RETRANSFORMED HEAVY-TAILED DENSITY ESTIMATES AND CLASSIFICATION PROBLEM

Natalia M. Markovitch * **Udo R. Krieger** **

* *Institute of Control Sciences, Russian Academy of Sciences, Profsoyuznay 65, 117997 Moscow, Russia*
** *T-Nova Deutsche Telekom, Technologiezentrum, Am Kavalleriesand 3, D-64295 Darmstadt, Germany*

Abstract: This report is devoted to the non-parametric probability density function (PDF) estimation for distributions with heavy tails. It is proposed to transform the data to a compact interval. The PDF of the original data is obtained by the "back-transform" of the PDF estimate of the transformed data. The transformation adapted to the data is explored to keep the tail decay rate of the true PDF and to provide a stable estimation regarding the deviations of the tail index estimates. The relative risk of the misclassification is proposed as a measure of quality for the PDF estimates.

Keywords: Non-parametric density estimation, Transformations, Classifiers, Risk of classification

1. INTRODUCTION

In the last few years heavy-tailed (Sigman, 1999) distributed random variables (r.v.) have been accepted as realistic models for various phenomena: the length of transmitted files, on and off periods of transmission, the claim sizes of catastrophes in insurance. The analysis of measurements of WWW-traffic by statistical methods also shows that WWW characteristics are often heavy-tailed distributed.

Here, the non-parametric estimation of heavy-tailed PDFs is considered. For the finite and light-tailed distributions (i.e. without heavy tails) the histogram is a good estimate of the PDF. But if the distribution is heavy-tailed, the histogram provides an absolutely misleading estimate in the "tail" domain. The same is true for most of the common non-parametric PDF estimates such as kernel, projection and spline estimates. In general, they have sharp peaks at "outliers" and don't provide the right rate of the decay at infinity.

The necessity of non-parametric PDF estimates with a good behavior at the tail domain is obvious. It is especially significant if PDFs of many populations are compared. Classification (pattern recognition) is another problem which is nearly related to the comparison of densities. If one uses an empirical Bayesian classification algorithm then the observation may be classified by the comparison of the PDF estimates of each class. Since the object can arise in the tail domain as well as in the body a good behavior of the estimate tails is principal for the classification.

To get more exact estimates at the tails it was proposed in (Devroye and Györfi, 1985) to use a preliminary transformation of the data which compresses them to a compact interval. Then one can estimate the PDF of the compressed sample and stretch this estimate by the inverse transformation. Based on this idea 1) a transformation, which is adapted to empirical data, 2) PDF estimates to keep the rate of the tail decay at infinity, and 3) the ratio of the empirical mean risk of the misclassification by an empirical Bayesian classifier to the risk of the misclassification by a Bayesian classifier as a quality measure of the PDF estimates are proposed.

2. ESTIMATE CONSTRUCTION

2.1 *Preliminary transformation*

Let $X_1,\ldots, X_n$ be i.i.d. r.v.s with the CDF F and the PDF f. Let us consider some monotone increasing continuous function $T : R_+ \to [0,1]$ and a sample $Y_j = T(X_j), j = 1,\ldots,n$. The CDF of Y_j is

$$G(x) = \Pr\{Y_j < x\} = \Pr\{T(X_j) < x\} \\ = \Pr\{X_j < T^{-1}(x)\} = F(T^{-1}(x)) \quad (1)$$

and the PDF is

$$g(x) = G'(x) = f(T^{-1}(x))(T^{-1}(x))', \quad (2)$$

analogously, since $X_j = T^{-1}(Y_j)$,

$$f(x) = g(T(x))T'(x).$$

One can estimate $g(x)$ by $Y_1,\ldots,Y_n$ using one of the non-parametric methods. Denote this estimate by $\breve{g}_n(x)$. Then one can use

