УДК 519.234.2:621.391

АППРОКСИМАЦИЯ ЭМПИРИЧЕСКИХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ПОЛИНОМАМИ БЕРНШТЕЙНА

Ф. В. Голик

Новгородский филиал Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации
173003, г. Великий Новгород, ул. Германа, 31

Статья поступила в редакцию 28 июня 2018 г.

Аннотация. Предмет исследования: аппроксимация полиномами Бернштейна (ПБ) выборочных распределений унимодального и полимодального типа, заданных на конечных интервалах. Объектом исследования являются методы оптимизации параметров аппроксимирующих ПБ. Цель работы состоит в разработке конструктивных методов оптимизации параметров ПБ и реализация этих методов средствами пакета прикладных программ MathCAD. Используемые методы: метод условного нелинейного программирования, метод наименьших квадратов, интервальное оценивание, численные расчеты в пакете прикладных программ MathCAD и имитационное компьютерное моделирование. Научная новизна работы заключается в применении метода наименьших квадратов при оптимизации весовых коэффициентов ПБ, что позволило упростить процедуру поиска решения: вместо задачи квадратичного программирования решается линейная оптимизационная задача с ограничениями. Результат: разработаны и апробированы методы поиска оптимальных параметров ПБ: 1) предложена конструктивная процедура оптимизации весовых коэффициентов; 2) проанализированы и обоснованы методы поиска «практически» оптимальных значений порядка ПБ. Практическая значимость: аппроксимация эмпирических распределений полиномами Бернштейна может успешно применяться при проектировании телекоммуникационных систем, моделировании стохастических систем и систем управления, решения задач статистической радиофизики и радиотехники и др., поскольку погрешность

аппроксимации мала, а методы расчета оптимальных параметров ПБ достаточно просто реализуются средствами пакета прикладных программ MathCAD.

Ключевые слова: эмпирическое распределение вероятностей, аппроксимация, полиномы Бернштейна, метод наименьших квадратов, компьютерное моделирование.

Abstract. The present paper is devoted to approximation of the empirical unimodal and multimodal distributions defined on a finite interval. The nonparametric approximation by Bernstein polynomials is studied. A comparative analysis of the optimality criteria is carried out. The criteria minimizing the root-mean-square error of approximation (L^2 metric), the uniform metric L^∞, the sigma-metric, the Kullback–Leibler divergence, the Anderson and Darling (AD) criterion, and the sum of the error squares are considered. Instead of AD statistics, which is used in a well-known work by Bradley C. Turnbull, Sujit K. Ghosh (2014), it is suggested to apply the criterion of the least squares method. This allowed to do without solving quadratic programming problems. Optimization of the weight coefficients of the Bernstein polynomial is reduced to solving a linear optimization problem with constraints. Stable and reliable solutions to this problem using the Mathcad software are obtained. Methods for choosing the order of Bernstein polynomial are considered. Search for the optimal order of Bernstein polynomials is carried out according to the generally accepted scheme comprising calculation of an optimal coefficients vector and assessment of approximation error under consistently increasing polynomial order values. The criteria for stopping computations are proposed, under which the order of polynomial is considered optimal. The issue concerning the necessary and sufficient accuracy of the approximation of empirical distributions is discussed. A statistical approach based on interval estimates of the histogram is proposed. The results of computer simulation are presented, which confirm the working ability and high efficiency of the proposed methods of approximation. The results of the work can be applied in solving a wide

range of scientific and practical problems related to the analysis of the distribution of empirical data.

Key words: empirical probability distribution; approximation; Bernstein polynomials; least square method; computer simulation.

1. Введение

При решении многих задач статистической радиотехники и радиофизики, теории управления и систем передачи информации возникает необходимость в аппроксимации выборочных законов распределения вероятностей некоторыми аналитическими моделями [1, 2]. Аппроксимационные задачи решаются, например, при проектировании телекоммуникационных систем аудиообмена [3, 4], систем подвижной связи и систем телекоммуникации [5], при оптимизации параметров процедуры обратимого сжатия цифровых данных [6], при создании математических моделей систем управления [7] и моделировании процессов, находящихся под воздействием случайных факторов [8].

Методы, которые используются при аппроксимации плотности распределения вероятностей (ПРВ) и функции распределения (ФР), можно разделить на две группы: параметрические и непараметрические. Параметрические используют априорные сведения о виде и/или параметрах генерального распределения. Непараметрические работают при большей неопределенности по априорной информации, вплоть до полного ее отсутствия, в связи, с чем имеют более широкую область применения.

Непараметрические методы с вычислительной точки зрения более трудоемкие, чем параметрические. До недавнего времени это сдерживало их применение. После появления компьютеров положение изменилось, и вычислительные проблемы больше не являются препятствием для их широкого использования.

В настоящее время существует два основных подхода к решению задач непараметрической аппроксимации вероятностных распределений: 1) оценки типа Розенблатта-Парзена [9] и 2) полиномиальные оценки, базирующиеся на

теореме К. Вейерштрасса, которая гласит: если $f(x)$ непрерывна на [0,1], то всякому $\varepsilon>0$ можно сопоставить многочлен $p \in \mathrm{P}_m$ для которого $\|f - p\| < \varepsilon$.

С.Н. Бернштейн в 1912 г. предложил изящное доказательство этой теоремы [10, 11]. В качестве приближающих многочленов P_m он использовал полиномы $B_m(f,x)$:

$$B_m(f,x) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) b_{m,k}(x), \text{ где } b_{m,k}(x) = C_m^k x^k (1-x)^{m-k},\ C_m^k = \frac{m!}{k!(m-k)!}.$$

Функцию $b_{n,k}(x)$ называют многочленом Бернштейна k-го порядка, а операторы $B_n(f,x)$ - полиномами Бернштейна (ПБ) порядка n функции $f(x)$. Очевидно, что $B_m(f) \in \mathrm{P}_m$. С.Н. Бернштейн доказал, что $\lim_{m\to\infty} \|f - B_m(f)\| = 0$.

Теория полиномов Бернштейна является важным разделом общей теории аппроксимации. Основные классические результаты, связанные с ПБ, представлены в [12, 13]. Возможности дальнейшего развития и новейшие результаты в области теории приближения функций ПБ приведены в [14, 15].

