# Nonparametric estimation of a heavy-tailed density

Markovich, Natalia
*Institute of Control Sciences, Russian Academy of Sciences*
*Profsoyuznaya str. 65*
*Moscow 117997, Russia*
*E-mail: markovic@ipu.rssi.ru*

## 1 Introduction

We discuss problems of the heavy-tailed pdf estimation. The following three approaches of the heavy-tailed pdf estimation are considered:

1. Combined parametric-nonparametric methods, where the tail domain of the pdf is fitted by some parametric model and the main part of the pdf (the body) is fitted by some nonparametric method like a histogram.

2. Re-transformed estimates, that use a preliminary transformation of an underlying random variable (rv) to a new one with a pdf that is more convenient for the restoration.

3. Kernel estimates with a variable bandwidth.

The estimators of the pdf contain smoothing parameters. The selection of the latter using samples of moderate sizes strongly impacts on the accuracy of the estimation. We consider some data-driven methods that are used for smoothing or regularization of the pdf estimates. Apart of well-known ones like the cross-validation methods we present an alternative discrepancy method proposed by the author. The discrepancy method is based on nonparametric statistics like the Kolmogorov-Smirnov or the von Mises-Smirnov statistics, and it uses quantiles of their limit distributions as the discrepancy value between an empirical df and a modeled df.
The convergence rates of the considered estimates are compared.

## 2 Specific features of the analysis of heavy-tailed distributions

Let $X_1, \dots, X_n$ be a sample of $n$ independent identically distributed (iid) rvs distributed with the heavy-tailed df $F(x)$.

**Definition 1** *A df $F(x)$ (or the rv $X$) is called heavy-tailed if its tail $\bar{F}(x) = 1 - F(x) > 0$, $x \geq 0$, satisfies $\forall y \geq 0$*

$$\lim_{x\to\infty} P\{X > x + y | X > x\} = \lim_{x\to\infty} \bar{F}(x+y)/\bar{F}(x) = 1.$$

Specific features of the analysis of heavy-tailed distributions are the following:

- a heavy tail goes to zero at $\infty$ slower than by an exponential rate;
- Cramér's condition, which states the existence of the moment generating function, is violated;
- sparse observations in the tail domain of the distribution.

We aim to obtain pdf estimators that fit the whole heavy-tailed pdf (the 'tail' and the 'body') well enough (that is the combined parametric-nonparametric method, Markovitch and Krieger (2002)), the nonparametric pdf estimators with an accurate tail behavior (that are the re-transformed nonparametric estimators, Maiboroda and Markovich (2004)). Such estimators are required particularly to

compare pdfs of different populations needed in classification. Since the object may have a property value belonging to the 'tail' or to the 'body' of the distribution, an accurate pdf estimate is of great importance. For example, in telecommunication systems one needs to classify measurements belonging to different sources such as mobile, fax, normal calls, Internet etc.

Due to the lack of observations beyond the sample maximum the only available information is given by an asymptotic limit distribution of the sample maximum $M_n = \max(X_1, X_2, \ldots, X_n)$. We use it as a parametric tail model of the distribution in the pdf estimators.

If $F(x)$ is such that the limit distribution of $M_n$ exists, then this limit distribution can only be of the following form for some normalizing constants $a_n, b_n$, Gnedenko (1943)

$$P\{(M_n - b_n)/a_n \le x\} \quad = \quad F^n(b_n + a_n x) \to_{n\to\infty} H_\gamma(x), x \in R,$$

and an extreme value df $H_\gamma(x)$ is of the following type:

$$H_\gamma(x) \quad = \quad \begin{cases} \exp(-x^{-1/\gamma}), & x > 0, \gamma > 0 \qquad \text{'Fréchet'}, \\ \exp(-(-x)^{-1/\gamma}), & x < 0, \gamma < 0 \qquad \text{'Weibull'}, \\ \exp(-e^{-x}), & \gamma = 0, x \in R \qquad \text{'Gumbel'}. \end{cases}$$

The parameter $\gamma$ is called the extreme value index (EVI) and defines the shape of the tail of the rv $X$.

Since kernel estimators with variable kernels are proposed in the literature as a good alternative to estimate the heavy-tailed pdfs, we compare the latter with other estimates.

