Global Journal of Researches in Engineering: J General Engineering
Volume 22 Issue 3 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4596 & Print ISSN: 0975-5861

Histogram Filter with Adjustment of the Smoothing Parameter based on the Minimization of the Chi-Square Test

By Ausiannikau Andrei V.
Belarusian State University

Abstract- For the formation of adequate models of objects of statistical research, with the possible high cost of a measuring experiment or the process of obtaining data, fast and "correct" identification (recognition) of the probability distribution density (PDD) based on the construction of simple histogram estimates is required. The requirement for rapid identification can be considered equivalent to having a limited and small amount of data. The article proposes a theoretically substantiated method for constructing a histogram filter (HF), which is a linear combination of the amount of data in adjacent intervals with constant weight coefficients, which can be expressed in terms of a single coefficient k - the smoothing parameter. The estimation of the smoothing coefficient is based on the minimization of the modified chi-square test. The theorem given in the article establishes that the value of the mathematical expectation of the chi-square test, after applying the HF, decreases by k times compared to the standard mathematical expectation of the criterion with a unit inclusion function.

Keywords: *distribution density, density identification, histogram filter, smoothing factor.*

GJRE-J Classification: *DDC Code: 020.3 LCC Code: Z1006*

Strictly as per the compliance and regulations of:

Histogram Filter with Adjustment of the Smoothing Parameter based on the Minimization of the Chi-Square Test

Гистограммныйфильтрснастройкойпараметрасглаживаниянаосновеминимизациик ритерияхи-квадрат

Ausiannikau Andrei V.

Реферат- **Для формирования адекватных моделей объектов статистических исследований, при возможной высокой стоимости измерительного эксперимента или процесса получения данных, требуется быстрая и «правильная» идентификация (распознавания) плотности распределения вероятности (ПРВ) на основе построения простых гистограммных оценок. Требование быстрой идентификации можно считать эквивалентным наличию ограниченного и малого объема данных. В статье предлагается теоретически обоснованная методика построения гистограммного фильтра (ГФ), представляющего собой линейную комбинацию количества данных на соседних интервалах с постоянными весовыми коэффициентами, которые могут быть выражены через один коэффициент *k*– параметр сглаживания. Оценка коэффициента сглаживания осуществляется на основе минимизации модифицированного критерия хи-квадрат. Приведенная в статье теорема устанавливает, что значение математического ожидания критерия хи-квадрат, после применения ГФ, уменьшается в *k*раз по сравнению со стандартным математическим ожиданием критерия с единичной функцией включения. Коэффициент сглаживания определяется сложной зависимостью числа данных, параметров идентифицируемой ПРВ (информационные коэффициенты Фишера первого и второго порядка) и ГФ (количество и ширина интервалов группирования). В статье показано, что взаимосвязь между числом данных, количеством и шириной интервалов группирования является нелинейной и имеет только численное решение. Рассмотренные примеры моделирования работы ГФ характеризуют эффективность идентификации ПРВ, целесообразность его применения в научных и прикладных статистических исследованиях.**

Ключевые слова: *плотность распределения, идентификация плотности, гистограммный фильтр, коэффициент сглаживания.*

Abstract- For the formation of adequate models of objects of statistical research, with the possible high cost of a measuring experiment or the process of obtaining data, fast and “correct” identification (recognition) of the probability distribution density (PDD) based on the construction of simple histogram estimates is required. The requirement for rapid identification can be considered equivalent to having a limited and small amount of data. The article proposes a theoretically substantiated method for constructing a histogram filter (HF), which is a linear combination of the amount of data in adjacent intervals with constant weight coefficients, which can be expressed in terms of a single coefficient k - the smoothing parameter. The estimation of the smoothing coefficient is based on the minimization of the modified chi-square test. The theorem given in the article establishes that the value of the mathematical expectation of the chi-square test, after applying the HF, decreases by k times compared to the standard mathematical expectation of the criterion with a unit inclusion function. The smoothing coefficient is determined by a complex dependence of the number of data, parameters of the identified PDD (Fisher information coefficients of the first and second order) and HF (number and width of grouping intervals). The article shows that the relationship between the number of data, the number and width of grouping intervals is non-linear and has only a numerical solution. The considered examples of modeling the work of the HF characterize the effectiveness of the identification of the PDD, the expediency of its application in scientific and applied statistical research.

Keywords: distribution density, density identification, histogram filter, smoothing factor.

I. Введение

Проблематика гистограммных оценок плотности распределения вероятности (ПРВ) хорошо известна: отсутствие единых взглядов на определение числа интервалов группирования данных (ГОСТ Р 50.1.033-2001 Прикладная статистика) и сильная изрезанность гистограммы при относительно малом числе наблюдений [1,2].

Особую важность и актуальность точные гистограммные оценки закона распределения приобретают в случае требований его быстрой идентификации (распознавания), возможной высокой стоимости измерительного эксперимента или процесса получения данных. Требование быстрой идентификации (распознавания) закона распределения в данном случае можно считать эквивалентным малому объему данных.

