This talk is dedicated to Ivette's 65th birthday

On parametric approach for extremes modeling in ecology

Milan Stehlík

Department of Applied Statistics
Johannes Kepler University in Linz
Email: Milan.Stehlik@jku.at

Talk addressing

Beran, J., Schell, D. and Stehlík, M. (2013a). The Harmonic Moment Tail Index Estimator: asymptotic distribution and robustness, *Annals of Inst. Stat. Math.*

Jordanova, P., Dušek, J. and Stehlík, M. (2013b). Modeling methane emission by the infinite moving average process, *Chemometrics and ILS*, 122, 40-49

Jordanova, P., Dušek, J. and Stehlík, M. (2013c). Microergodicity effects on ebullition of methane modelled by Mixed Poisson process with Pareto mixing variable, *Chemometrics and ILS*

Stehlík, M., Fabián Z. and Střelec L. (2012), Small sample robust testing for Normality against Pareto tails, *Communications in Statistics - Simulation and Computation*, 41:7, 1167-1194

Motivation: Ecological Modelling: methane

- **Greenhouse effect** increased by concentrations of CO_2 and CH_4

  Followed by extremes in weather (*IntPannelClimChange* 2007): floods, huricanes and tornados

- `CNN: recent massive tornado struck an area outside Oklahoma City on Monday afternoon, estimated peak wind ranged from 200 to 210 mph,Insurance claims will likely top $1 billion`, etc., Hong-Kong typhoons,..)

- **Cold period?** An eminent Mexican geophysicist Victor Manuel Velasco says that despite predictions of global warming based on computer models, the world may be on the verge of an eighty-year cold period similar to the "little ice age" experienced by Europe from 1300 to 1800 A.D..

- Víctor Manuel Velasco and some other geophysicists reject global warming theory, says world on verge of 'mini ice age' **Also:** *Hockey-stick-controversy*, Russian scientist Dr. H. Abdussamatov, of the St Petersburg Pulkovo Astronomical Observatory

Motivation: Earth & tropospheric methane-can we model that?

- **Tropospheric methane: can we model that? Design?** *Nature* 1991: Vaghjiani and Ravishankara, New measurement of the rate coefficient for the reaction of OH with methane.

- try to find design: Rodríguez-Díaz, Santos-Martín, Waldl and Stehlík (2012). Filling and D-optimal designs for the correlated Generalized Exponential models, *Chemom. and ILS* 114 10-18.

- **Earth CH_4: can we model it?..Questions posed by Jan Beirlant**

  In Jordanova, Dusek, Stehlík (2013) we model CH_4 flux "emissions"

  $$Em_\lambda(t) = a + bt + c.t^2 + X_\lambda(t) + Z_\lambda(t) - Z_{-,\lambda}(t) + \epsilon_\lambda(t).I_{[-c_1,c_2]}(\epsilon_\lambda(t)),\ t \geq 0,$$