## 3 Combined parametric-nonparametric methods

Let $X^n = \{X_1, \ldots, X_n\}$ be a sequence of positive iid rvs distributed with the heavy-tailed df $F(x)$ and the pdf $f(x) = F'(x)$. A combined parametric-nonparametric estimate employs a separate estimation of the 'tail' and the 'body' of the pdf, namely

$$\widetilde{f}(t, \gamma, N) = \begin{cases} f^N(t), & t \in [0, X_{(n-k)}], \\ f_\gamma(t), & t \in (X_{(n-k)}, \infty), \end{cases} \tag{1}$$

Here $X_{(n-k)}$ is some order statistic corresponding to the sample $X^n$,

$$f_\gamma(x) = (1/\gamma)x^{-1/\gamma-1} + (2/\gamma)x^{-2/\gamma-1},$$

is the parametric tail model of Pareto type and

$$f^N(t) = \frac{1}{X_{(n-k)}} \sum_{j=1}^{N} \lambda_j \varphi_j(\frac{t}{X_{(n-k)}}), \tag{2}$$

is the nonparametric estimate of the main part of the pdf. It is an expansion by basis functions $\varphi_k(t)$, $k = 1, 2, \ldots$, e.g., $\varphi_k(t) = \sqrt{4/\pi}\cos((2k-1)(\pi/2)t)$, $t \in [0, 1], k = 1, 2, \ldots$.

The EVI $\gamma$ is the most important parameter to describe the heaviness of the tail. It can be estimated by Hill's estimator (or by many other estimators) by the $k+1$ largest values of the order statistics $X_{(1)} < \ldots < X_{(n)}$ of the sample $X^n$. The parameter $k$ indicates $X_{(n-k)}$ and, therefore, the part of the distribution which controls the extreme values of the underlying rv. It can be estimated for instance by bootstrap methods, Markovich (2007). One has first to estimate $k$ to fit the 'tail' and then to adapt the 'body' of the pdf.

To provide all properties of the real pdf, the estimate (1) should be normalized, i.e. one can take

$$f^*(t, \gamma, N) = \begin{cases} \dfrac{\tilde{f}(t, \gamma, N)}{\int_0^\infty \tilde{f}(t, \gamma, N)\mathbf{1}(\tilde{f}(t, \gamma, N) > 0)dt}, & t \in A, \\ 0, & t \notin A \end{cases} \tag{3}$$

for $A = \left\{ t \in [0,\infty) : \tilde{f}(t,\gamma,N) > 0 \right\}$ instead of $\tilde{f}(t,\gamma,N)$ to get $\int_0^\infty f^*(t,\gamma,N)dt = 1$ and $f^*(t,\gamma,N) \geq 0$.

To avoid a gap between the two parts $f^N(t)$ and $f_\gamma(t)$ at the point $X_{(n-k)}$ one can undertake some smoothing in the neighborhood of $X_{(n-k)}$, Markovich (2007).

One is free to select many combinations of nonparametric and parametric estimators. The accuracy is then determined by the uncertainties of the estimates. The accuracy of $f_\gamma(t)$ depends on the selection of the appropriate parametric tail model and an accurate estimation of $\gamma$. Particularly, a histogram or a kernel estimate can be selected as $f^N(t)$.

The application of the structural minimization method (Vapnik (1982)) to estimate $N$ and the coefficients $\lambda = (\lambda_1, ..., \lambda_N)^T$ in (2) allows us to estimate a multimodal pdf defined at the compact interval better than the kernel estimates, Vapnik and Stephanyuk (1979). Such estimate (2) is used now to determine multimodal heavy-tailed pdfs, Markovich and Krieger (2002). The structural minimization method of the pdf estimation modifies the least squares method for correlated data and provides the minimum of a specific upper bound of the mean risk

$$J(N,\lambda)\left[\frac{l^{-1}\cdot(Y-\widehat{F}(\lambda))^t\cdot R_y^{-1}\cdot(Y-\widehat{F}(\lambda))}{1-\sqrt{l^{-1}\cdot[(N+1)(1+\ln l-\ln(N+1))-\ln\eta]}}\right]_\infty \tag{4}$$

with respect to parameters $(N,\lambda)$, where $\eta > 0$ is a confidence level, $[z]_\infty = \begin{cases} z, & z > 0 \\ \infty, & z \leq 0, \end{cases}$ and

$$Y = (Y_1,\ldots,Y_l)^T, \quad Y_i = y_i - \int_0^{\tau_i} \frac{\varphi_1(t)}{\psi_1}dt,$$

$R_y$ is a covariance matrix of the vector $y = (y_1,\ldots,y_l)^T$, (Markovich (2007), Sec. 3.2.1).