$$\breve{f}_n(x) = \breve{g}_n(T(x))T'(x) \quad (3)$$

as the estimate of f. In (Devroye and Györfi, 1985) the general theory of such estimates is considered. For kernel estimates with compact kernels the transformation to an isosceles triangular PDF $\phi^{\text{tr}}(x)$ on $[0,1]$ and for a histogram the transformation to a uniform PDF $\phi^{\text{un}}(x)$ are recommended. These $T(x)$ provide an optimal convergence rate in L_1. Since such $T(x)$ and, therefore, the distribution of $Y_j = T(X_j)$ depend on the unknown CDF F, it is impossible to obtain an absolute identity of g and ϕ^{tr} (or $\phi^{\text{un}}(x)$). Hence, Devroye et al. (1985) used instead of F some family of CDFs Ψ_γ which depends on the parameter γ and adapt γ to the sample. In (Wand, et al., 1991) this idea was extended, i.e. transformations not only to compactly supported PDFs are considered. A transformation from suitable parametric families is selected which provides the minimal mean-square error of the PDF estimation of a transformed r.v. by a kernel estimate. Here, as Ψ_γ some family of heavy-tailed distributions is proposed, where the tail index γ is estimated by Hill's estimate.

2.2 *Estimation algorithm*

The algorithm proceeds as follows:

- The tail index of X_j is estimated by the sample $\{X_1,\ldots,X_n\}$ using the Hill estimate $\hat{\gamma}_n$.
- The transformation $T = T_{\hat{\gamma}_n}$ is constructed such that if ξ has the guess CDF $\Psi_{\hat{\gamma}_n}$ then $T_{\hat{\gamma}_n}(\xi)$ has the target CDF Φ, i.e. a uniform or triangular one. (Here $\hat{\gamma}_n$ is considered as a fixed value).
- The transformed sample $Y_j = T_{\hat{\gamma}_n}(X_j), j = 1,\ldots,n$ is constructed.
- The PDF of $Y_1,\ldots,Y_n$ is estimated by some non-parametric estimate $\breve{g}_n(x)$.
- The PDF of X_j is estimated by (3).

2.3 *Analysis of the algorithm*

For the realization of the algorithm the guess family, the target distribution and the estimate $\breve{g}_n$ must be chosen. The guess family must be such that the transformation T and its derivative can be easily evaluated. If Ψ_γ is the guess and Φ is the target CDF, then by (1) $T(x) = \Phi^{-1}(\Psi_\gamma(x))$ and $T^{-1}(x) = \Psi_\gamma^{-1}(\Phi(x))$ hold. The generalized Pareto distribution

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

is assumed as a guess CDF. Evidently, the guess PDF is

$$\psi_\gamma(x) = \Psi_\gamma'(x) = (1+\gamma x)^{-(1/\gamma+1)} \quad (4)$$

Such an assumption is based on facts from extreme value theory, namely, that for a sufficiently high threshold u of the r.v. X, the distribution of $Y = X - u$ converges to the Generalized Pareto distribution (Pickands, 1975). As target PDFs the uniform PDF $\phi^{\text{un}}(x) = \mathbf{1}\{\mathbf{x} \in [\mathbf{0},\mathbf{1}]\}$ and the positive triangular PDF $\phi^{+\text{tr}}(x) = 2(1-x)\mathbf{1}\{\mathbf{x} \in [\mathbf{0},\mathbf{1}]\}$ are considered. The corresponding CDFs are

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

Two questions are clarified here:

- What transformation ensures a more stable estimation algorithm with respect to the deviations of a tail index estimate?
- What tail behavior is ensured by the inverse transformation?

