



Priority statement and some properties of t-IgHill estimator

Milan Stehlík^{1,2,3} · Jozef Kiseľák⁴ · Marijus Vaičiulis⁵ · Pavlina Jordanova⁶ · Ludy Núñez Soza⁷ · Zdeněk Fabián⁸ · Philipp Hermann⁹ · Luboš Střelec¹⁰ · Andrés Rivera^{11, 14} · Stéphane Girard¹² · Sebastián Torres¹³

Received: 19 February 2020 / Revised: 21 March 2020 / Accepted: 24 March 2020 /

Published online: 27 April 2020

© The Author(s) 2020

Abstract

We acknowledge the priority on the introduction of the formula of t-IgHill 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-IgHill 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-IgHill estimator · Hill estimator · T-score estimation · Asymptotic normality · Levels of arsenic in drinking water

AMS 2000 Subject Classifications 62F10 · 62F12

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.

✉ Milan Stehlík
mlnstehlik@gmail.com

Extended author information available on the last page of the article.

1 Theoretical motivation of t-IgHill estimator

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

$$H_{k_n,n}^L = \frac{M_{k_n,n}^{(2)} - (M_{k_n,n}^{(1)})^2}{M_{k_n,n}^{(1)}},$$

where

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

and $X_{1,n} \leq \dots \leq X_{n,n}$ are the order statistics of X_1, \dots, X_n . We recall that the statistics $M_{k_n,n}^{(j)}$, $j = 1, 2$ are introduced in Dekkers et al. (1989), while $\hat{\gamma}_{k_n,n}^{(H)} = M_{k_n,n}^{(1)}$ is nothing else but the Hill (1975) estimator. We also recall that the estimator $H_{k_n,n}^L$, as well as, the Hill estimator, is applicable for distributions F , which generalized inverse F^{\leftarrow} satisfies condition: $U(t) := F^{\leftarrow}(1 - t^{-1})$, $t > 1$ varies regularly at infinity with positive index γ . $H_{k_n,n}^L$ estimates parameter γ , but for better readability we also use the tail index $\alpha = 1/\gamma$.

We have found that the estimator $H_{k_n,n}^L$ was firstly developed in Gomes et al. (2000), see (2.10) therein. We have not been aware of this fact in time of publication of Jordanova et al. (2016), where we developed the estimator $H_{k_n,n}^L$, see Theorem 1 in Jordanova et al. (2016).

The main purpose of this priority letter is to demonstrate that the estimator $H_{k_n,n}^L$ was introduced in Gomes et al. (2000) and Jordanova et al. (2016) by applying different approaches. In Gomes et al. (2000) the estimator $H_{k_n,n}^L$ has been built by assuming the following second regular variation condition:

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

where $\rho < 0$ is the second order parameter, while $A(t)$ is a measurable function with the constant sign near infinity and $A(t) \rightarrow 0$ as $t \rightarrow \infty$. Precisely, the authors of the paper (Gomes et al. 2000) considered the generalized Jackknife statistic

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

where $\hat{\gamma}_{k_n,n}^{(MR)}$ is the so-called moment ratio estimator (see De Haan and Peng 1998) and q_n is the ratio of asymptotic biases of Hill and moment ratio estimators. From Theorem 1 in De Haan and Peng (1998) 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

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

This estimator was introduced firstly in Peng (1998). Motivated by the fact that asymptotic bias and variance of existing estimators of ρ are high, the authors of

Gomes et al. (2000) considered the estimator $\gamma_{k_n,n}^G(-1) \equiv H_{k_n,n}^L$. We observed that Theorem 1 in Jordanova et al. (2016) is a direct consequence of the distributional representation of $H_{k_n,n}^L$ provided in Gomes et al. (2000), see (2.11) therein. We also recognize a numerical typo in variance in Theorem 1 (Jordanova et al. 2016), where the asymptotic variance 8 should be replaced by 5.