As $y_i$ one can take the estimate $\Phi_{n^*}(t)$ of the unknown df $F(t)$ at $l$ uniformly distributed points $\tau_i = i/(l+1),\ i = 1,\ldots,l$ determined in (6). We use:

$$\begin{aligned} F(\lambda) &= \left(F_1^\lambda,\ldots,F_l^\lambda\right)^T, \quad F_i^\lambda = \int_0^{\tau_i}\left(\sum_{j=2}^N \lambda_j(\varphi_j(t) - \frac{\psi_j}{\psi_1}\varphi_1(t))\right)dt \\ F(\lambda) &= A\cdot\lambda 1 \end{aligned}$$

Here the elements of the $l\times(N-1)$ matrix $A$ are given by

$$\begin{aligned} A_{i,j} &= \int_0^{\tau_i}\left(\varphi_j(t) - \frac{\psi_j}{\psi_1}\varphi_1(t)\right)dt, \\ &\psi_j = \int_0^1 \varphi_j(t)dt, \quad i = 1,\ldots,l; j = 2,\ldots,N. \end{aligned}$$

$\Lambda_1$ is the $(N-1)\times 1$ vector of the parameters $\lambda_j, j = 2,\ldots,N$. The matrix

$$R_y^{-1} = \begin{pmatrix} r_1 & \rho_1 & 0 & \ldots & \ldots & 0 \\ \rho_1 & r_2 & \rho_2 & 0 & \ldots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ldots & 0 & \rho_{l-2} & r_{l-1} & \rho_{l-1} \\ 0 & \ldots & \ldots & 0 & \rho_{l-1} & r_l \end{pmatrix} \tag{5}$$

with

$$r_1 = \frac{n^*F(\tau_2)}{F(\tau_1)\left(F(\tau_2)-F(\tau_1)\right)}, \qquad r_l = \frac{n^*(1-F(\tau_{l-1}))}{(1-F(\tau_l))\left(F(\tau_l)-F(\tau_{l-1})\right)}$$

$$r_{i-1} = \frac{n^* \left(F(\tau_i) - F(\tau_{i-2})\right)}{\left(F(\tau_i) - F(\tau_{i-1})\right)\left(F(\tau_{i-1}) - F(\tau_{i-2})\right)}, \qquad i = 3, 4, \ldots, l;$$
$$\rho_i = -\frac{n^*}{\left(F(\tau_{i+1}) - F(\tau_i)\right)}, \qquad i = 1, 2, \ldots, l-1,$$

is used. However, the following estimate $\Phi_{n^*}(t)$ is used instead of the unknown df $F(t)$:

$$\Phi_{n^*}(t) = \begin{cases} \frac{1}{2n^*}\left(\frac{t}{t_1}\right), & 0 < t \le t_1 \\ \frac{m-0.5}{n^*} + \frac{1}{n^*}\left(\frac{t-t_m}{t_{m+1}-t_m}\right), & t_m < t \le t_{m+1}, \quad m = 1, \ldots, n^*-1 \\ \frac{n^*-0.5}{n^*} + \frac{1}{2n^*}\left(\frac{t-t_{n^*}}{1-t_{n^*}}\right), & t_{n^*} < t \le 1 \end{cases} \tag{6}$$

**The minimization algorithm has two stages:**

1. The df $F(t)$ and $R_y^{-1}$ are estimated by the sample using (6) and (5).
2. In (4) $R_y^{-1}$ is replaced by its estimate and the parameters of the pdf estimate $g^N(t)$ are obtained by the minimization of $J(N, \lambda)$ over $N$ and $\lambda$.
   The method provides $\int_0^1 \sum_{j=1}^N \lambda_j \, \varphi_j(t) dt = 1$.

