{s=10^l+1}^{\infty} \log \frac{s}{s-1} (10^l/s)^\alpha \int_{(10^l/s)^\alpha}^1 \beta(k-1, n-k+1)(x)\, dx\,.
\end{aligned}
$$

We also derive the formula of $\mathrm{Var}(\widehat{\alpha}_{k,l}^{-1})$ together with the figures (see Sec. 2 of Matsui et al. (2013)).

Figure 3 exhibits $E\widehat{\alpha}_{k,l}^{-1}$ as a function of k for $\alpha = 1$, $l = 0, 1, 2$, and sample size $n = 10\,000$. Evidently, the erratic behavior of the Hill estimators $\widehat{\alpha}_{k,l}^{-1}$ is also inherited by its mean value function. It shows significant deviations from the value α^{-1}, in particular for large k. This fact is a clear warning against using the maximum likelihood estimator $\widehat{\alpha}_n$ of α.

6 Concluding remarks

The estimation of the tail index α is a complicated statistical problem. The results and graphs above show that the estimation also depends on round-off effects which often are neglected, e.g. by assuming that the data have a Lebesgue density.

There exist numerous applied papers where power law behavior of the tails of the data has been postulated (e.g. in the literature on Zipf's law or on fractal dimensions of real-life data). The tail index α is often estimated by ordinary least squares (OLS) based on a plot of $-\log \overline{F}_n(x)$ (F_n is the empirical distribution function) against $\log x$, where x is chosen from the whole range of the data or from a "far-out" x-region where the plot is "roughly linear" . The round-off effect leads to undesirable oscillations of the log-log plot and, in turn, yields unreliable estimates of α.

For the Hill estimator $\alpha_{k,l}^{-1}$ of Pareto variables one can calculate the expectation and variance explicitly; numerical calculations and simulations show that these moments and the estimator itself may oscillate strongly, depending on the size of the round-off error. The results of this paper indicate that the region of k-values where the Hill and related estimators are asymptotically

