Estimation of heavy-tailed density functions with application to WWW-traffic

Natalia M. Markovich, *Institute of Control Sciences*
Russian Academy of Sciences,
Profsoyuznay 65, 117997 Moscow, Russia
markovic@ipu.rssi.ru
Traffic characterization

Abstract— **The estimation of heavy-tailed probability density function is an important tool for the description of the Web-traffic data and the solution of applied problems such as classification. The paper is devoted to the non-parametric estimation of a heavy-tailed probability density function by a variable bandwidth kernel estimator. Two approaches are used: (1) a preliminary transformation of the data to provide more accurate estimation of the density at the tail domain; (2) the discrepancy method based on the Kolmogorov-Smirnov statistic to evaluate the bandwidth of the kernel estimator. It is proved that the discrepancy method may provide the fastest achievable order of the mean squared error. An application to Web data analysis is presented.**

Index Terms—**Heavy-tailed distribution, discrepancy method, tail index, Web-traffic.**

I. Introduction

Measurements of Web-traffic shows that some WWW-traffic characteristics like file sizes, sizes and durations of sub-sessions are independent and heavy-tailed distributed. The latter implies that the "outliers" (or measurements those differ strongly from the main part of observations) play in these data a significant role and cannot be extracted from consideration.

In this paper, the non-parametric estimation of heavy-tailed densities by empirical sample $X^n = \{X_1, \ldots, X_n\}$ of independent identically distributed (i.i.d.) random variables with the density function $f(x)$, is considered. The problem is that a histogram cannot be directly applied to heavy-tailed densities since it is defined on the bounded interval. It provides an absolutely misleading estimation of the density at the "tail" domain.

The kernel estimator

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \qquad (1)$$

where K is a kernel function from R to R (as a rule, $K(x)$ is some probability density, e.g., a normal one), h is a smoothing parameter or a bandwidth (it has approximately the same meaning as a bin width of the histogram estimate), is defined on the whole real axes. However, it may provide sharp peaks at "outliers" or over-smoothes the density due to the constant bandwidth. The good fitting of heavy-tailed densities requires different amount of smoothing in different locations of the distribution. Roughly speaking, a tail domain of a heavy-tailed density containing sparse observations requires a flat estimate therewith the body of the density requires a sharper estimate. That is why, a variable bandwidth kernel estimator

$$\hat{f}^A(x|h) = \frac{1}{nh}\sum_{i=1}^{n} f(X_i)^{1/2} K\left((x - X_i) f(X_i)^{1/2}/h\right) \qquad (2)$$

approximates such densities better [1]. Since $f(X_i)$ is unknown, the following estimator

$$\widetilde{f}^A(x|h_1, h) = \frac{1}{nh}\sum_{i=1}^{n} \hat{f}_{h_1}(X_i)^{1/2} \qquad (3)$$

$$\cdot K\left(\frac{x - X_i}{h}\hat{f}_{h_1}(X_i)^{1/2}\right)$$

is used in practice. Usually, the non-variable bandwidth kernel estimator (1) is used as a pilot estimator $\hat{f}_h(x)$. The variable bandwidth kernel estimator is not reliable for the estimation of the density outside the range of the sample at least for compactly supported kernels like Epanechnikov's kernel $K(x) = 3/4(1 - x^2)\mathbb{1}(\mid x \mid \leq 1)$. In order to estimate the density outside the range of the sample better and particularly to apply the estimators defined on bounded intervals (e.g., a histogram) to heavy-tailed densities, the preliminary transformation of the data may be useful. The survey of transformations is given in Section II.

Another question is how to evaluate the bandwidth h in (2) and (3). The theory shows that if f has four bounded derivatives, the kernel is non-negative and symmetric (e.g. Epanechnikov's kernel) and if h is chosen asymptotic to any constant multiple of $n^{-1/9}$, where n is the sample size, then the mean squared error ($MSE = \mathbb{E}\left(\hat{f}^A(x|h) - f(x)\right)^2$) of (2) has the fastest achievable rate $n^{-8/9}$ [2]. Indeed, the MSE reflects the accuracy of the estimation at the main part of the density better but not at the tail where values are small.

For the practice the data-dependent selection methods of h (e.g., the cross-validation, the discrepancy method) work better. Apart of computational problems caused by the search of maximum of the maximum likelihood functional, the cross-validation method has slow convergence rates and high sampling variability [3]. A weighted version of the cross-validation was proposed in [4] for estimator (3). However, it was not proved that this method provides the optimal order $n^{-1/9}$ of h and consequently the fastest rate $n^{-8/9}$ of MSE.

In Section III the discrepancy method is presented. It is proved, that it may provide the variable bandwidth kernel estimator (3) with $MSE \sim n^{-8/9}$ using the samples of moderate sizes.

Further, we focus on estimate (3). We shall combine the advantages of transformations of the data and the discrepancy method to improve the behavior of the estimate outside of the sample and provide the $MSE \sim n^{-8/9}$.

Let us explain why non-parametric density estimates with good behavior at the tail domain are required. Apart of the visual data analysis and the estimation of moments of the distribution, this feature is very significant if densities of many populations are compared. Such comparison is required in the classification (pattern recognition). If one uses an empirical Bayesian classification algorithm, then observations will be classified by the comparison of the corresponding density estimates of each class. Since the object can arise in the tail domain as well as in the body, a tail estimator with good properties is principal for the classification. Application of classification techniques to Web data analysis is given in [5].

