Priority statement and some properties of t-lgHill estimator

Milan Stehlik, Jozef Kiseľák, Marijus Vaičiulis, Pavlina Kalcheva Jordanova, Ludy Nunez Soza, Zdeněk Fabián, Philipp Hermann, Luboš Střelec, Andres Rivera, Stéphane Girard, et al.

Noname manuscript No.
(will be inserted by the editor)

Priority statement and some properties of t-lgHill estimator

Milan Stehlík · Jozef Kiseľák · Marijus Vaičiulis · Pavlina Jordanova · Ludy Núñez Soza · Zdeněk Fabián · Philipp Hermann · Luboš Střelec · Andrés Rivera · Stéphane Girard · Sebastián Torres

Received: date / Accepted: date

We acknowledge support of research grants LIT-2016-1-SEE-023, the bilateral project Bulgaria - Austria, 2016–2019, 'Feasible statistical modeling for extremes in ecology and finance', BNSF, Contract number 01/8, 23/08/2017 and WTZ Project No. BG 09/2017, https:// pavlinakj.wordpress.com/. The authors are grateful also to bilateral project HU 11/2016, Proyecto UTA MAYOR 4746-19, the Slovak Research and Development Agency under the contract No. APVV-17-0568, the Czech Science Foundation under the project No. GA16-07089S and FONDECYT N 1171832.

M. Stehlík (Corresponding author)
Department of Applied Statistics & Linz Institute of Technology, Johannes Kepler University in Linz, Austria
Altenbergerstrasse 69, 4040 Linz, Austria
Tel.: +43 732 2468 6806, Fax: +43 732 2468 6800, E-mail: mlnstehlik@gmail.com
Institute of Statistics, Universidad de Valparaíso, Valparaíso, Chile
Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, USA

J. Kiseľák
Institute of Mathematics, Faculty of Science, P. J. Šafárik University in Košice, Slovakia

Marijus Vaičiulis
Vilnius University Institute of Data Science and Digital Technologies, Akademijos 4, LT-08663 Vilnius, Lithuania

P. Jordanova
Faculty of Mathematics and Informatics, Shumen University, Universitetska Str. 115, 9700 Shumen, Bulgaria

Ludy Núñez Soza
Departamento de Matemática, Facultad de Ciencias, Universidad de Tarapacá, Avenida 18 de Septiembre 2222, Casilla 7-D, Arica, Chile

Z. Fabián
Institute of Computer Science of the Czech Academy of Sciences, Pod vodárenskou věží 2, 18200 Prague, Czech Republic

P. Hermann
Department of Applied Statistics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria

L. Střelec
Department of Statistics and Operation Analysis (FBE), Mendel University, Zemědelská 1,

Abstract We acknowledge the priority on the introduction of the formula of t-lgHill estimator for the positive extreme value index. We provide a novel motivation for this estimator based on ecologically driven dynamical systems. Another motivation is given directly by applying the general t-Hill procedure to log-gamma distribution. We illustrate the good quality of t-lgHill estimator in comparison to classical Hill estimator on the novel data of the concentration of arsenic in drinking water in the rural area of the Arica and Parinacota Region, Chile.

Keywords t-lgHill estimator, Hill estimator, t-score estimation, asymptotic normality, levels of arsenic in drinking water.

1 Theoretical motivation of t-lgHill estimator

This is a priority letter on the first introduction of t-lgHill estimator for the positive extreme value index $\gamma > 0$:

$$H^L_{k_n,n} = \frac{M^{(2)}_{k_n,n} - \left(M^{(1)}_{k_n,n}\right)^2}{M^{(1)}_{k_n,n}},$$

where

$$M^{(j)}_{k_n,n} = \frac{1}{k_n}\sum_{i=1}^{k_n}\left(\ln\left(\frac{X_{n-i+1,n}}{X_{n-k_n,n}}\right)\right)^j, \quad j = 1, 2$$

and $X_{1,n} \leq \cdots \leq X_{n,n}$ are the order statistics of $X_1, \ldots, X_n$. We recall that the statistics $M^{(j)}_{k_n,n}$, $j = 1, 2$ are introduced in [2], while $\hat{\gamma}^{(H)}_{k_n,n} = M^{(1)}_{k_n,n}$ is nothing else but the Hill [6] estimator. We also recall that the estimator $H^L_{k_n,n}$, as well as, the Hill estimator, is applicable for distributions F, which generalized inverse $F^{\leftarrow}$ satisfies condition: $U(t) := F^{\leftarrow}\left(1 - t^{-1}\right)$, $t > 1$ varies regularly at infinity with positive index γ. $H^L_{k_n,n}$ estimates parameter γ, but for better readability we also use the tail index $\alpha = 1/\gamma$.

