Оценка плотности в условиях мультипликативного шума

Е. С. ФИЛАТОВА
Московский государственный университет им. М. В. Ломоносова
e-mail: evdokiapotapova@yandex.ru

А. В. ШКЛЯЕВ
Московский государственный университет им. М. В. Ломоносова
e-mail: ashklyaev@gmail.com

УДК 519.234.2

Ключевые слова: оценки плотности распределения, метод Фурье, зашумлённые данные.

Аннотация

В работе рассматривается задача оценивания плотности распределения величин, сосредоточенных на отрезке. Рассматриваемая выборка представляет собой выборку из исходного распределения, зашумлённую независимыми мультипликативными величинами. Используемый метод основывается на оценке коэффициентов разложения плотности в ряд Фурье.

Abstract

E. S. Filatova, A. V. Shklyaev, Estimation of probability density function in the case of multiplicative noise, Fundamentalnaya i prikladnaya matematika, vol. 23 (2020), no. 1, pp. 259—267.

In the present paper, we consider the problem of probability density function estimation. Our data has multiplicative noise; therefore, we cannot use direct methods. Our method is based on estimation of coefficients of the Fourier transform for the density function.

1. Введение

В работе рассматривается задача, возникающая при исследовании состава примесей в жидкости. Наблюдаемыми объектами в опытах выступают движущиеся частицы примесей в жидкости, имеющие размеры порядка 10^{-8} *м*. Жидкость просвечивают лучом лазера, попадающие в луч частицы фиксируются камерой.

В каждый момент времени мы видим на диафрагме изображение, для каждой частицы совокупность таких изображений составляет проекцию движения

Фундаментальная и прикладная математика, 2020, том 23, № 1, с. 259—267.

частицы на площадь камеры. Размер изображений напрямую не связан с размерами частиц.

Наблюдая за движением изображений на диафрагме, мы можем оценить векторы перемещений частиц в плоскости камеры за одинаковые промежутки времени. Для каждой отдельно взятой частицы эти перемещения образуют двумерное броуновское движение с нулевым сносом и дисперсией $\sigma^2 = c/d$, где c — некоторая константа, d — размер частицы.

Задача состоит в оценке плотности распределения линейных размеров d частиц примесей на основе данных камеры. Мы будем рассматривать равносильную задачу оценивания плотности распределения σ^2.

Рассмотрим n случайно выбранных наблюдаемых частиц $E_1, \ldots, E_n$. Набор дисперсий броуновского движения для этих частиц обозначим $\sigma_1^2, \ldots, \sigma_n^2$. Для частицы с номером i мы имеем два k-мерных вектора перемещений $A_1^i, \ldots, A_k^i$ и $A_{k+1}^i, \ldots, A_{2k}^i$ по осям абсцисс и ординат соответственно. $A_1^i, \ldots, A_{2k}^i$ условно независимы при условии σ_i^2 с условным распределением $\mathcal{N}(0, \sigma_i^2)$. Будем рассматривать вместо выборки $A_1^i, \ldots, A_{2k}^i$ достаточную статистику

$$\sum_{j=1}^{2k} (A_j^i)^2.$$

После деления на σ_i^2 она будет иметь распределение χ_{2k}^2. Таким образом,

$$\sum_{j=1}^{2k} (A_j^i)^2 = \sigma_i^2 \cdot Y_i,$$

где $Y_i \sim \chi_{2k}^2$, $i = 1, \ldots, n$, Y_i независимы и не зависят от дисперсии σ_i^2.

В данной работе рассмотрен частный случай, когда длина траектории k одна и та же для всех частиц.

Пусть

$$Z_i = \sum_{j=1}^{2k} (A_j^i)^2$$

и $X_i = \sigma_i^2$. Тогда задачу можно переформулировать следующим образом: $X_1, \ldots, X_n$ — независимые одинаково распределённые случайные величины с неизвестным распределением, $Y_1, \ldots, Y_n$ — независимые от них независимые одинаково распределённые случайные величины, имеющие распределение χ_{2k}^2, а $Z_i = X_i Y_i$, $i = 1, \ldots, n$, — наблюдаемые величины. Наша задача состоит в том, чтобы по наблюдениям Z_i, $i = 1, \ldots, n$, оценить плотность распределения величины X_1.

