УДК 519.272

© 2002 г. Н.М.МАРКОВИЧ

(Институт проблем управления РАН им.В.А.Трапезникова, Москва)

# ТРАНСФОРМИРОВАННЫЕ ОЦЕНКИ ПЛОТНОСТЕЙ РАСПРЕДЕЛЕНИЯ С ТЯЖЕЛЫМИ ХВОСТАМИ И КЛАССИФИКАЦИЯ

Рассматривается непараметрическое оценивание плотности распределения вероятности (ПР) с тяжелыми хвостами. Предлагается трансформировать исходные данные к ограниченному интервалу. Тогда ПР находится путем обратного преобразования оценки ПР трансформированных данных. Исследуется трансформация, адаптивная к данным, при которой сохраняется порядок убывания хвоста истинной ПР и обеспечивается стабильное оценивание относительно уклонений в оценках хвостового индекса. В задачах классификации в качестве меры качества оценок ПР предлагается эмпирический риск ошибочной классификации эмпирического байесовского классификатора.

## 1. Введение

В последние годы случайные величины (СВ), распределенные с тяжелыми хвостами [1], приняты как реалистичные модели различных явлений: длины передаваемых по Интернету файлов, размеры страховок при катастрофах, наблюдаемые очень низкие или высокие температуры, наводнения, сильные ветры, концентрации редких полезных ископаемых могут быть описаны тяжело-хвостовыми распределениями.

Определение тяжело-хвостовых распределений может быть дано из теории экстремальных величин. Пусть $x_1, x_2, \ldots, x_n$ - независимые одинаково распределенные СВ с функцией распределения (ФР) $F(x)$ и $M_n = \max(x_1, x_2, \ldots, x_n)$. Известно [2], что если предельное распределение $M_n$ существует, то при некотором удобном выборе констант $a_n, b_n$

$$\lim_{n\to\infty} P\{(M_n - a_n)/b_n \le x\} = H_\gamma(x),$$

где

$$H_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), & \gamma \neq 0 \\ \exp(-e^{-x}), & \gamma = 0. \end{cases}$$

Распределение $F$ называется тяжело-хвостовым, если параметр EVI ("extreme value index") $\gamma > 0$. Грубо говоря, тяжело-хвостовые распределения - это такие распределения, чьи хвосты убывают к 0 медленнее, чем экспоненциальный хвост. Примерами таких распределений являются распределения Парето, логнормальное, Коши, Вейбулла с $0 < \gamma < 1$.

Здесь рассматривается непараметрическое оценивание тяжело-хвостовых ПР.

Известно, что гистограмма является хорошей оценкой для финитных и легко-хвостовых (т.е. без тяжелых хвостов) ПР, но на хвостах дает оценки, вводящие в заблуждение. Это верно и для других известных непараметрических оценок таких, как ядерные, проекционные, сплайн оценки. Как правило, они показывают острые пики на точках, соответствующих редким событиям ("outliers"), и не обеспечивают правильного порядка убывания на бесконечности.

Необходимость непараметрических оценок ПР с правильным поведением в хвостовой области очевидна. Это особенно существенно, если сравниваются ПР нескольких популяций. Одной из проблем, связанных с оцениванием ПР, является классификация. При использовании эмпирического байесовского классификатора наблюдения классифицируются путем сравнения оценок плотностей каждого класса. Так как наблюдение может появиться как в области хвоста, так и в основной области ПР ("body"), точность оценивания на хвостах принципиальна для классификации.

Для улучшения оценивания на хвостах в [3] было предложено трансформировать данные к ограниченному интервалу, оценить ПР по трансформированным данным и растянуть эту оценку обратным преобразованием. Основываясь на этой идее, предлагается трансформация, адаптивная к эмпирическим данным. Выбираются оценки ПР, сохраняющие порядок убывания на хвостах, как у истинной ПР. В задачах классификации в качестве меры качества оценок ПР предлагается эмпирический риск ошибки классификации эмпирического байесовского классификатора.

## 2. Предварительная трансформация

Пусть $X_1, \ldots, X_n$ - независимые одинаково распределенные СВ с ФР $F$ и ПР $f$. Рассмотрим некоторую монотонно возрастающую непрерывную функцию $T : R_+ \to [0,1]$ и выборку $Y_j = T(X_j), j = 1, \ldots, n$. ФР для СВ $Y_j$ определяется как

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

а ПР как

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

Аналогично, так как $X_j = T^{-1}(Y_j)$,

(3) $$f(x) = g(T(x))T^{'}(x).$$

Можно оценить $g(x)$ по $Y_1, \ldots, Y_n$, используя один из непараметрических методов. Обозначим эту оценку через $\widehat{g}_n(x)$. Тогда

(4) $$\widehat{f}_n(x) = \widehat{g}_n(T(x))T^{'}(x)$$

может быть использована как оценка $f$. В [3] развита общая теория таких оценок. Для ядерных оценок с компактным ядрами рекомендована трансформация к треугольной ПР $\phi^{\mathrm{tr}}(x)$ на $[0,1]$, а для гистограммы - к равномерной ПР $\phi^{\mathrm{un}}(x)$. Эти трансформации $T(x)$ обеспечивают оптимальную скорость сходимости оценок ПР в $L_1$. Так как такие $T(x)$ и, следовательно, распределение $Y_j = T(X_j)$ зависят от неизвестной ФР $F$, невозможно получить $g$ абсолютно идентичной $\phi^{\mathrm{tr}}$ (или $\phi^{\mathrm{un}}(x)$ ). Поэтому в [3] вместо $F$ используются некоторые семейства ФР $\Psi_\gamma$, которые зависят от параметра $\gamma$, определяемого по выборке. В [4] рассматриваются трансформации не только к ПР с компактным носителем; выбирается трансформация из подходящего параметрического семейства, которая обеспечивает минимальную средне-квадратичную ошибку оценивания ПР трансформированной СВ с помощью ядерной оценки.