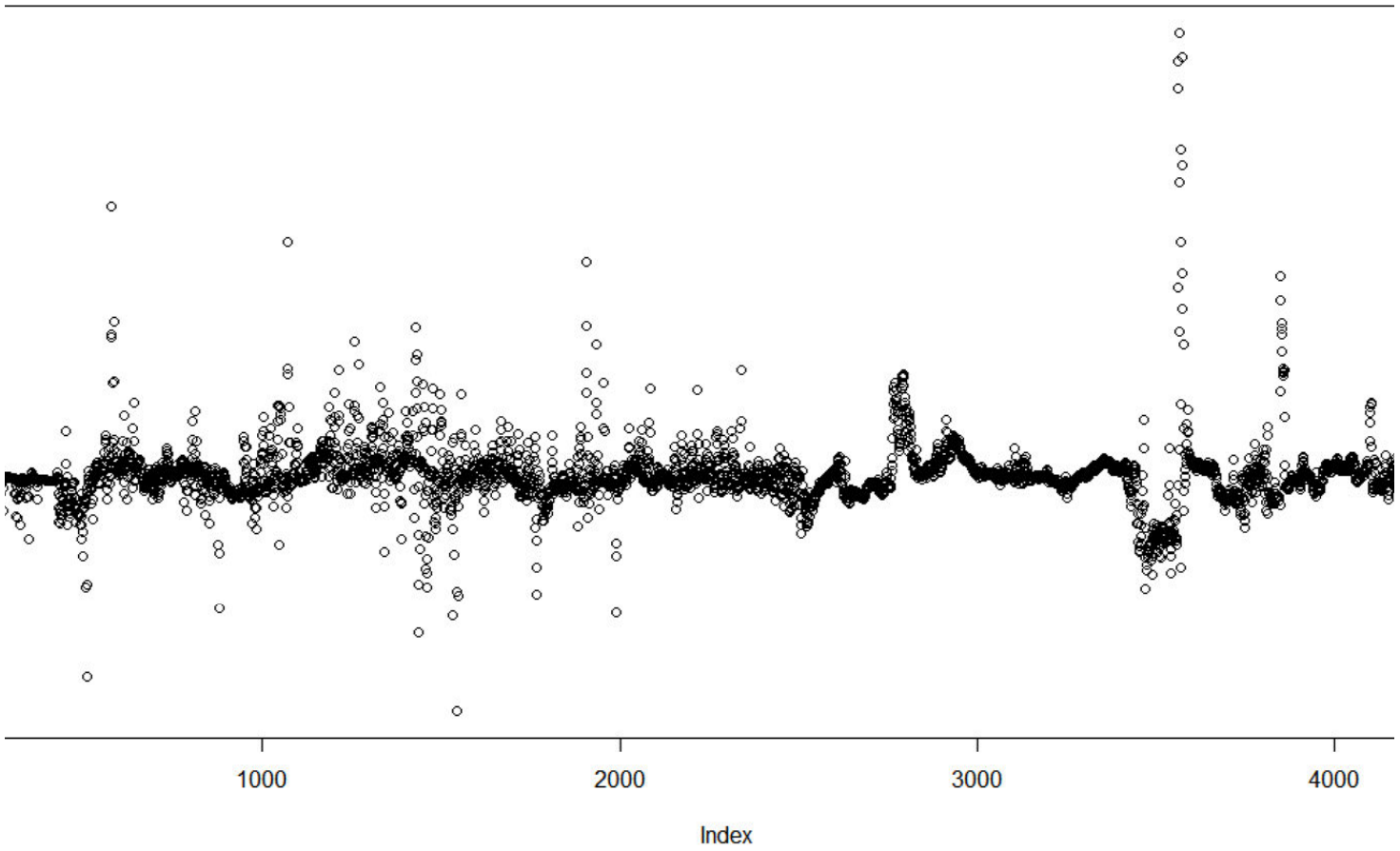
  $X_\lambda(t)$ is a moving average process,

  $\epsilon_\lambda(t) \sim N(0, \sigma^2)$ relates to standard diffusion,

  $Z_\lambda(t) = Pareto(\alpha_1)I_{Pareto(\alpha_1)>c_2}(t), \quad Z_{-,\lambda}(t) = Pareto(\alpha_2)I_{Pareto(\alpha_2)>c_1}(t).$

  Here λ is the intensity of the point process of exceedances of the process

  $$Em_{S,\lambda}(t) = Em_\lambda(t) - a - bt - c.t^2,\ t \geq 0$$


1000
2000
3000
4000
Index

Estimation: Standard score?

(Pearson and Filon (1898), Edgeworth (1908), Fisher (1925), Lehmann and Casella, 1998)

- The score is the gradient wrt θ of the logarithm of the likelihood function. $S(\theta, X) = \frac{\partial}{\partial \theta} \log L(\theta; X)$

  The score indicates the sensitivity of L (its derivative normalized by its value)

- The variance of the score=Fisher information, $\mathcal{I}(\theta) = E_\theta(S^2)$

- Systematic/axiomatic approach to score function? (categorial theory)

- **Example 1: Normal distribution** (an archetypical distribution: Gauss 1830, Pearson 1890, Fisher 1935) $suppf = R$, standard estimation $\eta = id$

  $S(x; \theta) = \frac{d}{dx} \log f(x; \theta)$

  Mle $\widehat{\theta}$: $X_1, .., X_n$ iid, then $\sum_i S(X_i; \widehat{\theta}) = 0$

t-score (Fabián 2001), robust tails estimation (Jürg Hüsler)?