Аппроксимировать полиномами Бернштейна плотности распределения первым предложил Vitale [16]. Babu, Canty и Chaubey [17] исследовали асимптотические свойства полиномов и разработали методы их адаптации для оценки ФР и ПРВ, заданных на ограниченном интервале. Они также показали, что ПБ могут быть в ряде случаев предпочтительней ядерных оценок Розенблатта-Парзена.

Отечественные ученые также внесли вклад развитие теории аппроксимации вероятностных распределений полиномами Бернштейна [18, 19]. Вместе с тем, надо признать, что оценки ПРВ и ФР полиномами Бернштейна до сих пор не нашли широкого применения в прикладных исследованиях, несмотря на то что они обеспечивают высокую точность аппроксимации как унимодальных таки полимодальных распределений, заданных на конечных интервалах. Возможно, что это связано с вычислительными трудностями, обусловленными необходимостью решать

условные задачи нелинейного программирования при поиске оптимальных параметров полиномов.

Цель настоящей работы состоит в разработке конструктивных методов аппроксимации эмпирических ФР и ПРВ полиномами Бернштейна и реализации этих методов средствами пакета прикладных программ MathCAD.

2. Обоснование применимости полиномов Бернштейна для аппроксимации вероятностных распределений и эквивалентной замены их функцией плотности бэта-распределения

2.1. Аппроксимация известной функции ПРВ

Пусть $f(x)$ ПРВ случайной величины X известна и нужно найти аппроксимирующую функцию $f^*(x)$, которая удовлетворяет следующим условиям:

(i) $f^*(x) \geq 0,\ x \in [0,1]$

(ii) $\int f^*(x)dx = 1$

Запишем выражение для полинома Бернштейна порядка m-1

$$B_m(x, f) = \sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right) C_{m-1}^{k-1} x^{k-1}(1-x)^{m-k}. \quad (2)$$

Чтобы полином (2) удовлетворял условию (ii), необходимо выполнить нормировку:

$$f_m(x) = \frac{1}{\sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right)} \sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right) C_{m-1}^{k-1} x^{k-1}(1-x)^{m-k}. \quad (3)$$

Нормировка не потребуется, если вместо функции ПРВ использовать вероятности $F\left(\frac{k}{m-1}\right) - F\left(\frac{k-1}{m-1}\right)$, где $F(x) = \int f(x)dx$ - ФР случайной величины X.

Многочлен Бернштейна

$$C_{m-1}^{k-1} x^{k--1}(1-x)^{m-k} = b_{m-1,k--1}(x), \quad (4)$$

входящий в формулу (3), для упрощения численных расчетов полезно заменить функцией ПРВ бэта-распределения:

$$b(x,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ x\in[0,1].$$

При $\alpha=k$, $\beta=m-k+1$ получим:

$$f_b(x;k,m-k+1)=\frac{(m+1)!}{k!(m-k+1)!}x^{k-1}(1-x)^{m-k}, \qquad (5)$$

что в точности совпадает с (4).

Тогда

$$f_m(x)=\frac{1}{\sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right)}\sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right) f_b(x;k,m-k+1),\ x\in[0,1]. \qquad (6)$$

Итак, выражение (6) является аппроксимацией $f^*(x)$ известной плотности распределения $f(x)$. Практическое применение полученного результата ограничено, поскольку значения функции $f_m(x)$ по формуле (6) можно найти только численными методами. Но, то же самое можно сделать и для большинства функций $f(x)$, какими бы сложными выражениями они ни описывались. Гораздо интересней рассмотреть ситуацию, когда плотность распределения неизвестна, и нужно найти ее оценку по экспериментальным данным.

2.2. Аппроксимация эмпирической ПРВ

В этом случае генеральная ПРВ случайной величины *X* неизвестна и для исследования доступна лишь выборка $X_{(n)}=\{X_1,X_2,...,X_n\}$ объема *n*. Для аппроксимации ПРВ выражение (6) непригодно, поскольку множители $f\left(\frac{k-1}{m-1}\right)$ неизвестны. Заменим их весовыми коэффициентами ω_j и представим аппроксимирующую функцию в виде взвешенной суммы бэта-распределений:

$$f_N(x,\omega)=\sum_{j=1}^{N}\omega_k\, f_b(x;\, j, N-j+1),\ x\in[0,1], \tag{7}$$

где $\omega\in S^N\equiv\{(\omega_1,\omega_2,...,\omega_N)\in[0,1]^N\}$ - вектор весовых коэффициентов, удовлетворяющий следующим ограничениям:

(a) $\omega_j\ge 0,\ j=1,2,...,N$,

(b) $\sum_{j=1}^{N}\omega_j=1$.

Ограничение (a) гарантирует выполнение свойства (i), поскольку $f_b(x)$ неотрицательна для всех значений x в [0, 1]. Ограничение (b) необходимо для выполнения свойства (ii).

Применим к переменной x линейное преобразования $x\to u=\dfrac{x-a}{b-a},\ a<b$, $u\in[0,1]$. Тогда оценку (7) можно распространить на произвольный интервал конечной длины:

$$f_N(x,\omega)=\frac{1}{b-a}\sum_{j=1}^{N}\omega_j\, f_b\left(\frac{x-a}{b-a};\, j, N-j+1\right),\ x\in[a,b]. \tag{8}$$

В дальнейшем нам понадобится оценка функции распределения. Очевидно, она равна

$$F_N(x,\omega)=\sum_{j=1}^{N}\omega_j F_b\left(\frac{x-a}{b-a};\, j, N-j+1\right),\ \ x\in[a,b] \tag{9}$$

где $F_b(x;\alpha,\beta)=\int_a^x f_b(z;\alpha,\beta)dz$, $x\in[a,b]$ - ФР бэта-распределения.