Здесь предлагается рассматривать некоторое семейство тяжело-хвостовых распределений $\Psi_\gamma$, где EVI $\gamma$ оценивается широко используемой оценкой Хилла [5]:

$$\widehat{\gamma}^H(n,k) = \frac{1}{k}\sum_{i=1}^{k} \log x_{(i)} - \log x_{(k+1)},$$

где $x_{(1)} \geq \ldots \geq x_{(n)}$ - порядковые статистики выборки $x_1, x_2, \ldots, x_n$, а $k$ - сглаживающий параметр. Оценка Хилла состоятельна при $k \to \infty$, $k/n \to 0$ $(n \to \infty)$ и асимптотически нормальна (при некоторых дополнительных условиях) со средним $\gamma$ и дисперсией $\gamma^2/k$ [11].

### 3. Алгоритм оценивания

Предлагается следующий алгоритм оценивания тяжело-хвостовых ПР:

1. EVI $\gamma$ СВ $X_j$ оценивается по выборке $\{X_1,\ldots,X_n\}$ с использованием оценки Хилла $\widehat{\gamma_n}$.

2. Конструируется трансформация $T = T_{\widehat{\gamma_n}}$ так, что если $\xi$ имеет предполагаемую ФР $\Psi_{\widehat{\gamma_n}}$, то $T_{\widehat{\gamma_n}}(\xi)$ имеет целевую ФР $\Phi$, например равномерную или треугольную ($\widehat{\gamma_n}$ предполагается фиксированной величиной).

3. Формируется трансформированная выборка $Y_j = T_{\widehat{\gamma_n}}(X_j)$.

4. ПР СВ $Y_1,\ldots,Y_n$ оценивается некоторой непараметрической оценкой $\widehat{g}_n(x)$.

5. ПР СВ $X_j$ оценивается согласно (3).

### 4. Анализ алгоритма

Для реализации алгоритма необходимо выбрать предполагаемое семейство ФР, целевое распределение и оценку $\widehat{g}_n$. Семейство ФР должно быть таким, чтобы трансформация $T$ и ее производная были легко вычисляемыми. Если $\Psi_\gamma$ - предполагаемое семейство и $\Phi$ - целевая ФР, то согласно (1) $T(x) = \Phi^{-1}(\Psi_\gamma(x))$ и $T^{-1}(x) = \Psi_\gamma^{-1}(\Phi(x))$. Обобщенное распределение Парето (Generalized Pareto distribution)

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

рассматривается здесь как предполагаемое семейство ФР. Тогда соответствующая ПР равна

$$\psi_\gamma(x) = \Psi_\gamma^{'}(x) = (1+\gamma x)^{-(1/\gamma+1)} \tag{5}$$

Это предположение основывается на результате теории экстремальных значений, а именно: для достаточно высокого порога $u$ для СВ $X$ распределение разности $Y = X - u$ сходится к обобщенному распределению Парето [6].

В качестве целевой ПР рассматриваются равномерная ПР $\phi^{\mathrm{un}}(x) = \mathbf{1}\{x \in [0,1]\}$ и треугольная ПР $\phi^{+\mathrm{tr}}(x) = 2(1-x)\mathbf{1}\{x \in [0,1]\}$. Соответствующие ФР определяются как

$$\Phi^{\mathrm{un}}(x) = x\mathbf{1}\{x \in [0,1]\} + \mathbf{1}\{x > 1\},$$

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

Далее решаются два вопроса:

1. Какая трансформация обеспечивает наиболее стабильное оценивание относительно ошибок оценивания $\gamma$?
2. Какая скорость убывания хвоста получается после обратной трансформации?

Предположим сначала, что $\phi = \phi^{\mathrm{un}}, \Phi = \Phi^{\mathrm{un}}$ - целевые ПР и ФР. Тогда

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

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

$$T_\gamma^{-1}(x) = ((1-x)^{-\gamma} - 1)/\gamma,$$

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

Предположим, что

$$f(x) = \ell(x)\psi_\alpha(x) \tag{6}$$

- истинная ПР $X$, где $\ell(x)$ - ограниченная функция, удовлетворяющая условию

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

(EVI этой ПР равняется $\alpha$). Тогда из (2) ПР для $Y = T_\gamma(X)$ при $x \in [0,1)$ будет:

$$g(x) = \left(1 + ((1-x)^{-\gamma} - 1)\alpha/\gamma\right)^{-(1/\alpha+1)} (1-x)^{-\gamma-1}\ell\left(((1-x)^{-\gamma} - 1)/\gamma\right).$$

Эта ПР будет оцениваться оценкой $\widehat{g}_n$ для некоторой величины $\gamma$. Поэтому поведение этой ПР должно быть удобным для оценивания на ее интервале определения $x \in$

$[0,1)$. Если $x$ близок к нулю, то $g(x)$ ограничена. Рассмотрим $g(x)$ для $x \to 1$. Заметим, что в этом случае $\ell(T_\gamma^{-1}(x)) \to \ell_\infty < \infty$, так как $T_\gamma^{-1}(x) \to \infty$. Тогда при $x \to 1$

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

Если $\alpha = \gamma$, то $g(x) \simeq \ell_\infty$, т.е. $g(x)$ ведет себя как равномерная ПР в окрестности 1 (с точностью до множителя $\ell_\infty$). Пусть $\alpha \neq \gamma$. Для $x \to 1$ и произвольной константы $C < \infty$ выполняется $(1-x)^{-\gamma} + C \sim (1-x)^{-\gamma}$. Отсюда

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

Если $\gamma$ больше истинного индекса $\alpha$, т.е. $\gamma/\alpha - 1 > 0$, то ПР $g(x)$ удобна для оценивания в окрестности 1. Но если $\gamma < \alpha$, то $g(x) \to \infty$ для $x \to 1$ и оценивание ПР не будет успешным.