**Practical recommendations**

1. Let $\eta = 0.05$.
2. Stefanyuk (1984) recommended selecting $l = 5n/\ln n$ to provide the asymptotic minimum of the $L_2$ error as $n \to \infty$.
3. To avoid division by zero in the formula (6), the points $\{t_m, m = 1, ..., n^*\}$ cannot repeat each other.[1]
4. $\lambda_1$ is calculated by $\lambda_1 = (1 - \sum_{j=2}^N \lambda_j \psi_j)/\psi_1$.
5. One minimizes the empirical risk $l^{-1}(Y - A\Lambda_1)^T R_y^{-1}(Y - A\Lambda_1)$ over $\Lambda_1 = (\lambda_2, \ldots, \lambda_N)^T$ for each fixed $N$. The minimum gives the following estimate:

   $$\lambda_N^* = \left(A^T R_y^{-1} A\right)^{-1} A^T R_y^{-1} Y$$

   Among the vectors $\lambda_N^*, N = 2, 3, \ldots, N_{max}$ (where $N_{max}$ is the maximum value of $N$ considered, usually $N_{max} = 20$) one selects those corresponding to the minimum of $J(N, \lambda)$.
6. The empirical risk (the numerator of (4)) has to decrease with increasing $N$. If this risk increases, then the matrix of the system is nearly singular.
7. The minimum of (4) is not necessarily reached for a maximal $N$. For such $N$ the empirical risk is minimal, but the inverse denominator of (4) is maximal.
8. Finally, the 'body' estimate of the pdf is calculated by the formula (2).
9. One can use another complete system of basis functions $\varphi_k(t), k = 1, 2, \ldots,$ instead of the trigonometric functions.

[1] For continuous $F(x)$ the repetitions are impossible.

## 4 Re-transformed nonparametric estimators

We consider the transform-retransform scheme to improve the behavior of the pdf estimate at the tail. The background of the transformation idea is given by the necessity of a different amount of smoothing at different locations of a heavy-tailed pdf. Then re-transformed pdf estimates with fixed smoothing parameters work like estimates with location-adaptive variable parameters. Hence, such estimates may be stretched at the tail of heavy-tailed pdfs.

Let $X$-space data are transformed via a monotone increasing continuously differentiable "one-to-one" transformation function $T(x)$ to obtain $Y_1, \dots, Y_n$ ($Y_i = T(X_i)$). The derivative of the inverse function $T^{-1}$ is assumed to be continuous. The df of $Y_j$ is given by

$$G(x) = P\{Y_i \le x\} = P\{T(X_i) \le x\} = F(T^{-1}(x)) \tag{7}$$

The re-transformed estimate of the pdf of $X_i$ is given by

$$\hat{f}(x) = \hat{g}(T(x))T'(x),$$

where $g(x)$ is the pdf of the rv $Y_i$ and $\hat{g}(x)$ is its estimate. To fit the pdf $f(x)$, one has first to select the transformation and to estimate $g(x)$.

The selection of $T(x)$ is determined by a 'target' df $G(x)$ and by the unknown df $F(x)$ of the rv $X_1$. The 'target' df can be selected in such a way that the pdf $g(x)$ should be convenient for the estimation. *Fixed transformations* like $\ln x$, $2/\pi \arctan x$ do not require any knowledge about the distribution of $X$ and are simpler for practice. But they can lead to a non-predictable pdf of the transformed rv $Y_j$ that has discontinuity and which is difficult to estimate.

There are the following problems of the application of the transform-re-transform scheme to heavy-tailed pdfs. The df $F(x)$ is unknown and it is impossible to transform it to a desirable pdf $g(x)$. One has to select a parametric or non-parametric family of distributions as a guess df $F(x)$ as well as a target pdf $g(x)$ to provide the stability of the re-transformed estimates to minor perturbations in the EVI estimates. Moreover, the selected pdf estimate has to keep the tail decay rate of the true pdf after the inverse transformation.

The *adaptive transformation*, Maiboroda and Markovich (2004),

$$T_{\hat{\gamma}}(x) = \Phi^{-1}(\Psi_{\hat{\gamma}}(x)) = 1 - (1 + \hat{\gamma}x)^{-1/(2\hat{\gamma})}$$

is obtained from (7) assuming that the fitted df $F(x)$ of $X_i$ is the Generalized Pareto distribution

$$\Psi_{\hat{\gamma}}(x) = \begin{cases} 1 - (1 + \hat{\gamma}x)^{-1/\hat{\gamma}}, & x \ge 0, \\ 0, & x < 0, \end{cases}$$

and the target df $G(x)$ of $Y_i$ is the positive triangular one

$$\Phi(x) = (2x - x^2)\mathbf{1}\{x \in [0,1]\} + \mathbf{1}\{x > 1\}.$$

The choice of a Generalized Pareto distribution is widespread and motivated by Pickands theorem. It states that for a certain class of distributions and for a sufficiently high threshold $u$ of the rv $X$ the conditional distribution of the overshoot $Y = X - u$, provided that $X$ exceeds $u$, converges to a Generalized Pareto distribution, Markovich (2007).