3. Постановка задачи. Критерии оптимальности

Пусть дана выборка $X_{(n)}=\{X_1,X_2,...,X_n\}$ случайной величины $X\in[a,b]$ с неизвестной ФР $F(x)$. Зададим границы m полуинтервалов на интервале $[a,b]: c_0<c_1<\cdots<c_{m-1}\le c_m$, $c_0=a, c_m=b$. Найдем эмпирическую ФР:

$$F_n(c_k)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(X_i\le c_k),\ k=1,2,...,m \tag{10}$$

где $\mathrm{I}(X_i \le c_k) = \begin{cases} 1, \textit{если } X_i \le c_k \\ 0, \textit{иначе} \end{cases}$.

Гистограмму (эмпирическую ПРВ) рассчитаем по ФР (10):

$$p_n(c_k) = \frac{F_n(c_k) - F_n(c_{k-1})}{c_k - c_{k-1}},\ k = 1,2,...,m \tag{11}$$

Необходимо найти функцию $F_0(x, N^*, \omega^*) = F_{N^*}(x, \omega^*)$, аппроксимирующую наилучшим образом эмпирическую ФР (10). Для этого нужно определить неизвестные параметры функции (9): порядок N^* полинома Бернштейна (ППБ) и вектор весовых коэффициентов (ВК) ω^*. Одновременная оптимизация по двум параметрам затруднительна. Поэтому задачу обычно [20, 21] решают в два этапа: 1) при заданном N находят вектор оптимальных ВК ω_N^*; 2) повторяют пункт 1) для набора значений N и выбирают вариант N^*, при котором критерий оптимальности принимает наименьшее или приемлемое для практических расчетов значение.

Рассмотрим постановку задачи первого этапа: поиск вектора весовых коэффициентов. Требуется найти вектор ω^* ВК, доставляющий минимум критерию $W\left[F_n(x), F_N(x,\omega)\right]$, при условии, что выполняются ограничения (a) и (b):

$$\begin{cases} \omega_N^* = \arg\min\limits_{\omega \in D} W\left[F_n(x), F_N(x,\omega)\right] \\ D = \left\{\omega \,|\, \sum_{j=1}^{N} \omega_j = 1\right\} \subset S^N \end{cases} \tag{12}$$

В качестве критериев оптимальности в теории вероятностей и ее приложениях используют метрические расстояния [22]. Рассмотрим основные.

1) Общий класс L^{p}-норм, в нашей конкретной постановке задается формулой

$$W_p(N,\omega) = \left(\sum_{k=1}^{m} \left| F_n(c_k) - F_N(c_k,\omega) \right|^p \right)^{\frac{1}{p}},\ p \ge 1. \tag{13}$$

При $p = 2$ - минимизируется среднеквадратическая ошибка аппроксимации.

При $p = \infty$

$$W_{\infty}(N,\omega) = \max_{\omega \in D}\{|F_n(c_k) - F_N(c_k,\omega)|\} \quad (14)$$

- равномерная метрика [22] широко применяется в задачах асимптотической аппроксимации распределений.

2) Сигма-метрика или расстояние полной вариации. Это «одна из самых сильных метрик, используемых в теории вероятностей» [23]

$$W_{\sigma}(N,\omega) = \frac{1}{2}\sum_{k=1}^{m}\left|p_n(c_k) - p_N(c_k,\omega)\right| \quad (15)$$

где $p_n(c_k) = F_n(c_k) - F_n(c_{k-1})$, $p_N(c_k,\omega) = F_N(c_k,\omega) - F_N(c_{k-1},m)$, $k = 1,2,...,m$.

3) Дивергенция Кульбака-Лейблера для дискретных вероятностных распределений F_n и F_m вычисляется по формуле [24]

$$W_{KL}(N,\omega) = \sum_{k=1}^{m} p_n(c_k)\log\frac{p_n(c_k)}{p_N(c_k,\omega)}. \quad (16)$$

Эта мера расстояния интерпретируется как величина потерь информации при замене истинного распределения F_n на распределение F_N.

4) Критерий AD [25] является одним из самых мощных тестов. Его модификация предложена в [26]:

$$W_{AD}(N,\omega) = m\sum_{k=1}^{m}\frac{[F_n(c_k) - F_N(c_k,\omega)]^2}{(F_n(c_k)+\varepsilon_n)(1+\varepsilon_n - F_N(c_k,\omega))}, \quad (17)$$

где $\varepsilon_n = \frac{3}{8n}$ - поправка, обеспечивающая устойчивость численных расчетов [27].

5) Сумма квадратов отклонений используется в качестве меры в методе наименьших квадратов (МНК):

$$W_{MHK}(N,\omega) = \sum_{k=1}^{m}\left[F_n(c_k) - F_N(c_k,\omega)\right]^2. \quad (18)$$

Методы решения задачи (12) зависят от выбранного критерия *W*. Задачи с критериями (13 – 16) относятся к классу условных задач нелинейного программирования. Оптимизация по критерию (17) является безусловной

задачей квадратичного программирования [20]. Оптимизация по критерию МНК сводится к решению системы линейных уравнений с ограничениями.

С точки зрения вычислительной сложности предпочтение следует отдать критериям (17) и (18). В [20] приведен алгоритм оптимизации по критерию AD (17). Авторы предоставляют по запросу программный код процедуры квадратичного программирования для расчетов в статистическом пакете прикладных программ R. Мы рассмотрим решение оптимизационной задачи (12) методом наименьших квадратов, реализованном в программе MathCAD.

4. Оптимизация весовых коэффициентов. Метод наименьших квадратов

Для сокращения последующих записей обозначим: $F_b\left(\frac{c_k-a}{b-a}; j, N-j+1\right) = F_b(c_k, j)$.