Предположим теперь треугольную ПР в качестве целевой. Тогда

$$\Phi^{-1}(x) = (\Phi^{+\mathrm{tr}})^{-1}(x) = 1 - \sqrt{1-x},$$

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

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

$$T_\gamma^{'}(x) = 0,5(1+\gamma x)^{-1/(2\gamma)-1},$$

$$(T_\gamma^{-1}(x))^{'} = 2(1-x)^{-2\gamma-1}.$$

Если ПР для $X$ имеет форму (6), то ПР СВ $Y = T_\gamma(X)$ будет

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

при $x \to 1$. Для $\gamma = \alpha$ ПР $g(x) \to 0$ при $x \to 1$ и скорость ее убывания будет такой же, как у треугольной ПР ($\sim C(1-x)$). Если $\gamma \neq \alpha$ и $\gamma > \alpha/2$, то $g(x)$ стремится к 0 для $x \to 1$. Скорость ее убывания не будет порядка $\sim C(1-x)$, но оценки ПР должны работать лучше в этом случае. Проблемы оценивания возникают, только если $\gamma < \alpha/2$. Однако, если $\alpha$ оценивается состоятельной оценкой (например оценкой Хилла), то такое грубое занижение оценки маловероятно. Действительно, вероятность получить $\gamma$, далекое от $\alpha$, стремится к 0, если размер выборки $n \to \infty$. Неасимптотический $(1-\epsilon)$ доверительный интервал

$$I_\epsilon = [\widehat{\gamma}_n/(1 + y_\epsilon k^{-0,5}); \widehat{\gamma}_n/(1 - y_\epsilon k^{-0,5})]$$

для индекса $\alpha$ был получен в [7], где $\Phi(-y_\epsilon) = (\epsilon/2 - c_* k^{-0,5})_+$ и $C_* < 0,8$.

Предположим теперь, что $g(x)$ оценивается некоторой оценкой $\widehat{g}_n$. Какое значение примет EVI для оценки $\widehat{f}_n$ ПР (4)?

Пусть $\widehat{g}_n$ - гистограммная оценка. Так как для обеих трансформаций $T_\gamma(x) \to 1$ при $x \to \infty$, то из (4) следует, что $\widehat{f}_n(x) \simeq \widehat{g}_n(1) T_\gamma^{'}(x)$ при $x \to \infty$ . Заметим, что $\widehat{g}_n(1)$ - это высота крайней правой ступеньки гистограммы и $\widehat{g}_n(1) > 0$ почти наверное для достаточно больших $n$, если $X$ определено на $R_+$.

Если равномерная ПР - целевая, то $\widehat{f}_n(x) \simeq \widehat{g}_n(1)(1+\gamma x)^{-(1/\gamma+1)}$, т.е. оценка имеет тот же индекс $\widehat{\gamma}_n$, что был выбран для предполагаемой ПР. Для ядерной оценки с компактным ядром (например ядром Епанечникова: $K(x) = 0,75(1 - x^2), |x| \leq 1$) $\widehat{g}_n(x)$ зависит от параметра сглаживания $h$ и $\widehat{g}_n(1,h) \sim O(h^{-1})$ для достаточно больших $h > 1 - T_\gamma(x_{(n)})$, $\widehat{g}_n(1,h) = 0$ для других $h$. Для такой ядерной оценки выполняется $\widehat{f}_n(x,h) \simeq \widehat{g}_n(1,h)(1+\gamma x)^{-(1/\gamma+1)}$. Однако для ядерной оценки, применяемой к ПР с компактным носителем, наблюдаются краевые эффекты вследствие усечения ядра около границ.

Если треугольная ПР - целевая, то

$$\widehat{f}_n(x) \simeq 0,5\widehat{g}_n(1)(1+\gamma x)^{-(1/(2\gamma)+1)}.$$

В этом случае индекс в 2 раза больше, чем истинный. Это также выполняется и для ядерной оценки, в которой вместо $\widehat{g}_n(1)$ используется $\widehat{g}_n(1,h)$. Для устранения этого эффекта может быть использована сглаженная гистограмма (или полиграмма, т.е. гистограмма с переменным размером интервалов [8]). Для такой оценки $\widehat{g}_n$ и для $x \simeq 1$ $\widehat{g}_n(x) = C_n(1-x)$, где $C_n$ - это наклон линии, соединяющей центр вершины самой правой ступеньки гистограммы с точкой $(1,0)$. Тогда для треугольной целевой ПР

$$\begin{aligned}\widehat{f}_n(x) &\simeq 0,5C_n(1-(1-(1+\gamma x)^{-1/(2\gamma)}))(1+\gamma x)^{-(1+1/(2\gamma))}\\ &\simeq 0,5C_n(1+\gamma x)^{-(1+1/\gamma)},\end{aligned}$$

т.е. в этом случае индекс оценки совпадает с индексом предполагаемой ПР.

Полиграмма предпочтительнее гистограммы в этом случае. Действительно, оценивается ПР $g$ такая, что $g(x) \to 0$ при $x \to 1$. Если длины интервалов равны (как у гистограммы), то только немногие наблюдения попадут в самый правый интервал. Поэтому гистограмма не будет устойчива на хвосте. Напротив, полиграмма адаптирует длину интервалов к данным и работает лучше.

Рисунок показывает плохое оценивание хвоста ядерной оценкой с ядром Епанечникова по сравнению со сглаженной полиграммой для ПР Парето с $\gamma = 0,5$. Оценки были получены с использованием трансформации $T_{\widehat{\gamma}}$ к треугольной ПР с $\widehat{\gamma} = 0,59$. Объем выборки $n = 100$, параметр сглаживания $h = 0,1 < 1 - T_{\gamma}(x_{(n)})$, число интервалов в полиграмме равно 16.

Из анализа можно заключить, что лучшей комбинацией среди рассмотренных является треугольная ПР в качестве целевой ПР и сглаженная полиграмма (или ядерная оценка с компактным ядром) в качестве оценки ПР на [0,1].

### 5. Классификация и качество оценивания ПР

Важная проблема - качество оценивания ПР. Обычно математическое ожидание разницы между оценкой и истинной ПР в пространствах $L_1$, $L_2$, $C$ рассматривается как мера качества оценки ПР. Состоятельность оценок может рассматриваться также в метрике Кульбака, в смысле "total variation"[3] или "$\chi^2$-divergence"[9]. Однако для практики способность оценки ПР решать какую-либо прикладную задачу гораздо важнее, чем точность оценки в каком-то смысле. Поэтому предлагается сравнивать точность оценок ПР с помощью альтернативной меры, а именно, по оценке вероятности ошибочной классификации.