The transformation to a uniform distribution at $[0,1]$ with the df $\Phi^{uni}(x) = x\mathbf{1}\{x \in [0,1]\} + \mathbf{1}\{x > 1\}$, as a target df has been checked also. However, it leads to $g(x) \to \infty$ as $x \to 1$ when the $\hat{\gamma}$ underestimates the true EVI $\gamma$ ($\hat{\gamma} < \gamma$). This may cause problems for the estimation of $g(x)$.

The transformation $T_{\hat{\gamma}}(x)$ provides a pdf $g(x)$ at $[0,1]$ which is continuous in the neighborhood of $x = 1$ for typical distributions (with regularly varying tails, lognormal-type tails and Weibull-like tails) and for a consistent estimate $\hat{\gamma}$ of EVI $\gamma$.

The polygram (i.e., a histogram with variable bin width) and kernel estimate (8) are taken as estimators $\hat{g}(x)$. Let us explain why the polygram is preferable compared the histogram. The polygram dynamically adapts the bin width to the data. Therefore, it works better than a histogram. This is especially important due to sparse observations at the tail domain of the heavy-tailed distribution and hence, in the neighborhood of $x = 1$ after the data transform. If the lengths of the bins are equal, very few observations fall into the right-most bin. Hence, the histogram estimate is not stable at the tail.

However, without the assumption on the class of the tail, an accurate restoration of the tail by means of a nonparametric method is impossible.

## 5 Kernel estimators with variable bandwidth

If the distribution is heavy-tailed, the well-known kernel estimators

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{8}$$

with a bandwidth $h > 0$ which is fixed across the entire sample may provide misleading results in the tail domain or over-smooth the 'body' of the pdf. To overcome this problem one can use kernel estimators with kernels that vary from one point to another. A variable bandwidth kernel estimate is defined as follows, Abramson (1982),

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

Its practical version is determined by

$$\widetilde{f}^A(x|h_1, h) = (nh)^{-1}\sum_{i=1}^{n} \hat{f}_{h_1}(X_i)^{1/2} K\left((x - X_i)\hat{f}_{h_1}(X_i)^{1/2}/h\right), \tag{9}$$

where $\hat{f}_{h_1}(x)$ is a pilot kernel estimate (8). Main advantages of $\hat{f}^A(x|h)$ are its non-negativity and the best possible mean squared error (MSE). The MSE of a kernel estimate is determined as

$$MSE = E\int \left(\hat{f}_h(x) - f(x)\right)^2 dx$$

The standard kernel estimate gives $MSE(\hat{f}_h) \sim n^{-4/5}$ (with bias $\sim h^2$ and variance $\sim (nh)^{-1}$) if a second-order kernel is used, the bandwidth $h$ is taken proportional to $n^{-1/5}$ and the pdf $f(x)$ has two continuous derivatives.

The variable bandwidth kernel estimate provides $MSE(\hat{f}^A(x|h)) \sim n^{-8/9}$ (with bias $\sim h^4$ and variance $\sim (nh)^{-1}$) if a symmetric kernel such as $\int x^4|K(x)|dx < \infty$ is used, the bandwidth $h \sim n^{-1/9}$ and $f(x)$ has four continuous derivatives, Hall and Marron (1988). This implies that the improvement of the MSE arises due to the reduction of the bias whereas the variance cannot be reduced.

However, it does not imply that the estimation of $\hat{f}^A(x|h)$ at the tail domain will be good enough. Relatively large values of the pdf at the body generate the main contribution in the MSE in contrast to small values at the tail. Hence, the MSE and measures of the metric spaces $C$, $L_1$ and $L_2$ are not sensitive to the accuracy of the estimation at the tail.

The disadvantage of $\hat{f}^A(x|h)$ is that it cannot provide an accurate estimation of the PDF at infinity, at least by compactly supported kernels, because is is defined on a finite interval which is approximately the same as the range of the sample. This range can be extended by a long-tailed kernel. Then the accuracy of the estimate at the infinity will depend on the form of the kernel.

The variable bandwidth kernel estimates can be improved by means of the transform-retransform scheme. This allows us to estimate the tail of the pdf better, see an example in Fig. 1.