Устранение проблем изрезанности гистограммы заключается в применении гистограммных фильтров (ГФ), например, усредняющего, медианного, гауссовского и др. [1,3-5]. В то же время, их применение эмпирически

Author: PhD, Associate Professor, Associate Professor of Belarusian State University, Minsk, Republic of Belarus. e-mail: andovs@mail.ru

интуитивно и исходит, в основном, из практических соображений. В работе предлагается теоретически обоснованная методика реализации ГФ, работающего на небольшом количестве данных, устраняющего изрезанность гистограммы, дающего «правильную» идентификацию закона распределения, ослабляющего зависимость «правильной» идентификации от числа интервалов группирования данных.

В работе развиваются идеи, предложенные в [6]. Прежде всего, предполагается отказаться от единичной функции включения данных в интервал группирования: данные могут находиться вблизи границ интервала и при изменении числа интервалов оказаться в соседнем интервале; для относительно небольшого количества данных, устранение эффекта изрезанности гистограммы может быть осуществлено сглаживанием данных на нескольких соседних интервалах.

В этом случае целесообразно заменить единичную функцию включения взвешенной функцией, учитывающей возможный вес «ошибочно» попавших в соседние интервалы данных. Физический смысл такой функции может быть охарактеризован нечеткой принадлежностью данных конкретному интервалу группирования.

Наиболее простой, с точки зренияреализации, весовой функцией удобно выбрать ступенчатую функцию.Тогдаматематическая модель гистограммного фильтра может быть представлена в виде: $u_j = \alpha_j \nu_{j-1} + k_j \nu_j + \beta_j \nu_{j+1}$, $\alpha_j + k_j + \beta_j = 1$, где ν_j – число данных попавших в *j*-тый интервал группирования, $\{\alpha_j; k_j; \beta_j\}$ – весовые коэффициенты интервалов (параметры сглаживания). В простейшем случае весовые коэффициенты являются постоянными величинами и могут быть выражены через один коэффициент *k* – параметр сглаживания.

Введение весовых коэффициентов для малых объемов данных позволяет перегруппировать эти данные так, чтобы обеспечить меньшую изрезанность гистограммы, увеличив при этом ее сглаженность и быструю идентификацию.

Вычисление параметра сглаживания, очевидно, требует некоторой априорной информации о идентифицируемой ПРВ. В работе предполагается, что такая идентификация проводится с помощью критерия согласия хи-квадрат, использование которого также основано на предположении о возможном виде идентифицируемой ПРВ. Таким образом, априорная информация является естественным и необходимым элементом построения ГФ.

Цель работы состоит в реализации гистограммного фильтра с настройкой параметра сглаживания на основе минимизации критерия хи квадрат с учетом априорной информации об идентифицируемой ПРВ.

II. Определение коэффициента сглаживания гистограммного фильтра

Пусть имеется выборка случайных данных $\{x_i\}. i = \overline{1,n}$ и определено разбиение числовой прямой на *m* непересекающихся и примыкающих друг к другу интервалов A_j , $j = \overline{1,m}$ равной длины $\Delta_x = X_{j+1} - X_j = R/m$. $X_{m+1} = x_{\max} = \max\limits_i x_i$. $X_1 = x_{\min} = \min\limits_i x_i$. где X_j – границы интервалов, $R = x_{\max} - x_{\min} = m\Delta_x$ – размах диапазона данных. Заменим обычную индикаторную функцию, используемую при стандартном способе построение гистограммы,весовой ступенчатой функцией $\mu_j(x_i)$, $0 \le \mu_j \le 1$, с областью определения $\Delta_\mu = 3\Delta_x$ и которая будет характеризовать принадлежность данных интервалу группирования A_j . При этом выбор весовой функции должен осуществляться с учетом условий нормировки:

$$\begin{cases} \sum_{t=j-1}^{j+1} \mu_{j,t} = 1, \quad j,t = \overline{2,m-1}, \\ \sum_{t}^{(t-m)(m-3)/(m-1)+(m-1)} \mu_{j,t} = 1, \quad j,t = 1,m. \end{cases} \tag{1}$$

Положимвесовые значения $\mu_{j,t}$ постоянными, не зависящими от индекса номера интервала:

$$\begin{cases} \mu_j(x) = \{k \text{ для } A_j;\ \alpha = (1-k)/2 \text{ для } A_{j-1} \text{ и } A_{j+1}\}, \quad j = \overline{2,m-1}, \\ \mu_j(x) = \{(1-\alpha) \text{ для } A_j;\ \alpha \text{ для } A_{(j-m)(m-3)/(m-1)+(m-1)}\}, \ j = 1,m, \end{cases} \tag{2}$$

где параметр *k*– коэффициент сглаживания. Условия нормировки (1) при этом выполняются автоматически. Тогда уравнение, реализующее алгоритм ГФ имеет вид

$$\begin{cases} u_j = \alpha v_{j-1} + k v_j + \alpha v_{j+1}, \quad j = \overline{2, m-1}, \\ u_j = (1-\alpha) v_j + \alpha v_{(j-m)(m-3)/(m-1)+(m-1)}, \quad j = 1, m, \\ \alpha = (1-k)/2. \end{cases} \tag{3}$$

Таким образом, задача построения адаптивного ГФ сводится к вычислению коэффициента сглаживания по информации о числе данных и априорной информации об идентифицируемой ПРВ..