In Jordanova et al. (2016) the estimator $H_{k_n,n}^L$ was introduced by using the t-score methodology to log-gamma distribution with pdf

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

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

In particular, in Jordanova et al. (2016) we introduced a t-score moment estimator $\hat{\theta}$ as the solution of equations

$$\frac{1}{n} \sum_{i=1}^n S_F^k(x_i; \theta) = \mathbb{E}[S_F^k(\theta)], \quad k = 1, \dots, m,$$

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 Jordanova et al. (2016). This defines the t-lgHill estimator of $1/\alpha$ given by (19) in Jordanova et al. (2016).

The t-score (see Fabián 2001) of distribution F (with prescribed support) is defined as

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

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 $S_G(y; \theta) = -\frac{d}{dy} \ln(g(y))$ with g as a pdf of G . In Stehlik et al. (2017) 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 Teicher 1974). 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

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

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 Stehlik et al. (2017). By direct integration we get

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

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 Eq. 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-IgHill estimator

The good robustness quality of t-IgHill estimator was already illustrated in Jordanova et al. (2016). Namely, t-IgHill 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-IgHill 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 Secretary 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

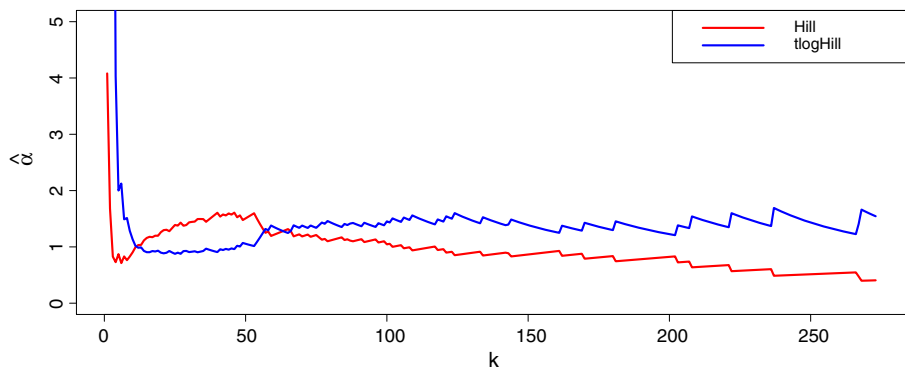


Fig. 1 Plots of t-IgHill and Hill estimators as functions of the sample fractions for the concentration of arsenic in drinking water

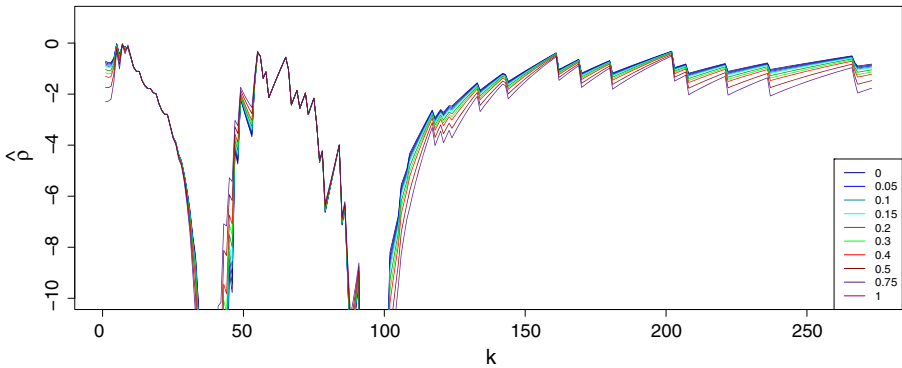


Fig. 2 $\hat{\rho}_n^{(\tau)}$ values for $\tau \in [0, 1]$

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 Fig. 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 $k \in \{20, \dots, 50\}$. By further analysis of Fig. 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.

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 Fig. 2, where we provide a set of sample path of the estimator $\hat{\rho}_n^{(\tau)}$, studied in Fraga Alves et al. (2003).