Для оценивания плотности распределения выборки используется ряд методов, наиболее популярным из которых является ядерное оценивание [3, гл. 3; 4, гл. 6.3]. Однако для восстановления плотности по зашумлённым данным более удобен метод Фурье оценки плотности распределения, основанный на оценке коэффициентов её разложения в ряд Фурье [3, гл. 2]. Полученные оценки

необходимо привести к виду вероятностной плотности, для чего используется подход работы [1]. Методы выбора количества используемых коэффициентов ряда рассматриваются в работе [2], мы используем модификацию одного из них.

В работе найдены несмещённые состоятельные оценки для коэффициентов разложения в ряд Фурье плотности величины X, построенные по величинам Z_i.

Работа построена следующим образом: в разделе 2.1 описан метод Фурье, модифицированный для зашумлённых данных. Раздел 2.2 посвящён изложению метода приведения оценки плотности распределения к вероятностному виду. В разделе 2.3 даны оценки погрешности используемого метода. В разделе 2.4 изложен метод оценки коэффициентов сглаживания. Раздел 2.5 описывает полученные практические результаты.

2. Основной метод

2.1. Метод Фурье для зашумлённых данных

Для удобства рассмотрим метод оценки плотности, сосредоточенной на отрезке $[-\pi, \pi]$; в случае произвольного отрезка оценки получаются из представленных линейной заменой.

Пусть f — функция из пространства $L^2[-\pi, \pi]$. В этом случае ряд Фурье функции f сходится к ней почти всюду:

$$f(x) = \sum_{\nu=0}^{\infty} f_\nu^{\cos} \cos \nu x + f_\nu^{\sin} \sin \nu x,$$

где

$$\begin{cases} f_\nu^{\cos} = \dfrac{1}{\pi} \displaystyle\int_{-\pi}^{\pi} f(x) \cos(\nu x)\, dx, \quad \nu = 1, 2, \dots, \\ f_0 = \dfrac{1}{2\pi} \displaystyle\int_{-\pi}^{\pi} f(x)\, dx, \\ f_\nu^{\sin} = \dfrac{1}{\pi} \displaystyle\int_{-\pi}^{\pi} f(x) \sin(\nu x)\, dx, \quad \nu = 1, 2, \dots . \end{cases} \tag{1}$$

Рассмотрим случайную величину X с плотностью распределения f с носителем на отрезке $[-\pi, \pi]$. Тогда (1) можно записать в виде

$$\begin{cases} f_\nu^{\cos} = \dfrac{1}{\pi} \mathbf{E} \cos \nu X, \quad \nu = 1, 2, \dots, \\ f_0 = \dfrac{1}{2\pi}, \\ f_\nu^{\sin} = \dfrac{1}{\pi} \mathbf{E} \sin \nu X, \quad \nu = 1, 2, \dots . \end{cases} \tag{2}$$

Построим функции $g_\nu^{\cos}(t)$ и $g_\nu^{\sin}(t)$, такие что

$$\begin{cases} \mathbf{E}\, g_\nu^{\cos}(Z) = \mathbf{E}\cos\nu X, & \nu = 1, 2, \dots, \\ \mathbf{E}\, g_\nu^{\sin}(Z) = \mathbf{E}\sin\nu X, & \nu = 1, 2, \dots . \end{cases} \tag{3}$$

Выразим из первого соотношения в (3) функцию $g_\nu^{\cos}(t)$, функция $g_\nu^{\sin}(t)$ может быть получена аналогичным образом из второго соотношения. Заметим, что

$$\mathbf{E}\, g_\nu^{\cos}(Z) = \int\limits_{-\infty}^{\infty}\int\limits_{0}^{\infty} g_\nu^{\cos}(xy) f_X(x) f_Y(y)\, dx\, dy = \\ = \mathbf{E}\cos(\nu X) = \int\limits_{-\infty}^{\infty} \cos(\nu x) f_X(x)\, dx, \quad \nu = 1, 2, \dots .$$