In Section IV variable bandwidth kernel estimator (3) with the discrepancy method as a smoothing tool are applied to WWW-traffic characteristics. Preliminary adapted transformation (7) of the data described in Section II is used.

II. Transformation approach

Estimators with data transformations are the alternative to variable bandwidth kernel estimators. The background of the transformation idea is the necessity of the different smoothing at different locations of a heavy-tailed density. Then back-transformed density estimates with fixed smoothing parameters work like location-adaptive estimates.

Let $T(x)$ be a monotone increasing continuously differentiable "one-to-one" transformation function (the derivative of the inverse function T^{-1} is assumed to be continuous). We apply it to our data $X_1, ..., X_n$ and obtain the new sample $Y_1, \ldots, Y_n$ ($Y_i = T(X_i)$). The distribution function of Y_i is given by

$$G(x) = \mathbb{P}\{Y_i < x\} = \mathbb{P}\{T(X_i) < x\} \qquad (4)$$
$$= \mathbb{P}\{X_i < T^{-1}(x)\} = F(T^{-1}(x)),$$

its density reads

$$g_0(x) = G^{'}(x) = f(T^{-1}(x))(T^{-1}(x))^{'}.$$

The density $g_0(x)$ of the r.v. Y_i is estimated by some estimator $\hat{g_0}(x)$ and after the re-transformation we get the density estimate of the X_i by the formula:

$$\hat{f}(x) = \hat{g_0}(T(x))T^{'}(x). \qquad (5)$$

One may take variable bandwidth kernel estimator (3) or standard kernel estimator (1) as $\hat{g_0}(x)$.

The selection of $T(x)$ is a principal problem. By (4) a transformation $T(x)$ is completely determined by the distribution functions $G(x)$ and $F(x)$. One can select any "target" $G(x)$, but $F(x)$ is unknown.

In [6] transformations $T : R_+ \to [0, 1]$ were proposed. It was proved that for kernel estimates with compact kernels the transformation to an isosceles triangular density $\phi^{tri}(x)$ on $[0, 1]$ and for a histogram the transformation to a uniform density $\phi^{uni}(x)$ provide the optimal convergence rate in the metric of space L_1. Since such $T(x)$ and, therefore, the distribution of $Y_j = T(X_j)$ depend on the unknown distribution function F, it is impossible to obtain an absolute identity of g_0 and ϕ^{tri} (or $\phi^{uni}(x)$). Hence, it is proposed in [6] to use instead of F some parametric models. However, the concrete models were not indicated and their influence on the decay rate at infinity of the re-transformed estimates was not discussed.

In [7], [8] the families of fixed transformations $T_\lambda(x)$ (independent on $F(x)$) given by

$$T_\lambda(x) = \begin{cases} x^\lambda sign(\lambda), & \text{if } \lambda \neq 0, \\ \ln x, & \text{if } \lambda = 0 \end{cases}$$

are considered. Here, λ is the parameter minimizing the functional $\int_R (g^{''}(y))^2 dy$, $g(x)$ is unknown density of the transformed r.v. $Y_1 = T_\lambda(X_1)$ that requires a preliminary estimation. Since the function $\int_R (g^{''}(y))^2 dy$ shows the curvature of the density then such transformations are applied for better restoration of curvy but not necessary heavy-tailed densities. In [9] the fixed transformation $T(x) = (2/\pi)\arctan x$, that provided a good accuracy for some heavy-tailed densities is considered. However, without the assumptions about the type of the distribution any transformation may lead to a density that is difficult to estimate by a limited sample and hence, one cannot provide an accurate estimation of the tails. In order to improve the estimation at the tails in [5] a transformation $T_{\hat{\gamma}}(x) : R_+ \to [0,1]$, which is adapted to the data (via the estimate $\hat{\gamma}$ of some parameter γ called the tail index[1]) is proposed. To construct $T_{\hat{\gamma}}(x)$ the distribution function of the triangular distribution $\Phi^{+tri}(x) = (2x - x^2)1\!\!\mathrm{I}\{x \in [0,1]\} + 1\!\!\mathrm{I}\{x > 1\}$ with the density $\phi^{+tri}(x) = 2(1-x)1\!\!\mathrm{I}\{x \in [0,1]\}$ is taken as the "target" distribution function $G(x)$ and the Pareto distribution function

$$\Psi_{\hat{\gamma}}(x) = \begin{cases} 1 - (1+\hat{\gamma}x)^{-1/\hat{\gamma}}, & \text{if} \quad x \geq 0, \\ 0, & \text{if} \quad x < 0. \end{cases} \tag{6}$$

is taken as the "fitted" distribution function $F(x)$. Then the transform from $\Psi_{\hat{\gamma}}(x)$ to $\Phi^{+tri}(x)$ is defined by the formulae

$$T_{\hat{\gamma}}(x) = (\Phi^{+tri})^{-1}(\Psi_{\hat{\gamma}}(x)) = 1 - \sqrt{1 - \Psi_{\hat{\gamma}}(x)} \tag{7}$$
$$= 1 - (1+\hat{\gamma}x)^{-1/(2\hat{\gamma})}.$$

The Pareto choice is widespread and motivated by a theorem, [10] which states that, for a certain class of distributions and for a sufficiently high threshold u of the r.v. X, the conditional distribution of the overshoot $Y = X - u$, provided that X exceeds u, converges to a Generalized Pareto distribution. The triangular "target" distribution function is selected in such a way to get the continuous density of the transformed r.v. $Y_1 = T_{\hat{\gamma}}(X_1)$, when the estimate $\hat{\gamma}$ deviates from γ. The choice of a uniform distribution function as the "target" distribution function leads to a discontinuity of the density of Y_1 at 1 and, hence, to the problem in the density estimation.