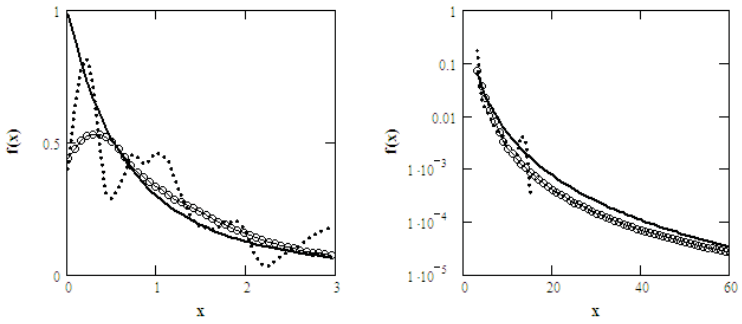


Figure 1: The re-transformed kernel estimate (8) (line marked by circles) and the variable bandwidth kernel estimate (9) without transformation (dotted line) with Epanechnikov's kernel $K(x) = 3/4(1 - x^2)\mathbf{1}\{|x| \leq 1\}$ for Pareto distribution (solid line): the pdf body (left) and the pdf tail (right). For both estimates $h$ is selected by $D$-method (10) ($h = 0.21$ for the first estimate and $h = 0.11$ for the second one), $h1 = 1.915$ is calculated by over-smoothing bandwidth selector (Wand and Jones (1995)) $\hat{h}_{OS} = \left(\frac{243R(K)}{35\mu_2(K)^2 n}\right)^{1/5} \cdot \sigma$, $\mu_2(K) = \int z^2 K(z)dz$, $R(K) = \int K^2(x)dx$, $\sigma$ is a standard deviation. The estimate (9) is interrupted before $x = 20$.

## 6 Smoothing methods

To improve the accuracy of re-transformed estimates, the selection of the smoothing parameter (e.g., the bandwidth of kernel estimates or the bin width of a polygram) constitutes the most important problem. For moderate sample sizes, a data-dependent choice of the smoothing parameter of the pdf estimate is more productive than one derived from the theory such as $h = cn^{-1/5}$, where $c$ is some positive constant.

A most popular data-dependent method is given by cross-validation. This method produces consistent nonvariable kernel estimates (8) in the $L_1$ metric only for distributions with a bounded support, Chow et al. (1983). For heavy-tailed pdfs it gives a $h$ which does not converge to 0 as $n \to \infty$. It implies inconsistency, Devroye and Györfi (1995). The cross-validation method that is based on the maximum likelihood idea has several disadvantages. These are a slow convergence rates, a high sampling variability (see Park and Marron (1990)) and a possibility to get a local extremum instead of a global one.

Further, we focus at the discrepancy method that is an alternative to the cross-validation method. It was proposed and investigated in Markovich (1989) and Vapnik et al. (1992). The smoothing parameter $h$ (bandwidth, bin width) is selected as a solution of the discrepancy equation

$$\rho(\widehat{F}_h, F_n) = \delta,$$

where, $\widehat{F}_h(x) = \int_{-\infty}^{x} \widehat{f}_h(t)dt$, $\hat{f}_h(t)$ is a nonparametric estimate of the pdf, $\delta$ is a uncertainty of the estimation of the df $F(x)$ by the empirical DF $F_n(t)$, i.e. $\delta = \rho(F, F_n)$, and $\rho(\cdot, \cdot)$ is a metric in the space of dfs.

Since $\delta$ is usually unknown, the quantiles of the limit distribution of the von Mises-Smirnov statistic

$$\omega_n^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x)$$

or Kolmogorov-Smirnov statistic

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

are used as $\delta$.[2]
As practical versions one can use the following equations regarding $h$:

$$\hat{\omega}_n^2(h) = 0.05 \qquad \text{as} \qquad \omega^2\text{-method},$$

$$\sqrt{n}\hat{D}_n(h) = 0.5 \qquad \text{as} \qquad D\text{-method}, \tag{10}$$

where

$$\hat{\omega}_n^2(h) = \sum_{i=1}^{n} \left( \widehat{F}_h(X_{(i)}) - \frac{i-0.5}{n} \right)^2 + \frac{1}{12n},$$

$$\sqrt{n}\hat{D}_n(h) = \sqrt{n}\max(\hat{D}_n^+, \hat{D}_n^-),$$

and

$$\hat{D}_n^+ = \max_{1\le i\le n} \left( \frac{i}{n} - \widehat{F}_h(X_{(i)}) \right), \qquad \hat{D}_n^- = \max_{1\le i\le n} \left( \widehat{F}_h(X_{(i)}) - \frac{i-1}{n} \right),$$

