Nonparametric estimation of the conditional tail index

Laurent Gardes, Stephane Girard

Nonparametric estimation of the conditional tail index

Laurent Gardes & Stéphane Girard

INRIA Rhône-Alpes (team MISTIS). 655, avenue de l'Europe, Montbonnot. 38334 Saint-Ismier Cedex, France

{Laurent.Gardes, Stephane.Girard}@inrialpes.fr

Abstract: We present a nonparametric family of estimators for the tail index of a Pareto-type distribution when covariate information is available. Our estimators are based on a weighted sum of the log-spacings between some selected observations. This selection is achieved through a moving window approach on the covariate domain and a random threshold on the variable of interest. Asymptotic normality is proved under mild regularity conditions and illustrated for some weight functions. Finite sample performances are presented on a real data study.

Key words and phrases: Tail index, extreme-values, nonparametric estimation, moving window.

1 Introduction

In extreme-value statistics, one of the main problems is the estimation of the tail index associated to a random variable Y. This parameter, denoted by γ, drives the distribution tail heaviness of Y. For instance, when γ is positive, the survival function of Y decreases to zero geometrically, and the larger γ is, the slower is the convergence. Here, we focus on the situation where some covariate information x is recorded simultaneously with the quantity of interest Y. In the general case, the tail heaviness of Y given x depends on x, and thus the tail index is a function $\gamma(x)$ of the covariate. Such situations occur for instance in climatology where one may be interested in how climate change over years might affect extreme temperatures. Here, the covariate is univariate (the time). Bivariate examples include the study of extremes rainfall as a function of the geographical location.

We investigate how to combine nonparametric smoothing techniques with extreme-value methods in order to obtain efficient estimators of $\gamma(x)$. The proposed estimator is based on a selection, thanks to a moving window approach, of the observations to be used in the estimator of the extreme-value index. This estimator is a weighted sum of the rescaled log-spacings between the selected largest observations. This approach has several advantages. From the theoretical point of view, very few assumptions are made on the regularity of $\gamma(x)$ and on the nature of the covariate

(in particular, we do not suppose that x is finite dimensional). From the practical point of view, the estimator is easy to compute since it is closed-form and thus does not require optimization procedures.

2 Estimators of the conditional tail index

Let E be a metric space associated to a metric d. We assume that the conditional distribution function of Y given $x \in E$ is

$$F(y,x) = \begin{cases} 1 - y^{-1/\gamma(x)} L(y,x) & \text{if } y \geq \theta > 0, \\ 0 & \text{if } y < \theta, \end{cases} \tag{2.1}$$

where $\gamma(.)$ is an unknown positive function of the covariate x and, for x fixed, $L(.,x)$ is a slowly varying function. Model (2.1) is well known to be equivalent to the so-called first order condition

$$U(y,x) \stackrel{def}{=} \inf\{s; F(s,x) \geq 1 - 1/y\} = y^{\gamma(x)} \ell(y,x),$$

where, for x fixed, $\ell(.,x)$ is a slowly varying function. Given a sample $(Y_1, x_1), \ldots, (Y_n, x_n)$ of independent observations from (2.1), our aim is to build a point-wise estimator of the function $\gamma(.)$. More precisely, for a given $t \in E$, we want to estimate $\gamma(t)$, focusing on the case where the design points $x_1, \ldots, x_n$ are non random. To this end, let us denote by $B(t, h_{n,t})$ the ball centered at point t and with radius $h_{n,t}$ defined by $B(t, h_{n,t}) = \{x \in E,\ d(x,t) \leq h_{n,t}\}$ where $h_{n,t}$ is a positive sequence tending to zero as n goes to infinity. The proposed estimate uses a moving window approach since it is based on the response variables $Y_i's$ for which the associated covariates $x_i's$ belong to the ball $B(t, h_{n,t})$. The proportion of such design points is thus defined by

$$\varphi(h_{n,t}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x_i \in B(t, h_{n,t})\}$$

and plays an important role in this study. It describes how the design points concentrate in the neighborhood of t when $h_{n,t}$ goes to zero. Thus, the nonrandom number of observations in $[\theta, \infty) \times B(t, h_{n,t})$ is given by $m_{n,t} = n\varphi(h_{n,t})$. Let $\{Z_i(t),\ i = 1, \ldots, m_{n,t}\}$ be the response variables $Y_i's$ for which the associated covariates $x_i's$ belong to the ball $B(t, h_{n,t})$ and let $Z_{1,m_{n,t}}(t) \leq \ldots \leq Z_{m_{n,t},m_{n,t}}(t)$ be the corresponding order statistics. Our family of estimators of $\gamma(t)$ is defined by

$$\hat{\gamma}_n(t, W) = \sum_{i=1}^{k_{n,t}} i \log\left(\frac{Z_{m_{n,t}-i+1,m_{n,t}}(t)}{Z_{m_{n,t}-i,m_{n,t}}(t)}\right) W\left(i/k_{n,t}, t\right) \Bigg/ \sum_{i=1}^{k_{n,t}} W\left(i/k_{n,t}, t\right),$$

where $k_{n,t}$ is a sequence of integers such that $1 \leq k_{n,t} < m_{n,t}$ and $W(.,t)$ a function defined on $(0,1)$ such that $\int_0^1 W(s,t)ds \neq 0$. Thus, without loss of generality, we can assume that $\int_0^1 W(s,t)ds = 1$.