As a quality measure of the re-transformed kernel estimate of $f(x)$ one may consider the mean integrated squared error ($MISE$ at the interval Ω)

$$\begin{aligned} MISE^h(\hat{\gamma}, \Omega) &= \mathbb{E}\int_{\Omega} (\hat{f}(x) - f(x))^2 dx \\ &= \mathbb{E}\int_{\Omega} (\hat{g}_h(T_{\hat{\gamma}}(x)) - g(T_{\hat{\gamma}}(x)))^2 T'_{\hat{\gamma}}(x) dT_{\hat{\gamma}}(x) \\ &= \mathbb{E}\int_{\Omega^*} (\hat{g}_h(y) - g(y))^2 T'_{\hat{\gamma}}(T^{-1}_{\hat{\gamma}}(y)) dy, \end{aligned}$$

where $\Omega^* = T_{\hat{\gamma}}(\Omega)$ and $g(x)$ is the density, which is actually estimated instead of $g_0(x) = f(T^{-1}_{\gamma}(x))(T^{-1}_{\gamma}(x))'$ (since $\hat{\gamma} \neq \gamma$),

$$g(x) = f(T^{-1}_{\hat{\gamma}}(x))(T^{-1}_{\hat{\gamma}}(x))'.$$

$\hat{g}_h(x)$ is some estimate of $g(x)$ with the smoothing parameter h.

Since for transformation (7) we have $T'_{\hat{\gamma}}(T^{-1}_{\hat{\gamma}}(x)) = 0.5\,(1-x)^{1+2\hat{\gamma}}$ at $\Omega^* = [0,1]$, i.e., $0 < T'_{\hat{\gamma}}(T^{-1}_{\hat{\gamma}}(x)) \leq c$ holds at Ω^* then we get

$$MISE^h(\Omega) \leq c \int_{\Omega^*} \mathbb{E}(\hat{g}_h(y) - g(y))^2 dy. \tag{8}$$

It means, that the order of the $MISE$ of re-transformed estimates at Ω is at least not worse than the order of the MSE of $\hat{g}_h(y)$.

III. Discrepancy method

The idea of the discrepancy method is to select h as a solution of the discrepancy equation

$$\rho(\hat{F}, F_n) = \delta.$$

Here, $\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(t)dt$, $\hat{f}(t)$ is some estimate of the density, δ is a known uncertainty of the estimation of the distribution function $F(x)$ by the empirical distribution function $F_n(t)$, i.e. $\delta = \rho(F, F_n)$, $\rho(\cdot,\cdot)$ is a metric in the space of distribution functions. The discrepancy method was proposed and investigated in [11], [12] for the smoothing of nonparametric density estimates. Since δ is usually unknown, in these papers some quantiles of the limit distribution of the Mises-Smirnov statistic and Kolmogorov-Smirnov statistic [2]

$$\sqrt{n}D_n = \sqrt{n} \sup_{-\infty < x < \infty} |F(x) - F_n(x)|$$

were used as δ. For Kolmogorov-Smirnov statistic one can take the value $\delta = 0.5$ corresponding to the mode of the distribution of the latter statistic [11]. Let h_* be a solution of the equation

$$\sup_{-\infty < x < \infty} |F_n(x) - F^A_{h,h_1}(x)| = \delta n^{-1/2}, \tag{9}$$

where $F^A_{h,h_1}(x) = \int_{-\infty}^{x} \widetilde{f}^A(t \mid h_1, h)dt$.

Further, we assume that the estimate (1) is taken as $\hat{f}_{h_1}$ in (3).

Theorem 1: Let $X^n = \{X_1, \ldots, X_n\}$ be i.i.d. r.v.s with a density $f(x)$. Let the non-random bandwidth $h_1 = cn^{-1/5}$ in $\hat{f}_{h_1}(x)$. We assume that $K(x)$, $x \in R$ is continuous, positive and satisfies

$$\sup_x K(x) < \infty, \qquad \int_R K(x)dx = 1.$$

Then any solution $h_* = h_*(n)$ of equation (9) obeys the condition

$$h_* \to 0, \qquad \text{as} \qquad n \to \infty.$$

[1] The tail index defines the shape of the tail.

[2] The distributions of these statistics do not depend on $F(x)$.

Theorem 2: Let the density $f(x)$ be estimated by variable bandwidth kernel estimate $\widetilde{f}^A(x|h_1,h)$ (3). Assume the conditions on $f(x)$ and $K(x)$ given in Theorem 1. In addition, we assume that $K(x)$ has the order $m+1$,[3] $f(x)$ has $m-1$ continuous derivatives and its mth derivative is bounded in the neighborhood of 0: $0 < \eta_1 \leq |f^{(m)}(x)| \leq \eta_2$, η_1 and η_2 are constants. Let the non-random bandwidth h_1 in $\hat{f}_{h_1}$ obeys the conditions: $h_1 \to 0$, $nh_1 \to \infty$ as $n \to \infty$. Then any solution $h_* = h_*(n)$ of equation (9) obeys the condition

$$\mathbb{P}\{h > \rho n^{-1/(\alpha(m+1))}\} < \exp\left(-2n^{1-2/\alpha}\right), \qquad (10)$$

where $\rho = (2(1+\delta)/G)^{1/(m+1)}$ is a constant, $G = \eta_1/(m+1)! \int_{-\infty}^{\infty} y^{m+1} K(y)dy$, for any $\alpha > 2$.