Значит, достаточно выполнения тождеств

$$\int\limits_{0}^{\infty} g_\nu^{\cos}(xy) f_Y(y)\, dy = \cos(\nu x)$$

при всех $\nu \geqslant 1$, x. Производя замену $xy = t$ и используя явное выражение для плотности $f_Y(y)$, получаем достаточное условие

$$\frac{1}{x}\int\limits_{0}^{\infty} g_\nu^{\cos}(t) \frac{(1/2)^k}{\Gamma(k)} \left(\frac{t}{x}\right)^{k-1} e^{-t/(2x)}\, dt = \cos(\nu x), \quad \nu = 1, 2, \dots, \tag{4}$$

где $2k$ — количество степеней свободы хи-квадрат распределения.

Предположим, что функция $g_\nu^{\cos}(t)$ представима рядом Тейлора:

$$g_\nu^{\cos}(t) = \sum_{j=0}^{\infty} a_j(\nu) \frac{t^j}{j!}, \quad \nu \geqslant 1.$$

Используя разложение в ряд Тейлора правой части равенства (4), имеем

$$\begin{aligned} &a_{2j+1}(\nu) = 0, && \nu = 1, 2, \dots, \\ &a_{2j}(\nu) = (-1)^j \left(\frac{\nu}{2}\right)^{2j} \frac{\Gamma(k)}{\Gamma(k+2j)}, && \nu = 1, 2, \dots . \end{aligned}$$

Аналогично получаются коэффициенты разложения $g_\nu^{\sin}(t)$ и формулы для них:

$$\begin{aligned} &b_{2j}(\nu) = 0, && \nu = 1, 2, \dots, \\ &b_{2j+1}(\nu) = (-1)^{j-1} \left(\frac{\nu}{2}\right)^{2j} \frac{\Gamma(k)}{\Gamma(k+2j)}, && \nu = 1, 2, \dots . \end{aligned}$$

Поскольку $Z_1, \dots, Z_n$ — независимые одинаково распределённые случайные величины, оценки

$$\begin{cases} \hat{f}_\nu^{\cos} = \dfrac{1}{\pi n} \displaystyle\sum_{m=1}^{n} g_\nu^{\cos}(Z_m), \quad \nu = 1, 2, \dots, \\ \hat{f}_\nu^{\sin} = \dfrac{1}{\pi n} \displaystyle\sum_{m=1}^{n} g_\nu^{\sin}(Z_m), \quad \nu = 1, 2, \dots, \end{cases} \tag{5}$$

являются несмещёнными и состоятельными оценками $f_\nu^{\cos}$ и $f_\nu^{\sin}$, $\nu \geqslant 1$ соответственно.

Таким образом, получаем следующую оценку для плотности f:

$$\hat{f}(x) = \frac{1}{2\pi} + \sum_{\nu=1}^{K_1} \hat{f}_\nu^{\cos} \cos \nu x + \sum_{\nu=1}^{K_2} \hat{f}_\nu^{\sin} \sin \nu x,$$

где $K_1 = K_1(Z_1, \dots, Z_n)$ и $K_2 = K_2(Z_1, \dots, Z_n)$ — статистики, которые мы будем называть коэффициентами сглаживания.