Let the target be $\phi = \phi^{\text{un}}, \Phi = \Phi^{\text{un}}$. Then

$$T_\gamma(x) = (1 - (1+\gamma x)^{-1/\gamma})\mathbf{1}\{\mathbf{x} \in [\mathbf{0},\mathbf{1}]\},$$
$$T_\gamma'(x) = (1+\gamma x)^{-(1/\gamma+1)}\mathbf{1}\{\mathbf{x} \in [\mathbf{0},\mathbf{1}]\},$$
$$T_\gamma^{-1}(x) = ((1-x)^{-\gamma} - 1)/\gamma,$$
$$(T_\gamma^{-1}(x))' = (1-x)^{-\gamma-1}, x \in [0,1).$$

Suppose that the true PDF of X is

$$f(x) = \ell(x)\psi_\alpha(x), \quad (5)$$

where $\ell(x)$ is a bounded function with

$$\lim_{x\to\infty} \ell(x) = \ell_\infty < \infty$$

(the tail index of this PDF is α). Then by (2) the PDF of $Y = T_\gamma(X)$ is

$$\begin{aligned} g(x) = & \left(1 + ((1-x)^{-\gamma} - 1)\alpha/\gamma\right)^{-(1/\alpha+1)} \\ & \cdot (1-x)^{-\gamma-1}\ell\left(((1-x)^{-\gamma} - 1)/\gamma\right) \end{aligned}$$

for $x \in [0,1)$. This PDF will be estimated by $\breve{g}_n$ for some value of γ. Hence, its behavior must be good at its support $x \in [0,1)$. If x is close to zero, then $g(x)$ is bounded. Let us consider $g(x)$ for $x \uparrow 1$. Note that in this case $\ell(T_\gamma^{-1}(x)) \to \ell_\infty < \infty$, since $T_\gamma^{-1}(x) \to \infty$. Then

$$\begin{aligned} g(x) \simeq & \ell_\infty \left(1 + ((1-x)^{-\gamma} - 1)\alpha/\gamma\right)^{-(1/\alpha+1)} \\ & \cdot (1-x)^{-\gamma-1} \text{ as } x \uparrow 1. \end{aligned}$$

If $\alpha = \gamma$ then $g(x) \simeq \ell_\infty$, i.e. $g(x)$ behaves as a uniform PDF in the neighborhood of 1 (up to an insignificant multiplier ℓ_∞). Let $\alpha \neq \gamma$. For $x \uparrow 1$ and any constant C it holds $(1-x)^{-\gamma} + C \sim (1-x)^{-\gamma}$. Hence,

$$g(x) \simeq \ell_\infty \left(\alpha/\gamma\right)^{-(1/\alpha+1)} (1-x)^{\gamma/\alpha-1}.$$

If γ overestimates the true tail index α, i.e. $\gamma/\alpha - 1 > 0$, then PDF $g(x)$ behaves nice in the neighborhood of 1 (while it is not like the uniform distribution). But if $\gamma < \alpha$ then $g(x) \to \infty$ for $x \uparrow 1$ and the performance of PDF estimates is bad since they are appointed for the estimation of finite PDFs.
Suppose now that the target PDF is a triangular one. Then

$$\begin{aligned} \Phi^{-1}(x) = (\Phi^{+\mathrm{tr}})^{-1}(x) &= 1 - \sqrt{1-x}, \\ T_\gamma(x) &= 1 - (1+\gamma x)^{-1/(2\gamma)}, \\ T_\gamma^{-1}(x) &= ((1-x)^{-2\gamma} - 1)/\gamma, \\ T_\gamma^{'}(x) &= 0.5(1+\gamma x)^{-1/(2\gamma)-1}, \\ (T_\gamma^{-1}(x))^{'} &= 2(1-x)^{-2\gamma-1}. \end{aligned}$$