Remark 1: Pareto distribution (6) gives an example of the distribution that satisfies the condition of Theorem 2. Let $\Re$ be a compact set of R. Given $\varepsilon > 0$, we use the following notation of [13]

$$\Re^{\varepsilon} \equiv \{x \in R : \text{for some} \quad y \in \Re, \quad \|x-y\| \leq \varepsilon\},$$

where $\|\cdot\|$ is the usual Euclidean norm.

Theorem 3: Let the density $f(x)$ be estimated by variable bandwidth kernel estimate $\widetilde{f}^A(x|h_1,h)$ (3). Assume the conditions on $f(x)$ and $K(x)$ given in Theorem 2 and $m = 3$. In addition, we assume that $f(x)$ and $1/f(x)$ have four continuous derivatives and $f(x)$ is bounded away from zero, on $\Re^{\varepsilon}$. Besides, we assume that $K(x)$ is symmetric. Let us assume, that a non-random bandwidth h_1 in (3) obeys $h_1 = c_* n^{-1/5}$, where c_* is some constant. Then for any solution h_* of (9) we have

$$\begin{aligned} &\mathbb{P}\{\mathbb{E}\widetilde{f}^A(x|h_1,h_*) - f(x) > \psi(x)n^{-4/9}\} \\ < \quad & 2\exp\left(-2n^{1/9}\right), \end{aligned}$$

where $\psi(x) = (K_3/24)\,(d/dx)^4\,(1/f(x))\rho^4$, ρ is defined in Theorem 2.

Corollary 1: Assume the conditions of Theorem 3. Let us assume, that $\mathbb{E}(Z \cdot \hat{f}^A(x|h)) = 0$, where Z is a standard normal r.v. Then, $MSE(\widetilde{f}^A(x|h_1,h_*))$ may reach the order $n^{-8/9}$ if a maximal solution of (9) h_* has the order $n^{-1/9}$.

Remark 2: Since the function of the r.v. X_1 (that is one term in the sum $\hat{f}^A(x|h)$) and the normal distributed r.v. Z are independent, the condition $\mathbb{E}(Z \cdot \hat{f}^A(x|h)) = 0$ is not rigorous.

IV. Application to Web-traffic Characteristics

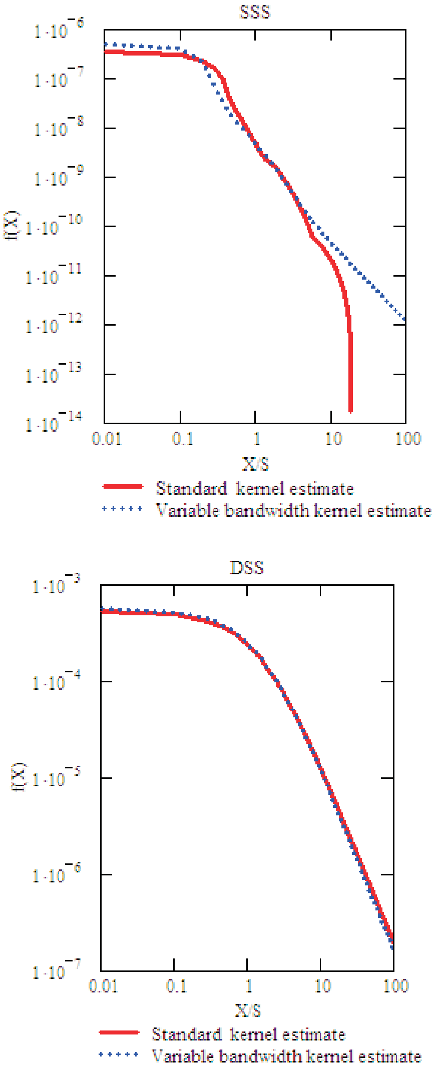

Fig. 1. The density estimation by standard kernel estimator (1) and variable bandwidth estimator (3) with the smoothing by the discrepancy method (9) for the data sets s.s.s., d.s.s..

We apply estimators (1) and (3), where h is estimated by discrepancy method (9) to the real Web-data. These data gathered in the Ethernet segment of the Department of Computer Science at the University of Würzburg were analyzed in papers [5], [9], [14]. The data describe the characteristics of sub-sessions, i.e., the size of a sub-session (s.s.s) in bytes and the duration of a sub-session (d.s.s.) in seconds, as well as the characteristics of the transferred Web-pages, i.e., the size of the response (s.r.) in bytes and the inter-response time (i.r.t.) in seconds. The description of all these r.v.s is presented in Table I. To simplify the calculation the data were scaled, i.e. all values were divided by the scaling parameter s (see Table I).

[3] A kernel has the order p if the conditions

$$\int K(x)dx = 1, \qquad \int x^i K(x)dx = 0, i = 1, ..., p-1;$$

$$\int x^p K(x)dx = K_{p-1} \neq 0$$

hold.