Используя в качестве критерия оценки коэффициента сглаживания критерий хи-квадрат и заменив число v_j в критерии $\chi^2(v)$ на число $u_j = \alpha v_{j-1} + k v_j + \alpha v_{j+1}$ для $j = \overline{2, m-1}$ и $u_j = (1-\alpha) v_j + \alpha v_{(j-m)(m-3)/(m-1)+(m-1)}$ для $j = 1, m$, получим

$$\chi^2_{\text{гф}}(u) = \sum_{j=1}^{m} \left[u_j - n p_j \right]^2 / n p_j \to \min_k \tag{4}$$

Решение оптимизационной задачи (4) приводит к выражению для коэффициента сглаживания по выборке данных

$$\begin{aligned} k_{\text{выб}} &= 1 + 2 \left[\sum_{j=1}^{m} U_j^{\,2} / n p_j \right]^{-1} \sum_{j=1}^{m} (v_j - n p_j) U_j / n p_j = \\ &= 1 + 2 \left[\sum_{j=1}^{m} U_j^{\,2} / n p_j \right]^{1-} \sum_{j=1}^{m} v_j U_j / n p_j, \end{aligned} \tag{5}$$

где $U_j = v_{j-1} - 2 v_j + v_{j+1}$ – конечная разность второго порядка для индексов $j = \overline{2, m-1,}$ и $U_j = -v_j + v_{(j-m)(m-3)/(m-1)+(m-1)}$ для индексов $j = 1, m$; $\sum_{j=1}^{m} U_j = 0$; p_j – гипотетические вероятности.

Статистическая конкретизация формулы (5) приводит к соотношению

$$k_0 = 1 - \frac{1}{1{,}5 + 0{,}5(\Delta_x^2 I_1 + 0{,}25 \Delta_x^4 I_2) + 0{,}25 \Delta_x^4 I_2 n (m-1)^{-1}}, \tag{6}$$

где $I_1^* = \int_R \left(f'^* / f^* \right)^2 f dx$, $I_2^* = \int_R \left(f''^* / f^* \right)^2 f dx$ –информационныекоэффициенты ПРВ, эквивалентные информации Фишера первого и второго порядка [7], $f = \lim_{m \to \infty, \Delta_x \to 0} \left[p_j / \Delta_x \right]$ – гипотетическая ПРВ, $f^* = \lim_{m \to \infty, \Delta_x \to 0} \left[v_j / n \Delta_x \right]$ – эквивалент идентифицируемой ПРВ, $\int_R f dx = \gamma$ – доверительная вероятность.

Проведем упрощённое обоснование формулы (6), для чего последовательно рассмотрим компоненты, входящие в (5). Совокупность статистическо-экспериментального метода, инженерного подхода и практических представлений приводит к следующим выражениям:

$$\lim_{\substack{m\to\infty\\ \Delta_x\to 0}} \sum_{j=1}^{m} U_j / np_j = (R/m)\int_R (f''^* / f)dx,$$

$$\mathbf{M}\left(\sum_{j=1}^{m}\frac{v_j U_j}{np_j}\right) = -2(n+m-1) + \mathbf{M}\left(\sum_{j=2}^{m-1}\frac{(v_{j-1}+v_{j+1})v_j}{np_j} + \frac{v_1 U_1}{np_1} + \frac{v_m U_m}{np_m}\right) = -2(m-1) \qquad (7)$$

$$a = \left(\frac{f^*_{j-1}}{f^*_j}\right) = \left(1 - \frac{f'^*}{f^*}\Delta_x + \frac{1}{2}\frac{f''^*}{f^*}\Delta_x{}^2\right) \qquad b = \left(\frac{f^*_{j+1}}{f^*_j}\right) = \left(1 + \frac{f'^*}{f^*}\Delta_x + \frac{1}{2}\frac{f''^*}{f^*}\Delta_x{}^2\right),$$

$$\lim_{\substack{m\to\infty\\ \Delta_x\to 0}} \mathbf{M}\left(\sum_{j=1}^{m}\frac{U_j{}^2}{np_j}\right) = \lim_{\substack{m\to\infty\\ \Delta_x\to 0}} \mathbf{M}\left(\sum_{j=2}^{m-1}\frac{v_{j-1}{}^2 + 4v_j{}^2 + v_{j+1}{}^2}{np_j}\right) -$$

$$-2\lim_{\substack{m\to\infty\\ \Delta_x\to 0}} \mathbf{M}\left(\sum_{j=2}^{m-1}\frac{2v_{j-1}v_j - v_{j-1}v_{j+1} + 2v_j v_{j+1}}{np_j}\right) + \lim_{\substack{m\to\infty\\ \Delta_x\to 0}} \mathbf{M}\left(\frac{(-v_1+v_2)^2}{np_1} + \frac{(-v_m+v_{m-1})^2}{np_m}\right) \approx$$

$$\approx \left[4 + M(a^2+b^2)\right](n+m-1) + \left[2M(ab) - 8\right]n =$$

$$= 6(m-1) + 2\Delta_x{}^2 I_1^*(m-1) + 0{,}5\Delta_x{}^4 I_2^*(m-1) + \Delta_x{}^4 I_2^* n$$