Acknowledgements Open access funding provided by Johannes Kepler University Linz. We acknowledge the very professional support of Editor-in-Chief Professor Thomas Mikosch, the unknown Associate Editor and Referee for their constructive comments.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- De Haan, L., Peng, L.: Comparison of tail index estimators. *Stat. Neerl.* **52**(1), 60–70 (1998)
- 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)
- Fabián, Z.: Induced cores and their use in robust parametric estimation. *Commun. Stat. Theory Methods* **30**(3), 537–555 (2001). <https://doi.org/10.1081/STA-100002096>
- 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)
- 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). <https://doi.org/10.1023/A:1011470010228>
- Hill, B.M.: A simple general approach to inference about the tail of a distribution. *Ann. Stat.* **3**, 1163–1174 (1975)
- Jordanova, P., Fabián, Z., Hermann, P., Štěpánek, L., Rivera, A., Girard, S., Torres, S., Stehlik, k., M.: Weak properties and robustness of t-hill estimators. *Extremes* **19**(4), 591–626 (2016). <https://doi.org/10.1007/s10687-016-0256-2>
- Stehlik, M., Aguirre, P., Girard, S., Jordanova, P., Kiselák, J., Torres, S., Sadovsky, Z., Rivera, A.: On ecosystems dynamics. *Ecol. Complex.* **29**, 10–29 (2017). <https://doi.org/10.1016/j.ecocom.2016.11.002>
- Peng, L.: Asymptotically unbiased estimators for the extreme-value index. *Stat. Probabil. Lett.* **38**(2), 107–115 (1998). [https://doi.org/10.1016/S0167-7152\(97\)00160-0](https://doi.org/10.1016/S0167-7152(97)00160-0)
- Teicher, H.: On the law of the iterated logarithm. *Ann. Probab.* **2**(4), 714–728 (1974). <https://doi.org/10.1214/aop/1176996614>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Milan Stehlík^{1,2,3} · Jozef Kiselák⁴ · Marijus Vaičiulis⁵ · Pavlina Jordanova⁶ · Ludy Núñez Soza⁷ · Zdeněk Fabián⁸ · Philipp Hermann⁹ · Luboš Štřelec¹⁰ · Andrés Rivera^{11, 14} · Stéphane Girard¹² · Sebastián Torres¹³

- ¹ Department of Applied Statistics, Linz Institute of Technology, Johannes Kepler University in Linz, Altenbergerstrasse 69, 4040, Linz, Austria
- ² Institute of Statistics, Universidad de Valparaíso, Valparaíso, Chile
- ³ Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, USA
- ⁴ Institute of Mathematics, Faculty of Science, P. J. Šafárik University in Košice, Košice, Slovakia
- ⁵ Vilnius University Institute of Data Science and Digital Technologies, Akademijos 4, LT-08663, Vilnius, Lithuania
- ⁶ Faculty of Mathematics and Informatics, Konstantin Preslavsky University of Shumen, Universitetska Str. 115, 9700, Shumen, Bulgaria
- ⁷ Departamento de Matemática, Facultad de Ciencias, Universidad de Tarapacá, Avenida 18 de Septiembre 2222, Casilla 7-D, Arica, Chile
- ⁸ Institute of Computer Science of the Czech Academy of Sciences, Pod vodárenskou věží 2, 18200, Prague, Czech Republic
- ⁹ Department of Applied Statistics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040, Linz, Austria
- ¹⁰ Department of Statistics and Operation Analysis (FBE), Mendel University in Brno, Zemědělská 1, 61300, Brno, Czech Republic
- ¹¹ Departamento de Geografía, Universidad de Chile, Avda. Portugal 84, Santiago, Chile
- ¹² Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000, Grenoble, France
- ¹³ Department of Mathematics, Federico Santa María Technical University, Valparaíso, Chile
- ¹⁴ Instituto de Conservación Biodiversidad y territorio, Facultad de Ciencias Forestales y Recursos NATurales, Universidad Austral de Chile, Valdivia, 5090000, Chile