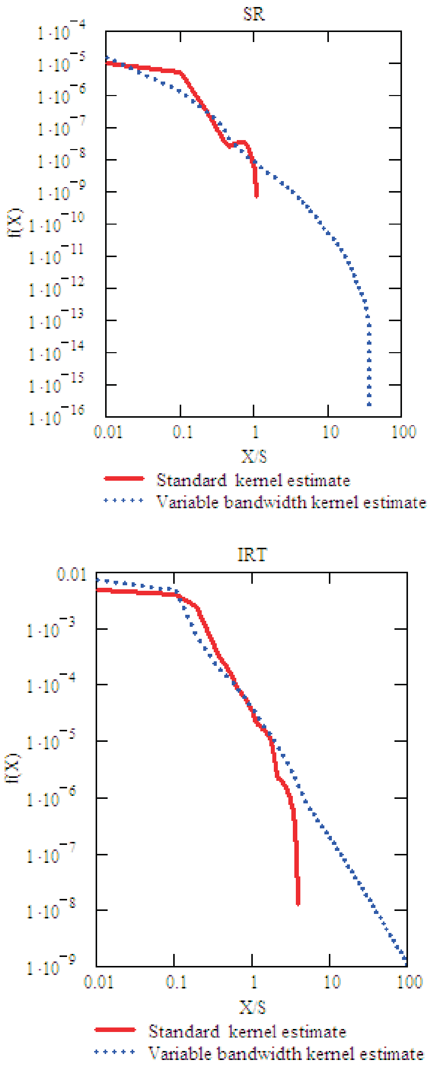

Fig. 2. The density estimation by standard kernel estimator (1) and variable bandwidth estimator (3) with the smoothing by the discrepancy method (9) for the data sets s.r., i.r.t..

In order to check whether the measurements corresponded to samples s.s.s., s.r., d.s.s. and i.r.t. are derived from heavy-tailed distributions, we estimated the tail index γ by the popular Hill's method

$$\gamma(k,n)=\frac{1}{k}\sum_{i=1}^{k}\ln X_{(n-i+1)}-\ln X_{(n-k)},$$

where $X_{(1)} \leq \ldots \leq X_{(n)}$ are order statistics of the sample X^n.

In Table II one can see the estimates $\gamma(k,n)$ and the values of the number of retained data k, for all data sets that are taken from [14]. Observing the estimates of γ one may conclude that the estimates of the tail index $\alpha = 1/\gamma$ are always less than 2 for all considered data sets. It follows from the extreme value theory [15], that at least βth moments, $\beta \geq 2$ of the distribution of s.s.s., d.s.s., s.r., i.r.t. are not finite. The distributions of considered Web-traffic characteristics are heavy-tailed.

TABLE I

DESCRIPTION OF THE DATA

	s.s.s.(B)	d.s.s.(sec)	s.r.(B)	i.r.t.(sec)
Sample Size	373	373	7107	7107
Minimum	128	2	0	$6.543 \cdot 10^{-3}$
Maximum	$5.884 \cdot 10^{7}$	$9.058 \cdot 10^{4}$	$2.052 \cdot 10^{7}$	$5.676 \cdot 10^{4}$
Mean	$1.283 \cdot 10^{6}$	$1.728 \cdot 10^{3}$	$5.395 \cdot 10^{4}$	80.908
StDev	$4.079 \cdot 10^{6}$	$5.206 \cdot 10^{3}$	$4.931 \cdot 10^{5}$	728.266
s	10^{7}	10^{3}	10^{6}	10^{3}

TABLE II

ESTIMATION OF THE TAIL INDEX AND THE BANDWIDTHS FOR WEB-TRAFFIC CHARACTERISTICS

r.v.	$\hat{\gamma}$	k	h_1	h_s	h_v	$1.01 - T_{\hat{\gamma}}(X_{(n)})$
s.s.s.	0.949	50	0.059	0.155	0.320	0.382
s.r.	0.898	211	0.020	0.059	0.175	0.75
i.r.t.	0.712	211	0.042	0.110	0.250	0.519
d.s.s.	0.601	50	0.170	1.000	1.100	0.063

Hence, we may transform the data by transformation (7). The density $g_0(x)$ of the new r.v. has been estimated by (1) and (3) with Epanechnikov's kernel. The re-transformed estimate of the unknown density $f(x)$ was calculated by (5):

$$\hat{f}(x) = 0.5\hat{g}_0(1-(1+\hat{\gamma}x)^{-1/(2\hat{\gamma})})(1+\hat{\gamma}x)^{-1/(2\hat{\gamma})-1}.$$

Bandwidths h_s and h_v in Table II have been selected by the discrepancy method (9) with $\delta = 0.5$ and correspond to estimates (1) and (3), respectively. The value h_1 of the non-variable kernel estimate $\hat{f}_{h_1}(x)$ in (3) is calculated by the formula

$$\hat{h}_{OS} = \left(\frac{243K_2}{35\mu_2(K)^2 n}\right)^{1/5} \cdot s,$$

where s is the sample standard derivation, $\mu_2(K) = \int z^2K(z)dz$ (the over-smoothing bandwidth selection [17]). For Epanechnikov's kernel $K_2 = 3/5$, $\mu_2 = 1/5$. This formula provides the minimal upper bound of the theoretical value of h that corresponds to the optimal $MSE \sim n^{-4/5}$ of estimate (1).