Далее, подставляя выражения (7) и (8) в (5), получим непосредственно формулу (6).

Формула (6) позволяет сделать ряд важных выводов.

Во-первых, при неограниченно возрастающем числе данных $n \to \infty$, очевидно, коэффициент сглаживания должен стремиться к единице, что и следует из формулы (6). В этом случае целесообразность применения ГФ исчезает. При значении компоненты знаменателя $\delta = 0{,}5(\Delta_x^2 I_1^* + 0{,}25\Delta_x^4 I_2^*) + 0{,}25\Delta_x^4 I_2^* n(m-1)^{-1}$ меньше единицы или $\delta \to 0$ коэффициент сглаживания стремиться к 1/3. Такое значение коэффициента сглаживания отвечает случаю сильной изрезанности гистограммы, возможно вследствие неправильно выбранного (относительно большого) значения количества интервалов при относительно небольшом количестве данных. ГФ, в этом случае, преобразуется в обычный усредняющий фильтр. Таким образом, диапазон изменения значений коэффициента сглаживания лежит в пределах $1/3 \le k \le 1$.

Во-вторых, подставляя значение коэффициента сглаживания (5) в формулу (4) для критерия согласия хи-квадрат получаемвыражение $\chi^2_{\Gamma\Phi}(u) = \chi^2(v) - \left(\sum_{j=1}^{m} U_j{}^2 / np_j\right)^{-1}\left(\sum_{j=1}^{m} v_j U_j / np_j\right)^2$, из которого следует соотношение между математическими ожиданиями критерия хи-квадрат: $\mathbf{M}\left(\chi^2_{\text{гф}}\right) = k\mathbf{M}\left(\chi^2\right)$, $\mathbf{M}\left(\chi^2\right) = m-1$.

Таким образом, применение ГФ позволяет уменьшить значение стандартного критерия согласия в *k* раз. Соотношение входящих в коэффициент параметров характеризует целесообразность применения и эффективность идентификации ГФ. Так, при небольших значениях компоненты знаменателя $\delta < 1$, значение критерия хи-квадрат после применения фильтра практически уменьшается в 3 раза, в противном случае при $n \to \infty$ ($k \to 1$) значение критерия хи-квадрат стремится к стандартному $\mathbf{M}\left(\chi^2_{\text{гф}}\right) \to (m-1)$ и применение ГФ нецелесообразно.

Следовательно, эффективность ГФ можно оценивать величиной обратной значению коэффициента сглаживания: $Э_{гф} = k^{-1}$.

В-третьих, предположив высокую апостериорную точность оценки ПРВ, плотность f^* винформационныхкоэффициентах формально можно заменить гипотетической f и, следовательно, величины I_1^* и I_2^* будут совпадать с вычисляемыми теоретически информациями Фишера первого и второго порядка $I_1^* = I_{1\gamma}$, $I_2^* = I_{2\gamma}$ для диапазона *R*(таблица 1). В этом случае, формула (6) становятся полностью определенной. Замечаем, что вычисление коэффициентов $I_{1\gamma}, I_{2\gamma}$ требует существование первой и второй производной ПРВ. Однако, если такой производной не существует, следует воспользоваться инженерными соображениями практической реализации. В частности, для равномерной ПРВ можно принять $f' = 0$, $f'' = 0$ и, следовательно, $I_{1\gamma} = I_{2\gamma} = 0$. Тогда численное значение коэффициента сглаживания будет равно 1/3 и ГФ преобразуется в обычный усредняющий фильтр, что в случае идентифицируемой равномерной ПРВ вполне очевидно.

В таблице 1 приведены также теоретические значения информационных коэффициентов I_1, I_2, вычисленных по области определения аргумента ПРВ.