2.2. Методы выравнивания оценок плотностей распределения

Может оказаться, что получившаяся в результате оценивания плотности функция $\hat{f}$ принимает отрицательные значения или $\int \hat{f}(x)\,dx \neq 1$. Возможны два случая:

1) интеграл от положительной части функции больше или равен 1, т. е.

$$\int \max(\hat{f}(x), 0)\,dx \geqslant 1;$$

2) интеграл от положительной части функции меньше 1, т. е.

$$\int \max(\hat{f}(x), 0)\,dx < 1.$$

Будем пользоваться методами, описанными в [1]. Для сравнения исходной и исправленной оценки будем использовать функционал среднего накопленного квадрата ошибки (MISE):

$$\mathrm{MISE}(\hat{f}) = \mathbf{E} \int_{-\infty}^{\infty} \left(\hat{f}(x) - f(x)\right)^2 dx,$$

где $\hat{f}$ — оценка плотности, f — оцениваемая плотность.

Рассмотрим первый случай. Будем использовать исправленную оценку плотности

$$\tilde{f}(x) = \max(0, \hat{f} - \xi),$$

где ξ выбрана так, чтобы

$$\int\limits_{-\infty}^{\infty} \tilde{f}(x)\,dx = 1.$$

Рассматриваемая величина ξ существует и единственна. В [1] показано, что полученная оценка $\tilde{f}$ имеет меньший показатель среднего накопленного квадрата ошибки, чем изначальная оценка $\hat{f}$:

$$\mathrm{MISE}(\tilde{f}) \leqslant \mathrm{MISE}(\hat{f}).$$

Рассмотрим второй случай. Будем использовать оценку

$$\tilde{f}(x) = \tilde{f}(x; a, b) = \begin{cases} \max\bigl(0, \hat{f}(x)\bigr) + \nu_{a,b}, & x \in [a, b], \\ 0, & x \notin [a, b], \end{cases}$$

где $[a, b]$ — отрезок, на котором сосредоточена плотность,

$$\nu_{a,b} = \frac{1}{b-a}\biggl(1 - \int\limits_a^b \max\bigl(0, \hat{f}(x)\bigr)\,dx\biggr).$$

В этом случае

$$\int\limits_a^b \bigl(\tilde{f}(x) - f(x)\bigr)^2\,dx = \int\limits_a^b \bigl(\max\bigl(0, \hat{f}(x)\bigr) - f(x)\bigr)^2\,dx + \nu_{a,b}\int\limits_a^b \bigl(\tilde{f}(x) - f(x)\bigr)\,dx + \\ + \nu_{a,b}\int\limits_a^b \bigl(\max\bigl(0, \hat{f}(x)\bigr) - f(x)\bigr)\,dx \leqslant \int\limits_a^b \bigl(\hat{f}(x) - f(x)\bigr)^2\,dx,$$

где последнее неравенство вытекает из соотношений

$$\int\limits_a^b \bigl(\tilde{f}(x) - f(x)\bigr)\,dx = 0, \quad \int\limits_a^b \bigl(\max\bigl(0, \hat{f}(x)\bigr) - f(x)\bigr)\,dx \leqslant 0,$$

$$\bigl(\max\bigl(0, \hat{f}(x)\bigr) - f(x)\bigr)^2 \leqslant \bigl(\hat{f}(x) - f(x)\bigr)^2.$$

Таким образом, почти наверное выполняется неравенство

$$\mathrm{MISE}(\tilde{f}) \leqslant \mathrm{MISE}(\hat{f}).$$

2.3. Оценка погрешности метода Фурье для зашумлённых данных

Оценим погрешность полученной оценки $\hat{f}$. Общая погрешность метода складывается из погрешности перехода от математического ожидания $\mathbf{E}\,g_\nu^{\cos}(Z)$

к его несмещённой состоятельной оценке $\sum\limits_{i=1}^{n} g_\nu^{\cos}(Z_i)/n$ и из замены ряда Тейлора функции $g_\nu^{\cos}$ её частичной суммой для всех $\nu \geqslant 1$, а также из аналогичных переходов для $g_\nu^{\sin}$.