Предположим, что наблюдаемый объект $O$ принадлежит к одной из $M$ различных популяций $P_k$, $k = \{1, \dots, M\}$. Номер этой популяции неизвестен. Пусть $p_k = \Pr\{O \in P_k\}$ - априорная вероятность популяции и наблюдается некоторая случайная характеристика (свойство) $X \in R_+$ объекта $O$. Обозначим ПР $X$, если $O \in P_k$, как $f_k(x)$. Необходимо оценить номер популяции $k$, к которой принадлежит $O$ по наблюдению $X$. Решением этой проблемы является классификатор, т.е. функция вида $\eta : R_+ \to \{1, \dots, M\}$, которая соотносит номер популяции $k \in \{1, \dots, M\}$ величине характеристики $X = x$.

Предположим, что штраф за ошибку классификации $q_k(x)$ зависит от номера истинной популяции $k$ и величины характеристики $x$. Риск ошибочной классификации определяется как средняя потеря

$$L(\eta) = E \sum_{i=1}^{M} \mathbf{1}\{O \in P_i, \eta(x) \neq i\} q_i(x)$$

$$= \sum_{i=1}^{M} p_i \int_0^\infty q_i(x) f_i(x) dx - \sum_{i=1}^{M} p_i \int_{\eta(x)=i} q_i(x) f_i(x) dx.$$

Минимальный риск достигается на байесовском классификаторе $\eta_B(x)$ :

$$L(\eta_B) = \sum_{i=1}^{M} p_i \int_0^\infty q_i(x) f_i(x) dx - \int_0^\infty \max_i p_i q_i(x) f_i(x) dx,$$

$$\text{где } \eta_B(x) = k, \text{ если } p_k q_k(x) f_k(x) \geq p_i q_i(x) f_i(x) \text{ для всех } i \neq k.$$

Как правило, истинные $f_k$ и $p_k$ неизвестны, но имеются их оценки. Эмпирический байесовский классификатор

$$\eta_{EB}(x) = k, \text{ если } \widehat{p}_k q_k(x) \widehat{f}_k(x) \geq \widehat{p}_i q_i(x) \widehat{f}_i(x) \text{ для всех } i \neq k.$$

Для него риск ошибочной классификации определяется как:

$$L(\eta_{EB}) = \sum_{i=1}^{M} p_i \int_0^\infty q_i(x) f_i(x) dx - \sum_{i=1}^{M} p_i \int_{\eta_{EB}(x)=i} q_i(x) f_i(x) dx.$$

Очевидно, $L(\eta_{EB}) > L(\eta_B)$. $L(\eta_{EB})$ отражает качество классификатора $\eta_{EB}$ и, следовательно, способность оценок $\widehat{f}_n$ решать проблему классификации.

Веса $q_i$ могут определяться как потери от ошибочной классификации в конкретной задаче классификации. Если поведение на хвосте не важно, то можно положить $q_k(x) = 1$ для всех $k$. Тогда вследствие того, что наблюдения в хвостовой области редки, улучшение классификации на хвосте незначительно уменьшает риск ошибочной классификации. В задачах, где оценивание хвоста важно, $q_k(x)$ должны быть больше на хвосте и меньше в основной части ПР, т.е. $q_k(x) \to \infty$ при $x \to \infty$. При использовании байесовского классификатора должно выполняться следующее условие:

$$\int_0^\infty q_k(x) f_k(x) dx < \infty. \tag{7}$$

Для задачи оценки ПР классификация означает оценивание $M$ различных ПР и использование их в эмпирическом байесовском классификаторе для объектов с этими

ПР. Однако классификатор $\eta_{EB}$ не чувствителен к качеству оценок ПР, например, оценки ПР различной точности, относящие объекты к одним и тем же классам, могут иметь одну величину риска $L(\eta_{EB})$. Поэтому в качестве критерия качества оценок ПР предлагается следующая оценка $L(\eta_{EB})$:

$$\widetilde{L(\eta_{EB})} = \sum_{i=1}^{M} \widehat{p}_i \int_0^\infty q_i(x)\widehat{f}_i(x)dx - \int_0^\infty \max_i \widehat{p}_i q_i(x)\widehat{f}_i(x)dx.$$

$\widetilde{L(\eta_{EB})}$ является риском байесовского классификатора, когда $\widehat{f}_i(x)$ и $\widehat{p}_i$, $i = 1,\ldots,M$ являются истинными ПР и априорными вероятностями классов соответственно. Поэтому $\widetilde{L(\eta_{EB})}$ флуктуирует вокруг минимального байесовского риска. Чем точнее оценка ПР, тем ближе $\widetilde{L(\eta_{EB})}$ к $L(\eta_B)$. Отсюда

$$|\widetilde{L}(\eta_{EB}) - L(\eta_B)| = |\sum_{i=1}^{M}\int_0^\infty (\widehat{p}_i q_i(x)\widehat{f}_i(x) - p_i q_i(x) f_i(x))dx$$

$$+ \int_0^\infty (\max_i p_i q_i(x) f_i(x) - \max_i \widehat{p}_i q_i(x)\widehat{f}_i(x))dx|$$

$$\leq \sum_{i=1}^{M} |\int_0^\infty q_i(x)\left(\widehat{p}_i\widehat{f}_i(x) - p_i f_i(x)\right)dx|. \tag{8}$$

Пусть $T : R^1 \to [0,1]$ - строго монотонно возрастающее взаимно однозначное преобразование. Обратная трансформация $T^{-1}$ и производные $T'$, $(T^{-1})'$ предполагаются непрерывными.