Тогда

$$W_{MHK}(N,\omega) = \sum_{k=1}^{m}\left[F_n(c_k) - F_N(c_k,\omega)\right]^2 = \sum_{k=1}^{m}\left[F_n(c_k) - \sum_{j=1}^{N}\omega_j F_b(c_k, j)\right]^2 \quad (19)$$

Частные производные образуют систему N линейных уравнений:

$$\frac{\partial}{\partial\omega_i} W_{MHK}(N,\omega)\Big|_{\omega_i^*} = 0, i = 1,2,...,N. \quad (20)$$

После дифференцирования получаем

$$\sum_{j=1}^{N}\sum_{k=1}^{m}\omega_j F_b(c_k,i)F_b(c_k,j) - \sum_{k=1}^{m}F_n(c_k)F_b(c_k,i) = 0, i = 1,2,...,N.$$

Обозначим:

$$\mathrm{A} = (a_{i,j})_{i=1,j=1}^{N,N},\ a_{i,j} = \sum_{k=1}^{m}F_b(c_k,i)F_b(c_k,j),$$

$$\mathrm{B} = (b_i)_{i=1}^{N},\ b_i = \sum_{k=1}^{m}F_n(c_k)F_b(c_k,i)$$

$$x = (\omega_i)_{i=1}^{N}$$

Запишем в матричной форме систему уравнений (20):

$$\mathrm{A}x = \mathrm{B} \quad (21)$$

В результате решения системы (21) может оказаться, что найденные весовые коэффициенты ω^* не удовлетворяют ограничениям (a) и /или (b), что исключает возможность их использования при аппроксимации распределений. При решении прикладных задач нас может устроить решение не совсем «точное», но обладающее необходимыми свойствами. Запишем новое уравнение $\mathrm{A}x = \mathrm{B}^*$, отличающееся от исходного вектором свободных членов и удовлетворяющее ограничениям (a), (b). Если теперь потребовать, чтобы вектор B^* «несильно» отличался от вектора B, например, в смысле минимума суммы квадратов отклонений, то прейдем к следующей условной задаче нелинейного программирования

$$\begin{cases} \omega_N^* = \arg\min\limits_{\omega \in D} \sum\limits_{i=1}^{N} (b_i^* - b_i)^2 \\ D = \left\{ \omega \,|\, \sum\nolimits_{j=1}^{N} \omega_j = 1 \right\} \subset S^N \end{cases}, \qquad (22)$$

где b_i^* - элементы вектора B^*

Решение (22) может быть получено численными итерационными методами. Ниже мы рассмотрим порядок вычислений в программе MathCAD с использованием функции Minerr.

5. Выбор порядка *N* полинома Бернштейна

Обоснование метода выбора ППБ N (размерности вектора ω) является проблемой как теоретического, так и вычислительного характера, поскольку погрешность аппроксимации зависит от N не в меньшей степени, чем от правильности выбора весовых коэффициентов. И если задачу оптимизации коэффициентов можно считать решенной, то задача поиска оптимального ППБ до сих пор не имеет конструктивного решения. Известные подходы сводятся к последовательным расчетам вектора ω_N^* при разных N и выбору лучшего, в некотором смысле, варианта. Так [20] использует информационный критерий Akaike (AIC), байесовский информационный критерий (BIC) и критерий минимума среднеквадратической ошибки аппроксимации. Критерии рассчитываются по данным, полученным методом компьютерного

моделирования, на основании которых делаются достаточно общие рекомендации по выбору N. Аналогичный подход реализован в [21] с использованием L^p-метрики. Показано, что расстояния при p=1, p=2 и p=∞ монотонно убывают с ростом N, в связи с чем авторы рекомендуют в качестве оптимального ППБ N принимать максимально возможное, с вычислительной точки зрения, значения.

В настоящей работе также используется описанный выше подход. В качестве критериев рассчитываются расстояния в L^p-метрике при p=∞ (14), сигма-метрика (15) и дивергенция Кульбака-Лейблера (16). Кроме того, рассчитывался дополнительный критерий: евклидово расстояние $R(N)$ в пространстве коэффициентов Пирсона:

$$R(N) = \sqrt{(\beta 1_n - \beta 1_N)^2 + (\beta 2_n - \beta 2_N)^2} \qquad (23)$$

где $\beta 1_n = \frac{\mu_{3,n}^2}{\mu_{2,n}^3}, \beta 2_n = \frac{\mu_{4,n}}{\mu_{2,n}^2}$, $\beta 1_N = \frac{\mu_{3,N}^2}{\mu_{2,N}^3}, \beta 2_N = \frac{\mu_{4,N}}{\mu_{2,N}^2}$ - коэффициенты Пирсона,

$\mu_{r,n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ - центральные выборочные моменты,

$\mu_{r,N} = \int_{c_0}^{c_m} (x - m_1)^r f_N(x, \omega_N^*)dx$, $m_1 = \int_{c_0}^{c_m} x \cdot f_N(x, \omega_N^*)dx$ - центральные моменты,

рассчитанные по ПРВ (8). Критерий показывает, насколько далеко отстоит точка $(\beta 1_N, \beta 2_N)$, соответствующая аппроксимирующему распределению, относительно точки $(\beta 1_n, \beta 2_n)$ эмпирического распределения.

Описанные выше методы оптимизации ППБ N ориентированы на получение как можно более точной аппроксимации. Но возникает вопрос: насколько необходима высокая точность? Дело в том, что при оптимизации весовых коэффициентов решалась оптимизационная задача (12), которая относится к классу детерминированных нелинейных задач программирования: функции $F_n(x)$ и $F_N(x, \omega)$ считаются неслучайными. На самом же деле функция $F_n(x)$ случайна. Следовательно, условия оптимальной аппроксимации

выполняются относительно конкретной выборочной функции, а не генеральной ФР. Поэтому стремление получить очень точную оценку для «неточной», случайной функции, излишне и может быть оправдано, если не требует чрезмерных вычислительных или иных затрат. Очевидно, что нас должно устроить любое решение, для которого выполняется условие:

$$Ver[z_1 < F_N(c_k,\omega) - F_N(c_{k-1},\omega) = p_N(c_k,\omega) \le z_2] > 1-\alpha, \qquad (24)$$

где z_1, z_2 - границы доверительного интервала на уровне значимости α.

При расчете доверительного интервала для вероятности обычно полагают допустимой аппроксимацию биномиального распределения нормальным. В этом случае границы рассчитывают по формулам [28]:

$$z_{1,2} = \frac{2np_n(c_k) + t^2}{2(n+t^2)} \mp \frac{\sqrt{4nt^2(1-p_n(c_k))p_n(c_k) + t^4}}{2(n+t^2)} \qquad (25)$$

где t – квантиль нормального распределения на уровне $1-\alpha/2$.