We have found that the estimator $H^L_{k_n,n}$ was firstly developed in [5], see (2.10) therein. We have not been aware of this fact in time of publication of [7], where we developed the estimator $H^L_{k_n,n}$, see Theorem 1 in [7].

The main purpose of this priority letter is to demonstrate that the estimator $H^L_{k_n,n}$ was introduced in [5] and [7] by applying different approaches. In [5]

61300, Brno, Czech Republic

A. Rivera
Departamento de Geografia, Universidad de Chile, Avda. Portugal 84, Santiago, Chile

S. Girard
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

S. Torres
Department of Mathematics, Federico Santa María Technical University, Valparaíso, Chile

the estimator $H^L_{k_n,n}$ has been built by assuming the following second regular variation condition:

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \quad x > 0,$$

where $\rho < 0$ is the second order parameter, while $A(t)$ is a measurable function with the constant sign near infinity and $A(t) \to 0$ as $t \to \infty$. Precisely, the authors of the paper [5] considered the generalized Jackknife statistic

$$\gamma^G_{k_n,n} = \frac{\hat{\gamma}^{(H)}_{k_n,n} - q_n \hat{\gamma}^{(MR)}_{k_n,n}}{1 - q_n},$$

where $\hat{\gamma}^{(MR)}_{k_n,n}$ is the so-called moment ratio estimator (see [1]) and q_n is the ratio of asymptotic biases of Hill and moment ratio estimators. From Theorem 1 in [1] it follows that $q_n = 1 - \rho$. Substituting q_n by $1 - \hat{\rho}_n$, where $\hat{\rho}_n$ is a weakly consistent estimator of the parameter ρ, we obtain the estimator

$$\gamma^G_{k_n,n}(\hat{\rho}_n) = \frac{\hat{\gamma}^{(H)}_{k_n,n} - (1 - \hat{\rho}_n)\,\hat{\gamma}^{(MR)}_{k_n,n}}{\hat{\rho}_n}.$$

This estimator was introduced firstly in [9]. Motivated by the fact that asymptotic bias and variance of existing estimators of ρ are high, the authors of [5] considered the estimator $\gamma^G_{k_n,n}(-1) \equiv H^L_{k_n,n}$. We observed that Theorem 1 in [7] is a direct consequence of the distributional representation of $H^L_{k_n,n}$ provided in [5], see (2.11) therein. We also recognize a numerical typo in variance in Theorem 1 ([7]), where the asymptotic variance 8 should be replaced by 5.

In [7] the estimator $H^L_{k_n,n}$ was introduced by using the t-score methodology to log-gamma distribution with pdf

$$f(x;\theta) = \begin{cases} \dfrac{\alpha^c \ln^{c-1}(x)}{\Gamma(c)\, x^{\alpha+1}}, & x \geq 1, \\ 0, & x < 1, \end{cases} \tag{1}$$

where θ consists of a pair of positive parameters (c, α).

In particular, in [7] we introduced a t-score moment estimator $\hat{\theta}$ as the solution of equations

$$\frac{1}{n} \sum_{i=1}^{n} S^k_F(x_i;\theta) = \mathbb{E}[S^k_F(\theta)],\ k = 1, \ldots, m,$$

where S_F is a scalar score of a random variable with distribution F (and smooth density f).

Since the log-gamma distribution in (1) has two parameters, we need $m = 2$ yielding two t-score equations, namely equations (15) and (16) in [7]. This defines the t-lgHill estimator of $1/\alpha$ given by (19) in [7].