Таблица 1: Значения информационных коэффициентов

	№1. Гауссовская плотность: $e^{-\frac{x^2}{2D}} / \sqrt{2\pi D}$
I_1	D^{-1}
$I_{1\gamma}$	$\frac{Erf\left[\frac{Erf^{-1}(\gamma)}{\sqrt{D}}\right]}{D} - \frac{2e^{-\frac{Erf^{-1}(\gamma)^2}{D}} Erf^{-1}(\gamma)}{D^{3/2}\sqrt{\pi}}$
I_2	D^{-2}
$I_{2\gamma}$	$\frac{2Erf\left[\frac{Erf^{-1}(\gamma)}{\sqrt{D}}\right]}{D^2} - \frac{2\ e^{-\frac{Erf^{-1}(\gamma)^2}{D}} Erf^{-1}(\gamma)\left(D + 2Erf^{-1}(\gamma)^2\right)}{D^{7/2}\sqrt{\pi}}$
	№2. Лапласовская плотность: $\lambda e^{-\lambda\lvert x\rvert} / 2$
I_1	λ^2
$I_{1\gamma}$	$-\left(-1+(1-\gamma)^{\lambda}\right)\lambda^2$
I_2	λ^4
$I_{2\gamma}$	$-\left(-1+(1-\gamma)^{\lambda}\right)\lambda^4$

Продолжение Таблицы 1. Значения информационных коэффициентов

f	№3. Логистическая плотность: $\alpha \operatorname{sech}^2(\alpha x)/2$
I_1	$4\alpha^2/3$
$I_{1\gamma}$	$\frac{4}{3}\alpha^2 \tanh\left[\frac{2\sqrt{3}\alpha ArcTanh[\gamma]}{\pi}\right]^3$
I_2	$16\alpha^4/5$
$I_{2\gamma}$	$\frac{1}{5}\alpha^4 Sech\left[\frac{2\sqrt{3}\alpha ArcTanh[\gamma]}{\pi}\right]^5 \times \left(30\sinh\left[\frac{2\sqrt{3}\alpha ArcTanh[\gamma]}{\pi}\right] - 5\sinh\left[\frac{6\sqrt{3}\alpha ArcTanh[\gamma]}{\pi}\right] + \sinh\left[\frac{10\sqrt{3}\alpha ArcTanh[\gamma]}{\pi}\right]\right)$
	№4. Коши плотность: $s(s^2+x^2)^{-1}/\pi$
I_1	$0{,}5s^{-2}$
$I_{1\gamma}$	$\frac{4\arctan\left[\tan\left[\frac{\pi\gamma}{2}\right]\right] - \sin[2\pi\gamma]}{4\pi s^2}$
I_2	s^{-4}
$I_{2\gamma}$	$\frac{48\arctan\left[\tan\left[\frac{\pi\gamma}{2}\right]\right] + 24\sin[\pi\gamma] + 6\sin[2\pi\gamma] + 8\sin[3\pi\gamma] + 3\sin[4\pi\gamma]}{24\pi s^4}$

III. Соотношения между числом данных, числом интервалов группирования и их шириной

Показатель эффективности ГФ может быть использован в целях нахождения оптимального соотношения между числом данных, числом интервалов группирования и шириной этих интервалов. Зафиксировав некоторое желаемое значение эффективности фильтра $Э^0_{гф}$, на основании (6) получим

$$\Delta_x^2 I_1 + 0{,}25\Delta_x^4 I_2 + 0{,}5\Delta_x^4 I_2 n(m-1)^{-1} = K^0, \quad \text{где } \Delta_x = R/m, K^0 = (Э^0_{гф} - 1)^{-1}(3 - Э^0_{гф}). Э^0_{гф} = k_0^{-1} \tag{9}$$

В приведенную формулувходят точные теоретические значения информационных коэффициентов по всей области определения аргумента, однако для практических расчетов следует использовать значения с учетом реального диапазона данных *R*, т.е. $I_{1\gamma}$ и $I_{2\gamma}$.

Уравнения (9) нелинейное, требующие численных методов решения в общем случае.

Анализ уравнения (9) показываетсложную взаимозависимость параметров идентифицируемой ПРВ и ГФ. Это, в частности, объясняет большое количество работ, посвященных тематике взаимосвязи этих параметров и рассматривающих проблему их выбора с тех или иных позиций.В этом смысле данная работа расширят подход [9,10], учитывая информацию Фишера первого и второго порядка относительно идентифицируемой ПРВ.

В некоторых частных случаях, с целью получения простых аналитических выражений взаимосвязи параметров ПРВ и ГФ, уравнение (9) возможно упростить.

Во-первых, если выполняется соотношение между информационными коэффициентами $I_2 = cI_1^2$, $c = \text{const}$ и число интервалов группирования значительно больше единицы, то обозначив $x = \Delta_x^2 I_1$, уравнения (9) можно записать в виде

$$a(n)x^{5/2} + bx^2 + x - K^0 = 0, b = 0,25c,\ a(n) = 2b(R\sqrt{I_1})^{-1}n \tag{10}$$

Уравнение (10) компактно и позволяет получать семейства зависимостей связывающих параметры идентифицируемой ПРВ и ГФ.

Во-вторых, анализируя вклад компонент знаменателя (6) в коэффициент сглаживания замечаем, что с увеличением числа данных возрастает влияние компоненты, содержащей параметр *n*. Тогда уравнение (9) можем преобразовать к виду

$$\frac{n}{m^4(m-1)} = \frac{2K^0}{R^4 I_2} \tag{11}$$

Для случая $m >> 1$ возможно приближенное аналитическое решение нелинейного уравнения (11): $m = \sqrt[5]{nR^4 I_2 / 2K^0}$. Последняя формула близка к выражениям вида $m \sim n^{0,2}$, приведенным в [1,8] с коэффициентом пропорциональности зависящим от параметров ПРВ и априорных установок ГФ по его эффективности $Э_{гф}^0$.