Очевидно, что существует множество N_α значений N, для которых выполняется условие (24):

$$\mathrm{N}_\alpha = \{N \mid z_1 < p_N(c_k,\omega_N^*) \le z_2, k = 1,2,...,m\}. \qquad (26)$$

Тогда процедуру выбора оптимального значения N ППБ можно свести к проверке принадлежности N множеству N_α: $N \in \mathrm{N}_\alpha$. Имитационное моделирование показало, что оптимальное значение $N^* \ge m$ и погрешность аппроксимации убывает с ростом N, что не противоречит выводам [21]. Следовательно, в качестве алгоритма поиска N^* можно использовать пошаговый расчет оптимальных ВК $\omega_{N=m+l}^*, l = 0,1,2,...$ и проверку на каждом шаге l выполнение условия (24). Как только условие выполнится, расчет прекращается и оптимальным считают $N^* = m + l^*$, где l^* - номер шага остановки.

6. Расчет оптимальных весовых коэффициентов в программе MathCAD

Рассмотрим процедуру поиска оптимальных ВК, полагая, что известны:

1) m –количество интервалов разбиения гистограммы;

2) границы интервала $[a,b]$ распределения случайной величины. Если границы не известны, то их можно заменить оценками [20]:

$$a^* = \min(X_{(n)}) - \frac{s}{\sqrt{n}},\ b^* = \max(X_{(n)}) + \frac{s}{\sqrt{n}},\ s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2},\ \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

3) $c_0 < \cdots < c_k < \cdots \le c_m$, $c_0 = a, c_m = b$ - границы полуинтервалов гистограммы;

4) $F_n(c_k)$ - выборочная ФР;

5) $P_n(c_k) = \frac{1}{c_k - c_{k-1}}\left[F_n(c_k) - F_n(c_{k-1})\right]$ - гистограмма.

Для предварительного расчета зададим ППБ $N=m$ и введем формулы для коэффициентов матрицы A и вектора B (21):

$$i := 0...N-1 \quad j := 0...N-1$$

$$A_{i,j} := \sum_{k=1}^{m} \text{pbeta}\left(\frac{c_k - c_0}{c_m - c_0}, i+1, N-i\right) \cdot \text{pbeta}\left(\frac{c_k - c_0}{c_m - c_0}, j+1, N-j\right)$$

$$B_i := \sum_{k=1}^{m} F_n(c_k) \cdot \text{pbeta}\left(\frac{c_k - c_0}{c_m - c_0}, i+1, N-i\right)$$

Примечание: pbeta(x,s_1,s_2) – функция MathCAD - возвращает кумулятивное бета-распределение вероятностей с параметрами формы s_1 и s_2.

Блок решения системы уравнений (21) при ограничениях (a), (b):

$X_i := \frac{1}{N}$ начальное приближение

Given

$A \cdot X = B$ уравнение (20)

$X \ge 0$ ограничение (a)

$\sum X = 1$ ограничение (b)

$\omega := \text{Minerr}(X)$ решение – вектор весовых коэффициентов ω^*

$B1 := A \cdot \omega$

$$W := \left[\frac{1}{N}\sum_{i=0}^{N-1}(B1_i - B_i)^2\right]^{\frac{1}{2}}$$ среднеквадратическая ошибка

Проверка выполнения ограничений:

$\sum \omega =$ должно быть равно 1

$\min(\omega) =$ должно быть больше или равно 0

Количество весовых коэффициентов, отличных от 0

$N_0 := \sum_i \text{if}(\omega_i > 0, 1, 0)$

Примечание: Функция Minerr (var1, var2, ...) возвращает значения var1, var2 ..., которые удовлетворяют уравнениям и неравенствам в блоке уравнений.

Многократные расчеты показали, что результат решения не зависит от вектора начального приближения, что говорит об устойчивости процедуры вычислений.

Оптимальные оценки ФР и ПРВ:

$$k := 1...m$$

$$x := c_0, c_0 + 0.1...c_m$$

$$F_0(x) = \sum_{i=0}^{N-1} \omega_i \cdot pbeta\left(\frac{x - c_0}{c_m - c_0}, i+1, N-1\right) \quad \text{ФР}$$

$$f_0(x) = \frac{1}{c_m - c_0} \sum_{i=0}^{N-1} \omega_i \cdot dbeta\left(\frac{x - c_0}{c_m - c_0}, i+1, N-1\right) \quad \text{ПРВ}$$

Примечание: Функция dbeta(x, s_1, s_2) возвращает плотность вероятности для бета-распределения с параметрами формы s_1 и s_2.

7. Результаты компьютерного имитационного моделирования

Моделирование проводилось с целью 1) подтверждения работоспособности предложенной процедуры аппроксимации эмпирических распределений и 2) оценки зависимости погрешности аппроксимации от порядка *N* полинома Бернштейна.

Генерировались выборки случайных величин, распределенных нормально и по бета-распределению. При каждом виде распределения формировались векторы X1, X2 с разными или одинаковыми параметрами, объема n_1 и n_2 соответственно. Затем посредством функции stack(X1,X2) выборки объединялись, что позволило изменять вид гистограммы в широких пределах. Число интервалов разбиения гистограмм рассчитывалось по формуле Стерджеса: m=1+floor($\log_2(n_1+n_2)$), где floor(z) – функция MathCAD возвращает наибольшее целое, меньшее или равное z.

Случайные числа с нормальным распределение генерировались в программе MathCAD функцией rnorm(n, μ, σ), где n – объем выборки, μ и σ – среднее и среднеквадратическое значения соответственно. Векторы случайных

чисел с бэта-распределением формировались функцией rbeta(n, s, u), где s, u – параметры формы. Доверительные интервалы рассчитаны на уровне значимости α=0,05.

Графики рис. 1…3 подтверждают хорошее качество аппроксимации как унимодальных, так и полимодальных распределений.

На рис. 4 приведены графики зависимостей значений критериев оптимизации от *N*, рассчитанные по выборке нормальных случайных величин с параметрами *n*=500, μ=0, σ=2. Видно, что оптимальным можно считать *N*=20.