- Let $\mathcal{X}$ be support of distribution F with density f, continuously differentiable according to $x \in \mathcal{X}$ and let $\eta : \mathcal{X} \to \mathbb{R}$ be given by (*Johnson 1949 and gen.*)

$$\eta(x) = \begin{cases} x & \text{if} \quad \mathcal{X} = \mathbb{R} \\ \log(x-a) & \text{if} \quad \mathcal{X} = (a, \infty) \\ \log \frac{x}{1-x} & \text{if} \quad \mathcal{X} = (0,1). \end{cases} \tag{1}$$

  Then the *transformation-based score* or shortly the *t-score* is defined by

$$T(x) = -\frac{1}{f(x)} \frac{d}{dx} \left(\frac{1}{\eta'(x)} f(x) \right). \tag{2}$$

  (2) expresses a relative change of a 'basic component of the density', the density divided by the Jacobian of mapping (1).

t-estimation

- As a measure of central tendency of distribution F has been suggested the zero of the t-score *t-mean*, x^* : $T(x^*) = 0$,

- The t-score moments $E_f T^k = \int_{\mathcal{X}} T^k(x) dF(x), \quad k = 1, 2, ...$

- 't-score moment estimate' of θ in the form

$$\widehat{\theta}_n : \qquad \frac{1}{n} \sum_{i=1}^{n} T^k(x_i; \theta) = E T^k(\theta), \qquad k = 1, ..., m,$$

  which turns out to be consistent and asymptotically normal.

The Pareto distribution:

- **Example 2: standard score estimation**: $\eta = id, f(x,\theta) = \theta x^{-\theta-1}; x > 1$

$$S(x;\theta) = \frac{1}{\theta} - \log x \quad (unbounded\ x \to \infty)$$

$$MLE:\ \widehat{\theta} = \frac{n}{\sum \log X_i}$$

- **Example 3: t-score estimation** (Fabian and Stehlík 2008, Stehlík et al. 2010,2012): $\eta(x) = \log(x-1)$

$$S(x;\theta) = \theta(1 - \frac{\theta+1}{\theta x}) \quad (bounded\ x \to \infty)$$

$$t - estimator\ (specialM - est.):\ \widehat{\theta} = \frac{1}{\bar{X}_H - 1},$$

where $\bar{X}_H = \frac{n}{\sum_i 1/X_i}$ is a harmonic mean.

Robust testing for normality against heavy tails

- There exist more than 1000 tests for normality..(Thode 2002) But which of them are robust? And is there a systematic way how to apply them against heavy tails?

- How to efficiently test for a normality against European Pareto distribution, possibly contaminated, with the density $\frac{\alpha c^\alpha}{x^{\alpha+1}}, x > c$.

  Two-step procedure Stehlík *et al.* (2011):

  Step 1) To estimate the nominal value of Pareto tail α under alternative

  Step 2) Based on Step 1 choose test for normality in the RT class.

- For that reason we relax the form of j-th theoretical moment $\mu_j = E(X - E(X))^j$ estimator by taking $M_j(T(F_n), r) = \frac{1}{n-2r}\sum_{m=1+r}^{n-r} \varphi_j(X_{m:n} - T(F_n))$, where $X_{1:n} < X_{2:n} < \ldots < X_{n:n}$ be the order statistics, $j \in \{0, 1, 2, 3, 4\}$ and φ_j is tractable and continuous function $\varphi_0(x) = \sqrt{\pi/2}|x|$, $\varphi_1(x) = x$, $\varphi_2(x) = x^2$, $\varphi_3(x) = x^3$ and $\varphi_4(x) = x^4$. The RT class is defined by

  $$\text{RT} = \frac{k_1(n)}{C_1}\left(\frac{M_{j_1}^{\alpha_1}(T_1,r_1)}{M_{j_2}^{\alpha_2}(T_2,r_2)} - K_1\right)^2 + \frac{k_2(n)}{C_2}\left(\frac{M_{j_3}^{\alpha_3}(T_3,r_3)}{M_{j_4}^{\alpha_4}(T_4,r_4)} - K_2\right)^2.$$

Robust testing for normality

- **Location Functional**: Bickel and Lehman (1975)

  Let $T(F)$ be a function defined on the set of distribution functions. We say that $T(F)$ is a location functional if the following conditions hold:

  1. if G is stochastically larger than F then $T(G) \geq T(F)$
  2. $T(F_{aX+b}) = aT(F_X) + b$
  3. $T(F_{-X}) = -T(F_X)$

- **RT class relevant location functionals:** mean, median, trimmed mean, pseudo-median $M_3 = H_n^{-1}(1/2)$, where $H_n(y) = \int F_n(2y - x)h(x)dx$.