Представим ряд Тейлора функции $g_\nu^{\cos}(t)$ в виде

$$g_\nu^{\cos}(t) = \sum_{j=0}^{s} a_j(\nu)\frac{t^j}{j!} + R_s(t),$$

где

$$R_s(t) = \sum_{j=s+1}^{\infty} a_j(\nu)\frac{t^j}{j!}.$$

При этом

$$\left|\frac{a_j(\nu)t^j}{j!}\right| = \frac{(\nu/2)^j\Gamma(k)t^j}{j!\,\Gamma(k+j)} = \frac{(\nu/2)^j t^j}{j!\,\big(k\cdot\ldots\cdot(k+j-1)\big)}.$$

Ввиду неравенств $j! > (j/e)^j$ и $k(k+1)\ldots(k+j-1) > k^j$ имеем

$$|R_s(t)| < \sum_{j=s+1}^{\infty}\left(\frac{\big(\nu/(2k)\big)e}{j}\right)^j t^j.$$

При $t \in [-r, r]$ получим следующую оценку $|R_s(t)|$:

$$|R_s(t)| < \sum_{j=s+1}^{\infty}\left(\frac{\nu r e}{2kj}\right)^j < \sum_{j=s+1}^{\infty}\left(\frac{\nu r e}{2k(s+1)}\right)^j. \tag{6}$$

При $\nu r e < 2k(s+1)$ оценка (6) приобретает вид

$$|R_s(t)| < \frac{\nu r e}{2k(s+1) - \nu r e}.$$

Оценим погрешность перехода от математического ожидания $\mathbf{E}\, g_\nu^{\cos}(Z)$ к выборочному среднему $\sum\limits_{i=1}^{n} g_\nu^{\cos}(Z_i)/n$. В силу центральной предельной теоремы

$$\sqrt{n}\left(\frac{\sum\limits_{i=1}^{n} g_\nu^{\cos}(Z_i)}{n} - \mathbf{E}\, g_\nu^{\cos}(Z)\right) \Big/ \sqrt{\mathbf{D}\, g_\nu^{\cos}(Z)} \xrightarrow{d} N \sim \mathcal{N}(0,1), \quad n \to \infty.$$

Отсюда с учётом состоятельности

$$S_\nu^2 = \frac{1}{n}\sum_{i=1}^{n}\left(g_\nu^{\cos}(Z_i) - \frac{\sum\limits_{i=1}^{n} g_\nu^{\cos}(Z_i)}{n}\right)^2,$$

как оценки дисперсии с вероятностью, стремящейся к 95 % при $n \to \infty$, погрешность U_ν оценивается следующим образом:

$$U_\nu := \left| \sum_{\nu=1}^{n} \frac{\sum\limits_{i=1}^{n} g_\nu^{\cos}(Z_i)}{n} - \mathbf{E}\, g_\nu^{\cos}(Z) \right| \leqslant \sum_{\nu=1}^{n} \frac{1.96\sqrt{S_\nu^2}}{\sqrt{n}} =: B_\nu.$$

2.4. Оценка коэффициентов сглаживания

Используя полученную в разделе 2.3 оценку B_ν погрешности перехода от математического ожидания $\mathbf{E}\, g_\nu^{\cos}(Z)$ к выборочному среднему $\Big(\sum\limits_{i=1}^{n} g_\nu^{\cos}(Z_i)\Big)/n$, получаем оценки для K_1 и K_2. Далее проведены рассуждения для K_1; оценка для K_2 может быть получена из аналогичных соображений.

Для каждого $\nu \geqslant 1$ будем сравнивать оценку погрешности B_ν с выборочным средним

$$M_\nu := \frac{\sum\limits_{i=1}^{n} g_\nu^{\cos}(Z_i)}{n}$$

Если $B_\nu / M_\nu \leqslant 1$, то это означает, что соответствующий коэффициент $\mathbf{E}\, g_\nu^{\cos}(Z_1)$ значимо отличается от нуля, и мы должны использовать этот коэффициент в наших оценках. В противном случае мы будем считать коэффициент незначимым и исключим его из суммы.

Алгоритм состоит в том, чтобы выбрать минимальное K_1, такое что

$$B_{K_1+1} > M_{K_1+1} \quad \text{и} \quad B_{K_1+2} > M_{K_1+2}.$$

2.5. Моделирование метода

В ходе исследования мы проверяли работу методов на данных, моделированных из различных распределений.