Пусть $g_i(x)$ есть ПР трансформированной СВ $Y = T(X)$, если $O \in P_i$, $G^i(x)$ и $G^i_n(x)$ соответствующие ФР и эмпирическая ФР. Заметим, что $L(\eta)$ инвариантна относительно любых трансформаций с перечисленными выше свойствами. В силу (3) выполняется

$$|\int_0^\infty q_i(x)\left(p_i f_i(x) - \widehat{p}_i\widehat{f}_i(x)\right)dx| = |\int_0^1 q_i(T^{-1}(x))(p_i g_i(x) - \widehat{p}_i\widehat{g}_i(x))dx| \tag{9}$$

$$\leq \int_0^1 |q_i(T^{-1}(x))(p_i g_i(x) - \mathbf{E}(\widehat{p}_i\widehat{g}_i(x)))|dx + \int_0^1 |q_i(T^{-1}(x))(\mathbf{E}(\widehat{p}_i\widehat{g}_i(x)) - \widehat{p}_i\widehat{g}_i(x))|dx.$$

Таким образом, уклонение $\widetilde{L(\eta_{EB})}$ от байесовского риска определяется точностью оценивания ПР трансформированной СВ в $L_1$.

## 6. Скорость сходимости $\widetilde{L(\eta_{EB})}$ к $L(\eta_B)$

Предположим, что

1. Рассматриваются только положительные СВ.

2. Штрафы $q_k(x)$ удовлетворяют кроме (7) следующему условию:

$$(10) \qquad \left(\int_0^\infty q_i^2(x)dT(x)\right)^{1/2} < \infty.$$

Например, если $q_i(x) = (d_i + x)^\sigma, f_i(x) \simeq (1+\alpha x)^{-(1+1/\alpha)}, \widehat{f_i}(x) \simeq (1+\gamma x)^{-(1+1/\gamma)}$, $T(x) = 1-(1+\gamma x)^{-(2\gamma)^{-1}}$ и $\sigma < \gamma^{-1}$, эти условия выполняются.

**Т е о р е м а** 1. Пусть $x_1, \ldots, x_n$ - независимые одинаково распределенные СВ с ПР $f(x)$, рассматривается трансформация к треугольной ПР $g(x) = 2(1-x)\mathbf{1}\{x \in [0,1]\}$. Если ПР трансформированной СВ $Y = T(X)$ оценивается ядерной оценкой

$$\widehat{g}(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right),$$

где $|K(x)| \le c$, $K(x) \neq 0$ для $|x| \le 1$ и $K(x) = 0$ для остальных $x$, $K_h(x) = 1/hK(x/h)\mathbf{1}\{|x| \le h\}$, $\int_{-h}^h K_h(x)dx = \int_{-1}^1 K(u)du = 1$ для $0 < h < 1$, параметр сглаживания $h = n^{-\beta}, \beta > 0, 0 < d < \min(0,5;\beta)$, штрафные функции удовлетворяют (10), то верно

$$P\{\overline{\lim}_{n\to\infty} n^d |L(\widetilde{\eta_{EB}}) - L(\eta_B)| \le c\} = 1,$$

где $c$ - константа, не зависящая от $n$.

**Т е о р е м а** 2. Пусть $x_1, \ldots, x_n$ - независимые одинаково распределенные СВ с ПР $f(x)$. Если ПР $g(x)$ трансформированной СВ $Y = T(X)$ такая, что $g(x) \le c < \infty, |g^{'}(x)| < \infty$ для $\{x : g(x) > 0\}$, оценивается полиграммой [8]

$$(11) \qquad \widehat{g}(x) = \frac{1}{m\lambda(\Delta_{rL})} \text{ для } \qquad x \in \Delta_{rL},$$

где обозначены $\Delta_{1L} = [x_{(1)}, x_{(L)}]$, $\Delta_{2L} = (x_{(L)}, x_{(2L)}]$, $\Delta_{3L} = (x_{(2L)}, x_{(3L)}], \ldots$, порядковые статистики выборки через $x_{(1)}, \ldots, x_{(n)}$, $\lambda(\Delta)$ длина интервала $\Delta$ (число наблюдений внутри каждого интервала $\Delta_{rL}$ меньше или равно $L$), $m = \frac{n+1}{L}, L = n^\beta, 0 < \beta < 1$ - параметр сглаживания, $0 < d < \frac{\beta}{2}$, штрафные функции удовлетворяют (10), то верно

$$P\{\overline{\lim}_{n\to\infty} n^d |L(\widetilde{\eta_{EB}}) - L(\eta_B)| \le c\} = 1,$$

где $c$ - константа, не зависящая от $n$.

**З а м е ч а н и е**. Асимптотические скорости сходимости для полиграммы и ядерной оценки одинаковы: $d < 0,5$.

### 7. Моделирование

Близость $L(\widetilde{\eta_{EB}})$ к $L(\eta_B)$ сравнивается для ядерной оценки

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

где $h$ - параметр сглаживания, $K(x)$ - ядерная функция, и для сглаженной полиграммы (11). Фиксированная трансформация $T(x) = (2/\pi)\arctan(x)$ и трансформация, адаптированная к эмпирическим данным $T_\gamma$ (см. раздел 4) применялись к обеим оценкам. Полученные оценки сравнивались для различных распределений посредством моделирования. Рассматривались пары популяций с ПР $f_1(x)$ и $f_2(x)$. Генерировались выборки с известными ПР: ПР Парето (5) с индексом $\gamma = \{0,3; 1,2\}$, ПР Бурра Burr$(\gamma, \rho)$ с параметрами $\gamma = 1, \rho = -1$ и $\gamma = 4, \rho = -2$, $f(x) = \lambda\tau x^{\tau-1}(1 + x^\tau)^{-\lambda-1}$, где EVI $\gamma$ и параметр $\rho$ определяются как $\gamma = 1/(\lambda\tau), \rho = -1/\lambda$. Рассматривались также смешанные пары распределений Burr(1,-1)-Pareto(1,2).

Средний риск ошибочной классификации для двух классов определяется как

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

Здесь $\eta : R_+ \to \{1, 2\}$ - классификатор. Так как при известных ПР можно вычислить точный байесовский риск ошибочной классификации $L(\eta_B)$ (наименьший среди возможных), то в качестве меры относительной точности оценок ПР рассматривалось

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

где

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

Среднее бралось над $m$ смоделированными выборками. Чем хуже оценка ПР, тем больше величина $J(\eta_{EB})$. Пропорции классов были взяты равными: $p_1 = p_2 = 0,5$.