The t-score (see [3]) of distribution F (with prescribed support) is defined as

$$T_F(x;\theta) = -\frac{1}{f(x;\theta)}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{\eta'(x)}f(x;\theta)\right),$$

where f is the pdf of F and η is a strictly increasing smooth function. It expresses a relative change of a "basic component of the density", i.e., density divided by the Jacobian of mapping η. Notice that we have the corresponding relation between t-score and Fisher score $S_G(x;\theta) = T_F(x;\theta)$, where $F(x) = G(\eta(x))$ and $S_G(y;\theta) = -\frac{\mathrm{d}}{\mathrm{d}y}\ln(g(y))$ with g as a pdf of G. In [8] it is shown how a dynamical system given by a t-score function for some class of monotonic data transformations generates consistent extreme value estimators.

We suppose that the solution x^* of $T_F(x;\theta) = 0$ is unique (t-score mean). Notice that the choice of η may change the value of t-score mean. Since our distribution has the support $(1,\infty)$, we want to find a function, which maps it into a whole real line. This can be done e.g. by choosing $\eta(x) = \ln(\ln(x))$, motivated by the well-known iterated logarithm law (see [10]). However first or third iteration cannot be used due to the demanded mapping between the support and the real line. This choice implies the form of t-score mean $x^* = \exp(\frac{c}{\alpha}) > 0$.

Clearly, T_F maps η uniquely on S_G. But, does there exist another smooth function $\eta(x)$ different from $\ln(\ln(x))$ such that $S_G(x;\theta) = T_F(x;\theta)$? This relation implies the exact second-order differential equation of the form

$$h(x;\theta) + \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{f(x;\theta)}{\eta'(x)}\right) = 0,$$

where $h(x;\theta) = S_G(x;\theta)f(x;\theta)$. If we allow η to depend on the parameters, then we have several different functions, see [8]. By direct integration we get

$$\eta'(x) = \frac{f(x;\theta)}{K(\theta) - H(x;\theta)}, \tag{2}$$

where H is a primitive function of h w.r.t. x. One can check that for $S_G(x;\theta) = \alpha\ln(x) - c$, the right hand side of equation (2) does not depend on parameters if and only if $K(\theta) = 0$. Therefore, $\eta(x) = \ln(\ln(x)) + k, k \in \mathbb{R}$, which implies the uniqueness.

2 Environmental applications of t-lgHill estimator

The good robustness quality of t-lgHill estimator was already illustrated in [7]. Namely, t-lgHill and t-Hill estimators are robust and also reasonably efficient and thus convenient for mass balance modelling of glaciers and threshold estimators for lava eruptions.

In this paper we illustrate the good quality of t-lgHill estimator on the example of the concentration of arsenic in drinking water in the rural area of the Arica and Parinacota Region, Chile. These data are novel, yet unpublished

and provided by the Regional Ministry of Health of the Region of Arica and Parinacota, corresponding to the measurements made in drinking water provided by the Rural Potable Water System (APR) or by a Precarious System (SP). Because these systems do not always provide drinking water according to the Chilean norm, the Regional Ministerial Secretary (SEREMI) of Health of Arica and Parinacota, periodically performs measurements of water quality in the rural area of the above mentioned Region. In the period 2017–2018 they reported 274 measurements of the concentration of arsenic in drinking water in various locations in the rural area. Many of these measurements are above the allowed standard 0.01 mg of arsenic per liter. In Figure 1 we plot a comparison of t-lgHill and Hill estimators. As we can see t-lgHill is very stable in comparison with Hill estimator in the range $k \in \{20, \ldots, 50\}$. By further analysis of Figure 1 we can conclude that:

(i) Considering a heuristic rule: a first constant flat area (from the left) in the plot gives a reasonable estimate of the tail index α. The t-lgHill plot is almost constant in the range $10 \leq k \leq 50$, while for the Hill plot we have $1 < k \leq 10$. Hence, keeping in mind a small number of observations, we can conclude that both estimators yield an estimate of the tail index approximately equal to 1.

(ii) From the t-lgHill plot we can see that estimates of α are close to 1 when the sample fraction k grows from 1 to $n-1$. The same conclusion also holds for Hill plot.