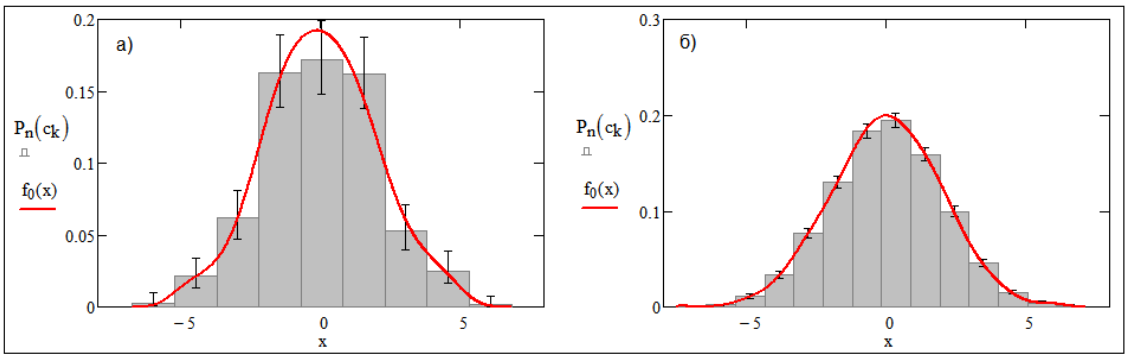

Рис. 1. Аппроксимация гистограммы, построенной по объединенной выборке нормальных случайных величин $\mu_1=\mu_2=0$, $\sigma_1=\sigma_2=2$: а) $n_1=n_2=250$, б) $n_1=n_2=5000$

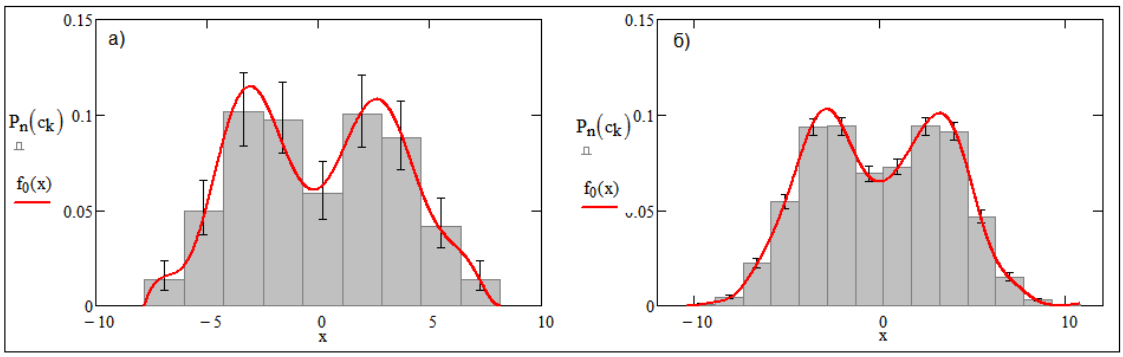

Рис. 2. Аппроксимация гистограммы, построенной по объединенной выборке нормальных случайных величин $\mu_1=-3$, $\mu_2=3$, $\sigma_1=\sigma_2=2$: а) $n_1=n_2=250$, б) $n_1=n_2=5000$

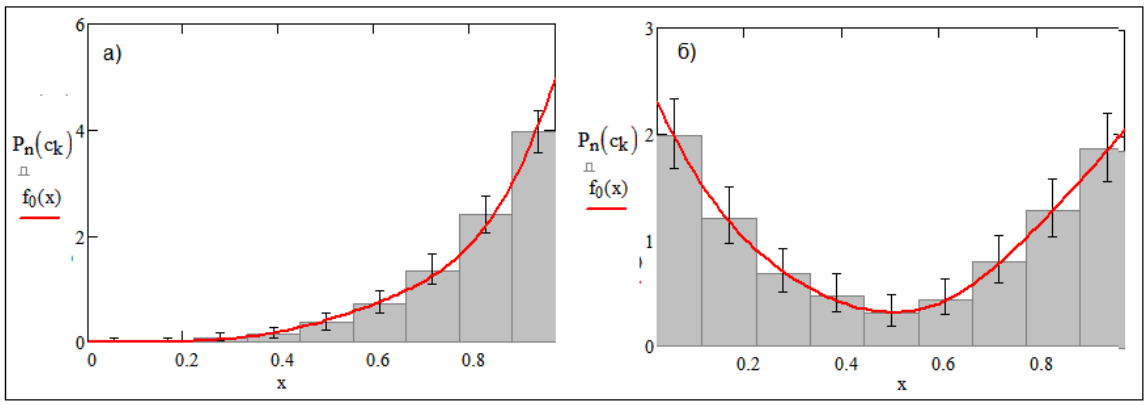

Рис. 3. Аппроксимации гистограммы, построенной по объединенной выборке случайных величин с бэта-распределением, n_1=n_2=250: а) s_1=s_2=5, u_1=u_2=1; б) s_1=5, u_1=1 и s_2=1, u_2=5