Штрафы $q_k(x)$ были определены так, чтобы переключение байесовского классификатора с одной популяции на другую происходило в хвостовой области, т.е.

для достаточно больших $x$ $q_i f_i(x) = q_j f_j(x), i \neq j$. Это делалось для того, чтобы исключить ситуацию, когда некоторые $q_i f_i(x)$ доминируют другие на хвостовых точках. В противном случае байесовский классификатор будет классифицировать эти точки как точки из одного класса и эффект от улучшения точности оценивания ПР будет незначительный. Использовались штрафные функции, удовлетворяющие (7): $q_1(x) = (d_1 + x)^\delta$ и $q_2(x) = (d_2 + x)^\delta$, где $d_1, d_2 > 0$, $\delta = 0,5 \min\{1/\gamma_1, 1/\gamma_2\}$. Здесь $\gamma_1$ и $\gamma_2$ - оценки EVI для пары популяций. Константы $d_k$ для двух популяций $(k = 1, 2)$ выбирались из условия переключения байесовского критерия в точке $(1-\alpha)-$квантили одной из популяций, т.е. $q_1(x)f_1(x)$ и $q_2(x)f_2(x)$ пересекают друг друга в точке $x^\star = F_1^{-1}(1-\alpha)$ ($F_1(x)$ - ФР одной из популяций). Для обобщенного распределения Парето

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

Величины $d_1$ и $d_2$ определяются из уравнения:

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

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

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

Выбиралась такая величина $\alpha$, чтобы обеспечить положительные величины для $d_1, d_2$ и приблизительно равные значения штрафов $q_1(x)$ и $q_2(x)$.

В таблице $J(\eta_{EB})$ приведен для двух типов трансформированных сглаженных полиграмм ($Pl$ и $PlF$) и трансформированных ядерных оценок с ядром Епанечникова ($Ke$ и $KeF$). А именно, для $Pl$ и $Ke$ использовалась трансформация $T_\gamma$, а для $PlF$ и $KeF$ - фиксированная трансформация $T(x)$.

Параметр сглаживания $h$ вычислялся по формуле $h = \sigma n^{-1/5}$ (где $\sigma^2$ - дисперсия, вычисленная по трансформированным данным), рекомендованной в [3] для ядерных оценок с ядром Епанечникова. Число интервалов для полиграммы выбиралось равным 10, 16, 40 соответственно для выборок объемов 50, 100 и 500. Это соответствует выбору числа точек $L$ внутри каждого интервала полиграммы по методу невязки, использующему статистику Колмогорова-Смирнова [10]. Число генерируемых выборок $m = 25$. Штрафные функции вычислялись для $\alpha = 0,25$.

Для оценки EVI использовалась оценка Хилла, в которой сглаживающий параметр $k$ выбирался методом бутстреп (bootstrap) [11].
Моделирование показывает, что:

1. трансформация, адаптивная к EVI, обеспечивает лучшее оценивание, чем фиксированная трансформация, несмотря на грубое оценивание EVI.

2. ядерная оценка хуже, чем полиграмма, для относительно малых выборок и лучше для больших.

### 8. Заключение

Рассматривается непараметрическое оценивание тяжело-хвостовых ПР. Для получения состоятельных и точных оценок ПР исследуется сжатие данных к ограниченному интервалу таким образом, чтобы ПР новой СВ была бы удобной для оценивания, восстановление этой ПР и обратная трансформация. Поскольку истинное распределение неизвестно, то невозможно трансформировать исходную СВ к новой СВ с определенной, точно известной ПР. Предложена трансформация, адаптивная к эмпирическим данным в предположении, что исходное распределение есть обобщенное распределение Парето и его параметр (EVI) оценивается по оценке Хилла. Такая трансформация обеспечивает лучшее оценивание тяжело-хвостовых распределений (особенно в хвостовой области), чем фиксированная трансформация (не зависящая от данных), особенно если оценка Хилла достаточно точна. Установлено, что наилучшей комбинацией для оценивания является использование сглаженной полиграммы (или ядерной оценки) и трансформации к треугольной ПР. Это позволяет: 1) получить устойчивую оценку ПР относительно небольших отклонений трансформации вследствии грубого оценивания EVI; 2) сохранить порядок убывания хвоста такой же, как у истинной ПР, после обратной трансформации. В отличие от полиграммы, ядерная оценка не обеспечивает правильного порядка убывания хвоста. Кроме того, ядерная оценка восстанавливает ПР трансформированной СВ плохо на границах ограниченного интервала вследствие усечения ядер в этих точках. Ввиду этого, полиграмма предпочтительнее для оценивания хвостов, особенно для сравнительно малых объемов выборок. Для достаточно больших выборок ядерная оценка восстанавливает "тело"ПР лучше.

В качестве меры качества оценки тяжело-хвостовой ПР предлагается эмпирический риск ошибочной классификации эмпирическим байесовским классификатором. Доказывается, что скорость сходимости этого риска к байесовскому риску одинакова для полиграммы и ядерной оценки .