- RT test mimics around 600 tests, a known special cases of RT class: JB (Jarque and Bera 1980), Urzua's Jarque-Bera (1996), RJB (Gel and Gastwirth 2008), Geary's test (1935), Uthoff's test (1973), skewness test, kurtosis test

score modified Hill estimator

- Gomes,I. M. and Stehlík, M., 2013, On the Hill estimator and its modifications, *Springer Ed. Volume.*

- many authors tried to robustify the Hill estimator, but core is MLE based (Fraga Alves (2001), Gomes and Oliveira (2003), Li et al. (2010)- introduced powers of original statistics, Beran and Schell 2011).

- the influence function of Hill estimator is slowly increasing, but unbounded.

- **t-Hill estimator: Fabián and Stehlík (2009)** - the distribution sensitive and robust:

$$\widehat{\gamma}_k = \frac{1}{\widehat{\alpha}_k} = H^*_{k,n} = \frac{1}{\frac{1}{k}\sum_{j=1}^{k} \frac{X_{n-k,n}}{X_{n-j+1,n}}} - 1, \tag{3}$$

- **t-Hill consistency: Stehlík et al. (2011)**

- **Harmonic Moment tail Index Estimator: Beran, Schell and Stehlík (2012)**

  –

$$\widehat{\gamma}_{n,k}(\beta) = \frac{1}{\widehat{\theta}_k} = \frac{1}{\beta - 1} \left\{ \left[\frac{1}{k} \sum_{j=1}^{k} \left(\frac{X_{n-k,n}}{X_{n-j+1,n}} \right)^{\beta-1} \right]^{-1} - 1 \right\},$$

  $\beta > 0$ is tuning parameter. For $\beta = 2$ t-Hill, $\beta = 1$ Hill estimator. Related to MOP estimator of Brilhante, Gomes and Pestana (2013).

  – As it turns out the tuning parameter β allows for regulating the trade-off between efficiency and robustness.

  For $\beta > 1$ the effect of large contaminations is bounded, since the Harmonic Moment Tail Index Estimator benefits from the properties of the harmonic mean.

  However, a larger value of β also implies an increased variance.

  For $\beta < 1$ the Harmonic Moment Tail Index Estimator also has a higher variance than Hill's estimator.

– However, in some situations, it possesses a smaller asymptotic bias such that the AMSE is smaller. The class of Harmonic Moment Tail Index Estimators is introduced in Henry III (2009), using the tuning parameter $1/(\beta - 1)$. Asymptotic results are obtained however only under the trivial and very restrictive assumption of an exact Pareto tail beyond a fixed finite threshold u.

– **t-Hill is distribution sensitive**

  * The score moment estimator is usually simple so that it makes possible for many distributions to apply a score moment Hill-like estimator.

  * **log-gamma distribution** on support $\mathcal{X} = (1, \infty)$ and density

$$f(z) = \frac{c^{\alpha}}{\Gamma(\alpha)} (\log z)^{\alpha - 1} z^{-(c+1)}. \tag{4}$$