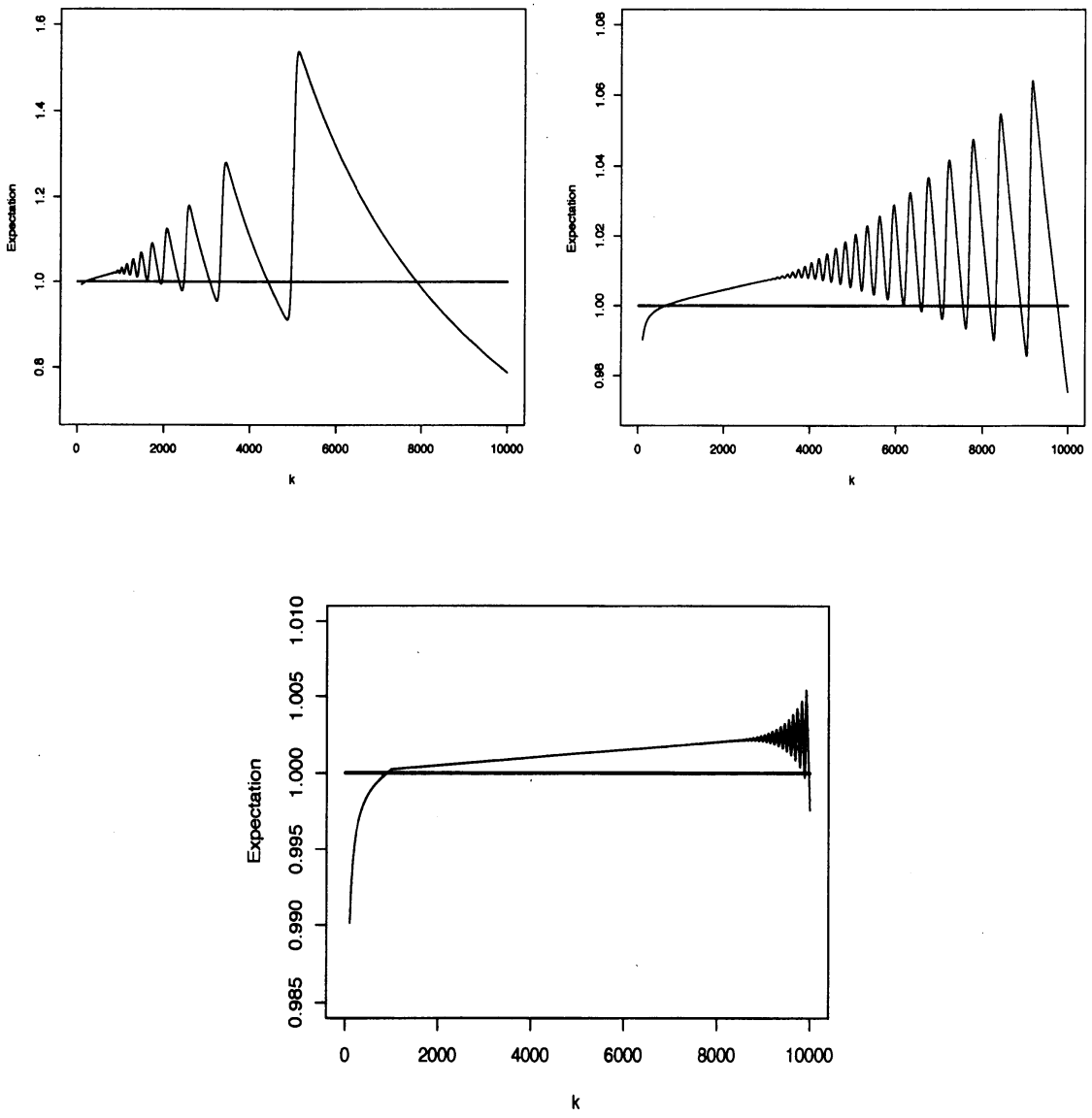

Figure 3: The mean value function of the Hill estimator $\widehat{\alpha}_{k,l}^{-1}$ for a sample of size $n = 10\,000$ of discretized Pareto distributed random variables $X_1^{(l)}, \ldots, X_n^{(l)}$ with parameter $\alpha = 1$. Top left: $l = 0$. Top right: $l = 1$. Bottom: $l = 2$.

normal is rather small and strongly depends on the size of the round-off error described by the parameter l. On the positive side, even under round-off effects the classical estimators are reliable (i.e. satisfy the usual asymptotic properties) in these k-regions and, in contrast to the estimation of α based on OLS, a body of standard theory is applicable.

REFERENCES

[1] DEHEUVELS, P., HÄUSLER, E. AND MASON, D.M. (1988) Almost sure convergence of the Hill estimator. *Math. Proc. Cambridge Philos. Soc.* **104**, 371–381.

[2] EMBRECHTS, P., KLÜPPELBERG, C. AND MIKOSCH, T. (1997) *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

[3] HAAN, L. DE AND FERREIRA, A. (2006) *Extreme Value Theory. An Introduction.* Springer Series in Operations Research and Financial Engineering. Springer, New York.

[4] MASON, D.M. (1982) Laws of large numbers for sums of extreme values. *Ann. Probab.* **10**, 756–764.

[5] MATSUI, M, MIKOSCH, T. AND TAFAKORI, L. (2013) Estimation of the tail index for lattice-valued sequences. *Extremes* **16**, 429–455.

[6] WEYL, H. (1916) Über die Gleichverteilung von Zahlen mod. Eins. *Math. Ann.* **77**, 313–352.

Correspondence to Department of Business Administration, Nanzan University, 18 Yamazato-cho Showa-ku Nagoya, 466-8673, Japan.
E-mail: mmuneya@nanzan-u.ac.jp

南山大学　松井　宗也