If the PDF of X has the form (5) then the PDF of $Y = T_\gamma(X)$ is

$$g(x) \simeq \ell_\infty \left(\alpha/\gamma\right)^{-(1/\alpha+1)} (1-x)^{2\gamma/\alpha-1},$$

as $x \uparrow 1$. For $\gamma = \alpha$ the PDF $g(x) \to 0$ as $x \uparrow 1$ and the rate of its decay is the same as for the triangular PDF ($\sim C(1-x)$). If $\gamma \neq \alpha$ and $\gamma > \alpha/2$ then $g(x)$ tends to 0 for $x \uparrow 1$. The rate of this decay is not $\sim C(1-x)$, but the PDF estimates will work well in this case. These problems can arise only if $\gamma < \alpha/2$. If α is estimated by a consistent estimate (e.g. by Hill's one) such a rough underestimation will be a rare event. Really, the probability to get γ, which is far from α tends to 0 as the sample size $n \to \infty$. The non-asymptotic confidence interval

$$I_\epsilon = [\hat{\gamma}_n/(1 + y_\epsilon k^{-0.5}); \hat{\gamma}_n/(1 - y_\epsilon k^{-0.5})]$$

of level $1-\epsilon$ for the index α has been proved in (Novak, 1998). Here, $\Phi(-y_\epsilon) = (\epsilon/2 - c_* k^{-0.5})_+$ and $C_* < 0.8$.
Suppose now that $g(x)$ is estimated by some estimate $\breve{g}_n$. What value will attain the tail index of the PDF estimate $\breve{f}_n$ defined by (3)?
Let $\breve{g}_n$ be a histogram estimate. Since for both transformations $T_\gamma(x) \to 1$ as $x \to \infty$ follows, then by (3) $\breve{f}_n(x) \simeq \breve{g}_n(1)T_\gamma^{'}(x)$ as $x \to \infty$ holds. Note that $\breve{g}_n(1)$ is the height of the right-most bar of the histogram and $\breve{g}_n(1) > 0$ holds a.s. for large n if the support of X is R_+.
If the target PDF is a uniform one, it means that $\breve{f}_n(x) \simeq \breve{g}_n(1)(1+\gamma x)^{-(1/\gamma+1)}$, i.e. the estimate has the tail index $\hat{\gamma}_n$ which has been chosen for the guess PDF. For the kernel estimate with a compactly supported kernel (e.g., Epanechnikov's kernel: $K(x) = 0.75(1-x^2), |x| \leq 1$) $\breve{g}_n(x)$ depends on the smoothing parameter h and $\breve{g}_n(1,h) \sim O(h^{-1})$ for sufficiently large $h > 1 - T_\gamma(x_{(n)})$, $\breve{g}_n(1,h) = 0$ for other h. For a such kernel estimate $\breve{f}_n(x,h) \simeq \breve{g}_n(1,h)(1+\gamma x)^{-(1/\gamma+1)}$ holds. Here, boundary effects of a kernel estimate applied to the PDF with a compact support are observed due to the truncation of a kernel near the boundaries.
If the target PDF is triangular then

$$\breve{f}_n(x) \simeq 0.5\breve{g}_n(1)(1+\gamma x)^{-(1/(2\gamma)+1)}.$$

holds. In this case the tail index is two times larger than the true one. The same is true for a kernel estimate with $\breve{g}_n(1)$ replaced by $\breve{g}_n(1,h)$. To remove this effect a smoothed histogram (or a polygram, i.e. a histogram with a variable bin width (Tarasenko, 1968)) may be used. For such $\breve{g}_n$ and for $x \simeq 1$ $\breve{g}_n(x) = C_n(1-x)$ holds, where C_n is the slope of the line which connects the center of the top of the right-most histogram bar with the point $(1,0)$. Then for the triangular target PDF

$$\begin{aligned} \breve{f}_n(x) \simeq & 0.5C_n(1 - (1 - (1+\gamma x)^{-1/(2\gamma)})) \\ & \cdot (1+\gamma x)^{-(1+1/(2\gamma))} \\ \simeq & 0.5C_n(1+\gamma x)^{-(1+1/\gamma)}, \end{aligned}$$

holds, i.e. in this case the tail index of the estimate coincides with its estimate for the guess PDF. The polygram is preferable to the histogram in this case. Really, a density g with $g(x) \to 0$ as $x \to 1$ is estimated. If the bin lengths are equal, then very few observations fall into the right-most bin. Hence, the histogram is unstable at the tail. The polygram adapts the bin widths to the data and works better. Figure 1 shows the poor behavior of