В-третьих, перераспределение части данных между соседними интервалами, не только уменьшает изрезанность гистограммы, но и способствует ослаблению требований к выбору числа интервалов группирования. Фиксируя некоторую нижнюю границу значения эффективности ГФ ($Э_{гф}^0$), можем определить значение числа интервалов группирования по формулам (11) из условия $m^4(m-1) \geq 0,5nR^4 I_2 / K^0$.

IV. Рекомендации по реализации гистограммного фильтра

Полученные теоретические результаты показывают целесообразность применения ГФ с целью эффективной и быстрой (на малых объемах данных) идентификации изменяющихся законов распределения в описательной статистике, при обработке гистограмм изображений. Программная реализация ГФ легко встраивается в существующие открытые алгоритмы построения гистограмм, например, в функции hist, histfit платформы Matlab.

Структура алгоритма идентификации (распознавания) ПРВ следующая.

1. Получение выборки данных, объемом *n*, определение размаха выборки *R*.
2. На оснований предположений о идентифицируемой ПРВ, вычисление информационных коэффициентов $I_{1\gamma}$ и $I_{2\gamma}$.
3. На основании выбранного числа интервалов группирования данных, размаха выборки, объема данных, информационных коэффициентов вычисляется значение коэффициента сглаживания (6).
4. Применение ГФ (3).
5. Вычисление критерия согласия хи-квадрат. На основании заданного уровня значимости принятие решения о идентификации.

Заметим, процедуру идентификации ПРВ можно сделать многоканальной, где каждый канал будет ориентирован на определенный заранее возможный вид ПРВ. Принятие решения о идентификации в этом случае может быть реализовано различными методами, например, простым или взвешенным голосованием.

V. Моделирование гистограммного фильтра

На рис.1 ($n = 100$, $m = 9$, количество серий экспериментов 100) приведены примеры работы ГФ для ПРВ: нормальной (рис.1,а, $k_{выб} = 0.64$, $Э_{выб} = 1,56$), логистической (рис.1,б, $k_{выб} = 0.74$, $Э_{выб} = 1,35$), экспоненциальной (рис.1,в, $k_{выб} = 0.53$, $Э_{выб} = 1,89$), равномерной (рис.1,г, $k_{выб} = 0,36$, $Э_{выб} = 2,78$). На рис.1 верхняя часть соответствует обычной гистограмме, нижняя – результат обработки ГФ. Во всех

приведенных на рис.1 случаях выполняется соотношение $\chi^2_{гф} < \chi^2_{кр} \leq \chi^2$, где $\chi^2_{кр}$ – критическое значения критерия согласия при заданном уровне значимости (0,05). Результаты моделирования, наглядно подтверждают идею применения ГФ. Эффективность применения ГФ на отмеченных плотностях указывает на существенное его превосходство перед стандартной гистограммной оценкой.