3 Asymptotic normality

Let us first briefly describe the conditions required to obtain the asymptotic normality of our estimators. First, we need the classical second order condition on the slowly varying function $\ell(.,t)$ which relies on a bias function denoted by $b(.,t)$ satisfying $b(y,t) \to 0$ as $y \to \infty$ and on a second order parameter $\rho(t) < 0$. This function drives the asymptotic bias of most tail index estimators. Second, we assume that the function $U(z,.)$ is α-Lipschitzian for $\alpha \leq 1$. Finally, conditions on the weight function W are the same as the ones introduced in [1] in the case where no covariate information is available. Denoting by

$$b_{n,t} \stackrel{def}{=} b\left(\frac{n\varphi(h_{n,t})}{k_{n,t}}, t\right),$$

our main result is the following: If $n\varphi(h_{n,t})/k_{n,t} \to \infty$, $k_{n,t} \to \infty$, $k_{n,t}^{1/2} b_{n,t} \to \lambda(t) \in \mathbb{R}$ and $k_{n,t}^{1/2} h_{n,t}^{\alpha} \to 0$, then

$$k_{n,t}^{1/2}\left(\hat{\gamma}_n(t,W) - \gamma(t) - b_{n,t}\mathcal{AB}(t,W)\right) \stackrel{d}{\to} \mathcal{N}\left(0, \gamma^2(t)\mathcal{AV}(t,W)\right),$$

where we have defined

$$\mathcal{AB}(t,W) = \int_0^1 W(s,t)s^{-\rho(t)}ds \text{ and } \mathcal{AV}(t,W) = \int_0^1 W^2(s,t)ds.$$

If we consider the constant weight function $W^{\mathrm{H}}(s,t) = 1$ for all $s \in [0,1]$, the expression of the obtained estimator is clearly the same as the Hill [2] estimator. Furthermore, we can show that the choice $W^{\mathrm{Z}}(s,t) = -\log(s)$ leads to an estimator similar to the Zipf [3, 4] estimator.

4 Illustration on real data

In this section, we propose to illustrate our approach on the daily mean discharges (in cubic meters per second) of the Chelmer river collected by the Springfield gauging station, from 1969 to 2005. These data are provided by the Centre for Ecology and Hydrology (United Kingdom) and are available at `http://www.ceh.ac.uk/data/nrfa`. In this context, the variable of interest Y is the daily flow of the river and the bi-dimensional covariate $x = (x_1, x_2)$ is built as follows: $x_1 \in \{1969, 1970, \ldots, 2005\}$ is the year of measurement and $x_2 \in \{1, 2, \ldots, 365\}$ is the day. The size of the dataset is $n = 13,505$. The smoothing parameter $h_{n,t}$ as well as the number of upper order statistics $k_{n,t}$ are assumed to be independent of t, they are thus denoted by h_n and k_n respectively. They are selected by minimizing the distance between conditional Hill and Zipf estimators discretized on the grid $T = \{1969, 1970, \ldots, 2005\} \times \{15, 45, \ldots, 345\}$. This heuristics is commonly used in functional estimation and relies on the idea that, for a properly chosen pair (h_n, k_n) both estimates should approximately yield the same value. The selected value of h_n corresponds to a smoothing over 4 years on x_1 and 2 months on x_2. Each ball $B(t, h_n)$, $t \in T$

contains $m_n = n\varphi(h_n) = 1089$ points and $k_n = 54$ rescaled log-spacings are used. The resulting conditional Zipf estimator is presented on Figure 4.1. The obtained values are located in the interval $[0.2, 0.7]$. It appears that the estimated tail index is almost independent of the year but strongly dependent of the day. The heaviest tails are obtained in September, which means that, during this month extreme flows are more likely than during the rest of year.

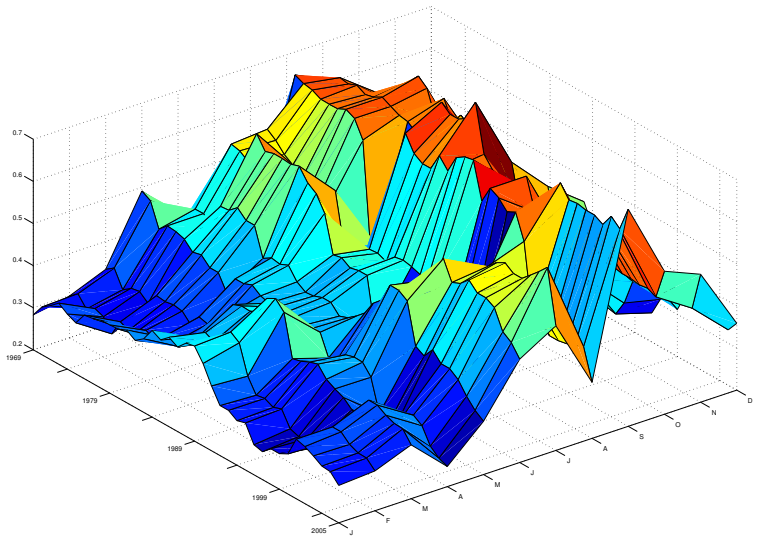

Figure 4.1: Conditional Zipf estimator of the tail index computed on the real dataset. Two covariates are available: The year ranging from 1969 to 2005 and the day ranging from 1 to 365. For the sake of readability, only the first letter of the corresponding month is represented.

References

[1] Beirlant, J., Dierckx, G., Guillou, A. and Stărică, C. (2002). On exponential representations of log-spacings of extreme order statistics, *Extremes*, **5**, 157–180.

[2] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.

[3] Kratz, M. and Resnick, S. (1996). The QQ-estimator and heavy tails, *Stochastic Models*, **12**, 699–724.

[4] Schultze, J. and Steinebach, J. (1996). On least squares estimates of an exponential tail coefficient, *Statistics and Decisions*, **14**, 353–372.