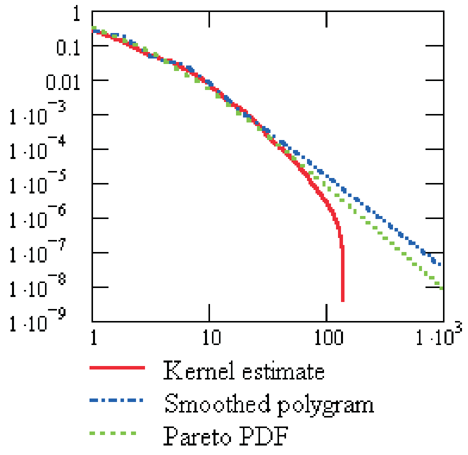

Fig. 1. Comparison of the tail estimates

the kernel estimate with Epanechnikov's kernel at the tail and its vanishing for large x in comparison to the smoothed polygram for a Pareto PDF with $\gamma = 0.5$. The estimates have been obtained using the transformation T_γ to a triangular PDF with $\gamma = 0.59$. The sample size is $n = 100$, the parameter $h = 0.1 < 1 - T_\gamma(x_{(n)})$, the number of bins for a polygram is 16.

Summarizing one may conclude that the best combination of the algorithm parameters among the considered ones is the triangular PDF as a target PDF and a smoothed polygram (or a kernel estimate with compactly supported kernel) as the PDF estimate on [0,1].

3. CLASSIFICATION PROBLEM AND QUALITY OF THE PDF ESTIMATION

Here, the problem of a quality measure of the PDF estimate is discussed. Normally, an expectation of a divergence in some functional space such as L_1, L_2, C is considered. The consistency of PDF estimates in the Kullback metric, in the total variation or χ^2-divergence is often considered, too. From the practical viewpoint the ability of an estimate to solve a specific applied problem is a much more important characteristic than its deviation from the estimated PDF in some sense. Here, the probability of the classification (pattern recognition) error is defined as a measure of the quality of the PDF estimate.

Suppose that the observed object O belongs to one of M different populations P_k, $k = \{1,\ldots,M\}$. The number of this population is unknown. Let $p_k = \Pr\{O \in P_k\}$ be an a priori probability of a population and some random characteristic (feature) $X \in R_+$ of O is observed. Denote the PDF of X if $O \in P_k$ as $f_k(x)$. The aim is to estimate the number k of the population to which O belongs by the observation X. The solution of this problem is a classifier, i.e. a function $\eta : R_+ \to \{1,\ldots,M\}$ which assigns a value to the characteristic $X = x$ corresponding to the estimated number $k \in \{1,\ldots,M\}$ of the population.

Suppose that the penalty of the mistake $q_k(x)$ depends on the number of the true population k and the value of the characteristic x. The risk of the misclassification is the mean of the loss

$$L(\eta) = E\sum_{k=1}^{M} \mathbf{1}\{\mathbf{O} \in \mathbf{P_k}, \eta(\mathbf{X}) \neq \mathbf{k}\}\mathbf{q_k}(\mathbf{x})$$

$$= \sum_{k=1}^{M} p_k \int\limits_{\eta(X)\neq k} q_k(x) f_k(x) dx \qquad (6)$$

The minimal risk is attained at the Bayesian classifier

$$\eta_B(x) = k \quad \text{if} \quad p_k q_k(x) f_k(x) \geq p_i q_i(x) f_i(x) \quad \text{for all} \quad i \neq k$$

Normally, the true f_k and p_k are unknown, but there are some estimates $\hat{f}_k, \hat{p}_k$ of them. The empirical Bayesian estimate is used

$$\eta_{EB}(x) = k \quad \text{if} \quad \hat{p}_k q_k(x) \hat{f}_k(x) \geq \hat{p}_i q_i(x) \hat{f}_i(x) \quad \text{for all} \quad i \neq k$$

Obviously, $L(\eta_{EB}) > L(\eta_B)$. Moreover $L(\eta_{EB})$ defined by (6) is a random variable since η_{EB} is a random function which depends on the data by which the estimates $\hat{f}_n$ were evaluated.