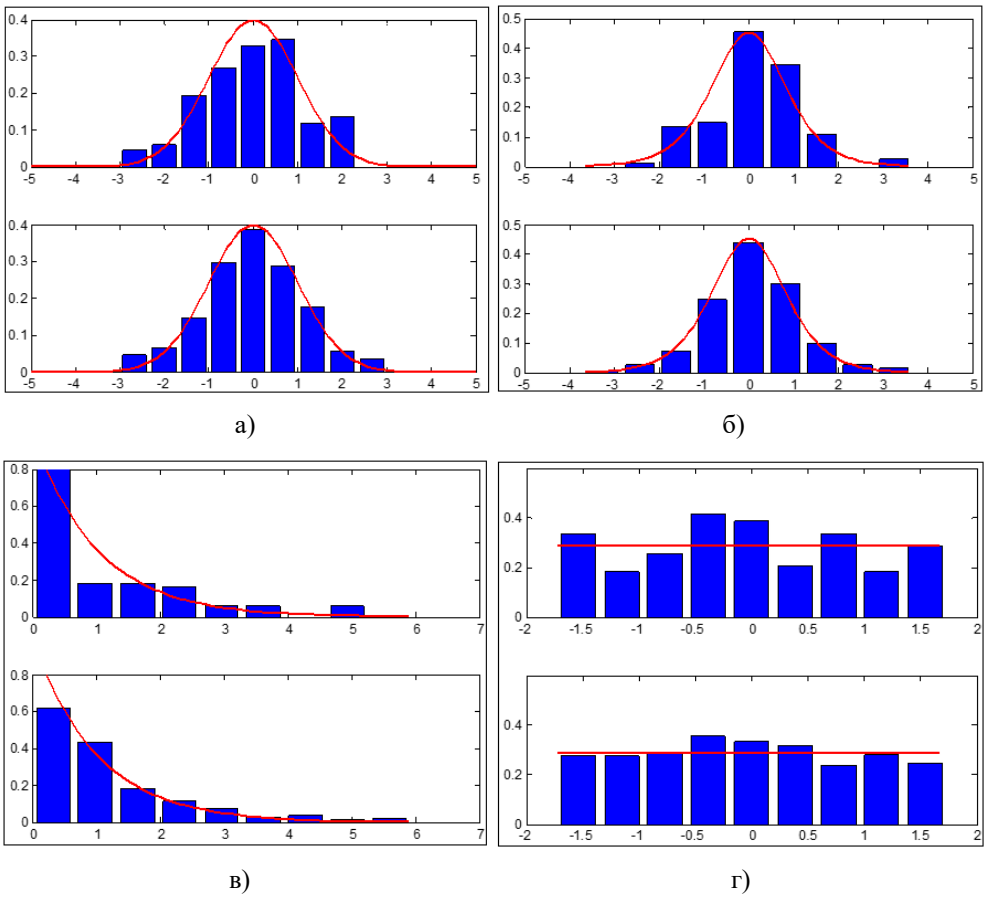


Рис.1: Результаты работы ГФ

В табл.1 (количество серий экспериментов 100) содержаться некоторые результаты моделирования работы ГФ (4) в сопоставлении с теоретическими результатами, полученными на основе формулы (6) для трех ПРВ: нормальной, логистической, лапласовской. На рис.2 приведены зависимости коэффициента сглаживания (6) от количества интервалов группирования для двух ПРВ: нормальной (кривая 1, $n=100$; кривая 2, $n=500$) и лапласовской (кривая 3, $n=100$; кривая 4, $n=500$). Численные (таблица 2) и графические результаты (рис.2) позволяют сделать вывод о том, что значение коэффициента сглаживания нелинейно уменьшается с уменьшением объема данных. Это объясняется тем, что при уменьшающемся объёме данных увеличивается изрезанность обычной гистограммной оценки ПРВ и, следовательно, требуется ее большая сглаженность, стремящаяся к равномерному (усредняющему) сглаживанию ($k \to 1/3$) при одном и том же числе интервалов группирования.

На рис.3 (количество серий экспериментов 100) на основе формулы (9) (кривые 1.1-1.3), приведены зависимости меду числом данных и числом интервалов их группирования для двух плотностей: гауссовской – рис.3,а и логистической – рис.3,б для различных значений коэффициента эффективности: $Э^0_{гф}=1,5$ ($k^0=0,(6)$) – кривая 1.1, $Э^0_{гф}=1,3$ ($k^0=0,77$) – кривая 1.2, $Э^0_{гф}=1,1$ ($k^0=0,91$) – кривая 1.3. На тех же рисунках приведены для сопоставления стандартно используемые формулы Старджеса $m=1+\log_2 n$ (кривая 2) и формулы, приведенной в [1,8] $m=C(E_x)n^{0,4}$ (кривая 3), где $C(E_x)=(E_x+4,5)/6$, E_x– коэффициент эксцесса ПРВ.

Таблица 2: Коэффициенты сглаживания и эффективность ГФ, $n = 100$

№	m	Нормальная ПРВ, $A(f) = 1,73$			Логистическая ПРВ, $A(f) = 2,14$			Лапласовская ПРВ, $A(f) = 0,99$		
		$k_{выб}$	k_0	$Э_{гф}$	$k_{выб}$	k_0	$Э_{гф}$	$k_{выб}$	k_0	$Э_{гф}$
1	5	0,8	0,96	1,04	0,96	0,98	1,02	0,98	0,99	1,01
2	7	0,81	0,81	1,23	0,90	0,92	1,09	0,94	0,97	1,03
3	9	0,66	0,61	1,64	0,75	0,77	1,30	0,86	0,91	1,10