  * $\eta : (1, \infty) \to \mathbb{R}, \eta(x) = \log(\log x)$

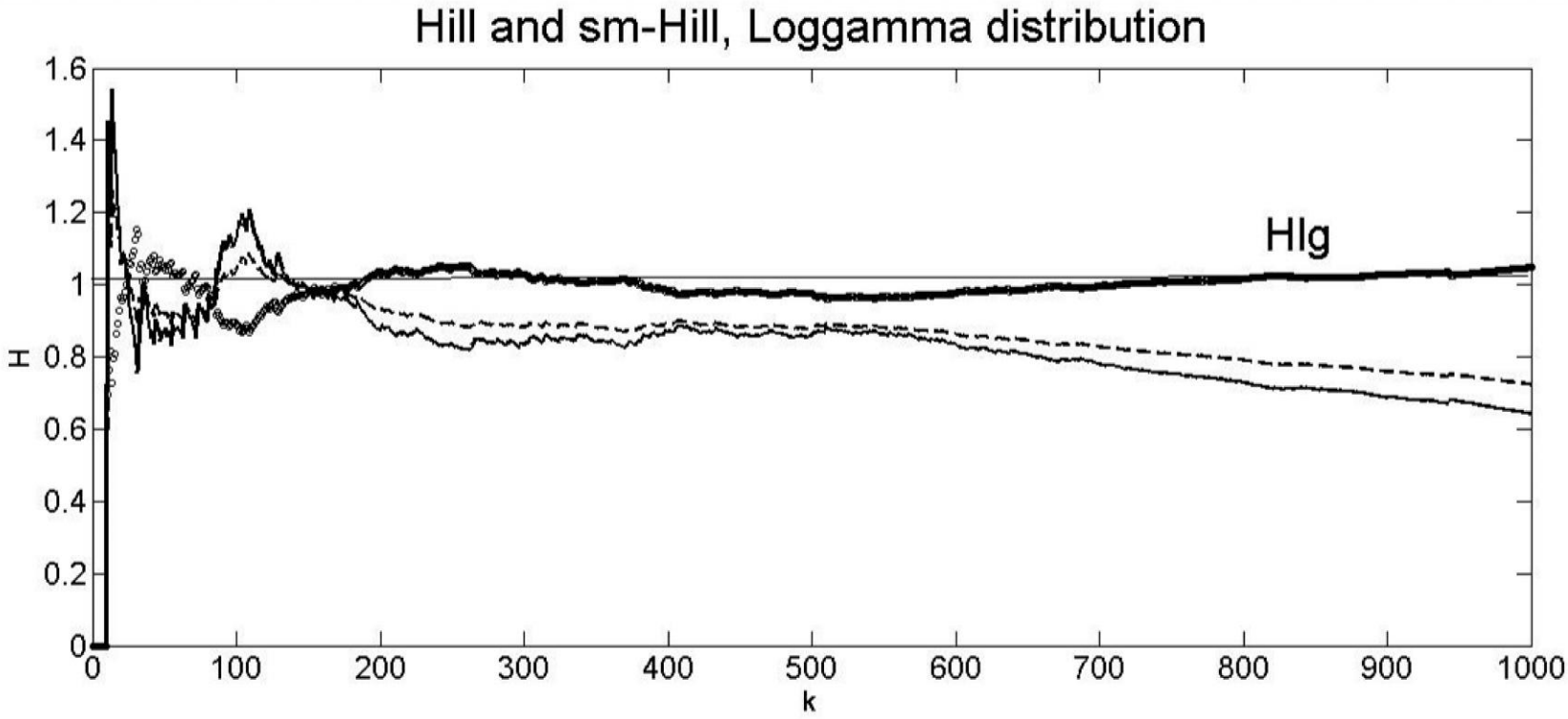
$$T(x) = \frac{1}{f(x)} \frac{d}{dx} \left(-(\log x)^c \alpha x^{-c} \right) = c \log x - \alpha$$

so that the 'loglog' t-mean is $x^* = e^{\alpha/c}$. As the 'second log-log moment' $ET^2 = E[c^2 \log^2(x/x^*)] = \alpha$, the estimation equations are

$$\sum_{i=1}^{n} c \log x_i - \alpha = 0, \quad \sum_{i=1}^{n} (c \log x_i - \alpha)^2 = \alpha$$

By setting $\widehat{s}_1 = \frac{1}{k}\sum_{i=1}^{k} \log x_i$ and $\widehat{s}_2 = \frac{1}{k}\sum_{i=1}^{k} \log^2 x_i$, it follows $\widehat{\alpha} = \widehat{s}_1 \widehat{c}$ and $\widehat{c}(\widehat{s}_2 - \widehat{s}_1^2) = \widehat{s}_1$ so that the Hill-like estimate of the tail index is given by closed-form expression

$$\widehat{\gamma}_k = \frac{1}{\widehat{c}}_k = \frac{\widehat{s}_2}{\widehat{s}_1^2} - 1.$$


Hill and sm-Hill, Loggamma distribution
Hlg
H
k
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
0
100
200
300
400
500
600
700
800
900
1000

Inverse Problem: Related transformations $\widehat{\gamma}_{n,k}(\beta)$ how looks a bijective support map $\eta : (1; \infty) \to R,\ \eta \in C^2$? Then we receive ODE for the η which is a Bernoulli-type ODE.

Can be transformed to linear 1st order differential equation solvable by quadrature defining map

$$\eta(x) = \int_{x0}^{x} \frac{t}{-\frac{\theta t}{\theta+2} + \frac{(\theta+1)t^{1-\beta}}{\theta(\theta-\beta+2)} + c_1 t^{-\theta-1}} dt$$

Easy solutions:

Hill estimator ($\beta = 1$): $\eta(x) = x$

t-Hill estimator: ($\beta = 2$): $\eta(x) = \log(x - 1)$

Mixed Poisson process with Pareto mixing variable: Microergodicity effects

- *Jordanova-Dusek-Stehlik 2013c, Chemom. & ILS* Modeling the dependence of the ebullition of methane on time is useful for ecology.