The values 0.05 and 0.7 are the quantiles corresponding to a maximum of the pdf of the $\omega_n^2$ and $D_n$ statistics, $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ are order statistics of the sample $X^n$. The corrected value $\delta = 0.5$ for $D_n$ is found for moderate samples, Markovich (1989).
For some heavy-tailed distributions the discrepancy equations may have no solutions. It implies that may be higher quantiles of distributions of statistics $\omega_n^2$ and $\sqrt{n}D_n$ are required.
Hence, it is better to transform first the data to a compact interval and then to estimate the smoothing parameter of the pdf $g(x)$ if the latter is sufficiently smooth.

**$D-$method for a variable bandwidth kernel estimator**

Consider the estimate (9). Let $h_*$ be a solution of the equation

$$\sup_{-\infty<x<\infty} |F_n(x) - F_{h,h_1}^A(x)| = \delta n^{-1/2}, \tag{11}$$

where $F_{h,h_1}^A(x) = \int_{-\infty}^{x} \widetilde{f}^A(t \mid h_1, h)dt$, $\delta > 0$ is some constant. The following theorems contain properties of the $D$-method (10), Markovich (2007).

**Theorem 1** *Let $X^n = \{X_1, \ldots, X_n\}$ be iid rvs with pdf $f(x)$. Select the bandwidth $h_1 = cn^{-1/5}$, $c > 0$ in $\hat{f}_{h_1}(x)$. We assume that for $x \in R$, $K(x)$ is continuous and satisfies*

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

*Then any solution $h_* = h_*(n)$ of (11) obeys the condition*

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

**Theorem 2** *Suppose that the pdf $f(x)$ has $m-1$ continuous derivatives and its $m$th derivative is bounded for a positive integer $m$. Let $f(x)$ be estimated by a variable bandwidth kernel estimate $\widetilde{f}^A(x|h_1, h)$ (9). Assume that the conditions on $K(x)$ given in Theorem 1 hold. In addition, we assume that $K(x)$ has the order $m+1$ and $\int_R |K(x)|dx = A < \infty$ holds. Let the non-random bandwidth $h_1$ in $\hat{f}_{h_1}(x)$ obey the conditions: $h_1 \to 0$, $nh_1 \to \infty$ as $n \to \infty$. Then any solution $h_* = h_*(n)$ of equation (11) obeys the condition*

$$\mathbf{P}\{h_* > \rho n^{-1/(\alpha(m+1))}\} < \exp\left(-2n^{1-2/\alpha}\right), \tag{12}$$

*where $\rho = (2(1+A\delta)/G)^{1/(m+1)}$ is a constant, $G = 1/(m+1)! \sup_x |\int_{-\infty}^{\infty} f^{(m)}(x - hy\theta)y^{m+1}K(y)dy|$, $0 < \theta < 1$, for any $\alpha > 2$.*

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

A kernel $K(x)$ has an order $r$ when a kernel function is chosen such that

$$\int u^k K(u)du = \begin{cases} 1, & k=0 \\ 0, & 1 \le k \le r-1, \\ K_{r-1} \neq 0, & k=r, \end{cases}$$

holds.

**Theorem 3** *Let $f(x)$ and $1/f(x)$ have four continuous derivatives and $f(x)$ be bounded away from zero on $\Re^{\varepsilon} \equiv \{x \in R : \text{for some} \quad y \in \Re, \|x-y\| \le \varepsilon\}$, $\varepsilon > 0$ ($\|\cdot\|$ is the usual Euclidean norm). Let the pdf $f(x)$ be estimated by a variable bandwidth kernel estimate $\widetilde{f}^A(x|h_1,h)$ (9). Assume the conditions on $K(x)$ given in Theorem 2 hold for $m=3$. We assume that $K(x)$ is symmetric, has two bounded derivatives and vanishes outside a compact set. Assume, that the non-random bandwidth $h_1$ in (9) obeys $h_1 = c_* n^{-1/5}$, where $c_* > 0$ is some constant. Then for any solution $h_*$ of (11) we have*

$$\mathbf{P}\{\overline{lim}_{n\to\infty} n^{4/9} \left(\mathbf{E}\widetilde{f}^A(x|h_1,h_*) - f(x)\right) \le \psi(x)\} = 1,$$