The re-transformed kernel estimates (1) and (3) have been calculated for samples d.s.s. and s.s.s., s.r. and i.r.t. (Figs. 1, 2). The estimate $f(x) = g(x/s)/s$ is shown, where $g(x/s)$ is the re-transformed estimate constructed by scaled data. A logarithmic scale both for the X and Y axes is used.

The curves of re-transformed kernel estimate (1) corresponding to all sets apart of d.s.s. and of re-transformed kernel estimate (3) for the sample s.r. are truncated for large values of x/s because the kernel is not wide enough. Such boundary effects are typical for kernel estimates that are used for finite densities. In this case, the kernel estimate of the density $g_0(x)$ located on $[0,1]$ may equal to 0 at the neighborhood of 1 beyond the maximal observation of the sample. It reflects on the re-transformed estimate. It becomes equal to 0 at the tail and logarithms of these values go to $-\infty$. In [5] it was shown that the choice $h = 1.01 - T_{\hat{\gamma}}(X_{(n)})$, where $T_{\hat{\gamma}}(x)$ is transformation (7), $X_{(n)}$ is the maximal observation in the sample, may improve the boundary problems. One can compare the values of h_s, h_v and $1.01 - T_{\hat{\gamma}}(X_{(n)})$ in Table II. Obviously, the discrepancy method selects larger values h that are closer to $1.01 - T_{\hat{\gamma}}(X_{(n)})$, for estimate (3) rather than for estimate (1). Hence, the re-transformed estimate (3) provides better estimation of the density at the tail domain for Web-traffic characteristics.

Appendix I
Proof of Theorem 1.

Proof: Suppose, that $h_* \not\to 0$ as $n \to \infty$. It implies, that for any integer $N > 0$ $\exists n > N$ such as $h_* = h_*(n) > H_*$, where H_* is some positive constant. We shall prove, that for such h_* it holds $\sup_{-\infty<x<\infty} |F_n(x) - F^A_{h_*,h_1}(x)| \not\to 0$ as $n \to \infty$.

For any solution h_* one may represent the divergence in (9) using the replacement $u = (t - X_i)\,\hat{f}_{h_1}(X_i)^{1/2}/h_*$

$$\begin{aligned} & F_n(x) - F^A_{h_*,h_1}(x) \\ = \; & \frac{1}{n}\sum_{i=1}^{n}(\theta\,(x - X_i) \\ - \; & \frac{\hat{f}_{h_1}(X_i)^{1/2}}{h_*}\int_{-\infty}^{x} K\left(\frac{t - X_i}{h_*}\hat{f}^{1/2}_{h_1}(X_i)\right)dt) \\ = \; & \frac{1}{n}\sum_{i=1}^{n}\left(\theta\,(x - X_i) - \int_{-\infty}^{t_i} K(u)du\right), \end{aligned} \tag{11}$$

where we denote $t_i = t_i(h_*) = (x - X_i)\,\hat{f}_{h_1}(X_i)^{1/2}/h_*$,

$$\theta\,(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

We denote

$$\varphi\,(h_*, X_i, x) = \int_{-\infty}^{t_i} K(u)du,$$

$$\Phi\,(h_*, n, x) = \sum_{i=1}^{n}\left(\theta\,(x - X_i) - \int_{-\infty}^{t_i} K(u)du\right).$$

First, we assume that $h_* = h_*(n)$ is a constant.

Assume for simplicity, that a sample contains only one r.v. X_1. Since for any i

$$0 < \int_{-\infty}^{t_i} K(u)du < 1$$

holds, then

$$-1 < \theta\,(x - X_1) - \int_{-\infty}^{t_1} K(u)du < 1 \qquad \forall x.$$

Hence,

$$\begin{aligned} & \sup_{-\infty<x<\infty} |F_n(x) - F^A_{h_*,h_1}(x)| \\ = \; & \sup_{x} |\theta\,(x - X_1) - \int_{-\infty}^{t_1} K(u)du| = \xi, \end{aligned}$$

$0 < \xi < 1$. The same conclusion may be obtain when a sample has more than one r.v.

Let us assume that h_* is not a constant. Without loss of generality, one can consider the sequence

$$h^*_1 \leq h^*_2 \leq ... \leq h^*_j \leq ...,$$

where $h^*_j = h_*(n_j) = H_* + j\Delta$, Δ is some positive constant, and $N < n_1 \leq n_2 \leq ... \leq n_j \leq ...$.