The penalties q_i are defined by the loss of the missclassification in a real classification problem. If the tail behavior is not important, then one can define $q_k(x) = 1$ for all k. Then, since the observations in the tail domain are rare, the improvement of the classification at the tail provides a negligible decrease of the risk of the misclassification. In problems where the tail behavior is important (like the WWW-traffic analysis), $q_k(x)$ should be larger on the tail and less on the body, i.e. $q_k(x) \to \infty$ as $x \to \infty$. The Bayesian approach can be applied only if the condition holds:

$$\int\limits_0^\infty q_k(x) f_k(x) dx < \infty \qquad (7)$$

For the PDF estimation the classification means that M different PDFs are estimated and used in the empirical Bayesian classifier for objects with these PDFs. However, a classifier η_{EB} is not sensitive with respect to the accuracy of the PDF estimates, i.e. different estimates classifying the objects to the same classes may have the same value $L(\eta_{EB})$. Hence, the estimate of $L(\eta)$

$$\widetilde{L(\eta)} = \sum_{k=1}^{M} \hat{p}_k \int_{\eta(X)\neq k} q_k(x)\hat{f}_k(x)dx$$

may be used which reflects the quality of the PDF estimator. $\widetilde{L(\eta_{EB})}$ is the risk of Bayesian classifier when $\hat{f}_i(x)$ and $\hat{p}_i$, $i = 1,\ldots,M$ are the true PDFs and a priori probabilities of classes, respectively. Therefore, $\widetilde{L(\eta_{EB})}$ fluctuates around the minimal Bayesian risk. The more accurate the PDF estimates are the more close is $\widetilde{L(\eta_{EB})}$ to $L(\eta_B)$. Note, that $L(\eta)$ is invariant regarding any "one-to-one" transformation with the properties mentioned above.

4. SIMULATION STUDY

The closeness of $\widetilde{L(\eta_{EB})}$ to $L(\eta_B)$ is compared for a kernel estimate

$$\hat{f}_n^K(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right),$$

where h is a smoothing parameter, $K(x)$ is a kernel function and a smoothed polygram

$$f_{L,n}(t) = L((n+1)\lambda(\Delta_{rL}))^{-1} \text{ for } \quad t \in \Delta_{rL}.$$

Here $\lambda(\Delta)$ is the length of Δ, $\lambda(\Delta_{rk}) \to 0$, $L = o(n)$, the number of observations inside each interval Δ_{rL} is less or equal to L. The fixed transformation $T(x) = (2/\pi)\arctan(x)$ and the transformation T_γ adapted to the empirical data (see section 2.3) are applied to both estimates. The obtained estimates are compared for different distributions by means of a Monte Carlo study. Couples of populations with PDFs $f_1(x)$ and $f_2(x)$ were considered. Samples of the known PDFs of a Pareto distribution (4) with the extreme value index $\gamma = \{0.3, 1.2\}$, and of a Burr distribution Burr(γ,ρ) with $\gamma = 1, \rho = -1$ and $\gamma = 4, \rho = -2$, $f(x) = \lambda\tau x^{\tau-1}(1 + x^\tau)^{-\lambda-1}$, where the extreme value index γ and a parameter ρ are defined as $\gamma = 1/(\lambda\tau), \rho = -1/\lambda$ have been generated. The mixed couples Burr(1,-1)-Pareto(1.2) are also considered.