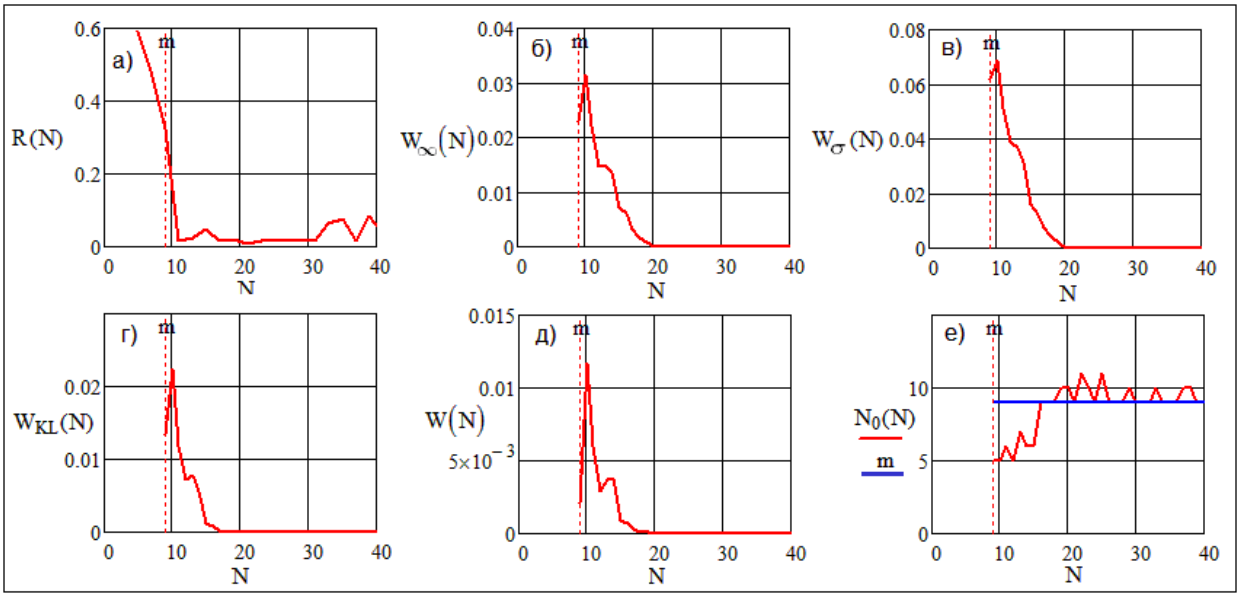

Рис. 4. Зависимости критериев оптимальности от ППБ N: а) эвклидово расстояние $R(N)$; б) расстояние равномерной метрики $W_\infty(N)$; в) расстояние в σ-метрике $W_\sigma(N)$; г) дивергенция Кульбака-Лейблера $W_{KL}(N)$; д) среднеквадратическая ошибка $W(N)$ решения уравнения (21); е) количество $N_B(N)$ ненулевых весовых коэффициентов.

Понятно, что для другой выборочной реализации результат будет иным, но общая тенденция сохранится: точность аппроксимации увеличивается с ростом N. Однако при слишком большом N появляется излишняя изрезанность

графика аппроксимирующей функции, что наглядно иллюстрируется диаграммами рис.5. Из этих же диаграмм следует, что при всех значениях N графики функции $f_0(x)$ лежат в пределах доверительных интервалов, то есть выполняется условие (24). Поэтому со статистической точки зрения все варианты равноценны. Окончательный выбор может быть чисто субъективным, основанным, например, на эстетических предпочтениях исследователя.

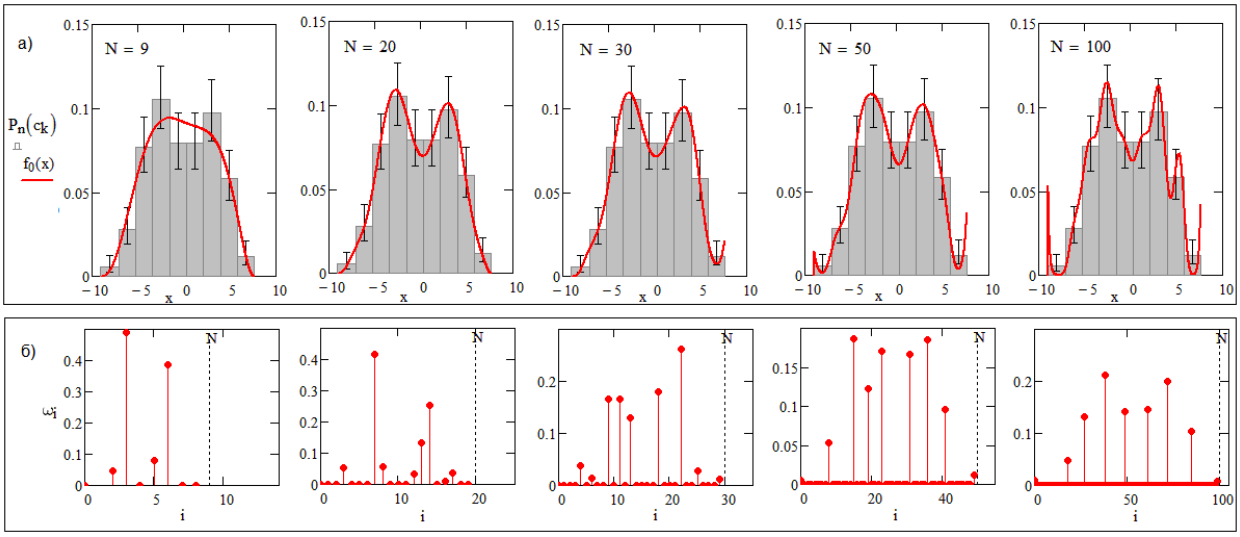

Рис. 5. Примеры аппроксимации при разных значения ППБ N: а) гистограммы и аппроксимирующие ПРВ; б) весовые коэффициенты ω_i^*.

Анализируя диаграммы е) рис. 4 и б) рис. 5 можно заметить, что число N_0 ненулевых ВК с ростом N сначала увеличивается, а затем, после превышения количества m интервалов гистограммы, колеблется в небольших пределах. Отсюда следует еще одно правило выбора ППБ: считать оптимальным ППБ значение $N = m + l, l = 0,1,2,...$ при котором число $N_0(N)$ ненулевых ВК впервые превысило m.

8. Заключение

Разработаны методы расчета оптимальных параметров полиномов Бернштейна, используемых при аппроксимации выборочных плотностей и функций распределения вероятностей.

1. Предложено оптимизировать весовые коэффициенты по критерию метода наименьших квадратов. Это позволило отказаться от решения задачи квадратичного программирования [20] и упростить процедуру расчета коэффициентов, сведя ее к решению системы линейных уравнений с ограничениями. Численные расчеты реализуются в программе MathCAD и обеспечивают получение надежного и устойчивого решения.

2. Поиск оптимального порядка полиномов Бернштейна осуществляется по общепринятой схеме, состоящей в расчете вектора оптимальных коэффициентов и погрешности аппроксимации при последовательно возрастающих значениях порядка полинома. Предложены критерии остановки вычислений, при которых порядок полинома считается оптимальным.