**Приложение**

**Доказательство теоремы** 1. Рассмотрим неравенство (9). Чтобы оценить первое слагаемое в правой части, заметим, что

$$\mathbf{E}(\widehat{p}_i\widehat{g}_i(x)) = p_iK_h * g_i(x),$$

где $*$ знак свертки. Отсюда

$$\int_0^1 |q_i(T^{-1}(x))(p_ig_i(x) - \mathbf{E}(\widehat{p}_i\widehat{g}_i(x))|dx$$

$$= \int_0^1 |q_i(T^{-1}(x)) \left( p_ig_i(x) \int_{-h}^h K_h(t)dt - p_iK_h * g_i(x) \right) |dx$$

$$\leq \int_0^1 q_i(T^{-1}(x)) \int_{-1}^1 |(p_ig_i(x-uh) - p_ig_i(x))K(u)|dudx$$

$$\leq \sup_{x\in[0,1],|u|\leq 1} |g_i(x-uh) - g_i(x)| \int_0^1 q_i(T^{-1}(x)) \int_{-1}^1 K(u)dudx$$

$$= \sup_{x\in[0,1],|u|\leq 1} |g_i(x-uh) - g_i(x)| \int_0^1 q_i(T^{-1}(x))dx$$

Так как $g_i(x)$ - треугольная ПР, т.е. $g_i(x) \simeq (1-x)\mathbf{1}(x \in [0,1])$, то выполняется

$$g_i(x-uh) - g_i(x) \simeq uh.$$

Тогда согласно предположению (10)

$$\int_0^1 |q_i(T^{-1}(x))(p_ig_i(x) - \mathbf{E}(\widehat{p}_i\widehat{g}_i(x)))|dx \simeq O(h)$$

Оценим теперь второе слагаемое в правой части (9). Так как для объектов из $i$-го класса

$$\widehat{p}_i\widehat{g}_i(x) = \frac{1}{nh}\sum_{j=1}^n K(\frac{x-y_j}{h})\mathbf{1}[\eta_j = i]\widehat{p}_i, 1 \leq i \leq M,$$

(где $\eta_j$ - метки объектов), мы можем представить $\widehat{p}_i\widehat{g}_i(x)$ как

$$\widehat{p}_i\widehat{g}_i(x) = \frac{1}{h}\int_{0,|x-t|\leq h}^1 K(\frac{x-t}{h})\widehat{p}_i dG_n^i(t),$$

где $G_n^i(t)$ - эмпирическая ФР, построенная по трансформированным наблюдениям $y_j = T(x_j), j = 1, \dots, n$, соответствующих $i$-му классу. Очевидно, что

$$\mathbf{E}(\widehat{p}_i \widehat{g}_i(x)) = \frac{1}{h} \int_{0,|x-t|\le h}^{1} K(\frac{x-t}{h}) p_i dG^i(t),$$

где $G^i(t)$ - ФР для $y_j = T(x_j), j = 1, \dots, n$, если $O \in P_i$.

Отсюда, имеем

$$\int_0^1 |q_i(T^{-1}(x))(\widehat{p}_i \widehat{g}_i(x) - \mathbf{E}(\widehat{p}_i \widehat{g}_i(x)))| dx$$
$$= \frac{1}{h} \int_0^1 |q_i(T^{-1}(x)) \int_{0,|x-t|\le h}^{1} K(\frac{x-t}{h}) d(G_n^i(t) - G^i(t))| dx$$
$$\le \int_0^1 |q_i(T^{-1}(x)) \int_{-1}^{1} K(u) d(G_n^i(x-uh) - G^i(x-uh))| dx$$
$$\le C_1 \int_0^1 |G_n^i(x-h) - G^i(x-h) - (G_n^i(x+h) - G^i(x+h))| dx$$
$$\le 2C_1 \sup_x |(G_n^i(x) - G^i(x)|.$$

В заключение из (9) имеем

$$\sum_{i=1}^{M} \int_0^\infty |q_i(x)(p_i f_i(x) - \widehat{p}_i \widehat{f}_i(x))| dx < 2C_1 M \sup_x |G_n^i(x) - G^i(x)| + MO(h).$$

Ввиду $h = n^{-\beta}$ из (8) следует

$$|L(\widetilde{\eta_{EB}}) - L(\eta_B)| n^d < 2C_1 M n^d \sup_x |(G_n^i(x) - G^i(x)| + MO(n^{d-\beta})$$

Обозначим первый член в правой части последнего неравенства через $B_n$. Для достаточно больших $n$ и $d < \beta$ второй член меньше 1. Отсюда, если $B_n \le 1$, то $n^d |L(\widetilde{\eta_{EB}}) - L(\eta_B)| \le 2$. Поэтому следует

$$P\{n^d |L(\widetilde{\eta_{EB}}) - L(\eta_B)| > 2\} < P\{B_n > 1\}.$$

Оценим $P\{B_n > 1\}$ из неравенства Колмогорова-Смирнова

$$P\{\sup_x |G_n^i(x) - G^i(x)| > \epsilon\} \le 2\exp(-2\epsilon^2 n).$$

Таким образом, выполняется

$$P\{B_n > 1\} \le 2\exp(-\frac{n^{1-2d}}{2M^2 C_1^2}).$$

Обозначим правую часть этого неравенства через $H(n,d)$. Если $d < \min(0,5;\beta)$, то ряд $\sum_{n=1}^{\infty} H(n,d)$ сходится, и согласно лемме Борелля-Кантелли следует заключение теоремы.

**Доказательство теоремы** 2.

Рассмотрим (9). Согласно неравенству Гельдера

$$|\int_0^1 q_i(T^{-1}(x))(p_i g_i(x) - \widehat{p}_i \widehat{g}_i(x))dx|$$

$$\le (\int_0^1 |q_i(T^{-1}(x))|^2 dx)^{\frac{1}{2}} (\int_0^1 |p_i g_i(x) - \widehat{p}_i \widehat{g}_i(x)|^2 dx)^{\frac{1}{2}}. \tag{12}$$

Предположим, что

$$(\int_0^1 |p_i g_i(x) - \widehat{p}_i \widehat{g}_i(x)|^2 dx)^{\frac{1}{2}} \le n^{-d}.$$

Тогда из (8), (9), (10) и (12) для $C_1 > 0$ выполняется

$$n^d |L(\widetilde{\eta_{EB}}) - L(\eta_B)| \le 2MC_1.$$

Отсюда имеем

$$P\{n^d |L(\widetilde{\eta_{EB}}) - L(\eta_B)| > 2MC_1\} < P\{(\int_0^1 |p_i g_i(x) - \widehat{p}_i \widehat{g}_i(x)|^2 dx)^{\frac{1}{2}} > n^{-d}\}$$

$$\le P\{sup_{x\in[0,1]}(g_i(x) - \widehat{g}_i(x))^2 > n^{-2d}\}$$

$$= P\{\widehat{g}_i(x_*) > g_i(x_*) + n^{-d}\} + P\{\widehat{g}_i(x_*) < g_i(x_*) - n^{-d}\},$$

где $x_* \in [0,1]$.