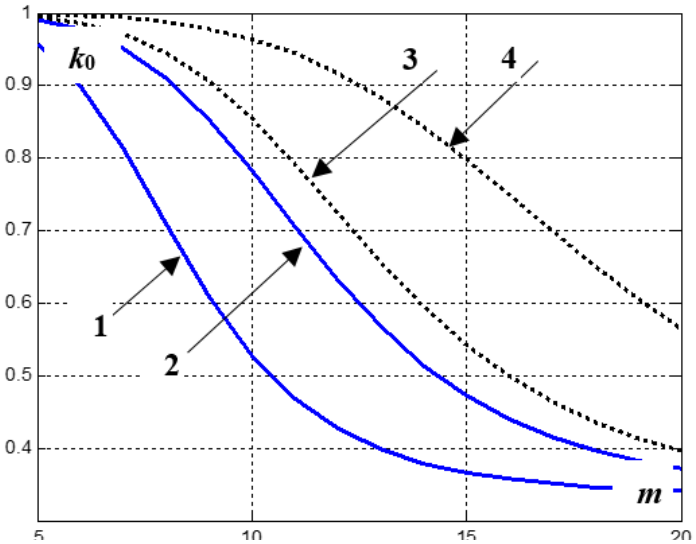
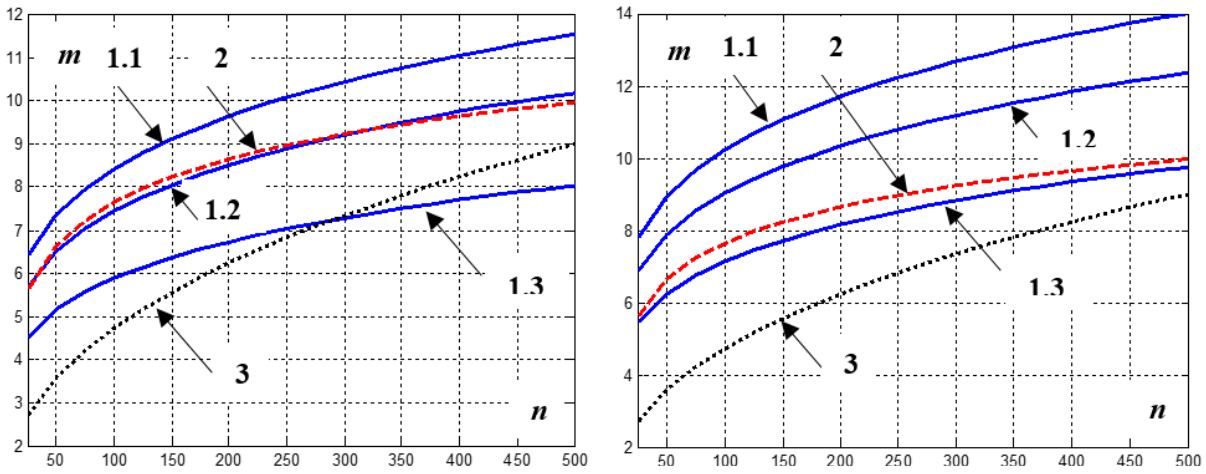


Рис.2: Коэффициенты сглаживания ГФ



а) б)

Рис.3: Зависимость числа интервалов группирования от объема данных

VI. Выводы

Рассмотренный в статье ГФ (3) с настройкой параметра сглаживания может быть эффективно применен в задачах идентификации (распознавания) ПРВ для малых объемов данных с учетом имеющейся в наличии априорной информации о предполагаемой ПРВ.

Установлено соотношение между математическими ожиданиями согласия критерия хи-квадрат при стандартном подходе построения гистограммной оценки и с использованием ГФ. Такое соотношение определяется коэффициентом сглаживания (5), (6). Численное значение коэффициента сглаживания зависит от параметров: объема данных, числа интервалов группирования, информационных характеристик ПРВ (таблица 1). Зависимость коэффициента сглаживания от указанных параметров позволяет определить взаимосвязь между количеством

интервалов группирования данных и их объемом. Эта зависимость нелинейная, не имеющая аналитического решения.

Гистограммный фильтр является простым для реализации инструментом, который легко может быть встроен в любой открытый алгоритм идентификации (распознавания) ПРВ.

Список литературы

1. Орлов Ю,Н, Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда, Препринты ИПМ им, М,В,Келдыша, 2013, № 14, 26с, URL: http://library,keldysh,ru/preprint,asp?id. 2013-14.
2. Chong Gu, Yongho Jeon and Yi Lin, Nonparametric density estimation in high-dimensions, Statistica Sinica 23 (2013), 1131-1153.
3. Devroye L. and Gyorfi L. Nonparametric Density Estimation: The L1 View. John Wiley Sons, New York. 1985/
4. Solomon C,J,, Breckon T,P, Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab, — Wiley-Blackwell, 2010, ISBN 978-0470844731, DOI:10.1002/9780470689776.
5. Gonzalez Rafael, Digital image processing, New York, NY: Pearson, 2018, ISBN 978-0-13-335672-4, OCLC 966609831.
6. Овсянников А.В., Козел В.М. Фильтрация гистограммной оценки плотности вероятности на основе нечеткой принадлежности данных интервалу группирования // Научный журнал «Доклады БГУИР» Том 19, № 4 (2021). С.13-21. https://doklady.bsuir.by/jour/article/view/3103.
7. Овсянников А.В. Статистические неравенства в сверхрегулярных статистических экспериментах теории оценивания // ВестінацыянальнайакадэміінавукБеларусі. №2, 2009. Сер фіз-мат. навук. С.106-110.
8. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов наблюдений. – Л.: Энергоатомиздат. Ленингр. отд-ние, 1991. – 304с.
9. Freedman D. and Diaconis P. On the histogram as a density estimatorr: L2 theory // Zeitschrift fur Wahrscheinlichkeitstheorie verw. Gebiete. 1981. Vol. 57. Pp. 453-476.
10. Scott D.W. On optimal and data-based histograms // Biometrika. 1979. V.66. P. 605-610.