We get

$$\begin{aligned} & \varphi\,(h^*_1, X_i, x) - \varphi\left(h^*_j, X_i, x\right) \\ = \; & \int_{-\infty}^{t_i(H_*+\Delta)} K(u)du - \int_{-\infty}^{t_i(H_*+j\Delta)} K(u)du \\ = \; & \int_{t_i(H_*+j\Delta)}^{t_i(H_*+\Delta)} K(u)du > 0, \end{aligned} \tag{12}$$

where $t_i(H_* + j\Delta) = t_i(h^*_j)$. It implies, that

$$\varphi\,(h^*_1, X_i, x) > \varphi\,(h^*_2, X_i, x) > ... > \varphi\left(h^*_j, X_i, x\right) > ...$$

for any fixed X_i and x. Hence, it follows from (12), that

$$\begin{aligned} & \Phi\left(h^*_j, n, x\right) - \Phi\,(h^*_1, n, x) \\ = \; & \sum_{i=1}^{n}\left(\varphi\,(h^*_1, X_i, x) - \varphi\left(h^*_j, X_i, x\right)\right) \\ = \; & \sum_{i=1}^{n}\left(\int_{t_i(H_*+j\Delta)}^{t_i(H_*+\Delta)} K(u)du\right) \\ \sim \; & j\Delta n. \end{aligned}$$

Evidently, for any x

$$0 < \Phi\left(h^*_j, n, x\right) - \Phi\,(h^*_1, n, x)\,,$$

$$\Phi\left(h^*_j, n, x\right) > \Phi\left(h^*_{j-1}, n, x\right) > ... > \Phi\,(h^*_1, n, x)$$

hold. Since for any constant h $(1/n)\Phi\left(h,n,x\right)=\xi$ for some $0<\xi<1$ then $(1/n)\Phi\left(h_j^*,n,x\right)\sim j\Delta$ for any fixed h_j^*. Hence, it follows

$$\sup_{x\in\Omega^*}|F_n(x)-F^A_{h_j^*,h_1}(x)|\sim j\Delta.$$

It implies, that the sequence $\{F_n(x)-F^A_{h_i^*,h_1}(x), i=1,2,...\}$, corresponding to $h_1^*,h_2^*,...,h_j^*,...$ does not go to 0 as h_i^* increases for any x. Hence,

$$\sup_{-\infty<x<\infty}|F_n(x)-F^A_{h_*,h_1}(x)|\not\to 0 \qquad \text{as} \qquad n\to\infty.$$

Therefore, $h_*\to 0$ as $n\to\infty$. ■

Appendix II
Proof of Theorem 2.

Proof: We denote

$$\begin{aligned}I(x,h)=&\int_{-\infty}^{\infty}[(F(x-hy)-F(x))\\&-(F_n(x-hy)-F_n(x))]K(y)dy.\end{aligned}$$

Using the fact that the kernel $K(x)$ has $m+1$th order and applying Taylor's expansion to $F(x-hy)$ up to the term of order h^{m+1}, we get for any x that

$$\begin{aligned}&\int_{-\infty}^{\infty}|F(x-hy)-F(x)|K(y)dy\\&=h^{m+1}\int_{-\infty}^{\infty}\frac{y^{m+1}}{(m+1)!}|F^{(m+1)}\left(\theta hy\right)|K(y)dy\\&\geq h^{m+1}G,\end{aligned}$$

where $G=\eta_1/(m+1)!\int_{-\infty}^{\infty}y^{m+1}K(y)dy$ is a positive constant.
Suppose, that for $\alpha>2$

$$\sup_x|F(x)-F_n(x)|\leq n^{-1/\alpha}. \tag{13}$$

Then, it follows

$$\begin{aligned}|I(x,h)| &\leq \int_{-\infty}^{\infty}|F(x-hy)-F_n(x-hy)|K(y)dy\\&+|F(x)-F_n(x)|\\&\leq 2n^{-1/\alpha}. \end{aligned}\tag{14}$$

Since h is selected from (9), we have from (13) and (14)

$$\begin{aligned}h^{m+1}G &\leq \sup_x\int_{-\infty}^{\infty}|F(x-hy)-F(x)|K(y)dy\\&\leq \sup_x|I(x,h)|\\&+\sup_x\int_{-\infty}^{\infty}|F_n(x-hy)-F_n(x)|K(y)dy\\&\leq 2n^{-1/\alpha}+2\delta n^{-1/2}\end{aligned}$$

as $n\to\infty$.
Hence, from (13) it follows $h\leq\rho n^{-1/(\alpha(m+1))}$, where $\rho=(2(1+\delta)/G)^{1/(m+1)}$, since $\alpha>2$. According to well-known inequality [16]

$$\mathbb{P}\{\sup_x|F_n(x)-F(x)|>\eta\}\leq 2\exp\left(-2n\eta^2\right)$$

holds. Then it follows

$$\begin{aligned}&\mathbb{P}\{h>\rho n^{-1/(\alpha(m+1))}\}\\<\;&\mathbb{P}\{\sup_x|F(x)-F_n(x)|>n^{-1/\alpha}\}\\\leq\;&2\exp\left(-2n^{1-2/\alpha}\right).\end{aligned}$$

■

Appendix III
Proof of Theorem 3.

Proof: Denote $\varphi(x)=(d/dx)^4\,1/f(x)$.
It was proved in [13] that for assumed $K(x)$ it holds

$$\widetilde{f}^A(x|h_1,h_*)=\hat{f}^A(x|h_*)+cZ(nh_*)^{-1/2}+o((nh_*)^{-1/2}), \tag{15}$$

where c is a constant, Z is a standard normal r.v., when $h_1\simeq n^{-1/5}$ was taken. The value $c=c(h_1)$ may be obtained from formula (4.5) of the latter paper and the application of Lindeberg's theorem to $\widetilde{f}^A(x|h_1,h_*)-\hat{f}^A(x|h_*)$ that is a sum of i.i.d. r.v.s.
Then the bias of $\widetilde{f}^A(x|h_1,h_*)$ is the same as for $\hat{f}^A(x|h_*)$, i.e.