The mean penalty of the misclassification for two classes is

$$L(\eta) = p_1 \int_0^\infty q_1(x)f_1(x)\mathbf{1}\{\eta(\mathbf{x}) = \mathbf{2}\}\mathbf{dx}$$
$$+ p_2 \int_0^\infty q_2(x)f_2(x)\mathbf{1}\{\eta(\mathbf{x}) = \mathbf{1}\}\mathbf{dx}.$$

Here, $\eta : R_+ \to \{1, 2\}$ is the classifier. Since for known PDFs one can calculate an exact Bayesian risk of the misclassification $L(\eta_B)$ (the best possible), then as a measure of the relative accuracy of the PDF estimate one can consider

Table 1. Loss of the misclassification

PDFs	n	$J(\eta_{EB}) \cdot 10^3$			
		Pl	PlF	Ke	KeF
Pareto	50	21	87	84	86
	100	34	46	11	1
	500	24	43	3.365	5.029
Burr	50	45	29	41	63
	100	46	51	24	58
	500	42	55	13	15
Burr-Pareto	50	133	279	201	242
	100	102	209	121	186
	500	89	156	41	131

$$J(\eta_{EB}) = \left|\frac{1/m\sum_{i=1}^{m} \widetilde{L_i(\eta_{EB})}}{L(\eta_B)} - 1\right|,$$

where

$$\widetilde{L(\eta_{EB})} = \hat{p}_1 \int_0^\infty q_1(x)\hat{f}_1(x)\mathbf{1}\{\eta_{\mathbf{EB}}(\mathbf{x}) = \mathbf{2}\}\mathbf{dx}$$
$$+ \hat{p}_2 \int_0^\infty q_2(x)\hat{f}_2(x)\mathbf{1}\{\eta_{\mathbf{EB}}(\mathbf{x}) = \mathbf{1}\}\mathbf{dx}.$$

The mean is taken over m samples of a Monte Carlo study. The worse is the PDF estimate the higher is $J(\eta_{EB})$. As proportions $p_1 = p_2 = 0.5$ were taken.

Penalties $q_k(x)$ are defined in such a manner that the switching of a Bayesian classifier from one population to another is in the tail domain, i.e. for some sufficiently large x $q_i f_i(x) = q_j f_j(x), i \neq j$. It was made to reject the situation where some $q_i f_i(x)$ dominates the other for tail points. Otherwise the empirical Bayesian classifier will classify such points as the points from one class and the effect of the improvement of the accuracy of the PDF estimation will be negligible. Penalty functions $q_1(x) = (d_1 + x)^\delta$ and $q_2(x) = (d_2 + x)^\delta$ have been used, where $d_1, d_2 > 0$, $\delta = 0.5\min\{1/\gamma_1, 1/\gamma_2\}$. Here γ_1 and γ_2 are the estimates of tail indexes of the population couple. They ensure the convergence of (7). The constants d_k for two populations ($k = 1, 2$) are selected from the assumption that switching of the Bayesian criterion occurs at the $(1-\alpha)-$quantile of an arbitrary population, i.e. $q_1(x)f_1(x)$ and $q_2(x)f_2(x)$ cross each other in the point $x^\star = F_1^{-1}(1-\alpha)$ ($F_1(x)$ is the CDF of one of the population). For the Generalized Pareto distribution

$$x^* = (\alpha^{-\gamma_1} - 1)/\gamma_1$$

holds. The values d_1 and d_2 are defined by the equation:

$$(d_1 + x^*)^\delta f_1(x^*) = (d_2 + x^*)^\delta f_2(x^*),$$

i.e. $d_1 = (d_2 + x^*)\left(f_2(x^*)/f_1(x^*)\right)^{1/\delta} - x^*$,

$$d_2 > x^*\left(\left(f_2(x^*)/f_1(x^*)\right)^{-1/\delta} - 1\right).$$

The quantile α is selected which provides positive values of d_1, d_2 and similar values of the penalties $q_1(x)$ and $q_2(x)$.
In the Table 1, $J(\eta_{EB})$ is shown for two types of transformed smoothed polygrams (Pl and PlF) and transformed kernel estimates with Epanechnikov's kernel (Ke and KeF). Namely, Pl and Ke are calculated by means of the transformation T_γ, but PlF and KeF are calculated by means of the fixed transformation $T(x)$. A smoothing parameter $h = \sigma n^{-1/5}$ is selected (σ^2 is the dispersion of the transformed data), which performs well with Epanechnikov's kernel (Devroye and Györfi, 1985). 10, 16, 40 were chosen as the number of quantile intervals in the polygram for the sample sizes 50, 100 and 500, respectively. The latter corresponds to the selection of the number of points inside each interval of a polygram by the discrepancy method $L = [(\sqrt{n}/2 + 1/n)(n + 1)]$, $[r]$ denotes the integer part of r (Markovitch and Krieger, 2000). The number of Monte Carlo repetition times is $m = 25$. Penalty functions corresponding to the quantile $\alpha = 0.25$ have been used.
Hill's estimate for the tail index with a bootstrap selection of the smoothing parameter k is used (Caers and Van Dyck, 1999).
The simulation study shows that:

- The transformation adapted to the tail index provides better estimates than the fixed transformation despite of the rough estimation of the tail indexes.
- A kernel estimate is worse than a polygram for relatively small samples and better for large samples.

5. CONCLUSIONS

In the paper the non-parametric estimation of a heavy-tailed PDF is considered. The compression of the data to a finite interval in such a way that the PDF of the new r.v. will be convenient for the estimation, the restoration of this PDF and an inverse transformation to get a reliable and accurate PDF estimate are explored. Since the true PDF is unknown, it is impossible to transform the r.v. to the r.v. with an exact (preliminary known) finite PDF. The transformation adapted to the data is constructed under the assumption that the PDF is Generalized Pareto and that its parameter (a tail index) is estimated by Hill's estimate. Such a transformation may provide a better estimation of the heavy-tailed PDFs than a fixed transformation (independent on the data), especially if Hill's estimate ensures a quite accurate estimate. The transformation to a triangular PDF and a smoothed polygram (or a kernel estimate) are established as the best combination for the estimation. It allows: 1) to get a stable estimation of the PDF regarding small perturbations of the transformation due to the rough tail index estimation; 2) to keep the tail decay (of the true PDF) after the inverse transformation. In contrast to a polygram, a kernel estimate does not ensure a correct tail decay rate. Besides, a kernel estimate restores the PDF of a transformed r.v. poor on the boundaries of the finite interval due to the truncation of the kernel. In this respect, a polygram is preferable, especially for quite limited sample sizes and a tail estimation. For sufficiently large samples kernel estimates restore the PDF body better.
As a quality measure of a heavy-tailed PDF estimate the relative risk of the misclassification is proposed. It is investigated for different estimates and heavy-tailed PDFs by a Monte Carlo study.

6. REFERENCES

[1] Caers, J. and J. Van Dyck (1999). Nonparametric tail estimation using a double bootstrap method. *Computational Statistics & Data Analysis*, **29**, 191–211.

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

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

[4] Novak, S.Y. (1998). Berry-Esseen inequalities for a ratio of sums of random variables. *University of Sussex: Research Report No 98/07 SMS/CSSM.*

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

[6] Sigman, K. (1999). Appendix: A primer on heavy-tailed distributions. *Queueing Systems*, **33**, 261-275.

[7] Tarasenko, F.P.(1968). On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable and the distribution-free test of goodness-of-fit In: *Proceedings IEEE*, Vol.56, pp. 2052–2053.

[8] Wand, M.P., J.S. Marron and D. Ruppert (1991). Transformations in density estimation. *Journal of American Statistical. Association, Theory and Methods*, **86(414)**, 343–353.