Для построения оценок плотности величины X мы выбирали отрезок $[X_{\min}, X_{\max}]$ и приводили его линейным преобразованием к отрезку $[-\pi, \pi]$ для применения описанных в статье методов.

Приведённые графики построены для следующих параметров выборки: количество наблюдений $n = 1000$, длина траекторий $k = 20$. Из графиков видно, что математическое ожидание оценённой плотности распределения в большинстве случаев близко к математическому ожиданию истинной плотности распределения. По графику *г* видно, что в случае бимодальной истинной плотности распределения оценка также получается бимодальной. Можно заметить, что в некоторых случаях оценка не слишком хороша в абсолютном смысле. На графике *в* видно, что из-за редких значений выборки оценка может приобрести дополнительные моды. По всей видимости для предотвращения таких ситуаций нужно устранять выбросы и использовать отрезок, отличный от $[X_{\min}, X_{\max}]$.

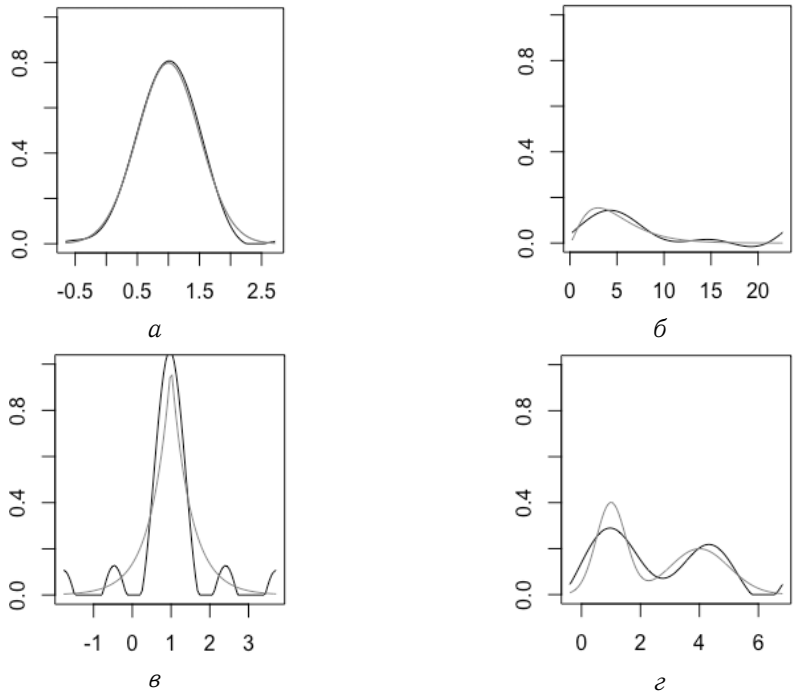

Рис. 1. Графики плотности и её оценка для распределений:
а — нормального, *б* — хи-квадрат, *в* — Лапласа, *г* — Тьюки

3. Дальнейшее развитие метода

Рассмотренный метод естественным образом обобщается на случай различных длин траекторий движения различных частиц. Необходимо также рассмотреть другие варианты оценки плотности, в частности не использующие знание носителя плотности. Требуется рассмотреть другие варианты оценки коэффициентов сглаживания K_1 и K_2.

Литература

[1] Glad I. K., Hjort N. L., Ushakov N. G. Correction of density estimators that are not densities // Scand. J. Statist. — 2003. — Vol. 30, no. 2. — P. 415—427.

[2] Hart J. D. On the choice of a truncation point in Fourier series density estimation // J. Statist. Comput. Simul. — 1985. — Vol. 21. — P. 95—116.

[3] Silverman B. W. Density Estimation for Statistics and Data Analysis. — London: Chapman & Hall, 1952

[4] Wasserman L. All of Nonparametric Statistics. — Berlin: Springer, 2006.