$$\mathbb{E}\widetilde{f}^A(x|h_1,h_*)-f(x)=\frac{K_3}{24}h_*^4\varphi(x)+o(h_*^4), \tag{16}$$

holds, [13]. Suppose, that $h_*\leq\rho n^{-1/(\alpha(m+1))}$, where ρ is defined in Theorem 2. Then, it follows, that

$$\begin{aligned}&\mathbb{E}\widetilde{f}^A(x|h_1,h_*)-f(x)\\\leq\;&\frac{K_3}{24}\varphi(x)\rho^4n^{-4/(\alpha(m+1))}+o(n^{-4/(\alpha(m+1))}).\end{aligned}$$

For $\alpha=9/(m+1)$ the bias of $\widetilde{f}^A(x|h_1,h_*)$ has the order $n^{-4/9}$ for any positive integer $m<3.5$, since $\alpha>2$. Then, for $m=3$ we have

$$\begin{aligned}&\mathbb{P}\{\mathbb{E}\widetilde{f}^A(x|h_1,h_*)-f(x)>\frac{K_3}{24}\varphi(x)\rho^4n^{-4/9}\}\\<\;&\mathbb{P}\{h_*>\rho n^{-1/(\alpha(m+1))}\}\\\leq\;&2\exp\left(-2n^{1-2(m+1)/9}\right)=2\exp\left(-2n^{1/9}\right).\end{aligned}$$

■

APPENDIX IV
PROOF OF COROLLARY 1.

Proof: Denote $K_2^* = \int K^2(t)dt$. From (15) and since $\mathbb{E}(Z \cdot \hat{f}^A(x|h_*)) = 0$ it holds then the variance of $\widetilde{f}^A(x|h_1, h_*)$ is

$$\begin{aligned} & var\left(\widetilde{f}^A(x|h_1, h_*)\right) \qquad (17) \\ = \; & var\left(\hat{f}^A(x|h_*)\right) + c^2(nh_*)^{-1} + o((nh_*)^{-1}) \\ & = (nh_*)^{-1}\left(c^2 + f(x)^{3/2}K_2^*\right) + o((nh_*)^{-1}). \end{aligned}$$

From Theorem 2 it follows that $h_* = O\left(n^{-1/9}\right)$ if $\alpha = 9/(m+1)$ and $m = 3$. Hence, from (16), (17) we have that

$$\begin{aligned} & MSE(\widetilde{f}^A(x|h_1, h_*)) = (K_3/24)^2\, h_*^8(\varphi(x))^2 \\ & +(nh_*)^{-1}\left(c^2 + f(x)^{3/2}K_2^*\right) + o(h_*^8) \sim n^{-8/9}, \end{aligned}$$

as $n \to \infty$, if a maximal solution h_* of (9) has the order $n^{-1/9}$. ■

REFERENCES

[1] I. S. Abramson, *On bandwidth estimation in kernel estimators - A square root law*, Annals of Statistics, 10, 1982, 1217–1223.

[2] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986.

[3] B. U. Park and J. S. Marron, *Comparison of data-driven bandwidth selectors*, J. Amer. Statist. Assoc., 85, 1990, 66–72.

[4] P. Hall, *A simple general approach to inference about the tail of a distribution*, Annals of Statistics, 20(2), 1992, 762–778.

[5] R. E. Maiboroda and N. M.Markovich, *Estimation of heavy-tailed probability density function with application to Web data*, Computational Statistics, 19, 2004, 569–592.

[6] L. Devroye and L. Györfi, *Nonparametric density estimation: The L_1 view*, Wiley, New York, 1985.

[7] M. P. Wand, J. S. Marron and D. Ruppert, *Transformations in density estimation*, J. Amer. Statist. Assoc. Theory and Methods, 86(414), 1991, 343–353.

[8] L. Yang and J. S. Marron, *Iterated transformation-kernel density estimation*, Journal of the American Statistical Association, 94(446), 1999, 580–589.

[9] N. M. Markovitch and U. R. Krieger, *Nonparametric estimation of long-tailed density functions and its application to the analysis of World Wide Web traffic*, Performance Evaluation, 42, 2000, 205–222.

[10] J. Pickands, *Statistical inference using extreme order statistics*, The Annals of Statistics, 3, 1975, 119–131.

[11] N. M. Markovich, *Experimental analysis of nonparametric probability density estimates and of methods for smoothing them*, Automation and Remote Control, 50, 1989, 941–948.

[12] V. N. Vapnik, N. M. Markovich and A. R. Stephanyuk, *Rate of convergence in L_2 of the projection estimator of the distribution density*, Automation and Remote Control, 53, 1992, 677–686.

[13] P. Hall, J. S. Marron *Variable window width kernel estimates of probability densities*, Probab. Theory Related Fields, 80, 1988, 37–49.

[14] N. M. Markovich, *High quantile estimation for heavy-tailed distributions*, Performance Evaluation, 62, 2005, 178–192.

[15] P. Embrechts, C. Klüppelberg and T. Mikosch, *Modelling Extremal Events for Finance and Insurance*, Springer, Berlin, 1997.

[16] B. L. S. Prakasa Rao, *Nonparametric Functional Estimation*, Academic, Orlando, Fla, 1983.

[17] D. W. Scott, *Multivariate Density Estimation Theory, Practice and Visualization*, Wiley, N.Y., 1992.