Обозначим $z = \widehat{g}_i(x)$ и $g_i = g_i(x)$. В предположениях теоремы полиграмма имеет асимптотическое (по $n$) распределение:

$$p(z) = \frac{(g_i/z)^{5/2}}{g_i\sqrt{2\pi/L}} \exp\{-\frac{1}{2}\left(\frac{(1-g_i/z)^2}{g_i/(Lz)} + \frac{L\,(g_i/z)^2}{m(1-g_i/(mz)}\right)\}\left(1 + O(1/\sqrt{n})\right)$$

[8]. Тогда следует

$$P\{\widehat{g}_i(x_*) > g_i(x_*) + n^{-d}\} = \sqrt{\frac{L}{2\pi}} \int_{g_i+1/n^d}^{\infty} \frac{(g_i/z)^{5/2}}{g_i} \times$$

$$\times \exp\{-\frac{1}{2}\left(\frac{(1-g_i/z)^2}{g_i/(Lz)} + \frac{L\,(g_i/z)^2}{m(1-g_i/(mz))}\right)\}\left(1 + O(1/\sqrt{n})\right) dz$$

После замены $u = g_i/z$ имеем

$$\ldots = \sqrt{\frac{L}{2\pi}} \int_0^{g_i/(g_i+1/n^d)} \sqrt{u}\exp\{-\frac{1}{2}\left(\frac{L(1-u)^2}{u} + \frac{Lu^2}{m(1-u/m)}\right)\} \left(1 + O(1/\sqrt{n})\right) du$$

$$\leq \sqrt{L/(2\pi)}\exp\{-\frac{L}{n^{2d}(g_i+n^{-d})g_i}\}$$

$$\simeq \sqrt{L}\exp(-L/n^{2d}) \simeq \sqrt{n^\beta}\exp(-n^{\beta-2d})$$

То же можно доказать для $P\{\widehat{g}_i(x_*) < g_i(x_*) - n^{-d}\}$.

Обозначим правую часть как $H(n,d,\beta)$. Так как $0 < d < \beta/2$, то $\sum_{n=1}^\infty H(n,d,\beta)$ сходится, и заключение теоремы следует из леммы Борелля-Кантелли.

## СПИСОК ЛИТЕРАТУРЫ


1. *Sigman K.* Appendix: A primer on heavy-tailed distributions//Queueing Systems. 1999. № 33, P. 261-275.
2. *Gnedenko B.V.* Sur la distribution limite du terme maximum d'une srie ale'atoire// Ann. Math. 1943. № 44. P. 423-453.
3. *Devroye L., Györfi L.* Nonparametric density estimation: The $L_1$ view. N.Y., Wiley, 1985.
4. *Wand M.P., Marron J.S., Ruppert D.* Transformations in density estimation//J. American Statist. Assoc. 1991. № 86(414). P.343-353.
5. *Hill B.M.* A simple general approach to inference about the tail of a distribution// Ann.Statist. 1975. № 3. P. 1163-1174.
6. *Pickands J.* Statistical inference using extreme order statistics//Ann. Statist. 1975. № 3. P. 119-131.
7. *Novak S.Y.* Berry-Esseen inequalities for a ratio of sums of random variables//University of Sussex: Research Report No 98/07 SMS/CSSM. 1998.
8. *Тарасенко Ф.П.* Непараметрическая статистика. Томск. Изд-во. ТГУ, 1976.
9. *Berlinet A., Vajda I., van der Meulen E.C.* About the asymptotic accuracy of Barron density estimates// IEEE Trans. Inf. Theory. 1998. № 44. P. 999-1009.
10. *Markovich N.M.* Experimental analysis of nonparametric probability density estimates and of methods for smoothing them// Autom. Remote Control. 1989. № 50. P. 941-948.
11. *Caers J., Van Dyck J.* Nonparametric tail estimation using a double bootstrap method//Computat. Statistics & Data Analysis. 1999. № 29. P. 191-211.

Потери от ошибочной классификации

| ПР | n | $J(\eta_{EB}) \cdot 10^3$ | | | |
|---|---|---|---|---|---|
| | | $Pl$ | $PlF$ | $Ke$ | $KeF$ |
| Парето | 50 | 21 | 87 | 84 | 86 |
| | 100 | 34 | 46 | 11 | 1 |
| | 500 | 24 | 43 | 3,365 | 5,029 |
| Бурр | 50 | 45 | 29 | 41 | 63 |
| | 100 | 46 | 51 | 24 | 58 |
| | 500 | 42 | 55 | 13 | 15 |
| Бурр-Парето | 50 | 133 | 279 | 201 | 242 |
| | 100 | 102 | 209 | 121 | 186 |
| | 500 | 89 | 156 | 41 | 131 |

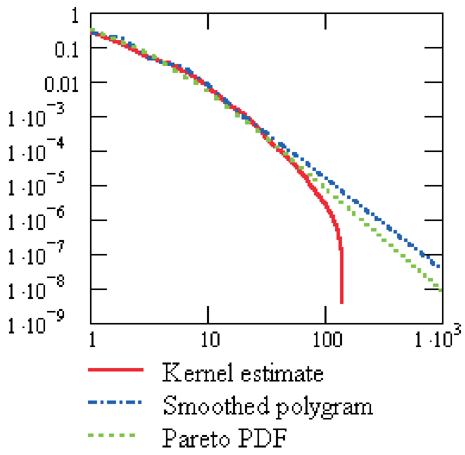


Рис. 1: Сравнение оценок хвоста:

Kernel estimate - ядерная оценка, Smoothed polygram - сглаженная полиграмма, Pareto PDF - ПР Парето.

Рис. Сравнение оценок хвоста:

Kernel estimate - ядерная оценка, Smoothed polygram - сглаженная полиграмма,

Pareto PDF - ПР Парето.