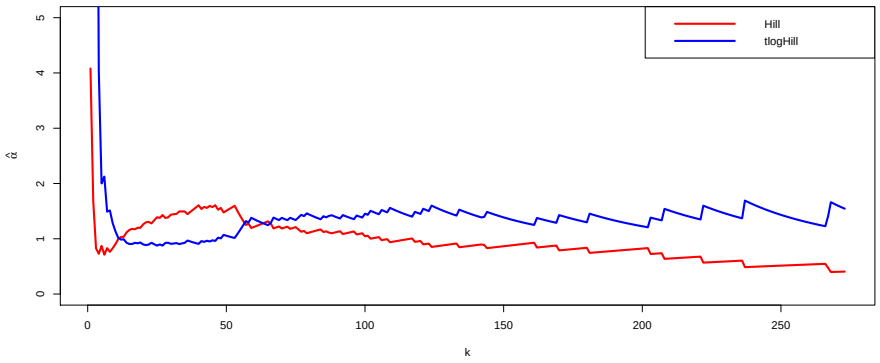

Fig. 1: Plots of t-lgHill and Hill estimators as functions of the sample fractions for the concentration of arsenic in drinking water.

We acknowledge the suggestion of the Referee to estimate the second-order parameter ρ, which in agreement with a Referee's comment is indeed close to -1 for large values of the sample fraction k. See also the range $k \in [266, 272]$ in Figure 2, where we provide a set of sample path of the estimator $\hat{\rho}_n^{(\tau)}$, studied in [4].

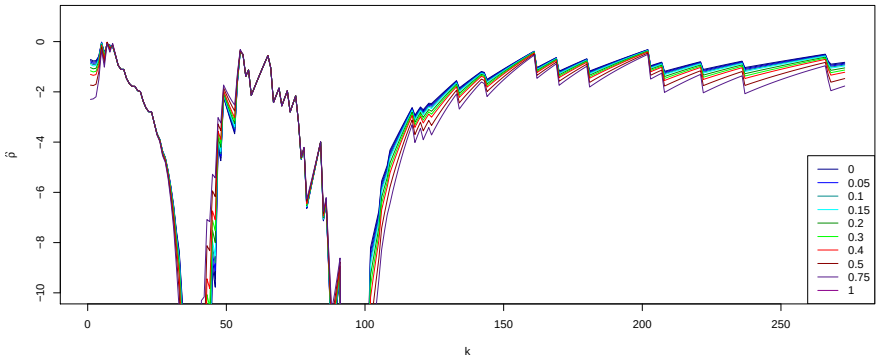

Fig. 2: $\hat{\rho}_n^{(\tau)}$ values for $\tau \in [0, 1]$

Acknowledgements We acknowledge the very professional support of Editor-in-Chief Professor Thomas Mikosch, the unknown Associate Editor and Referee for their constructive comments.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. De Haan, L., Peng, L.: Comparison of tail index estimators. Stat. Neerl. **52**(1), 60–70 (1998)
2. Dekkers, A.L.M., Einmahl, J.H.J., De Haan, L.: A moment estimator for the index of an extreme-value distribution. Ann. Stat. **17**(4), 1833–1855 (1989)
3. Fabián, Z.: Induced cores and their use in robust parametric estimation. Communications in Statistics - Theory and Methods **30**(3), 537–555 (2001). DOI 10.1081/STA-100002096
4. Fraga Alves, M.I., Gomes, M.I., de Haan, L.: A new class of semi-parametric estimators of the second order parameter. pp. 193–214. Portugaliae Mathematica (2003)
5. Gomes, I.M., Martins, J.M., Neves, M.: Alternatives to a semi-parametric estimator of parameters of rare events—the jackknife methodology*. Extremes **3**(3), 207–229 (2000). DOI 10.1023/A:1011470010228
6. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
7. Jordanova, P., Fabián, Z., Hermann, P., Střelec, L., Rivera, A., Girard, S., Torres, S., Stehlík, M.: Weak properties and robustness of t-hill estimators. Extremes **19**(4), 591–626 (2016). DOI 10.1007/s10687-016-0256-2
8. M.Stehlik, Aguirre, P., Girard, S., Jordanova, P., Kiselk, J., Torres, S., Sadovsky, Z., Rivera, A.: On ecosystems dynamics. Ecological Complexity **29**, 10 – 29 (2017). DOI https://doi.org/10.1016/j.ecocom.2016.11.002
9. Peng, L.: Asymptotically unbiased estimators for the extreme-value index. Statistics & Probability Letters **38**(2), 107 – 115 (1998). DOI https://doi.org/10.1016/S0167-7152(97)00160-0

10. Teicher, H.: On the law of the iterated logarithm. Ann. Probab. **2**(4), 714–728 (1974). DOI 10.1214/aop/1176996614