- ebullition is bubble transport of gasses from places with a high gas production or concentration to neighboring environment mainly in the soil-water-air interfaces. Ebullition is typical process for direct gas transport in wetland or aquatic ecosystems where accumulated gas in deeper sediments transferred directly to the atmosphere via gas bubbles. (Yamamoto 1976, Martens and Klump 1980).

- Assuming that λ is a constant, the properties of the corresponding point process of exceedences over a high threshold are well described in Davis and Resnick (1985).

- The intensity of the process of exceedences over a high threshold of $Em_{S,\lambda}(\frac{t}{\lambda})$ is 1.

Mixed Poisson process with Pareto mixing variable: Microergodicity effects

- The methodological novelty in Jordanova-Dusek-Stehlik 2013c: we show that the intensity of the observed Poisson processes of exceedances is not a constant.

- The intensity is usually considered as a constant because the parameters of the process are estimated using only one sample path.

- Our conclusions come from the fact that if we resample the data, in such a way that to preserve the dependence structure of the process and its main properties we obtain the process of the innovations that is an uncountable mixture of a moving average processes and Lomax (shifted Pareto) mixing random variable (r.v.) plus i.i.d. innovations. (*so not a constant* λ, *Pareto mixing: Ross Leadbetter's talk*)

Thank you for your attention!

References

Beran J. and Schell D. (2010). On robust tail index estimation, *Computational Statistics and Data Analysis,* doi:10.1016/j.csda.2010.05.028

Beran, J., Schell, D. and Stehlík, M. (2013). The Harmonic Moment Tail Index Estimator: asymptotic distribution and robustness, Annals of Inst. Stat. Math.

Bickel P.J. and Lehmann E.L. (1975). Descriptive Statistics for Nonparametric Models II. Location. *Ann. Statist.* **3**(5): 1045-1069.

Fabián Z. (2001) Induced cores and their use in robust parametric estimation, *Communication in Statistics, Theory Methods*, **30** (2001), pp.537-556.

Fabián Z, Stehlík M (2008) A note on favorable estimation when data is contaminated, *Communications in Dependability and Quality Management*, **11** (4): 36-43.

Fabián, Z. and Stehlík, M. (2009). On robust and distribution sensitive Hill like method, IFAS Technical report 43, April 2009.

Fraga Alves, M.I. (2001). A location invariant Hill-type estimator. *Extremes* 4, 199-217.

Hill, B. (1975). A Simple General Approach to Inference About the Tail of a Distribution, *The Annals of Statistics*, 3, 1163-1173.

Huisman, R. et al. (2001). Tail-Index Estimates in Small Samples, *Journal of Business & Economic Statistics,* vol. 19(2), 208-16.

Gomes,I. M. and Stehlík, M., 2013, On the Hill estimator and its modifications, submitted.

Johnson, N. L., Systems of frequency curves generated by methods of translations, *Biometrika* 36 (1949), pp.149-176.

Jordanova, P., Dušek, J. and Stehlík, M. (2013a). Modeling methane emission by the infinite moving average process, *Chemometrics and ILS*, 122, 40-49

Stehlík, M., Potocký, R., Waldl, H. and Fabián, Z. (2010). On the favourable estimation of fitting heavy tailed data, *Computational Statistics,* 25:485-503.

Stehlík, M., Fabián, Z. and Střelec, L. (2012). Small sample robust testing for Normality against Pareto tails, *Communications in Statistics - Simulation and Computation,* 41:7, 1167-1194

Vandewalle, B., Beirlant, J., Christmann, A., Hubert, M. (2007). A robust estimator for the tail index of Pareto-type distributions, *Computational Statistics & Data Analysis*, Vol. 51 (12): 6252-6268.

Vidmar G., Blagus R., Střelec L. and Stehlík M., 2012: "Business indicators of healthcare quality: outlier detection in small samples", *Applied Stochastic Models in Business and Industry*, 28, 282-295

Yamamoto S, Alcauskas JB, Crozier TE. Solubility of methane in distilled water and seawater. J Chem Eng Data, 21 (1976) 78-80.