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

**Corollary 1** *Assume that the conditions of Theorem 3 hold. Let us assume, that $\mathbf{E}(Z \cdot \hat{f}^A(x|h)) = 0$, where $Z$ is a standard normal rv. Then, $MSE(\widetilde{f}^A(x|h_1,h_*))$ may reach the order $n^{-8/9}$ if a maximal solution of (11) $h_*$ has the order $n^{-1/9}$.*

**Remark 1** *Since the function of the rv $X_1$ (that is one term in the sum $\hat{f}^A(x|h)$) and the normal distributed rv $Z$ are independent, the condition $\mathbf{E}(Z \cdot \hat{f}^A(x|h)) = 0$ is not rigorous.*

## REFERENCES


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

Chow Y.-S., Geman S. & Wu L.-D. (1983) Consistent cross-validated density estimation. *Annals of Statistics, 11, 25-38.*

Devroye L. & Györfi L. (1985) Nonparametric density estimation. The $L_1$ view., *Wiley, New York.*

Gnedenko B.V. (1943) Sur la Distribution Limite du Terme Maximum d'une Série AléatoirE *Annals of Mathematics, 44, 423-453.*

Hall P. & Marron J.S. (1988) Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields, 80, 1, 37-49.*

Markovich N.M. (1989) Experimental analysis of nonparametric probability density estimates and of methods for smoothing them. *Automation and Remote Control, 50, 941-948.*

Markovitch, N.M., & Krieger, U.R. (2002) The estimation of heavy-tailed probability density functions, their mixtures and quantiles. *Computer Networks, Vol. 40, Issue 3, 459-474.*

Markovich N.M. (2007) Nonparametric Estimation of Univariate Heavy-Tailed Data. *J. Wiley & Sons, Chichester.*

Maiboroda R.E., & Markovich N.M. (2004) Estimation of heavy-tailed probability density function with application to Web data. *Computational Statistics, 19:569-592.*

Park B.U., & Marron J.S. (1990) Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association, 85, 66-72.*

Stefanyuk A.R. (1984) Estimation of the probability density function. In V.N. Vapnik (ed.), Algorithms and Programs for Dependency Reconstruction, 688-706, *Nauka, Moscow (in Russian).*

Vapnik V.N. & Stephanyuk A.R. (1979) Nonparametric methods for probability density reconstruction. *Automation and Remote Control 39, 1127-1140.*

Vapnik V.N. (1982) Estimation of dependences based on empirical data. *Springer, New York Heidelberg Berlin XVI.*

Vapnik V.N., Markovich N.M. & Stephanyuk A.R. (1992) Rate of convergence in $L_2$ of the projection estimator of the distribution density. *Automation and Remote Control, 53, 677-686.*

Wand M.P. & Jones M.C. (1995) Kernel smoothing. *Chapman & Hall, New York.*

## ABSTRACT

*To estimate a heavy-tailed probability density function (pdf), different approaches are summarized: (1) a combined parametric - nonparametric method, (2) methods based on data transformations and, (3) a variable bandwidth kernel estimator. The first method implies a separate estimation of the 'tail' and 'body' of the pdf by parametric and nonparametric methods, respectively. We consider a Pareto-type model to fit the 'tail' and a finite series expansion in terms of trigonometric functions as 'body' estimate. To fit the body of a multi-modal pdf better, we use a structural risk minimization method for the selection of the parameters. The second approach requires a special data transformation which improves the estimation in the 'tails', namely, the transformation from a Generalized Pareto distribution function (df) which is assumed as a fitted df to a triangular df selected as the target df. The latter transformation is robust regarding the uncertainty of the tail index estimation. The triangular pdf can be estimated by a nonparametric estimator, e.g., a Parzen kernel estimator or a polygram. Regarding the heavy-tailed pdf estimation a kernel estimator with a variable bandwidth is usually recommended due to the variability of its bandwidth for each observation. It is demonstrated that this estimator works better if a preliminary data transformation is used. To select data-driven smoothing parameters for the mentioned estimators, a discrepancy method is considered as an alternative to the cross-validation method. The discrepancy method is based on nonparametric statistics like the Kolmogorov-Smirnov or the von Mises-Smirnov statistics, and it uses quantiles of their limit distributions as a unknown discrepancy between the fitted and empirical dfs. Moreover, the convergence rates of these estimates are discussed.*