3. Компьютерное моделирование подтвердило работоспособность предложенных методов, их применимость при аппроксимации унимодальных и полимодальных распределений, заданных на конечных интервалах.

Результаты работы могут найти применение при решении широкого круга научных и практических задач, связанных с анализом распределений эмпирических данных.

Литература

1. Левин, Б. Р. Вероятностные модели и методы в системах связи и управления. / Б. Р. Левин, В. Шварц. — М.: Радио и связь, 1985. — 312 с.

2. Тихонов, В. И. Статистическая радиотехника / В. И. Тихонов. — М.: Радио и связь, 1982. — 624 с.

3. Кропотов Ю. А Методы оценивания моделей плотности вероятностей акустических сигналов в телекоммуникациях аудиообмена. // Системы управления, связи и безопасности №1. 2017, с. 26 – 39

4. Кропотов Ю.А. Моделирование и методы исследования акустических сигналов, шумов и помех в ситемах телекоммуникаций: монография / Ю.А. Кропотов, В.А. Ермолаев. - Берлин: Директ-Медиа, 2016. – 251 с..

5. Киреева Н.В., Чупахина Л.Р Сравнение возможностей использования различных методов аппроксимации для анализа трафика с самоподобным

распределением // Международный журнал прикладных и фундаментальных исследований №12, 2016, с. 1287- 1289.

6. Сушко Д.В. Оптимальная аппроксимация частотных вероятностей // Информационные процессы, том 18 № 1, 2018, с. 40-54.

7. Новоселов А.А. Параметризация моделей управляемых систем // Вестник государственного аэрокосмического университета им. академика М.Ф. Решетнева, № 5 (31), 2010, с. 52-56.

8. Димаки А.В., СветлаковА.А. Аппроксимация плотностей распределений случайных величин с применением ортогональных полиномов Чебышева-Эрмита // Известия Томского политехнического университета. 2006. том № 8 с. 6 -11.

9. Parzen E. On the estimation of probability density function and the mode // Ann. Math. Stat., 1962. – Vol. 33. – P. 1065-1076.

10. Бернштейн С.Н. Экстремальные свойства полиномов и наилучшее приближение непрерывных функций одной вещественной переменной. Часть I М.: Главная редакция общетехнической литературы, 1937. – с. 203

11. Бернштейн С.Н. Доказательство теоремы Вейерштрасса, основанное на теории вероятностей / Собрание сочинений. Т. 1. М. : АН СССР, 1952.

12. Никольский С.М. Приближение многочленами функций действительного переменного. / Математика в СССР за тридцать лет 1917–1947, ОГИЗ ГИТТЛ, М.-Л., 1948, с. 288–318.

13. Миракьян Г.М. Приближение полиномами С. Н. Бернштейна непрерывных функций / Докл. АН СССР, 159:5 (1964), с. 982–984.

14. Тихонов И.В., Шерстюков Б.В. Новые результаты из теории полиномов Бернштейна / Международная конференция по функциональным пространствам и теории приближения функций, посвященная 110-летию со дня рождения академика С. М. Никольского, 2015, Приближения функций и гармонический анализ, г. Москва, МИАН.

15. Тихонов И.В., Шерстюков Б.В., Петросова М.А. Полиномы Бернштейна: старое и новое. ч. 1. Исследования по математическому анализу /

Математический форум, 8, ЮМИ ВНЦ РАН и РСО-А, Владикавказ, 2014, с. 126–175.

16. Vitale, Richard A. “A Bernstein polynomial approach to density function estimation.” Statistical Inference and Related Topics 2 (1975): p. 87-99.

17. Babu, G. Jogesh, Angelo J. Canty, and Yogendra P. Chaubey. “Application of Bernstein polynomials for smooth estimation of a distribution and density function.” Journal of Statistical Planning and Inference 105.2 (2002): p. 377-392.

18. Зубов В.И. Аппроксимация локализованных вероятностных распределений // В.И. Зубов. Процессы управления и устойчивость.- СПб., 1999-с. 294-299.

19. Тукачев П.А. Непараметрические методы нахождения плотности распределения вероятностей // Процессы управления и устойчивость: Тр. XXX науч. конф.- СПб., 1999.- С. 397-401.

20. Bradley C. Turnbull, Sujit K. Ghosh. Unimodal density estimation using Bernstein polynomials. Computational Statistics & Data Analysis, 2014, Vol. 72, April 2014, pp. 13-29, DOI https://doi.org/10.1016/j.csda.2013.10.021

21. Ghosh S.K., Burns C., Prager D., Zhang L., Hui G., On nonparam etricestimation of the latent distribution for ordinal data. Computational Statistics and Data Analysis (2017), https://doi.org/10.1016/j.csda.2017.10.001

22. Золотарев В.М. Метрические расстояния в пространствах случайных величин и их распределений / Матем. сб., 1976, том 101(143), номер 3(11), с. 416–451.

23. Золотарев В. М. Современная теория суммирования независимых случайных величин. – М.: Наука. Гл. ред. физ.-мат. лит., 1986. – 416 с.

24. MacKay, David J.C. Information Theory, Inference, and Learning Algorithms. First ed. Cambridge University Press, 2003. p. 34.

25. Anderson, Theodore W., and Donald A. Darling. Test of goodness of fit.” Journal of the American Statistical Association 49.268 (1954): p. 765-769.

26. Stephens, M., 1974. Components of goodness-of-fit statistics. Annales de l’Institute Henri Poincare, Section B (Calcul des Probabilites et Statistique) 10, p. 37–54.

27. Anscombe, F., Aumann, R., 1963. A definition of subjective probability. Annals of Mathematical Statistics 34, p. 199–205.

28. Кулешов Е.Л. Интервальная оценка функции распределения вероятностей // Автометрия. 2015. Т. 51, № 2 с. 23-26.

Для цитирования:
Ф. В. Голик. Аппроксимация эмпирических распределений вероятностей полиномами Бернштейна. Журнал радиоэлектроники [электронный журнал]. 2018. № 7. Режим доступа: http://jre.cplire.ru/jre/jul18/5/text.pdf
DOI 10.30898/1684-1719.2018.7.5